UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATION DE LA FRÉQUENCE DE FEU EN ASSURANCE COMMERCIALE PAR LE REBALANCEMENT DE DONNÉES ET L'ÉLICITATION D'AVIS D'EXPERTS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ANNE PAQUET

DÉCEMBRE 2025

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma sincère gratitude envers mon directeur de mémoire, Jean-Philippe Boucher, pour ses précieux conseils, son soutien constant et ses encouragements tout au long de ce projet. Sa rigueur académique et sa disponibilité ont grandement contribué à la réussite de ce travail.

Un remerciement tout particulier va aux autres professeurs et étudiants de la Chaire de recherche dont je fais partie. L'environnement stimulant et les discussions enrichissantes m'ont permis de progresser dans ma compréhension des enjeux actuariels et de mener à bien ce mémoire. Les échanges avec mes collègues chercheurs ont été précieux et ont grandement nourri ma réflexion. Je remercie également mes enseignants et intervenants, dont les cours et les échanges ont nourri ma réflexion et approfondi mes connaissances en actuariat. Leur expertise a été une source d'inspiration.

Je tiens également à remercier Co-Operators, qui a soutenu et financé une partie de mes études. Ce financement m'a permis de me consacrer pleinement à la recherche et de mener à bien mes travaux dans un cadre propice à l'épanouissement intellectuel. De plus, leur soutien m'a permis de réaliser l'importance des partenariats entre le milieu académique et l'industrie dans l'avancement de la discipline.

Je remercie aussi mes collègues et amis de la promotion, pour leurs discussions enrichissantes et leur soutien moral. Ils ont été des partenaires précieux dans ce parcours, partageant les mêmes défis et la même passion pour la discipline. Je n'oublie pas ma famille, dont l'amour, la patience et la compréhension ont été essentiels tout au long de mes études. Leur soutien indéfectible m'a permis de persévérer et de mener à bien ce projet.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont contribué à l'accomplissement de ce mémoire. Leur aide, qu'elle soit intellectuelle ou émotionnelle, a été indispensable.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Dans le monde de l'assurance, les différents assureurs font face à un défi de taille : obtenir la meilleure segmentation de risques possible. Cet enjeu peut compromettre la capacité d'un assureur à rester compétitif sur le marché, et ainsi diminuer sa part par rapport aux concurrents. De plus, en assurance commerciale spécifiquement, les différents risques contenus dans un portefeuille peuvent être très diversifiés. De ce fait, les méthodes de modélisation traditionnellement utilisées ne permettent pas toujours d'obtenir une solide segmentation. Cette réalité pousse donc les assureurs à constamment essayer d'innover leurs méthodes de segmentation. De plus, pour plusieurs lignes d'affaires dont l'assurance commerciale, la fréquence des réclamations est souvent très basse. De ce fait, la distribution de la variable d'intérêt se trouve être largement débalancée. Ce déséquilibre entraîne certains enjeux quant à l'utilisation des algorithmes de tarification traditionnels. Alors, pour faire face à ce problème, il peut être important de rebalancer l'échantillon utilisé.

Nous avons structuré ce mémoire en deux volets. Le premier volet consiste à rebalancer la base de données de façon à équilibrer les différentes classes de la variable d'intérêt. Nous utilisons une méthode hybride, en appliquant l'algorithme de suréchantillonnage SMOTENC et l'algorithme de nettoyage Tomek Link. Ensuite, nous faisons une analyse de l'échantillon obtenu pour voir comment ce balancement a modifié le portefeuille de l'assureur. Dans le second volet, nous ajustons 4 différents modèles à cette base de données : un modèle n'incluant pas l'avis d'expert et trois autres modèles dans lesquels différents niveaux de précision de l'avis des experts sont inclus. Dans ces trois modèles, l'élicitation des experts se fait par la troncation de la distribution *a posteriori* des effets aléatoires du modèle. Finalement, nous évaluons la performance de ces modèles par l'utilisation de deux différentes métriques : l'aire sous la courbe ROC et l'aire sous la courbe de précision-rappel.

Selon notre analyse, l'inclusion de l'avis des experts améliore la performance du modèle de référence, et ce pour les trois niveaux de précision. De plus, plus l'avis des experts est précis, plus la performance du modèle est élevée. Cette méthode a cependant ses limites. L'avis d'expert que nous utilisons ici est simulé. Il serait intéressant de recueillir une vraie élicitation d'expert et tester si les résultats sont toujours aussi concluants.

INTRODUCTION

Le but premier de l'assurance est de protéger les individus contre les conséquences financières négatives de certains événements aléatoires défavorables. Lorsqu'un individu assume complètement un risque par lui-même, advenant la réalisation de cet événement, les dommages peuvent être si importants qu'il ne sera pas en mesure de supporter la totalité des pertes financières. Par exemple, si la résidence d'un propriétaire est complètement anéantie par un incendie, sa situation financière ne lui permettra peut-être pas d'en acquérir une nouvelle. Cependant, si le risque d'incendie est partagé entre plusieurs propriétaires par les mécanismes de l'assurance, en cas d'incendie, le propriétaire est protégé et peut faire face aux pertes financières encourues.

L'assurance permet donc de garantir une certaine sécurité financière aux individus faisant face à l'adversité. Cette protection se traduit par la création d'un contrat d'assurance entre une compagnie d'assurance et un assuré. Il s'agit d'une entente selon laquelle l'assuré transfère un certain risque spécifique à l'assureur, moyennant le paiement d'une prime. Dans le cadre de notre recherche, le risque spécifique que nous modéliserons est la couverture de feu pour la structure, et ce pour la ligne d'affaires commerciale. Par ce contrat d'assurance, l'assureur s'engage donc à assumer les pertes financières associées à cet événement aléatoire spécifique. En regroupant un grand nombre de ces risques individuels, la distribution globale des pertes devient plus prévisible. Alors, la compagnie d'assurance peut en tirer avantage et générer du profit (Grize, 2015). Ce mécanisme est souvent référé comme le principe de l'assurance.

Ce principe se base sur la loi des grands nombres. Une hypothèse clé dans l'application de ce théorème est que les risques doivent être similaires ou se comporter de manière homogène. Ainsi, pour que les compagnies d'assurance puissent créer de la valeur en réduisant la volatilité des sinistres, leur portefeuille doit être composé de risques homogènes, ou du moins comprendre un nombre suffisamment important de risques similaires (Macedo, 2009). C'est pourquoi le concept de segmentation des risques est si important dans le domaine de l'assurance. Ce marché tend à différencier les taux de prime selon le degré de risque des assurés en les regroupant dans différentes catégories ayant un ensemble homogène de caractéristiques. Cela se résume généralement à classifier les risques dans différentes classes selon lesquelles les membres partagent le même profil de risque. Conséquemment, chaque catégorie a un certain nombre de réclamations d'assurance et de pertes accumulées, et les assurés de la même classe paieront le même taux de prime (El Kassimi et Zahi, 2021).

Pour certaines lignes d'affaires, comme par exemple l'assurance auto, les assurés qui composent le portefeuille sont, dans une certaine mesure, assez similaires les uns des autres ; ce sont tous des individus qui conduisent un véhicule. Nous pouvons segmenter ces assurés en différents groupes selon certaines caractéristiques, ces dernières étant corrélées au risque d'avoir un accident, et ainsi regrouper les risques similaires ensemble. Cependant, pour d'autres lignes d'affaires, comme par exemple en assurance commerciale, notre sujet de recherche, cette segmentation peut être difficile. Les différents risques qui composent ce genre de portefeuilles sont souvent très diversifiés. Cela peut s'expliquer par la nature très variée des occupations des différentes entreprises clientes d'un assureur. La méthode de tarification que nous proposons donc ici modélise le taux de prime selon l'occupation de l'entreprise.

Ce taux de prime représente le prix facturé à l'assuré pour le transfert du risque à la compagnie d'assurance et il remplit une double fonction. Premièrement, il doit produire des fonds totaux suffisants pour couvrir l'obligation de l'assureur. Deuxièmement, il doit partager équitablement le coût de l'assurance entre les assurés. Les assureurs sont particulièrement préoccupés par l'établissement d'un système de tarification qui répartisse ces sinistres et pertes entre les assurés de la manière la plus équitable et la plus raisonnable possible (El Kassimi et Zahi, 2021). Pour ce faire, cette prime est calculée selon la probabilité de survenance d'un sinistre, et les facteurs de risque pour chaque classe sont déterminés de façon à ce que chacun paie une prime qui reflète leur profil de risque. Nous pouvons nous référer à (El Kassimi et Zahi, 2021) pour une revue de littérature des différentes techniques de tarification en assurance de dommage.

Comme mentionné plus haut, un contrat d'assurance est une entente selon laquelle l'assuré transfère un certain risque spécifique à l'assureur, moyennant le paiement d'une prime. Ce type de transaction possède un caractère abstrait où le hasard joue un rôle car l'entreprise acquiert un risque et enlève au client l'incertitude d'une perte potentielle. C'est pour cette raison que les probabilités et les statistiques sont à la base de l'entreprise (Grize, 2015). Le marché de l'assurance est un secteur pour lequel les statistiques constituent la base fondamentale sur laquelle l'entreprise est construite, c'est-à-dire pourquoi elle peut exister et générer des profits.

Nous pouvons diviser le domaine des statistiques en deux principales écoles : l'approche fréquentiste et l'approche bayésienne. Ces deux philosophies s'opposent de façon fondamentale par leur interprétation de la probabilité. L'approche fréquentiste utilise des données pour évaluer la probabilité qu'un résultat spécifique provenant d'un certain événement se réalise en étudiant de nombreuses répétitions de ce même

événement (Harris et Rice, 2013). Cette approche voit la probabilité dans le sens de proportions à long terme. Cependant, dans certains cas où nous n'avons pas ou peu de données, comme par exemple pour la tarification d'une nouvelle ligne d'affaires ou d'un nouveau territoire, cette approche peut causer problème. Dans ce type de situation, l'approche bayésienne peut être mieux adaptée.

L'approche bayésienne se concentre sur l'estimation de la probabilité qu'une hypothèse soit vraie, compte tenu des résultats disponibles (Harris et Rice, 2013). Le raisonnement bayésien s'intéresse donc tout d'abord à formuler la probabilité que l'hypothèse soit vraie par l'utilisation des données existantes. Cette approche peut être très utile dans une situation où il existe très peu de données. D'autre part, dans le cas d'un nouveau secteur d'activité ou d'un nouveau territoire, les actuaires s'appuient souvent sur la disponibilité de données externes ou sur l'expérience d'un secteur d'activité similaire. Par la suite, lorsque de nouvelles données sont collectées, cette probabilité antérieure peut être révisée en fonction de la nouvelle information obtenue.

Jusqu'ici, il existe beaucoup de recherches sur l'élicitation de données historiques dans un contexte d'inférence bayésienne. L'article (Raina *et al.*, 2006) présente une analyse de différentes méthodes par lesquelles de l'information peut être formalisée dans une distribution *a priori*. Dans (Ibrahim et Chen, 2000), les auteurs proposent une classe de fonctions *a priori* générale, les *power priors*, ces dernières pouvant être utilisées dans plusieurs domaines afin d'inclure des données historiques dans différents modèles de régression. L'article (Chen *et al.*, 2003) propose aussi une méthode pour inclure les données historiques, mais ce dans le contexte de modèles linéaires mixtes généralisés.

Comme mentionné ci-dessus, nous pouvons intégrer les données disponibles comme information *a priori* dans un problème d'inférence bayésienne. Tout comme les données externes et l'expérience d'un secteur d'activité similaire, l'avis d'experts représente une forme de données existantes qui peut être intéressante à utiliser. Par experts, nous entendons une personne qui a des connaissances sur un certain sujet d'études, comme par exemple sur les différentes classes de risque d'un portefeuille, sans nécessairement tout connaître sur le sujet.

Dans le domaine de l'assurance, un souscripteur est un professionnel qui a la capacité de comprendre les risques d'un contrat d'assurance qui sont transférés à l'assureur (Macedo, 2009). Cette capacité s'acquiert non seulement grâce à des études théoriques, mais est également le résultat d'années d'expérience dans la gestion de risques similaires et dans le paiement des réclamations pour ces risques. En assurance com-

merciale, comme les différents assurés d'un portefeuille sont très diversifiés, la segmentation de ces risques peut devenir difficile. Les souscripteurs possèdent donc une connaissance intéressante des risques qui peut être utile à la modélisation d'un portefeuille. C'est pourquoi nous croyons que l'élicitation de leur connaissance et son intégration dans les algorithmes de tarification peut apporter une aide importante à la segmentation des différentes classes de risque, et ainsi améliorer la performance des modèles de tarification. Dans cette recherche, nous proposons donc une méthode de tarification bayésienne permettant d'inclure l'avis des souscripteurs en assurance commerciale comme information *a priori*.

L'élicitation d'experts consiste à obtenir de l'information de la part des experts et de la représenter sous forme de distributions *a priori*. Ces étapes clés de l'analyse bayésienne informative permettent d'obtenir les opinions préalables d'experts sur les valeurs possibles de paramètres en termes de probabilité. Cette méthode peut être approchée de deux différentes façons. D'un côté, l'élicitation directe consiste à demander à l'expert d'attribuer une valeur directement aux coefficients du modèle de régression. L'élicitation indirecte, quant à elle, implique plutôt d'obtenir des experts les valeurs de la variable réponse compte tenu des valeurs des covariables, ou vice versa (O'Leary *et al.*, 2008). Cette opinion est ensuite transformée mathématiquement en une distribution *a priori* pour les paramètres du modèle. Bien que l'avis d'expert représente une source d'information intéressante, cette dernière peut être biaisée et doit être utilisée avec vigilance. Il existe plusieurs ouvrages traitant ce sujet. (Kynn, 2008) représente une bonne référence pour la recherche psychologique sur l'évaluation des probabilités et donne des lignes directrices concrètes pour obtenir des connaissances d'experts.

Bien que dans la littérature scientifique, l'information *a priori* en inférence bayésienne provient majoritairement de données historiques, nous retrouvons plusieurs études sur l'élicitation des connaissances des experts. Dans (Low Choy *et al.*, 2009), les auteurs décrivent un cadre de conception statistique pour quantifier les connaissances d'experts sous une forme adaptée aux modèles bayésiens. Plusieurs différents exemples de modélisation incluant l'avis d'experts sous forme d'information *a priori* sont proposés dans (O'Leary *et al.*, 2008) et (Kuhnert, 2011). De façon plus spécifique au domaine de l'assurance, dans (Zhang et Miljkovic, 2018), les auteurs proposent d'inclure l'avis d'expert sous forme de *power priors* pour la tarification d'un nouveau territoire.

Dans les exemples cités ci-dessus, l'avis d'experts est inclus sous forme de fonction *a priori* pour les coefficients du modèle. Dans (Ni *et al.*, 2018), les auteurs proposent par simulation une méthode incluant plutôt

l'avis d'experts en tronquant la distribution des effets aléatoires d'un modèle linéaire mixte. Suite à leurs résultats concluants, ils appliquent leur méthode dans (Ni *et al.*, 2021) pour modéliser le diagnostic de cétose subclinique chez les vaches laitières. Suite à une analyse de ces études, nous croyons que cette méthode pourrait être utile dans le domaine de l'assurance pour améliorer les algorithmes de tarification. Dans le contexte de notre recherche, nous proposons donc des modèles linéaires mixtes généralisés incluant l'avis des souscripteurs sur le risque des différentes classes de risque par la troncation de la distribution des effets aléatoires.

Une partie importante de la modélisation de données est d'étudier la qualité de l'échantillon avec lequel nous travaillons. En assurance commerciale, la fréquence de réclamations d'un portefeuille est souvent très basse, résultant en une distribution déséquilibrée de la variable d'intérêt. Afin de maximiser le pouvoir prédictif des modèles que nous developpons, une partie importante de notre analyse est de rééquilibrer la base de données avec laquelle nous ajustons les modèles. Nous proposons ici une méthode de rebalancement de données hybride combinant un algorithme de suréchantillonnage, SMOTENC, et un algorithme de nettoyage, Tomek Link.

Le chapitre I présente notre exploration des différentes méthodes d'échantillonnage pour base de données débalancée, et notre évaluation de la méthode la mieux adaptée pour rebalancer la base de données avec laquelle nous travaillons. Par la suite, dans le chapitre II, nous appliquons l'algorithme de rebalancement élaboré dans le chapitre I et analysons la base de données obtenue. Ensuite, dans le chapitre III, nous présentons les différents modèles incluant l'avis d'expert et comparons leur performance par rapport à un modèle de référence. Finalement, nous terminons avec les conclusions que nous tirons des résultats de notre analyse et par différentes propositions qui pourraient améliorer la modélisation proposée.

CHAPITRE 1
PROCÉDURE DE REBALANCEMENT

Le but de notre recherche est d'inclure l'avis des experts dans les algorithmes de tarification commerciale afin d'améliorer la prédiction des réclamations de feu pour la structure. Après une analyse de la base de données avec laquelle nous travaillons, nous avons remarqué que l'échantillon contient une proportion de polices sans réclamation beaucoup plus importante que celle avec réclamation. Observant une fréquence de sinistres de moins de 0,2%, nous pouvons conclure que nous travaillons avec une base de données largement débalancée.

1.1 Base de données débalancée

Nous qualifions une base de données comme étant débalancée lorsque les différentes classes d'une variable ne sont pas également représentées. Bien que cette situation puisse aussi se présenter dans des problèmes de classification multinomiale, dans cette analyse, nous allons nous concentrer sur la classification binomiale. Lorsqu'une des classes de la variable d'intérêt est largement sous-représentée, les algorithmes d'apprentissage ont tendance à moins bien performer. Le modèle pourra plus difficilement développer des règles de classification qui considèrent les deux classes. Alors, les instances de la classe minoritaire risquent d'être ignorées, celles-ci étant noyées par celles de la classe majoritaire (Lunardon *et al.*, 2014). Cela s'explique par le fait que ces algorithmes utilisent typiquement l'exactitude comme mesure de performance (Chawla *et al.*, 2002).

Considérons la table de contingence suivante, où VP représente le nombre de vrai positif, FP le nombre de faux positif, FN le nombre de faux négatif et VN le nombre de vrai négatif :

	Prédiction positive	Prédiction négative
Instance positive	VP	FN
Instance négative	FP	VN

Figure 1.1 Table de contingence pour un problème de classification dichotomique

L'exactitude E se définit alors comme étant :

$$E = \frac{VP + VN}{VP + FP + VN + FN}$$

Considérons maintenant une base de données ayant une seule instance positive pour 100 observations. Nous avons donc 99 observations dans la classe majoritaire, et une seule observation dans la classe minoritaire. Supposons maintenant que l'algorithme ignore complètement la classe minoritaire et ne prédit que des instances négatives. Nous obtenons donc la table de contingence suivante :

	Prédiction positive	Prédiction négative
Instance positive	0	1
Instance négative	0	99

Figure 1.2 Table de contingence pour un problème de classification dichotomique

L'exactitude aura alors la valeur suivante :

$$E = \frac{0 + 99}{0 + 1 + 99 + 0} = 99\%$$

En ne prédisant que des instances négatives et en ignorant complètement la classe minoritaire, l'algorithme atteint une exactitude de 99%. En entraînant ce genre de modèles de classification sur une base de données débalancées, nous nous retrouvons à favoriser largement la classe majoritaire, au détriment de la classe minoritaire.

Cette situation pose un défi important dans les différents domaines où le principal intérêt est la classe minoritaire. C'est le cas, par exemple, dans la détection des déversements d'hydrocarbures, une menace

majeure pour les écosystèmes océaniques et côtiers, ainsi que pour diverses activités humaines liées à la marine. Ce type de pollution, pouvant être détectée par l'analyse d'images radar, représente une importante préoccupation (Krestenitis *et al.*, 2019).

Plus la base de données avec laquelle nous travaillons est débalancée, plus les risques d'une mauvaise performance des algorithmes sont grands. Cela pose problème dans les domaines comme l'analyse de données biomédicales, où les données sont extrêmement débalancées. Dans une base de données de mammographie, nous pouvons habituellement compter à peu près 98% de pixels normaux et 2% de pixels anormaux (Chawla *et al.*, 2002). Avec un modèle utilisant l'exactitude comme métrique de performance, nous risquons de ne pas détecter les cellules cancéreuses, ce qui est d'un intérêt capital pour la santé des patients.

Nous pouvons aussi retrouver des bases de données extrêmement débalancées en détection de fraude, sujet d'intérêt dans le domaine de l'assurance. Dans la plupart des bases de données, la fraude se produit typiquement dans moins de 0,5% des cas (Baesens *et al.*, 2021). Cela s'explique par le fait que seulement une petite fraction des instances est concernée par la fraude. De plus, parmi ces observations, seulement certaines d'entre elles sont connues comme étant frauduleuses. Cela résulte en des bases de données largement débalancées.

Il existe plusieurs autres sujets d'intérêt dans le domaine de l'assurance où nous pouvons observer des bases de données largement débalancées. C'est notamment le cas pour notre domaine de recherche, la tarification en assurance commerciale. Comme mentionné précédemment, la base de données que nous utiliserons pour notre analyse de la couverture de feu pour la structure présente une fréquence de réclamations de moins de 0,2%. Il est donc important de rebalancer notre base de données avant d'aller plus loin dans notre analyse.

Nous avons vu dans les exemples précédents que le but principal de l'analyse est de classifier aussi bien que possible chacune des observations. Ce sont donc des problèmes de classification. En analyse de données biomédicales, il est primordial de détecter toutes cellules cancéreuses rapidement afin de traiter le patient le plus tôt possible. Il est donc important de bien évaluer chacune des instances individuellement. De la même façon, en détection des déversements d'hydrocarbures, chacune des images radar doit être bien classifiée afin de détecter les déversements dans les océans et ainsi limiter leur impact écologique. En détection de fraude, la situation est similaire. Chacune des réclamations doit être classée comme étant

frauduleuse ou non afin d'éviter le paiement de fausses réclamations. Nous pouvons observer que, dans chacune de ces situations, notre but premier est de détecter la réalisation d'un événement passé.

En tarification en assurance, notre but est plutôt d'évaluer la probabilité qu'un événement futur se produise. Lorsqu'un assureur émet un nouveau contrat, il doit évaluer la valeur de ce contrat afin de déterminer la prime à charger à l'assuré. Typiquement, cette prime varie en fonction du profil de risque de l'assuré. Alors, l'assureur fait une segmentation de son portefeuille en regroupant les risques similaires ensemble, et évalue la prime à charger pour chacun de ces groupes. Cette prime contient deux différentes composantes : la fréquence et la sévérité. La fréquence se traduit par la probabilité que l'assuré fasse une réclamation à l'assureur. La sévérité, quant à elle, représente le montant de cette réclamation. La relation de ces deux composantes avec la prime chargée peut être décrite par la formule suivante :

$$\text{Prime} = \text{Fréquence} \times \text{Sévérité}$$

Dans notre analyse, nous allons nous concentrer sur la fréquence des réclamations. Comme mentionné plus haut, la portion d'observations ayant une réclamation dans l'échantillon est de moins de 0,2%. Alors, sans réajustement, les instances avec réclamation se retrouveront noyées par les instances sans réclamation, et les assurés risqués seront sous-représentés. Conséquemment, l'assureur ne sera pas en mesure de charger une prime adéquate lui permettant de couvrir toutes les pertes encourues. C'est pourquoi il est important de rebalancer la base de données avant d'aller plus loin dans notre analyse.

Il existe à ce jour un grand nombre de stratégies visant à gérer les problèmes liés aux bases de données débalancées. Nous pouvons considérer ces méthodes sous quatre angles différents : les stratégies d'échantillonnage, l'apprentissage avec pénalité, les méthodes par noyau et l'apprentissage actif. Dans notre analyse, nous allons nous concentrer sur les stratégies d'échantillonnage. Vous pouvez vous référer à (He et Garcia, 2009) pour une description détaillée des autres approches.

1.2 Stratégies d'échantillonnage

Les stratégies d'échantillonnage consistent à modifier la base de données afin de rebalancer la distribution des différentes classes. Parmi ces méthodes, citons notamment le sous-échantillonnage, le sur-échantillonnage,

le nettoyage de données et les méthodes hybrides.

1.2.1 Le sous-échantillonnage

Le sous-échantillonnage consiste à sélectionner des instances de la classe majoritaire et à les retirer de l'échantillon. Le nombre d'instances retirées est spécifié de façon arbitraire par l'utilisateur de l'algorithme. Typiquement, le nombre d'instances de la classe majoritaire est réduit au nombre d'instances de la classe minoritaire, pour ainsi obtenir un ratio de 1 :1 entre les deux classes, résultant en une fréquence de 50%.

Il existe plusieurs méthodes de sélection pour identifier les instances de la classe majoritaire que nous désirons retirer de l'échantillon. Le sous-échantillonnage aléatoire est la forme la plus simple de sous-échantillonnage car il sélectionne des instances de la classe majoritaire de façon complètement aléatoire, alors que dans le sous-échantillonnage informé, certaines connaissances statistiques sont utilisées pour sélectionner les instances majoritaires à retirer (Baesens *et al.*, 2021).

1.2.1.1 Le sous-échantillonnage aléatoire

Considérons un échantillon original S_0. Le sous-échantillonnage aléatoire consiste à sélectionner aléatoirement un certain nombre d'instances $|E|$ de la classe majoritaire S_{maj} à retirer de l'échantillon. Le nombre d'instances $|E|$ à retirer de l'échantillon original est pré-sélectionné par l'utilisateur afin d'atteindre le ratio désiré entre la classe minoritaire S_{min} et la classe majoritaire S_{maj}. Le nombre d'instances de notre échantillon final S se retrouve donc à être $|S| = |S_{min}| + |S_{maj}| - |E|$.

Un désavantage important de cette méthode est la suppression d'informations importantes provenant de la classe majoritaire. Le modèle de classification qui sera alors obtenu par cette base de données réduite peut manquer des concepts importants propres à la classe majoritaire (He et Garcia, 2009). Le sous-échantillonnage informé peut nous permettre d'atténuer ce problème.

1.2.1.2 Le sous-échantillonnage informé

Les méthodes de sous-échantillonnage informé se concentrent sur l'identification des instances qui devraient être conservées dans l'échantillon, plutôt que sur celles qui devraient être écartées de l'analyse. Un exemple de sous-échantillonnage informé est l'algorithme *NearMiss*. Cet algorithme suggère 3 différentes

méthodes de sélection pour choisir les instances majoritaires à conserver dans l'échantillon.

NearMiss-1 consiste à sélectionner, pour chacune des instances de la classe majoritaire, ses N plus proches voisins faisant partie de la classe minoritaire. Les instances retenues dans l'échantillon seront celles pour lesquelles la distance moyenne avec ces voisins est la plus petite. Par exemple, sur la figure 1.3, nous pouvons constater que la distance moyenne entre l'instance encerclée en vert et ses $N = 3$ plus proches voisins est plus petite que celle de l'instance encerclée en rouge. Entre ces deux points, ce sera donc le vert qui sera conservé par l'algorithme.

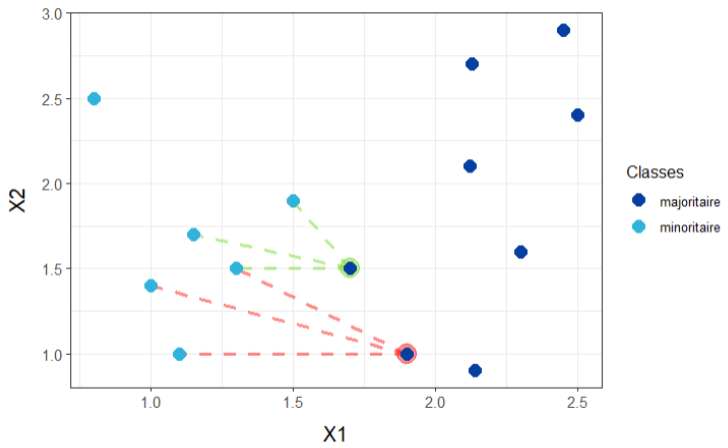

Figure 1.3 L'algorithme *NearMiss-1*

D'autre part, *NearMiss-2* consiste à sélectionner, pour chacune des instances de la classe majoritaire, ses N voisins les plus éloignés faisant partie de la classe minoritaire. Cependant, les instances retenues dans l'échantillon seront celles pour lesquelles la distance moyenne avec ces voisins est la plus grande. Par exemple, sur la figure 1.4, nous pouvons constater que la distance moyenne entre l'instance encerclée en vert et ses $N = 3$ voisins les plus éloignés est plus grande que la distance moyenne de l'instance encerclée en rouge. Entre ces deux points, ce sera donc le point en vert qui sera sélectionné par l'algorithme.

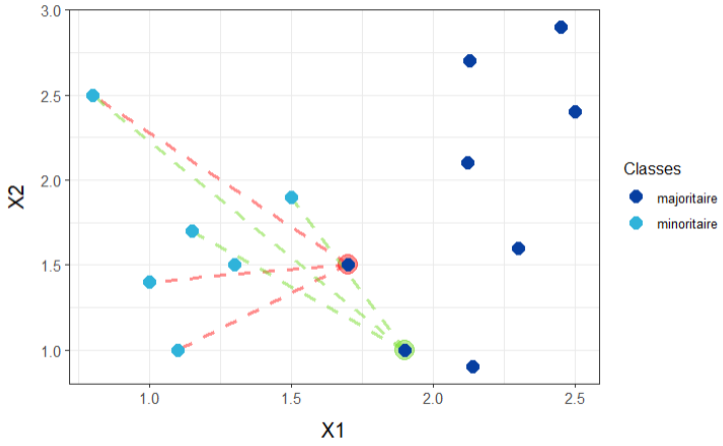

Figure 1.4 L'algorithme *NearMiss-2*

L'algorithme *NearMiss-3*, pour sa part, se fait en deux différentes étapes. Premièrement, pour chaque instance de la classe minoritaire, ses N plus proches voisins faisant partie de la classe majoritaire sont sélectionnés. Dans la figure 1.5, ces points sont représentés en jaune. Toutes les instances non sélectionnées sont supprimées. Ensuite, la deuxième étape consiste à sélectionner, pour chaque instance conservée de la classe majoritaire, ses N plus proches voisins faisant partie de la classe minoritaire. De la même façon qu'avec l'algorithme *NearMiss-1*, les instances retenues dans l'échantillon seront celles dont la distance moyenne avec ces voisins est la plus petite.

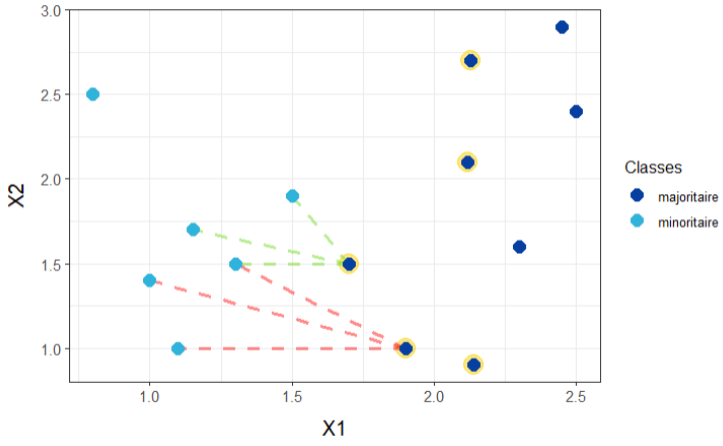

Figure 1.5 L'algorithme *NearMiss-3*

Le principal avantage des différentes méthodes de sous-échantillonnage est la réduction des ressources computationnelles requises. Cependant, comme mentionné plus haut, cela implique une suppression d'information potentiellement importante pour notre analyse. Bien qu'elles requièrent plus de temps de calcul, les méthodes de sur-échantillonnage peuvent, dans certains contextes, être mieux adaptées et produirent de meilleurs résultats.

1.2.2 Le sur-échantillonnage

Contrairement au sous-échantillonnage, les algorithmes de sur-échantillonnage ne suppriment aucune donnée de la classe majoritaire de l'échantillon. La stratégie est plutôt d'ajouter des instances de la classe minoritaire.

1.2.2.1 Le sur-échantillonnage aléatoire

Le sur-échantillonnage aléatoire consiste à créer un échantillon E en sélectionnant aléatoirement $|E|$ instances de la classe minoritaire S_{min}. Une première instance est d'abord sélectionnée, pour ensuite être dupliquée et ajoutée à l'échantillon E, avant d'être replacée dans la base de données originale. Cette étape est répétée $|E|$ fois afin de créer la base de données E. Cet échantillon est ensuite ajouté à l'échantillon original S_0. Le nombre d'instances $|E|$ ajouté à l'échantillon original est, de la même façon que pour le sous-échantillonnage, pré-sélectionné par l'utilisateur afin d'atteindre un certain ratio entre les deux classes. Le nombre d'instances de notre échantillon final S se retrouve donc à être $|S| = |S_{min}| + |S_{maj}| + |E|$.

Comme nous venons de le voir, le sur-échantillonnage aléatoire consiste simplement à dupliquer des instances de la classe minoritaire, ce qui peut amener des règles de décisions très spécifiques, et ainsi créer du surajustement (Baesens *et al.*, 2021). Différentes méthodes de génération de données synthétiques peuvent être utilisées afin d'atténuer ce problème.

1.2.2.2 Le sur-échantillonnage synthétique

Le sur-échantillonnage synthétique ajoute de l'information à l'échantillon original par la génération de nouvelles instances synthétiques à l'intérieur de la classe minoritaire. L'objectif de cette approche est d'agrandir la zone de décision de la classe minoritaire, et ainsi améliorer la performance de classification de l'algorithme. La génération de nouvelles données artificielles qui n'ont pas été observées précédemment ré-

duit le risque de surajustement et augmente notre habileté à généraliser, celle-ci compromise par le sur-échantillonnage aléatoire (Lunardon *et al.*, 2014).

Il existe plusieurs techniques de sur-échantillonnage synthétique. La *Synthetic Minority Oversampling Technique (SMOTE)* est un puissant algorithme ayant eu beaucoup de succès dans différentes applications (He et Garcia, 2009). Afin de générer une nouvelle donnée synthétique, cet algorithme sélectionne tout d'abord une des instances de la classe minoritaire comme point de référence. Ensuite, ses N plus proches voisins sont calculés, et l'un d'entre eux est sélectionné aléatoirement. Sur la figure 1.6, ce point correspond à celui relié à la donnée de référence par la ligne foncée. Après avoir calculé la distance entre ces deux points, cette distance est multipliée par un nombre aléatoire compris entre 0 et 1. En ajoutant ce résultat à la valeur de notre point de référence, nous obtenons une nouvelle donnée, représentée en vert sur la figure 1.6.

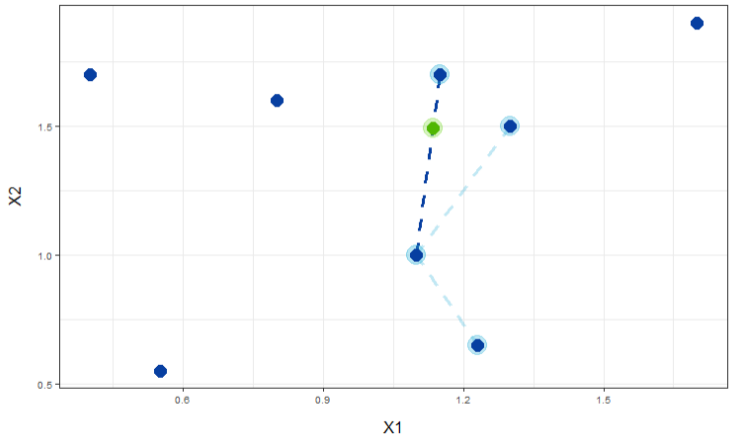

Figure 1.6 L'algorithme *SMOTE*

Après avoir appliqué ce même processus à chacune des observations de la classe minoritaire, nous nous retrouvons à avoir doublé la dimension de cette classe. Si nous sélectionnons plus d'un des N plus proches voisins à partir desquels une nouvelle instance est créée, nous pouvons obtenir plusieurs nouvelles données par instance, et ainsi augmenter le nombre d'observations de façon plus importante, et ce jusqu'à l'obtention de la proportion qui nous convient.

Une des limites de *SMOTE* est que cette méthode ne fonctionne qu'avec les variables numériques. Il existe cependant une variante de l'algorithme, *SMOTENC*, qui prend aussi en charge les variables catégorielles. De

la même manière que *SMOTE*, l'algorithme sélectionne tout d'abord les N plus proches voisins du point de référence. Cependant, au lieu de créer une nouvelle instance par l'interpolation de deux points, la catégorie de la nouvelle donnée correspondra à celle qui est le mieux représentée par les voisins du point de référence. Sur la figure 1.7, nous pouvons voir que parmi les points reliés entre eux, la majorité proviennent de la catégorie A. Alors, ce sera cette catégorie qui sera assignée à la nouvelle donnée synthétique, présentée en vert.

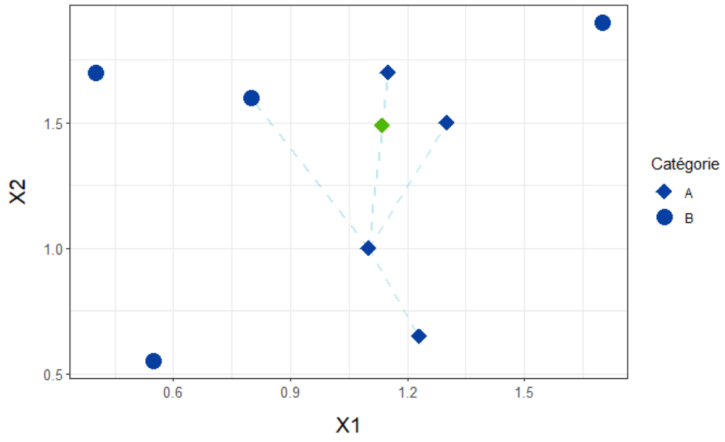

Figure 1.7 L'algorithme *SMOTENC*

SMOTE est une technique de rebalancement de base de données largement utilisée. Cependant, cette méthode comporte des désavantages importants. D'une part, comme les nouvelles données synthétiques ne sont qu'une représentation simplifiée de la vraie distribution de la classe minoritaire, certaines de ces dernières peuvent venir ajouter du bruit à la classe, et ainsi compromettre la qualité de l'analyse. Cela se produit lorsqu'une nouvelle donnée se retrouve dans la zone de la classe majoritaire. L'algorithme *KMeans-SMOTE* est une variante de *SMOTE* qui tente d'y remédier en regroupant les données de l'échantillon avec l'algorithme *KMeans*. Seuls les groupes ayant une certaine proportion d'instances de la classe minoritaire, typiquement ceux au-dessus de 50%, seront utilisés pour générer de nouvelles données. De plus, une plus grande portion des nouvelles données proviendront des groupes clairsemés.

Un autre problème de *SMOTE* sont les données synthétiques générées trop près des limites leur classe. Ces instances ont aussi tendance à déteriorer la qualité de la classification. L'algorithme *Borderline-SMOTE* a été développé pour régler ce problème en faisant une pré-sélection des points de l'échantillon original

auxquels *SMOTE* sera appliqué. Seules les instances qui se situent aux limites des classes de l'échantillon seront utilisées pour générer de nouvelles données synthétiques.

Une autre exemple de sur-échantillonnage synthétique est l'*adaptive synthetic (ADASYN) sampling*. L'idée principale de l'algorithme *ADASYN* est d'utiliser la distribution des données comme critère pour automatiquement décider le nombre de données synthétiques qui seront générées pour chacune des instances minoritaires (He *et al.*, 2008). Cette technique s'oppose à *SMOTE*, où un même nombre de données sythétiques sont créées pour chaque instance minoritaire. Un des avantages d'*ADASYN* par rapport à *SMOTE* est sa capacité à réduire le biais introduit par le débalancement des classes. De plus, l'algorithme permet de déplacer les frontières de décision, et ainsi permettre un plus grand focus sur les instances desquelles il est difficiles d'apprendre (He *et al.*, 2008).

1.2.3 Le nettoyage de données

Le nettoyage de données est une méthode de sous-échantillonnage qui tente d'identifier et supprimer les données qui réduisent la qualité de la classification. Certaines techniques ont pour but d'enlever les données ajoutant du bruit dans l'échantillon, alors que d'autres se concentrent plutôt sur la suppression d'instances trop faciles à classifier. Une des particularités des méthodes de nettoyage est que le ratio entre les classes ne peut pas être spécifié par l'utilisateur. Le nombre final d'instances dans chacune des classes varie selon l'algorithme de nettoyage choisi et la base de données sur laquelle il est appliqué.

1.2.3.1 L'algorithme *TomekLink*

Le but de l'algorithme *TomekLink* est de retirer les instances qui chevauchent les différentes classes de la base de données afin de nettoyer les frontières de décision. Les instances appartenant à deux différentes classes formant des liens Tomek sont retirées jusqu'à ce que toutes les distances minimales entre deux points soient formées d'observations appartenant à la même classe.

Les points formant des liens Tomek sont identifiés selon la méthode suivante. Soit deux instances (x_i, x_j), où $x_i \in S_{min}$, $x_j \in S_{max}$ et $d(x_i, x_j)$ est la distance entre x_i et x_j. La paire (x_i, x_j) forment un lien Tomek s'il n'y a pas d'autres instances x_k telles que $d(x_i, x_k) < d(x_i, x_j)$ ou $d(x_j, x_k) < d(x_i, x_j)$. Dans la figure 1.8, nous pouvons voir une représentation d'un lien Tomek par les deux points reliés en vert. L'algorithme supprimera ces deux observations de la base de données.

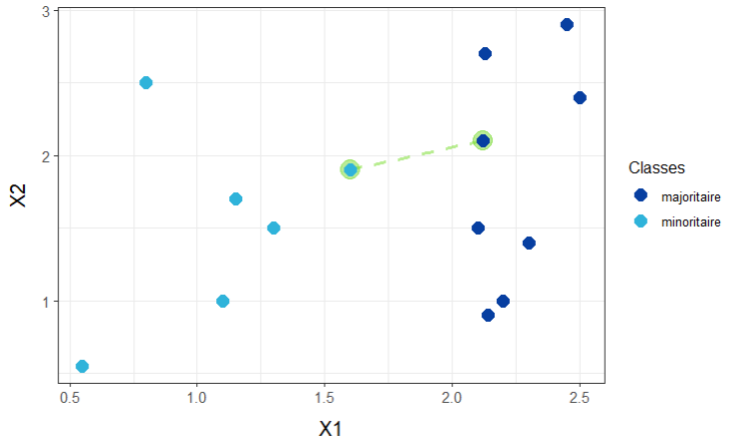

Figure 1.8 L'algorithme *TomekLink*

1.2.3.2 L'algorithme *Edited Nearest Neighbour (ENN)*

L'algorithme *ENN* consiste à supprimer les instances qui sont dans la classe sous-représentée par leurs plus proches voisins. La méthode se base sur l'idée que ces données nuisent à l'apprentissage, soit en étant mal classifiées ou en venant ajouter du bruit à l'échantillon. Pour identifier ces données, l'algorithme calcule les N plus proches voisins de chacune des instances de la classe majoritaire. Ensuite, la classe de la donnée est comparée avec celle qu'elle crée avec ses N plus proches voisins. Sur la figure 1.9, nous pouvons voir que le point d'intérêt et ses plus proches voisins forment un groupe comprenant principalement des instances de la classe minoritaire. Alors, puisque ce point fait partie de la classe sous-représentée, l'algorithme *ENN* le supprimera de l'échantillon. Un des avantages d'*ENN* est la réduction du bruit présent dans l'échantillon. Il permet aussi de réduire le surajustement des modèles de classifications et de préserver leurs frontières de décision.

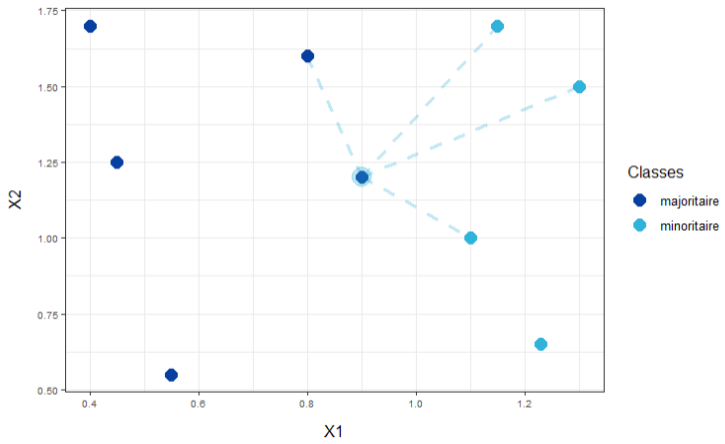

Figure 1.9 L'algorithme *ENN*

1.2.4 Les méthodes hybrides

Les méthodes hybrides consistent à combiner le sur-échantillonnage et le sous-échantillonnage, typiquement de nettoyage, afin d'améliorer la capacité de classification des modèles. Par exemple, l'algorithme *SMOTE-Tomek* utilise le sur-échantillonnage de données synthétiques *SMOTE*, et applique par la suite la méthode de nettoyage *TomekLink*. De la même façon, l'algorithme *SMOTE-ENN* utilise *SMOTE* pour rebalancer la base de données. Cependant, c'est plutôt la méthode *ENN* qui est ensuite appliquée comme technique de nettoyage.

1.3 Métriques de performance

Afin de pouvoir comparer les différentes techniques d'échantillonnage et ainsi sélectionner celle qui maximise la qualité de segmentation des modèles, nous devons choisir une métrique qui nous permettra d'évaluer la performance de ceux-ci. Ces mesures peuvent être divisées en deux groupes : les métriques d'évaluation singulière et les mesures de classement.

1.3.1 Métriques d'évaluation singulière

Afin de calculer les métriques d'évaluation singulière, nous devons tout d'abord assigner les instances dans la classe minoritaire ou majoritaire. Pour ce faire, nous devons fixer un certain seuil afin de délimiter les classes. Typiquement, ce seuil est fixé à 50%. Les données ayant une prédiction au-dessus de ce seuil sont

considérées comme des instances positives, et celles en dessous sont alors considérées comme des instances négatives. Ces métriques sont typiquement utilisées lorsque nous voulons minimiser le nombre d'erreurs. Elles sont donc surtout utiles pour des problèmes de classification. Dans cette famille, certaines de ces mesures sont plus appropriées pour les bases de données balancées ou débalancées, pour les signaux ou la détection de fraude, ou pour les tâches de récupération d'information (Ferri *et al.*, 2009). L'exactitude et le F-score sont des exemples de métriques d'évaluation singulière.

1.3.1.1 Exactitude

Comme nous l'avons déjà vu plus haut, l'exactitude E représente le pourcentage d'instances positives et négatives correctement classées. Elle se définit comme étant :

$$E = \frac{VP + VN}{VP + FP + VN + FN}$$

VP représente le nombre d'instances positives correctement classées, FP le nombre d'instances positives incorrectement classées, VN le nombre d'instances négatives correctement classées et FN le nombre d'instances négatives incorrectement classées. Cette métrique est largement utilisée pour mesurer la performance des modèles de classification. Cependant, le principal problème associé à l'exactitude est sa dépendance par rapport à la distribution des classes positives et négatives de la base de données (Ferri *et al.*, 2009). Lorsque les probabilités des classes sont très différentes, cette mesure peut être trompeuse. Elle n'est donc pas bien adaptée aux problèmes d'apprentissage de bases de données débalancées. Dans ce contexte, le F-score peut être un meilleur choix.

1.3.1.2 F-score

Le F-score est un autre exemple de métriques d'évaluation singulière. Cette mesure peut être interprétée comme étant une moyenne pondérée de la précision et du rappel. Parmi toutes les instances classées positives par le modèle, la précision représente la fraction de celles-ci correctement classifiées, et elle se définit comme suit :

$$P = \frac{VP}{VP + FP}$$

De son côté, le rappel représente plutôt la fraction des vraies instances positives correctement classifiées. Cette mesure est équivalente au taux de vrais positifs de la courbe ROC. Elle se définit alors par la formule suivante :

$$R = \frac{VP}{VP + FN}$$

Alors, en considérant une certaine précision P et un certain rappel R, le F-score F se définit comme étant :

$$F = 2 \cdot \left(\frac{P \cdot R}{P + R} \right)$$

Lors du calcul de ces métriques, la distance entre le seuil et les prédictions n'a aucune influence sur la mesure de performance du modèle. Comme il a été mentionné plus tôt, ces mesures sont utiles dans les problèmes de classification où nous nous intéressons au nombre d'erreurs. Cependant, dans un contexte de tarification en assurance, nous recherchons plutôt à optimiser la segmentation des assurés, et ainsi modéliser leur classe de risque. Dans ce contexte, les mesures de classement sont mieux adaptées.

1.3.2 Mesures de classement

Similairement aux métriques d'évaluation singulière, les mesures de classement ne dépendent pas de la valeur des prédictions. Elles évaluent plutôt l'ordre dans lequel les instances sont classées. Ces métriques mesurent comment les instances positives sont bien ordonnées par rapport aux instances négatives et peuvent être vues comme le sommaire de la performance du modèle pour toutes les différentes valeurs de seuil possibles. Ces mesures sont importantes dans plusieurs applications, comme la détection de fraude et la lutte anti-pourriel (Ferri *et al.*, 2009).

La mesure de classement la plus populaire est sans doute la fonction d'efficacité du récepteur, communément appelée la courbe ROC (Receiver operating characteristic).

1.3.2.1 Courbe ROC

La courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs, et ce pour toutes les différentes valeurs de seuil possibles. Si nous regardons la figure 1.10, le taux de vrais positifs est représenté par l'axe y, le taux de vrais positifs par l'axe x, et les différentes valeurs de seuils par les variations de bleu. La ligne diagonale pointillée représente la courbe ROC moyenne que nous pouvons obtenir par une classification complètement aléatoire. Lorsque nous travaillons sur un problème de classification, notre objectif est de maximiser le taux de vrais positifs, et ainsi s'éloigner de la courbe pointillée pour s'approcher le plus possible de la courbe noire. Cette courbe représente le modèle parfait, affichant un taux de vrais positifs de 100% pour toutes valeurs de seuil.

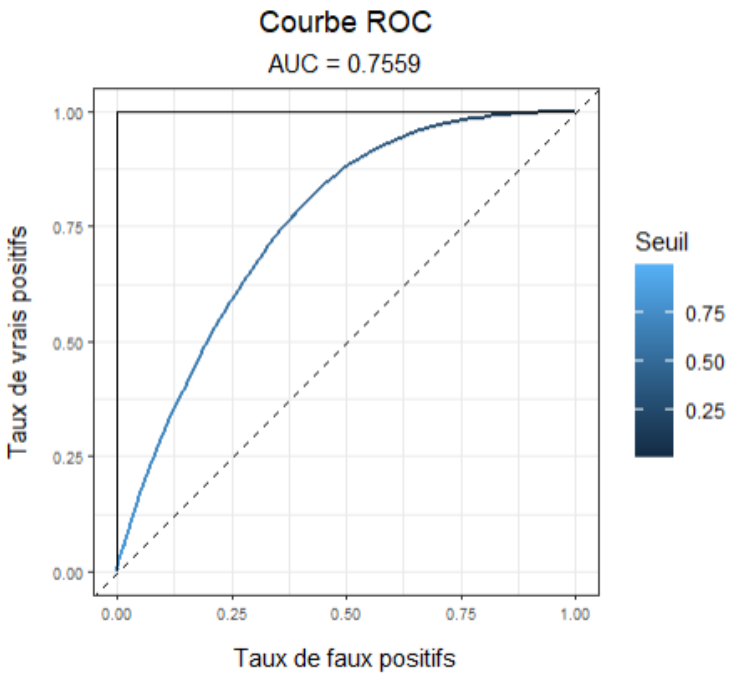

Figure 1.10 Exemple d'une courbe ROC

Afin d'évaluer la qualité d'un modèle, nous devons calculer l'aire sous la courbe ROC, communément appelée l'AUC (Area Under the Curve), et ainsi comparer ces résultats pour différents modèles. L'AUC peut être interprétée comme étant la probabilité qu'une instance de la classe minoritaire sélectionnée aléatoirement ait une prédiction plus élevée qu'une instance de la classe majoritaire (Baesens *et al.*, 2021). Alors, une AUC plus élevée indique une meilleure performance du modèle. Sur la figure 1.10, le modèle présente une AUC de 75,59%, située entre une AUC de 100% pour le modèle parfait et de 50% pour le modèle aléatoire.

Cependant, si nous travaillons avec une base de données débalancée, la courbe ROC et l'AUC peuvent être trop optimistes et surévaluer la qualité de la classification de l'algorithme. Ceci s'explique par le fait que le nombre d'instances de la classe majoritaire est beaucoup plus élevé que le nombre d'instances de la classe minoritaire. Alors, le nombre de vrais négatifs est typiquement beaucoup plus élevé que le nombre de faux positifs. Conséquemment, une importante augmentation ou diminution du nombre de faux positifs n'aura pratiquement aucun impact sur le taux de faux positifs de la courbe ROC. De ce fait, cette mesure n'est pas bien adaptée pour un problème présentant un fort débalancement de la variable d'intérêt. Alors, si nous travaillons avec une base de données largement débalancée, il est préférable d'explorer l'utilisation d'autres métriques. Par exemple, la courbe de précision-rappel (courbe PR) permet d'obtenir une meilleure représentation de la performance d'un algorithme en présence de données débalancées (Baesens *et al.*, 2021).

1.3.2.2 Courbe de précision-rappel

Comme nous pouvons le voir sur la figure 1.11, la courbe PR représente la précision en fonction du rappel. De la même manière que la courbe ROC, ces valeurs sont calculées pour les différentes valeurs de seuil. La mesure servant à comparer les modèles est l'aire sous cette courbe, l'AUPRC. Le modèle représenté par la figure 1.11 affiche une AUPRC de 70,94%. L'objectif d'un modèle de classification est de maximiser la précision pour les différentes valeurs de seuil. De ce fait, le modèle parfait, représenté par la courbe noire, affiche une précision de 100% pour chacune des différentes valeurs de seuil. Cela équivaut à une AUPRC de 100%. Comme la précision compare le nombre de faux positifs au nombre de vrais positifs, cette mesure est plus sensible au débalancement de la variable d'intérêt, et donc mieux adaptée pour mettre en lumière la différence de performance des différents modèles (Baesens *et al.*, 2021).

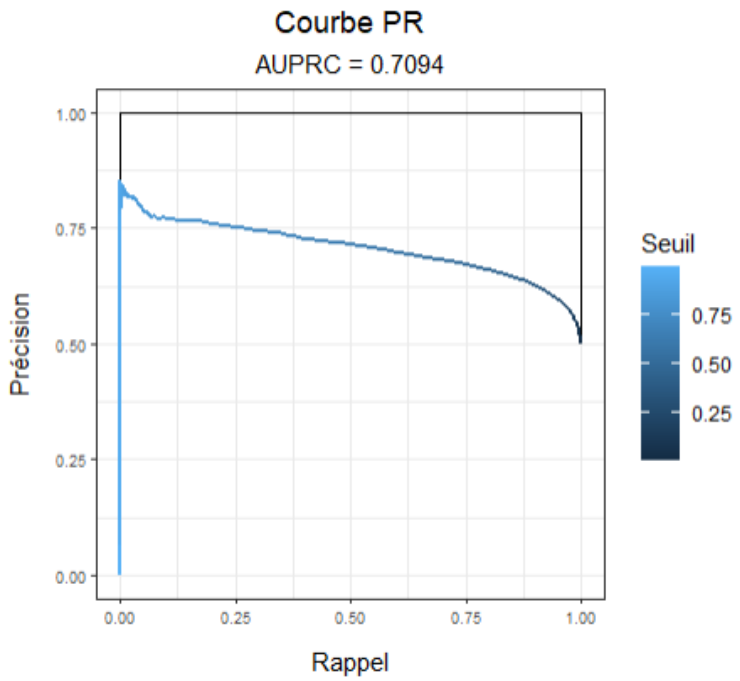

Figure 1.11 Exemple d'une courbe de précision-rappel (PR)

Dans (Jeni *et al.*, 2013), les auteurs ont testé différentes métriques, à la fois sur une base de données asymétrique et sur un échantillon balancé. Selon les résultats de leur recherche, presque toutes les métriques testées, incluant l'aire sous la courbe PR, indiquent une meilleure performance des algorithmes avec une base de données balancée qu'avec un échantillon ayant une distribution de la variable d'intérêt asymétrique. La seule exception est l'AUC ; la courbe ROC est très peu affectée par l'asymétrie de la distribution de la variable d'intérêt. Leurs résultats suggèrent que la courbe ROC peut masquer la piètre performance des modèles lorsque nous sommes en présence de données débalancées.

En faisant le même genre d'expérimentation, nous pouvons observer des résultats similaires. Sur la figure 1.12, nous pouvons voir que, lorsque nous passons d'un ratio de 1 :10 à un ratio de 1 :1 entre les classes de la variable d'intérêt, l'AUC de la courbe ROC reste à peu près la même, passant de 75,23% à 75,59%. Cela représente une augmentation de 0,36% seulement. Cependant, lorsque nous observons les résultats obtenus avec la courbe PR, nous pouvons observer une augmentation de l'AUPRC de 50,35%, celle-ci passant de 20,59% à 70,94%. Ceci représente une augmentation très importante, suggérant que la performance du modèle est beaucoup plus élevée lorsque nous travaillons sur une base de données rebalancée. De la même façon que les résultats dans l'article (Jeni *et al.*, 2013), nous pouvons voir que la courbe ROC n'est

pratiquement pas affectée par l'asymétrie de la base de données.

À la lumière de ces résultats, nous croyons que l'utilisation de l'aire sous la courbe ROC comme métrique de performance n'est pas suffisante pour évaluer la qualité de segmentation de nos modèles. Comme nous travaillons avec une base de données dont la distribution de la variable d'intérêt est asymétrique, nous allons aussi inclure l'AUPRC dans les résultats de notre analyse.

1.4 Comparaison des méthodes d'échantillonage

1.4.1 Étape 1 : Préparation des données

Afin de maximiser la performance de nos modèles, nous devons choisir la techniques d'échantillonnage optimale en vue du rebalancement de la base de données. Comme la plupart des algorithmes ne fonctionnent qu'avec les variables numériques, la première étape consiste à convertir les variables catégorielles de notre base de données en variables numériques. Pour ce faire, nous avons utilisé l'algorithme *target_encode* de la librairie *dataPreparation* dans R.

En premier lieu, cet algorithme calcule la moyenne de la variable d'intérêt pour chacune des classes de la variable catégorielle. Par exemple, pour la variable de genre, une fréquence de réclamations sera calculée pour l'ensemble des femmes, et une autre pour l'ensemble des hommes. Ensuite, cette valeur numérique sera assignée à chacune des observations de la base de données selon la classe à laquelle elle appartient. Si une observation est classée dans la catégorie des femmes, ce sera donc la moyenne de réclamation de l'ensemble des femmes de l'échantillon qui lui sera assignée. Comme mentionné plus haut, cette méthode n'est pas requise avec la méthode *SMOTENC* car cet algorithme a été développé afin de fonctionner autant avec les variables numériques que catégorielles.

1.4.2 Étape 2 : Rebalancement de la base de données

Ensuite, la deuxième étape consiste à appliquer une méthode d'échantillonnage afin de rééquilibrer la base de données à une fréquence de réclamation de 50%. Sur la figure 1.13, nous pouvons voir sur l'axe x les différentes techniques qui ont été explorées. Pour une description détaillée de chacune de ces techniques, vous pouvez vous référer au début de ce chapitre, où toutes ces différentes méthodes ont été décrites. Une fois ces méthodes appliquées à l'échantillon initial, nous obtenons un échantillon différent pour chacune des différentes méthodes d'échantillonnage.

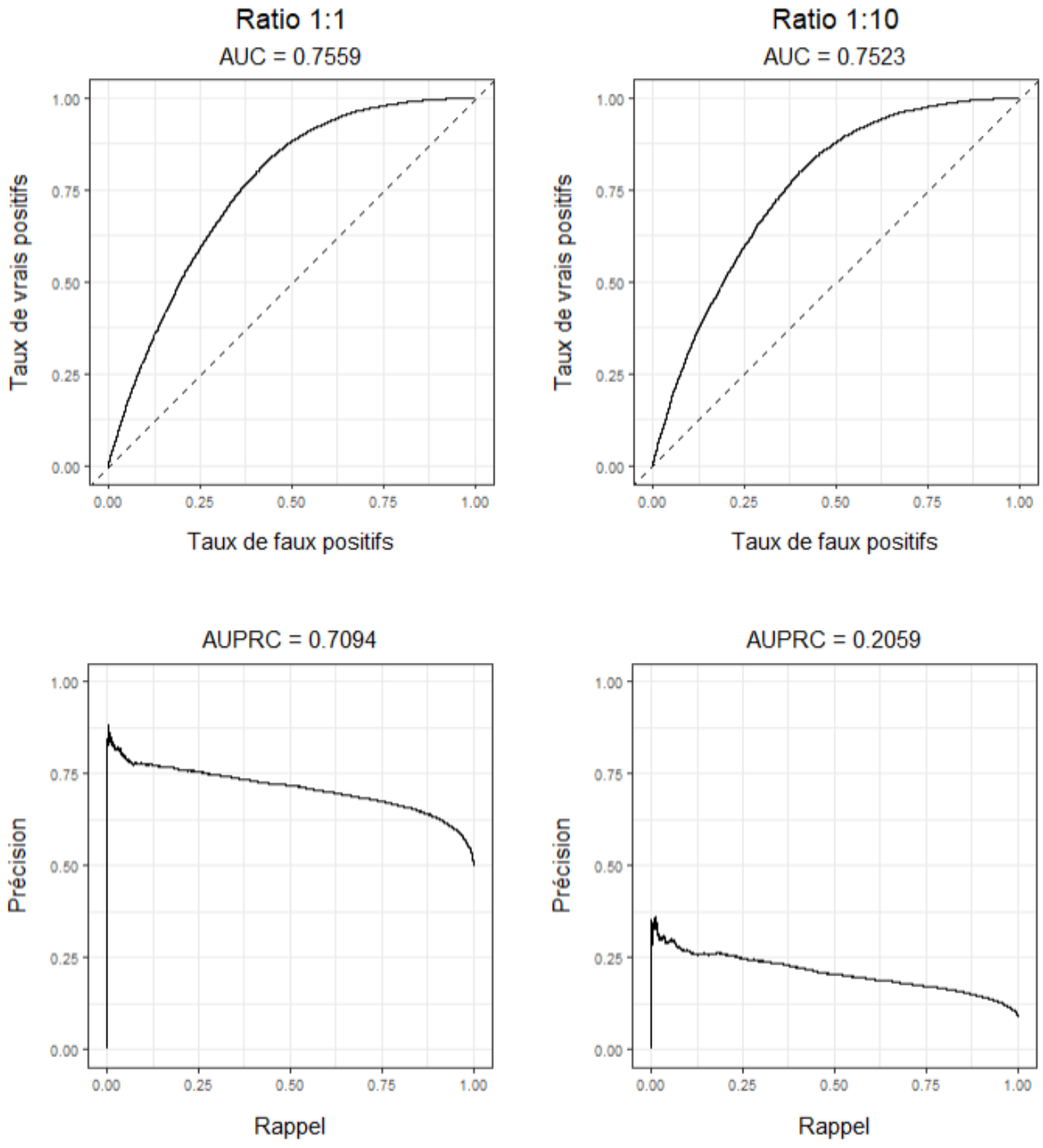

Figure 1.12 Comparaison des courbe ROC et PR pour différents ratios

1.4.3 Étape 3 : Calcul des métriques de performance

Une fois que nous avons obtenu nos différents échantillons rebalancés, la troisième étape consiste à calculer la performance du modèle de référence pour chacune de ces bases de données. Afin d'obtenir un écart-type et des intervalles de confiance de la performance du modèle, nous avons utilisé la méthode du bootstrap avec 1000 itérations.

Chaque itération consiste tout d'abord à créer une nouvelle base de données en sélectionnant aléatoirement, avec replacement, N observations de la base de données rebalancée. Ensuite, nous pouvons diviser ce nouvel échantillon en deux parties : les données d'entraînement et les données de validation. Parmi les 24 catégories de la variable « occupation », 12 d'entre elles seront utilisées dans la partie d'entraînement, et les autres serviront à valider le modèle. Une fois que nous avons ajusté le modèle de référence à partir des données d'entraînement, nous pouvons alors calculer les prédictions à l'aide des données de validation. À l'aide de ces prédictions, nous pouvons finalement calculer les courbes ROC et PR, et ainsi obtenir l'AUC et l'AUPRC du modèle.

1.4.4 Étape 4 : Analyse des résultats

Les résultats du bootstrap de l'AUC sont représentés par les boîtes à moustache de la figure 1.13. La première boîte à gauche représente la performance du modèle obtenue avec la base de données originale. Ensuite, juste à droite de celle-ci, nous pouvons observer les résultats obtenus par la méthode de sous-échantillonnage *NearMiss*. Nous pouvons observer que la médiane de l'AUC avec cette technique semble plus basse que celle résultant de l'échantillon original. Effectivement, si nous regardons la table 1.1, nous pouvons voir que la médiane des itérations est de 68,4% avec *NearMiss* contre 69,4% avec la base de données originale. De plus, nous pouvons observer que la performance du modèle présente une plus grande variabilité avec cette base de données. Nous pouvons donc conclure que, selon la courbe ROC, l'utilisation de la méthode de sous-échantillonnage *NearMiss* réduit la performance de notre modèle de référence.

Les autres boîtes à moustache de la figure 1.13 représentent les résultats des méthodes de sur-échantillonnage et des techniques hybrides. Pour chacune de ces méthodes, en plus d'observer une médiane de l'AUC plus élevée, la variabilité de la performance à chaque itération est plus basse, suggérant une meilleure stabilité des modèles.

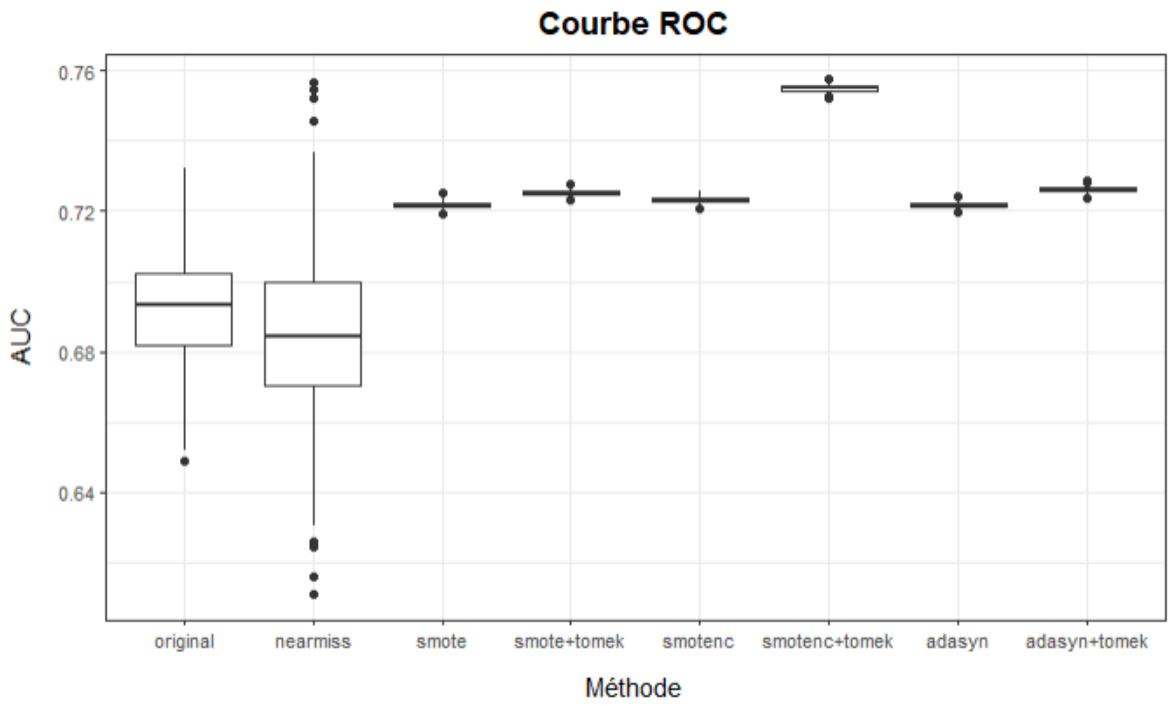

Figure 1.13 Bootstrap d'échantillonnage : courbe ROC

Sur la figure 1.14, si nous comparons les résultats de chacune des méthodes de sur-échantillonnage avec leur méthode hybride respective, nous pouvons voir que, dans tous les cas, l'application de l'algorithme *Tomek* améliore la performance des modèles. Parmi ces trois techniques hybrides, celle qui maximise l'AUC est SMOTENC+Tomek. À la lumière de ces résultats, selon la courbe ROC, la méthode d'échantillonnage qui maximise la performance du modèle de référence est la méthode hybride, combinant l'algorithme de sur-échantillonnage *SMOTENC* et l'algorithme de nettoyage *Tomek*.

Cependant, comme mentionné plus haut, la courbe ROC peut surestimer la performance des modèles lorsque nous travaillons avec une base de données débalancée. De ce fait, nous avons aussi calculé la performance des différentes méthodes d'échantillonnage selon la courbe PR.

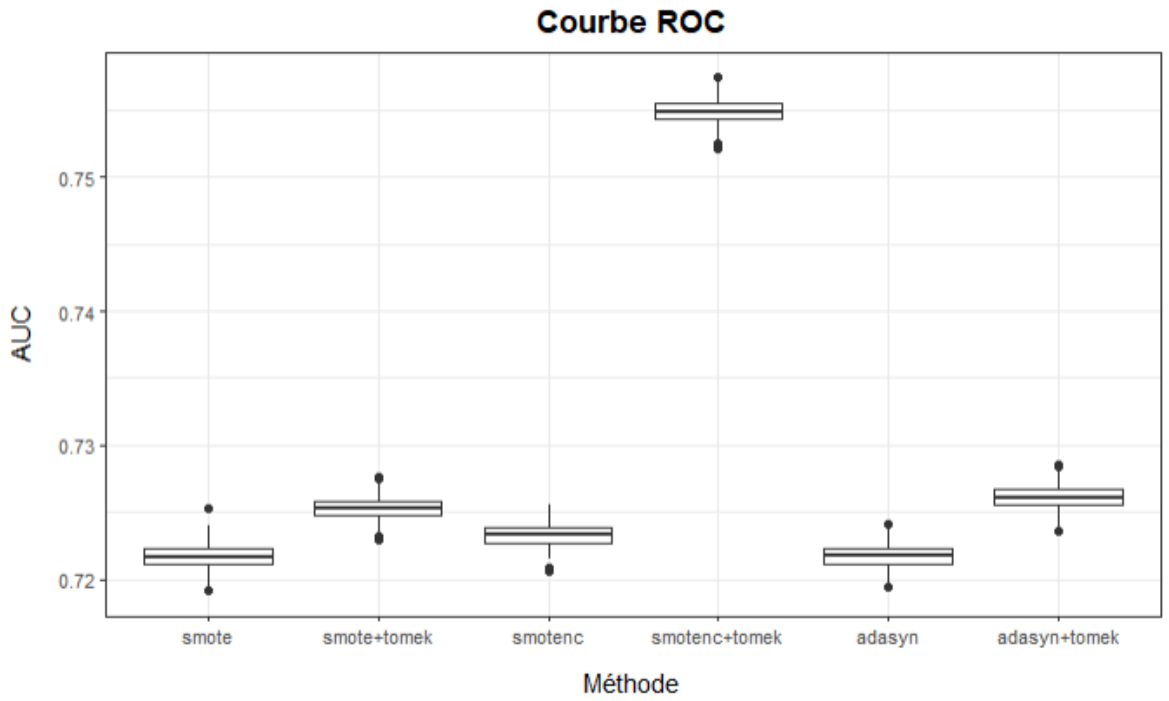

Figure 1.14 Bootstrap de sur-échantillonnage et hybrides : courbe ROC

Les résultats du bootstrap pour la courbe PR sont présentés sur la figure 1.15. Comme la fréquence des réclamations de la base de données originale est très basse, les résultats de l'AUPRC sont très près de zéro. Si nous comparons ces résultats avec les différentes bases de données ayant une fréquence rebalancée de 50%, nous pouvons voir que la courbe PR est très affectée par la basse fréquence. Aussi, si nous comparons les médianes de la figure 1.1, nous pouvons voir que, de la même façon qu'avec la courbe ROC, la médiane de l'AUPRC obtenue par la méthode de sous-échantillonnage *NearMiss* est plus basse que les autres, et présente une variabilité plus élevée. Nous pouvons donc conclure, avec cette métrique aussi, que l'utilisation de la méthode de sous-échantillonnage *NearMiss* ne maximise pas la performance de notre modèle de référence. Nous pouvons donc écarter cette méthode de notre analyse.

Si nous nous concentrons maintenant sur la figure 1.16, nous pouvons constater que, de la même manière qu'avec la courbe ROC, les méthodes hybrides présentent une meilleure performance que leur algorithme de sur-échantillonnage respectif. De plus, la technique qui maximise l'AUPRC est aussi la méthode hybride SMOTENC+TOMEK. Comme cette méthode est celle qui maximise à la fois l'AUC et l'AUPRC, c'est cette technique que nous allons utiliser pour rebalancer la distribution de la variable d'intérêt.

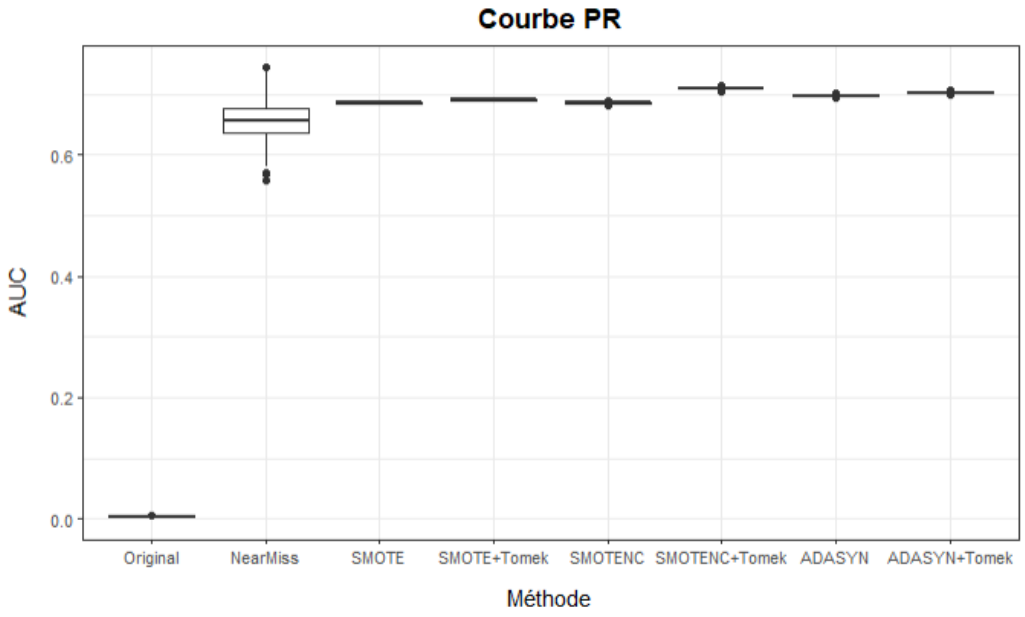

Figure 1.15 Bootstrap d'échantillonnage : courbe PR

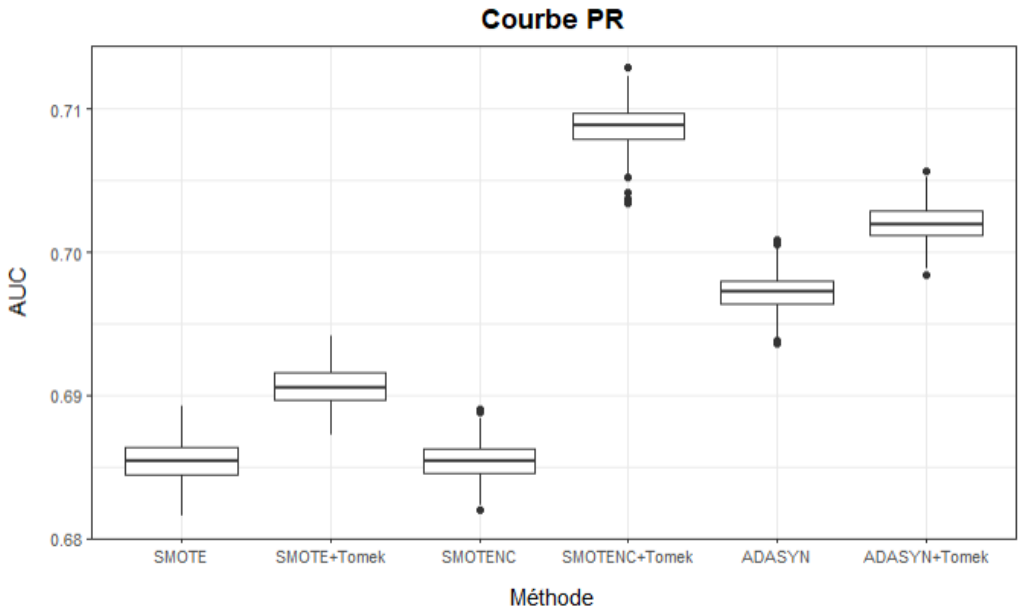

Figure 1.16 Bootstrap de sur-échantillonnage et hybrides : courbe PR

	Médiane	
Méthode	AUC	AUPRC
ORG	0,694	0,004
NM	0,684	0,657
SM	0,722	0,685
SMTK	0,725	0,691
SNC	0,723	0,685
SNCTK	0,755	0,709
AD	0,722	0,697
ADTK	0,726	0,702

Table 1.1 Médiane du bootstrap de l'échantillonnage

Maintenant que nous avons sélectionné la méthode d'échantillonnage optimale, nous pouvons nous demander si un rebalancement de la base de données à une fréquence de réclamation de 50% est optimal. Pour répondre à cette question, nous avons comparé la performance du modèle de référence ajusté à 20 bases de données différentes, ayant toutes des fréquences de réclamation variant entre 10% à 50%. D'après les résultats de la table 1.2, nous pouvons voir que la médiane de l'AUC reste assez stable pour les différents niveaux de fréquence. D'un autre côté, sur la figure 1.17, nous pouvons observer que la variabilité de la performance du modèle semble diminuer au fur et à la mesure que la fréquence augmente.

Pour ce qui est des résultats de l'AUPRC, nous pouvons voir sur la figure 1.18 et la table 1.2 qu'une augmentation de la fréquence fait aussi augmenter la médiane de la performance du modèle. De plus, la variabilité de la performance du modèle reste à peu près la même pour les différentes fréquences. En comparant les résultats des deux courbes, nous pouvons voir que la courbe PR semble beaucoup plus affectée par le débalancement des classes que la courbe ROC. Alors, nous pouvons conclure qu'il est plus judicieux d'utiliser la courbe PR pour sélectionner la fréquence optimale. Selon la table 1.2, une fréquence de 50% semble maximiser la performance du modèle de référence.

Pour conclure, nous avons sélectionné la méthode d'échantillonnage hybride combinant les algorithmes *SMOTENC* et *Tomek* afin de rebalancer la base de données à une fréquence de réclamation de 50% car c'est cette méthode qui maximise l'aire sous les courbes ROC et PR.

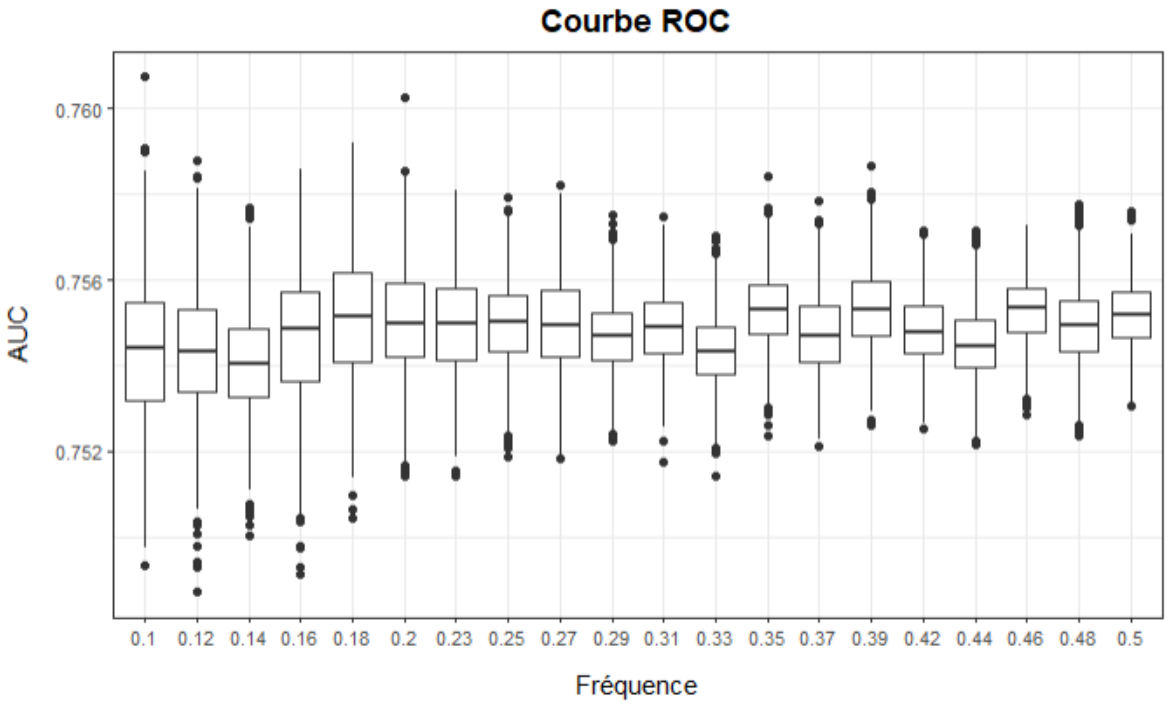

Figure 1.17 Bootstrap des fréquences : courbe ROC

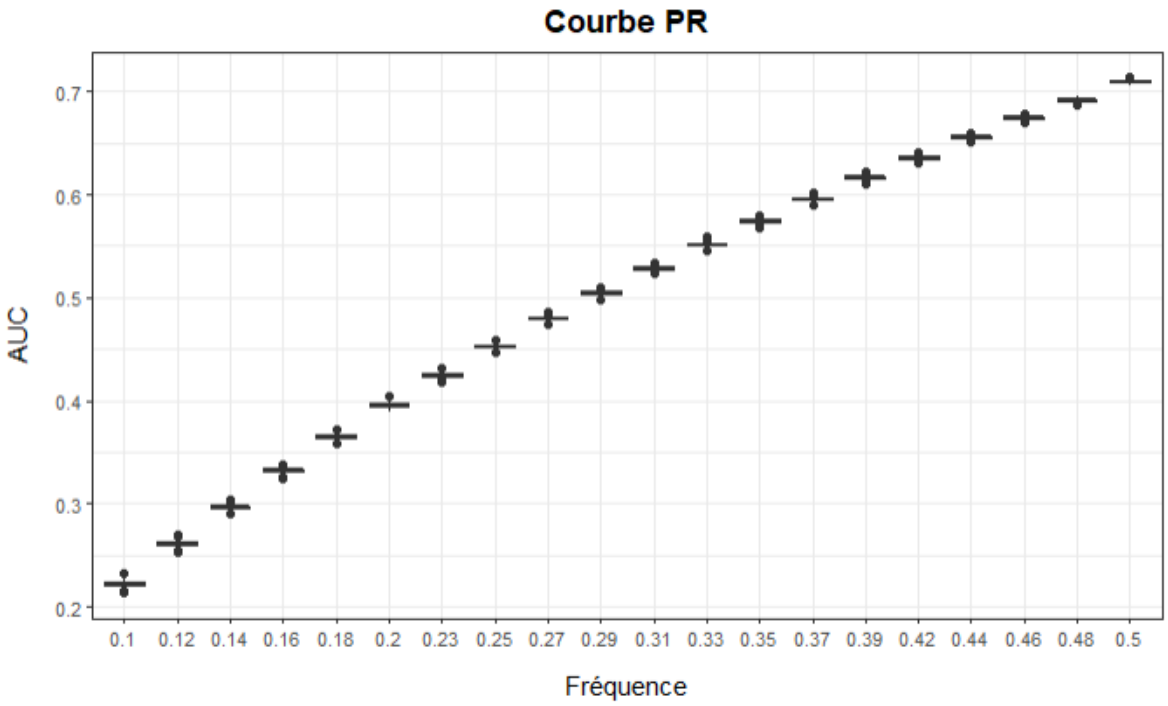

Figure 1.18 Bootstrap des fréquences : courbe PR

	Médiane			Médiane	
Fréquence	AUC	AUPRC	Fréquence	AUC	AUPRC
0,10	0,754	0,223	0,31	0,755	0,529
0,12	0,754	0,261	0,33	0,754	0,551
0,14	0,754	0,297	0,35	0,755	0,575
0,16	0,755	0,332	0,37	0,755	0,595
0,18	0,755	0,365	0,39	0,755	0,616
0,20	0,755	0,396	0,42	0,755	0,636
0,23	0,755	0,425	0,44	0,754	0,655
0,25	0,755	0,453	0,46	0,755	0,674
0,27	0,755	0,480	0,48	0,755	0,691
0,29	0,755	0,505	0,50	0,755	0,709

Table 1.2 Médiane du bootstrap des fréquences

CHAPITRE 2
ANALYSE DES DONNÉES REBALANCÉES

2.1 Comparaison des bases de données

Les méthodes d'échantillonnage consistent à modifier une base de données afin de rebalancer les différentes classes d'une variable. Dans notre cas, nous cherchons à ajuster la distribution de la variable d'intérêt de façon à réquilibrer la fréquence des réclamations. Nous avons donc élaboré une procédure d'échantillonnage adaptée à la base de données sur laquelle nous travaillons, qui a été présentée dans le chapitre précédent. Cette procédure est constituée de deux principales parties ; le suréchantillonnage *SMOTENC* et le nettoyage de données *TomekLink*.

Le but de ce chapitre est d'observer les données obtenues à chacune des étapes de la procédure de rebalancement, afin d'analyser comment ce rebalancement a modifié les données originalement contenues dans le portefeuille de l'assureur. En premier lieu, nous comparerons l'échantillon obtenu par la méthode de suréchantillonnage *SMOTENC* avec les observations originales. En deuxième lieu, nous comparerons ces données rebalancées avec l'échantillon généré par l'algorithme de nettoyage *TomekLink*. Dans les deux cas, nous ferons d'abord une analyse selon la base de données générale, et ensuite par rapport à chacune des covariables que nous inclurons dans nos modèles.

2.1.1 L'échantillon *SMOTENC*

2.1.1.1 Analyse générale

La première partie de la procédure consiste à appliquer l'algorithme de suréchantillonnage *SMOTENC* à la base de données originale afin de créer de nouvelles données synthétiques, et ainsi rééquilibrer les proportions des différentes classes de la variable d'intérêt. Sur la table 2.1, la colonne « Originale » présente l'état de la base de données de départ, avant qu'elle n'ait été modifiée par les différentes étapes de la procédure d'échantillonnage. Nous pouvons voir qu'originalement, sur un total de 421 462 observations, seulement 1176 d'entre elles affichaient une réclamation, résultant en une fréquence de seulement 0,28%. Tel que discuté précédemment, c'est ce débalancement qui, en premier lieu, nous a motivé à élaborer cette procédure de rebalancement.

La colonne « *SMOTENC* » présente, quant à elle, l'état de la base de données après l'application de l'algorithme *SMOTENC*. Le but de cet algorithme est de créer de nouvelles données dans la classe minoritaire de la variable d'intérêt, afin d'atteindre un certain ratio prédéfini. Nous avons fixé ce ratio à 1 :1 afin d'obtenir une fréquence de 50%. La colonne « # données sythétiques » répertorie les données synthétiques créées par le suréchantillonnage. Nous pouvons voir que l'algorithme a créé 419 110 nouvelles données synthétiques dans la classe minoritaire, atteignant ainsi un total de 420 286 réclamations, et une fréquence de 50%. Nous pouvons aussi observer qu'aucune donnée sans réclamation n'a été créée. Cela s'explique par le fait que l'algorithme *SMOTENC* se concentre sur la création d'instances minoritaires de façon à ce que cette classe atteigne la proportion désirée dans l'échantillon. Maintenant que la base de données a été rebalancée et que cette proportion est atteinte, nous obtenons un total de 840 572 observations dans la base de données.

Classe	Originale	*SMOTENC*	# données synthétiques
1 réclamation	1 176	420 286	419 110
0 réclamation	420 286	420 286	0
Total	421 462	840 572	419 110
Fréquence	0,28%	50%	-

Table 2.1 Comparaison des bases de données originale et *SMOTENC*

Tel que discuté précédemment, le but de notre analyse est d'intégrer l'avis des experts dans les algorithmes de tarification, et ce de façon à améliorer la segmentation des risques du portefeuille de l'assureur. Cette segmentation consiste à regrouper, selon certaines caractéristiques, les assurés représentant des risques similaires. Dans notre recherche, nous modéliserons la fréquence des réclamations selon deux différentes caractéristiques de risques : le secteur d'activité des entreprises, désignée par la covariable OCCUPATION, ainsi qu'une autre caractéristique de segmentation, désignée par la covariable X.

2.1.1.2 Analyse selon la covariable OCCUPATION

Maintenant que nous avons analysé la base de données de façon générale, observons maintenant les résultats du rebalancement selon la variable OCCUPATION. Sur la table 2.2, nous avons classé les 24 différentes

occupations selon leur importance en termes de nombre d'observations dans la base de données originale. La colonne « Occupation » représente les 24 différents secteurs d'activité des entreprises du portefeuille, et elles sont désignées par une lettre. Les deux colonnes sous « # observations » affichent la quantité d'instances dans les bases de données orignales, et dans l'échantillon obtenu par l'algorithme *SMOTENC*. Ensuite, les trois colonnes sous « # réclamations » répertorient le nombre de réclamations dans chaque échantillon, et la colonne « Différence » affiche le nombre de nouvelles données synthétiques créées par le suréchantillonnage. Finalement, les deux colonnes sous « Fréquence (%) » présentent la fréquence, en pourcentage, des réclamations des bases de données originale et rebalancée.

Si nous regardons la colonne du nombre d'instances originales, nous pouvons voir qu'il y a beaucoup plus de données pour l'occupation A que dans le reste des groupes. En comparant la quantité de données de ce groupe avec le nombre total d'instances, nous pouvons voir que cette occupation regroupe presque la moitié de l'échantillon original, avec une proportion de 49,36%. Une fois rebalancée, cette proportion augmente à 61,21%. Si nous analysons maintenant les plus petits groupes de l'échantillon, nous pouvons voir qu'il y en a 5 ayant moins de 1 000 instances, dont un comptant pas plus de 8 assurés.

Maintenant, si nous nous penchons sur la dernière ligne de la table 2.2, nous pouvons observer qu'au départ, il y avait 1 176 réclamations dans la base de données originale. L'algorithme de suréchantillonnage a créé 419 110 nouvelles données synthétiques, ce qui équivaut à une génération d'environ 356 instances pour chaque réclamation originale. Comme expliqué dans le chapitre précédent, lorsque l'algorithme *SMOTENC* crée une nouvelle donnée synthétique, il part d'une réclamation originale, calcule les N plus proches voisins, et assigne à cette nouvelle instance l'occupation la plus représentée parmi ces voisins. Cela veut dire que chacune des 1 176 réclamations originales génère environ 356 nouvelles réclamations synthétiques dans un même groupe d'occupation.

Dans la base de données de départ, les groupes P et Q ont originalement un nombre d'observations et de réclamations similaire. Cependant, aucune réclamation n'est générée dans le groupe P, alors que pour le groupe Q, il y en a 1069. En d'autres mots, pour 3 réclamations originales, les observations Q se situent majoritairement plus près que les observations des autres groupes, ce qui n'est jamais le cas pour les observations P.

Si nous regardons maintenant la figure 2.1, nous pouvons comparer le nombre de nouvelles données synthé-

	# observations		# réclamations			Fréquence (%)	
Occupation	Originale	*SMOTENC*	Originale	*SMOTENC*	Différence	Originale	*SMOTENC*
A	208 045	514 553	663	307 171	306 508	0,32	59,70
B	44 348	83 899	118	39 669	39 551	0,27	47,28
C	25 278	30 622	35	5 379	5 344	0,14	17,57
D	21 646	43 380	71	21 805	21 734	0,33	50,27
E	18 392	19 461	22	1 091	1 069	0,12	5,61
F	16 024	27 433	34	11 443	11 409	0,21	41,71
G	14 109	17 671	39	3 601	3 562	0,28	20,38
H	12 317	13 743	27	1 453	1 426	0,22	10,57
I	11 744	12 456	17	729	712	0,14	5,85
J	11 179	15 815	33	4 669	4 636	0,30	29,52
K	11 090	32 111	62	21 083	21 021	0,56	65,66
L	6 731	6 731	7	7	0	0,10	0,10
M	5 013	5 369	7	363	356	0,14	6,76
N	4 084	4 797	11	724	713	0,27	15,09
O	3 361	3 361	11	11	0	0,33	0,33
P	2 453	2 453	6	6	0	0,24	0,24
Q	2 156	3 225	2	1 071	1 069	0,09	33,21
R	1 258	1 258	5	5	0	0,40	0,40
S	1 141	1 141	4	4	0	0,35	0,35
T	469	469	0	0	0	0,00	0,00
U	279	279	1	1	0	0,36	0,36
V	218	218	0	0	0	0,00	0,00
W	119	119	1	1	0	0,84	0,84
X	8	8	0	0	0	0,00	0,00
Total	421 462	840 572	1 176	420 286	419 110	0,28%	50%

Table 2.2 Comparaison des bases de données originale et *SMOTENC* : covariable OCCUPATION

tiques par rapport à la quantité d'instances de la base de données originale. Nous observons que dans trois groupes, A, D et K, le nombre de nouvelles données synthétiques générées est plus élevé que le nombre d'instances originales. Sur la figure 2.2, nous avons limité l'ordonnée du graphique 2.1 de façon à faciliter l'analyse des plus petits groupes. Nous constatons qu'il n'y a pas présence de forte corrélation entre la proportion des groupes et le nombre de nouvelles données synthétiques. Certains groupes assez populeux n'ont que très peu de nouvelles réclamations, et vice versa. De plus, 10 groupes ne présentent aucune nouvelle observation. Ces groupes sont L, O, P, R, S, T, U, V, W et X. Selon la table 2.2, il s'agit principalement d'occupations ayant, à la base, très peu de réclamations.

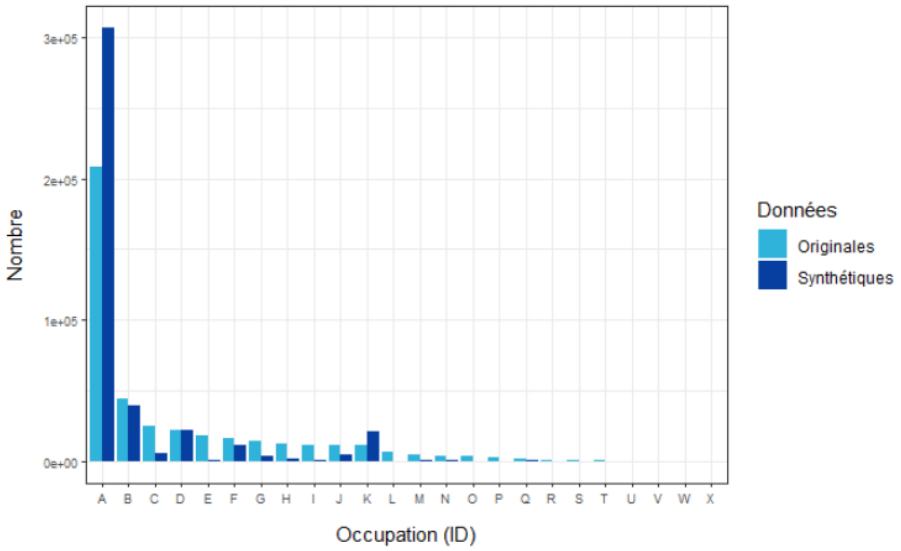

Figure 2.1 Nombre de données originales et synthétiques selon la covariable OCCUPATION

2.1.1.3 Analyse selon la covariable X

Maintenant que nous avons analysé la base de données selon la covariable OCCUPATION, observons maintenant les résultats du rebalancement selon la covariable X. Cette variable peut être interprétée comme une variable catégorielle ordonnée. Dans notre analyse, elle est utilisée comme variable numérique continue. De la même manière que la table 2.2, la table 2.3 compare les données originales avec la base de données obtenue par l'algorithme *SMOTENC*. Cependant, les observations sont cette fois-ci regroupées selon la covariable X.

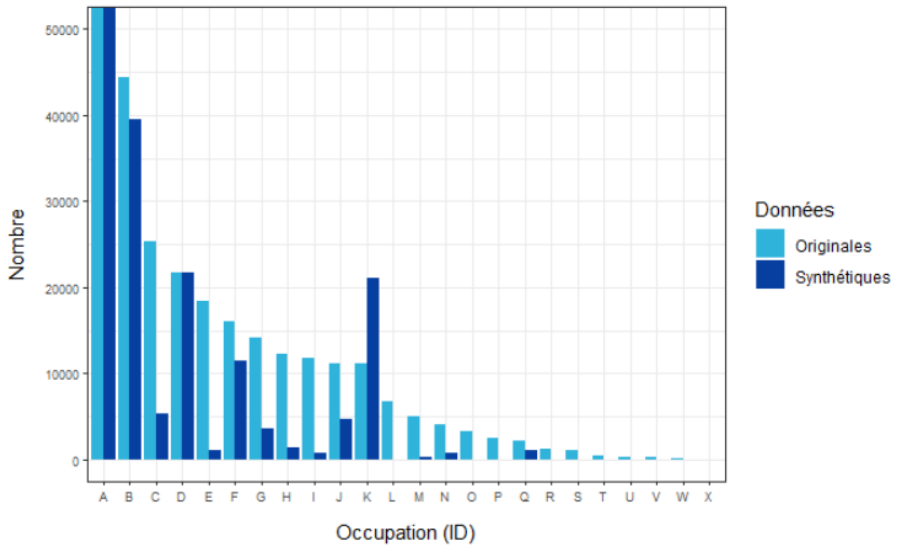

Figure 2.2 Zoom sur la figure 2.1

	# observations		# réclamations			Fréquence (%)	
X	Originale	*SMOTENC*	Originale	*SMOTENC*	Différence	Originale	*SMOTENC*
1	22 401	32 130	39	9 768	9 729	0,17	30,40
2	23 771	46 870	75	23 174	23 099	0,32	49,44
3	24 853	49 303	63	24 513	24 450	0,25	49,72
4	26 115	55 628	70	29 583	29 513	0,27	53,18
5	27 502	62 958	77	35 533	35 456	0,28	56,44
6	29 926	68 142	103	38 319	38 216	0,34	56,23
7	31 468	85 015	106	53 653	53 547	0,34	63,11
8	33 797	78 606	105	44 914	44 809	0,31	57,14
9	36 500	82 700	113	46 313	46 200	0,31	56,00
10	38 476	85 756	129	47 409	47 280	0,34	55,28
11	41 330	77 555	104	36 329	36 225	0,25	46,84
12	42 520	67 466	100	25 046	24 946	0,24	37,12
13	40 224	45 730	88	5 594	5 506	0,22	12,23
14	2 579	2 713	4	138	134	0,16	5,09
Total	421 462	840 572	1 176	420 286	419 110	0,28%	50%

Table 2.3 Comparaison des bases de données originale et *SMOTENC* : covariable X

Sur la table 2.3, si nous excluons la classe 14, nous pouvons remarquer une tendance à la hausse du nombre de polices émises. Effectivement, sur la figure 2.3, nous pouvons voir que le nombre d'observations originales, représentées par les bandes bleu pâle, augmente à partir du groupe 1 jusqu'au groupe 12, pour finalement redescendre légèrement dans la catégorie 13. Les bandes bleu foncé, quant à elles, représentent le nombre de données synthétiques générées par l'algorithme de suréchantillonnage. Le nombre de réclamations créées augmente à partir du groupe 1 jusqu'au groupe 10, pour finalement redescendre à partir du groupe 11. Nous pouvons voir une quantité de nouvelles données plus importante pour le groupe 7, avec 53 547 nouvelles observations. À l'opposé, cette quantité est particulièrement basse pour le groupe 13, avec seulement 5 506 nouvelles instances.

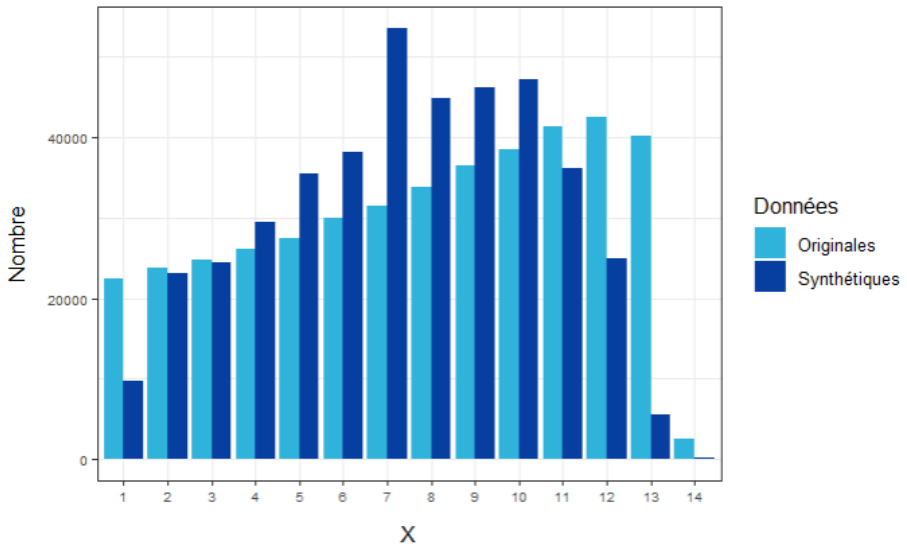

Figure 2.3 Nombre de données originales et synthétiques selon la covariable X

Les figures 2.4 et 2.5 représentent l'évolution de la fréquence de réclamations originale et après le suréchantillonnage. En les comparant, nous pouvons voir, par la graduation de leur ordonnée, que les fréquences se situent maintenant dans un intervalle plus large et plus élevé qu'originalement. De plus, la courbe de la figure 2.4 suggère une plus grande fluctuation des fréquences, alors que la courbe de la figure 2.5 semble plus lisse. Sur la figure 2.5, nous pouvons observer une certaine stabilité à partir du groupe 2 jusqu'au groupe 10. À l'exception du groupe 7, qui présente une fréquence soudainement plus élevée, les variations de fréquence restent petites. À partir du groupe 11, nous pouvons voir une baisse assez marquée des fréquences de réclamations.

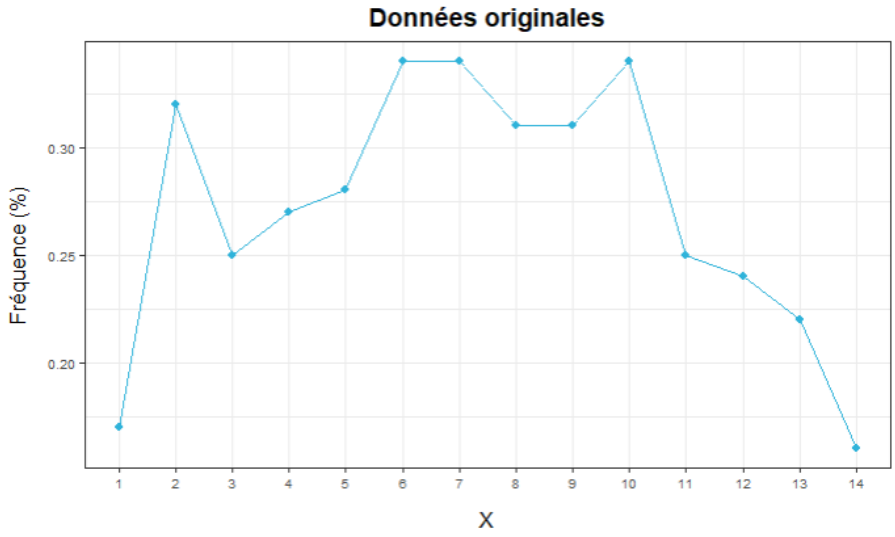

Figure 2.4 Évolution de la fréquence originale selon la variable X

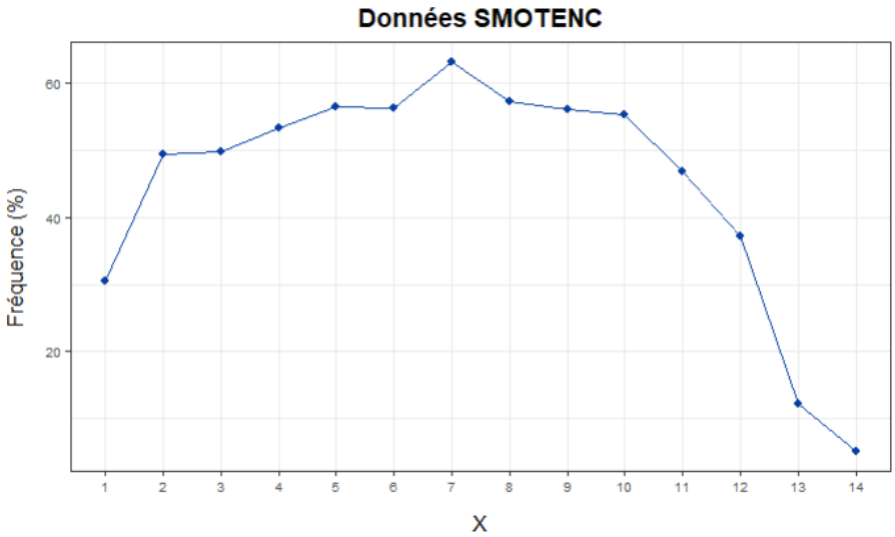

Figure 2.5 Évolution de la fréquence *SMOTENC* selon la variable X

2.1.2 L'échantillon *TomekLink*

Maintenant que la distribution des classes de la variable d'intérêt est rééquilibrée, nous appliquons la deuxième partie de la procédure de rebalancement : l'algorithme de nettoyage *TomekLink*. Cette méthode permet de retirer les instances qui réduisent la performance des modèles. Analysons maintenant l'effet que cet algorithme a eu sur la base de données.

2.1.2.1 Analyse générale

Sur la table 2.4, nous pouvons voir que le nombre d'instances écartées de la base de données est le même pour chacune des classes. Cela s'explique par le fait que ce sont les liens *Tomek* formés par des instances de classes différentes qui sont retirés de l'échantillon. Dans la base de données rebalancée, nous pouvons voir qu'il y avait 772 liens *Tomek*, donc 772 observations sans réclamation ont été retirées, ainsi que 772 observations avec une réclamation. De ce fait, la fréquence initiale de 50% est maintenue. Sur un total de 840 572 observations, 1 544 d'entre elles ont été retirées, ce qui représente 0,18% des instances de l'échantillon. Nous obtenons alors une base de données ayant un total de 839 028 observations.

# réclamations	*SMOTENC*	*TomekLink*	# données supprimées
1	420 286	419 514	772
0	420 286	419 514	772
Total	840 572	839 028	1 544
Fréquence	50%	50%	-

Table 2.4 Comparaison des bases de données *SMOTENC* et *TomekLink*

2.1.2.2 Analyse selon la covariable OCCUPATION

De la même façon que la table 2.4, les 24 différentes occupations des tables 2.5 et 2.6 ont été classées selon leur importance en termes de nombre d'observations. Comme le portefeuille a été rebalancé, certaines catégories d'occupation ont été repopulées de façon plus importante que d'autres, ce qui explique le changement d'ordre des différents groupes. Sur ces tables, nous comparons, d'une part, la base de données obtenue en première partie par l'algorithme *SMOTENC*, et d'autre part, l'échantillon généré par l'algorithme de nettoyage *TomekLink*.

Les premières colonnes de la table 2.5, sous « # observations », affichent la quantité d'instances dans les bases de données avant et après l'application de l'algorithme *TomekLink*. La colonne « Différence » représente le nombre d'observations de chaque groupe qui ont été retirées. Ensuite, les trois colonnes sous « # réclamations » répertorient le nombre initial de réclamations, le nombre restant après le nettoyage, ainsi que la quantité d'instances positives qui ont été écartées de l'échantillon. Sur la table 2.6, les deux colonnes sous « Fréquence (%) » présentent la fréquence, en pourcentage, de chaque groupe avant et après le nettoyage.

Sur cette table, nous pouvons observer que, à l'exception de l'occupation T, des réclamations ont été retirées dans tous les groupes de la base de données, mis à part ceux n'ayant initialement aucune instance positive. De plus, nous pouvons constater que la seule réclamation de l'occupation W a été écartée par l'agorithme. Sur la table 2.6, nous pouvons voir que la fréquence de ce groupe passe alors de 0,84% à zéro. Le nombre de groupes ayant une fréquence nulle passe donc de trois à quatre.

Le nombre d'instances retirées dans chaque groupe ne représente pas nécessairement le double du nombre de réclamations écartées par l'algorithme. Ceci s'explique par le fait qu'un lien *Tomek* n'est pas obligatoirement créé par des observations de la même occupation. Par exemple, on peut voir sur la table 2.5 que, dans le groupe T, une seule instance a été retirée, et comme aucune réclamation de ce groupe n'a été écartée, il s'agit donc d'une observation sans réclamation. Cela signifie que cette observation a créé un lien Tomek avec une réclamation d'un autre groupe de la base de données.

Finalement, la figure 2.6 compare les fréquences de réclamations des différentes occupations, et ce avant et après le nettoyage. Les lettres sur l'abcisse correspondent aux identifiants des tables 2.5 et 2.6. Nous pouvons observer qu'en général, la technique de nettoyage *TomekLink* n'a pas eu beaucoup d'impact sur la fréquence des différents groupes. Celle-ci a été principalement affectée dans les plus petits groupes, où la fréquence était déjà relativement basse. Dans les cas où le nombre d'observations est assez bas, le retrait d'une seule réclamation peut avoir un grand impact sur la fréquence.

2.1.2.3 Analyse selon la covariable X

La table 2.7 répertorie les observations et les réclamations des bases de données *SMOTENC* et *TomekLink* selon la covariable X, alors que la table 2.8 compare leur fréquence de réclamations. Le groupe pour lequel il y a eu le nettoyage le plus important est le 10, où 153 instances ont été retirées. La fréquence de récla-

	# observations			# réclamations		
Occupation	*SMOTENC*	*TomekLink*	Différence	*SMOTENC*	*TomekLink*	Différence
A	514 553	513 650	903	307 171	306 702	469
B	83 899	83 751	148	39 669	39 591	78
D	43 380	43 303	77	21 805	21 761	44
K	32 111	32 055	56	21 083	21 052	31
C	30 622	30 554	68	5 379	5 363	16
F	27 433	27 392	41	11 443	11 421	22
E	19 461	19 424	37	1 091	1 079	12
G	17 671	17 619	52	3 601	3 575	26
J	15 815	15 777	38	4 669	4 650	19
H	13 743	13 708	35	1 453	1 438	15
I	12 456	12 437	19	729	720	9
L	6 731	6 718	13	7	2	5
M	5 369	5 360	9	363	359	4
N	4 797	4 784	13	724	718	6
O	3 361	3 344	17	11	5	6
Q	3 225	3 224	1	1 071	1 070	1
P	2 453	2 447	6	6	3	3
R	1 258	1 254	4	5	3	2
S	1 141	1 137	4	4	1	3
T	469	468	1	0	0	0
U	279	279	0	1	1	0
V	218	218	0	0	0	0
W	119	117	2	1	0	1
X	8	8	0	0	0	0
Total	978 548	977 016	1 532	489 274	488 508	766

Table 2.5 Comparaison des bases de données *SMOTENC* et *TomekLink* : covariable OCCUPATION

	Fréquence (%)			Fréquence (%)	
Occupation	*SMOTENC*	*TomekLink*	Occupation	*SMOTENC*	*TomekLink*
A	59,70	59,71	M	6,76	6,70
B	47,28	47,27	N	15,09	15,01
D	50,27	50,25	O	0,33	0,15
K	65,66	65,67	Q	33,21	33,19
C	17,57	17,55	P	0,24	0,12
F	41,71	41,69	R	0,40	0,24
E	5,61	5,55	S	0,35	0,09
G	20,38	20,29	T	0,00	0,00
J	29,52	29,47	U	0,36	0,36
H	10,57	10,49	V	0,00	0,00
I	5,85	5,79	W	0,84	0,00
L	0,10	0,03	X	0,00	0,00

Table 2.6 Comparaison des fréquences *SMOTENC* et *TomekLink* : covariable OCCUPATION

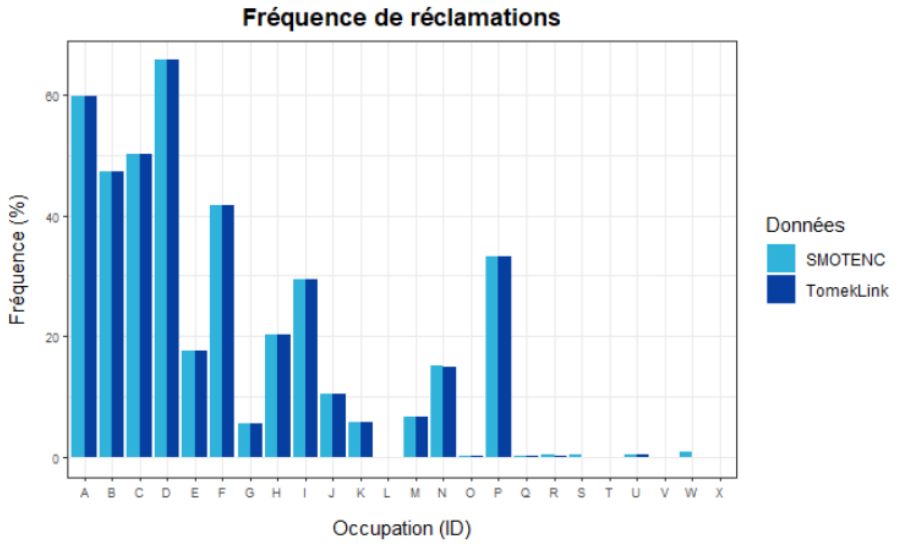

Figure 2.6 Comparaison des fréquence *SMOTENC* et *TomekLink* : covariable OCCUPATION

mations pour ce groupe a légèrement augmenté, passant de 55,28 % à 55,29 %. Le nombre le plus élevé de réclamations retirées est de 87, pour le groupe 6. Cela n'a pas eu d'impact sur la fréquence, celle-ci restée à 56,23 %. À l'opposé, c'est pour le groupe 14 que nous observons la plus petite quantité d'observations écartées, avec seulement 4 instances au total, dont seulement une représente une réclamation. Ce nettoyage a légèrement affecté la fréquence, passant de 5,09 % à 5,06 %. Finalement, de la même manière que pour la variable OCCUPATION, nous pouvons constater qu'en général, le nettoyage n'a pas eu beaucoup d'impact sur la fréquence de chacun des groupes.

	# observations			# réclamations		
X	*SMOTENC*	*TomekLink*	Différence	*SMOTENC*	*TomekLink*	Différence
1	32 130	32 076	54	9 768	9 750	18
2	46 870	46 763	107	23 174	23 122	52
3	49 303	49 211	92	24 513	24 468	45
4	55 628	55 510	118	29 583	29 523	60
5	62 958	62 843	115	35 533	35 471	62
6	68 142	67 996	146	38 319	38 232	87
7	85 015	84 882	133	53 653	53 592	61
8	78 606	78 472	134	44 914	44 849	65
9	82 700	82 556	144	46 313	46 243	70
10	85 756	85 603	153	47 409	47 328	81
11	77 555	77 436	119	36 329	36 272	57
12	67 466	67 342	124	25 046	24 986	60
13	45 730	45 629	101	5 594	5 541	53
14	2 713	2 709	4	138	137	1
Total	840 572	839 032	1 540	420 286	419 516	770

Table 2.7 Comparaison des bases de données *SMOTENC* et *TomekLink* : covariable X

2.2 Base de données finale

Une fois que nous avons rebalancé et nettoyé les données originales, nous obtenons la base de données finale à partir de laquelle nous allons modéliser les risques du portefeuille de l'assureur. Pour ce faire, nous avons séparé cette base de données en échantillon d'entraînement et de test. La base de données d'entraî-

	Fréquence (%)			Fréquence (%)	
X	*SMOTENC*	*TomekLink*	X	*SMOTENC*	*TomekLink*
1	30,40	30,40	8	57,14	57,15
2	49,44	49,45	9	56,00	56,01
3	49,72	49,72	10	55,28	55,29
4	53,18	53,19	11	46,84	46,84
5	56,44	56,44	12	37,12	37,10
6	56,23	56,23	13	12,23	12,14
7	63,11	63,14	14	5,09	5,06

Table 2.8 Comparaison des fréquences *SMOTENC* et *TomekLink* : covariable X

nement servira à ajuster le modèle de référence, celui auquel nous intégrerons l'avis des experts. L'échantillon de test, quant à lui, nous permettra d'estimer et de comparer la performance du modèle de référence avec ceux incluant l'élicitation des souscripteurs. Cela nous permettra d'évaluer si l'intégration de cette information dans notre modélisation améliore les prédictions des risques de nouvelles occupations jamais observées par l'assureur. Pour ce faire, nous devons donc diviser la base de données selon les différentes occupations afin d'en classer un certain nombre dans la base de données d'entraînement, et le reste dans la base de données de test. Les occupations de la base de données d'entraînement représenteront les secteurs d'activité déjà présents dans le portefeuille de l'assureur, et les occupations de la base de données de test représenteront les secteurs d'activité de nouveaux clients oeuvrant dans un domaine jamais observé par l'assureur.

Comme observé précédemment, l'occupation A est beaucoup plus importante que les autres, regroupant au total 60,96 % des observations de la base de données. Afin de créer les bases de données d'entraînement et de test de façon à obtenir des échantillons équilibrés, et ainsi maintenir une fréquence similaire entre les deux, nous avons d'abord divisé cette occupation de façon à obtenir deux plus petits groupes. Pour ce faire, nous avons utilisé la variable Y, une covariable dichotomique. A1 regroupe les entreprises faisant partie d'un des deux groupe de la variable Y, et le reste des assurés sont clasés dans le groupe A2. La table 2.9 répertorie le nombre d'observations, le nombre de réclamations, la proportion de chaque groupe dans la base de données, ainsi que leur fréquence. Nous pouvons voir que cette dernière est restée très près de la fréquence originale de 59,54 %, avec respectivement 59,71 % et 59,33 %.

Occupation	# observations	# réclamations	Proportion (%)	Fréquence (%)
A1	282 785	168 854	33,70	59,71
A2	228 724	135 700	27,26	59,33
Total	511 509	304 554	60,96	59,54

Table 2.9 Division de l'occupation A

Une fois que nous avons divisé ce groupe en deux, nous obtenons 25 différents niveaux d'occupation. Nous avons d'abord assigné aléatoirement chacun des groupes A1 et A2 de façon à ce qu'ils ne se retrouvent pas dans le même échantillon. Ensuite, 12 des 23 autres groupes ont été sélectionnés aléatoirement et assignés à la base de données d'entraînement, et les 11 autres groupes font partie de la base de données de test. Sur la table 2.10, nous pouvons voir les résultats de cette sélection, les occupations étant classées de façon décroissante selon leur fréquence de réclamations. La base de données d'entraînement contient un groupe de plus que la base de données de test.

Les figures 2.7 et 2.8 représentent la portion qu'occupe chacune des occupations des bases de données d'entraînement et de test. Dans les deux cas, l'occupation A regroupe au-dessus de la moitié des observations. Dans la base de données d'entraînement, le reste des instances est regroupé de façon plus équilibrée que dans la base de données de test, cette dernière contenant 4 des 5 groupes ayant moins de 1 000 observations.

La table 2.11 présente le nombre d'observations, le nombre de réclamations, la proportion, ainsi que la fréquence de réclamations des bases de données d'entraînement et de test. L'échantillon d'entraînement regroupe un plus grand nombre d'instances, représentant 75 % des observations, par rapport à l'échantillon de test qui regroupe seulement le quart de celles-ci. Leur fréquence, quant à elle, est restée plutôt stable, étant de 50,02 % pour la base de données d'entraînement, et 49,95 % pour la base de données de test. Bien que cette division a été faite selon une variable présentant des niveaux de fréquence très variés, leur fréquence globale est restée stable. C'est cette variation qui rend cette variable aussi intéressante pour la segmentation des risques, et l'élicitation des experts.

ENTRAÎNEMENT		TEST	
Occupation	Fréquence (%)	Occupation	Fréquence (%)
A2	59,28	K	65,67
F	41,69	A1	60,05
Q	33,19	D	50,25
J	29,47	B	47,27
C	17,55	G	20,29
N	15,01	I	5,79
H	10,49	U	0,36
M	6,70	R	0,24
E	5,55	L	0,03
O	0,15	X	0,00
P	0,12	V	0,00
S	0,09	T	0,00
W	0,00		
Total	50,02	Total	49,95

Table 2.10 Groupes d'occupation des bases de données d'entraînement et de test

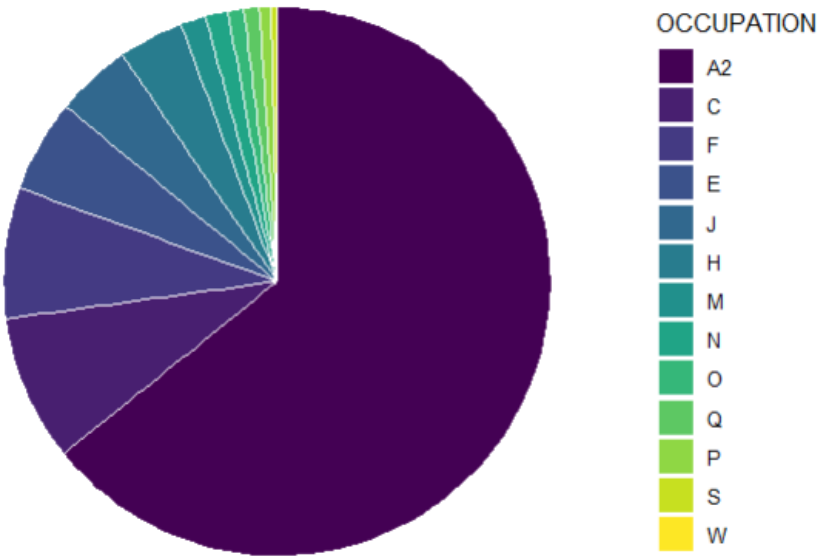

Figure 2.7 Proportion des types d'occupation de la base de données *Train*

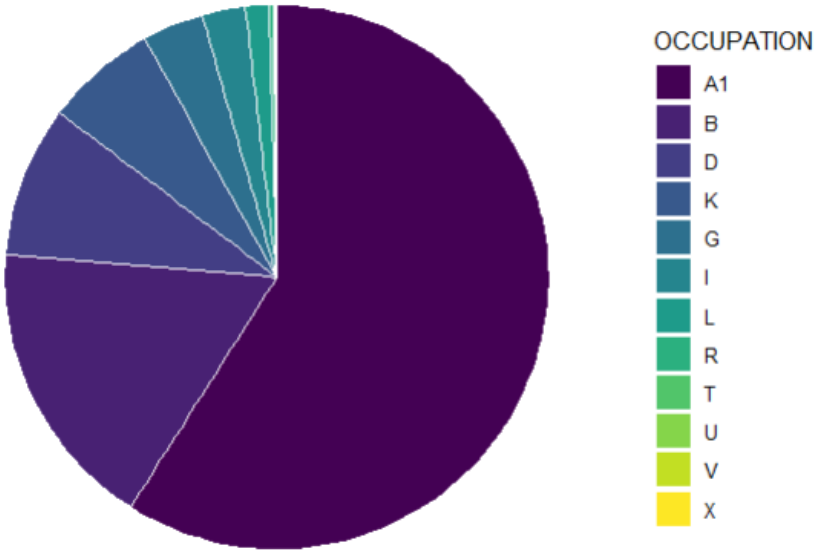

Figure 2.8 Proportion des types d'occupation de la base de données *Test*

Données	# observations	# réclamations	Proportion (%)	Fréquence (%)
TRAIN	629 274	314 745	75,00	50,02
TEST	209 758	104 771	25,00	49,95
Total	839 032	419 516	100,00	50,00

Table 2.11 Bases de données d'entraînement et de test

CHAPITRE 3
ÉLICITATION DE L'AVIS D'EXPERT

3.1 Modélisation en assurance

3.1.1 Modèles linéaires

Au 19^{e} siècle, les modèles linéaires initialement proposés par Legendre et Gauss étaient largement utilisés par les actuaires (Kassimi et Zahi, 2021). Cependant, certaines hypothèses de ces modèles ne sont pas compatibles avec le contexte de modélisation des risques de réclamations en assurance (David, 2015). C'est le cas notamment pour l'hypothèse de linéarité des prédicteurs, signifiant que la relation entre la variable d'intérêt et les covariables est linéaire. Nous pouvons aussi remettre en question l'hypothèse d'homoscédasticité, qui veut que la variance des erreurs stochastiques de la régression soit la même pour chaque observation.

Une autre hypothèse importante des modèles linéaires est que la variable d'intérêt suit une distribution gaussienne, cette dernière ayant un support dans les nombres réels. Cependant, les risques que nous nous retrouvons à modéliser en assurance ont fréquemment un support qui ne comprend pas tous les nombres réels. Comme par exemple, lorsque nous travaillons sur la sévérité des réclamations, il peut être préférable d'utiliser une distribution ayant un support dans les nombres réels positifs seulement. De la même façon, la modélisation de la fréquence des réclamations, notre sujet de recherche, peut imposer une hypothèse de distribution autre que gaussienne. Cela peut expliquer la popularité des modèles linéaires généralisés dans le domaine de l'assurance.

3.1.2 Modèles linéaires généralisés

La première utilisation des modèles linéaires généralisés dans le domaine des sciences actuarielles date de la fin du 20^{e} siècle (Kassimi et Zahi, 2021). Ces modèles, proposés par John Nelder et Robert Wedderburn, représentent une généralisation des modèles linéaires classiques. En effet, ces modèles vont au-delà de l'hypothèse de normalité, et permettent la modélisation de toute une gamme de distributions ; la famille exponentielle (Kassimi et Zahi, 2021). Cette famille regroupe, entre autres, les lois binomiales et de Poisson, ces dernières étant très utiles pour la modélisation de données dichotomiques et de comptage. De ce fait, les modèles linéaires généralisés représentent aujourd'hui une pratique statistique courante dans le

domaine de la tarification des risques en assurance non-vie (David, 2015).

Le sujet d'intérêt de notre recherche est la fréquence de réclamations de feu pour la structure, pour la ligne d'affaires commerciale. Nous considérons cette variable comme étant dichotomique, c'est-à-dire que les données ont été regroupées en deux classes. La première classe représente les polices sans réclamation, celle-ci s'opposant à la deuxième classe regroupant les instances ayant une réclamation. Autrement dit, nous voulons modéliser un événement ayant seulement deux réalisations possibles. Ce genre de problèmes peut être modélisé par une régression logistique, où l'hypothèse de distribution de la variable d'intérêt est la loi bernoulli, cette dernière faisant partie de la famille exponentielle.

La régression logistique fait partie des modèles linéaires généralisés. Ces modèles mettent en relation la variable d'intérêt et la régression linéaire par une fonction lien. Cela permet de modéliser une transformation de la moyenne de la variable d'intérêt, plutôt que sa moyenne directement. Dans le cas de la régression logistique, ce lien, la fonction logit, est défini par l'équation 3.1. Ce lien est représenté graphiquement sur la figure 3.1, où nous pouvons visualiser la relation entre la variable p et sa transformation logit.

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{3.1}$$

Une autre des hypothèses de la régression linéaire, celle-ci étant commune aux modèles linéaires généralisés, est l'indépendance des erreurs. Cela signifie que nous considérons que les observations de notre échantillon sont indépendantes les unes des autres. Cependant, cette hypothèse ne peut être vérifiée lorsque nous sommes en présence de données groupées. Lorsque nous travaillons avec une base de données présentant de la variabilité entre les groupes, ainsi que de la corrélation entre les observations à l'intérieur de ceux-ci, l'utilisation des modèles linéaires mixtes généralisés peut être considérée (Broström et Holmberg, 2011).

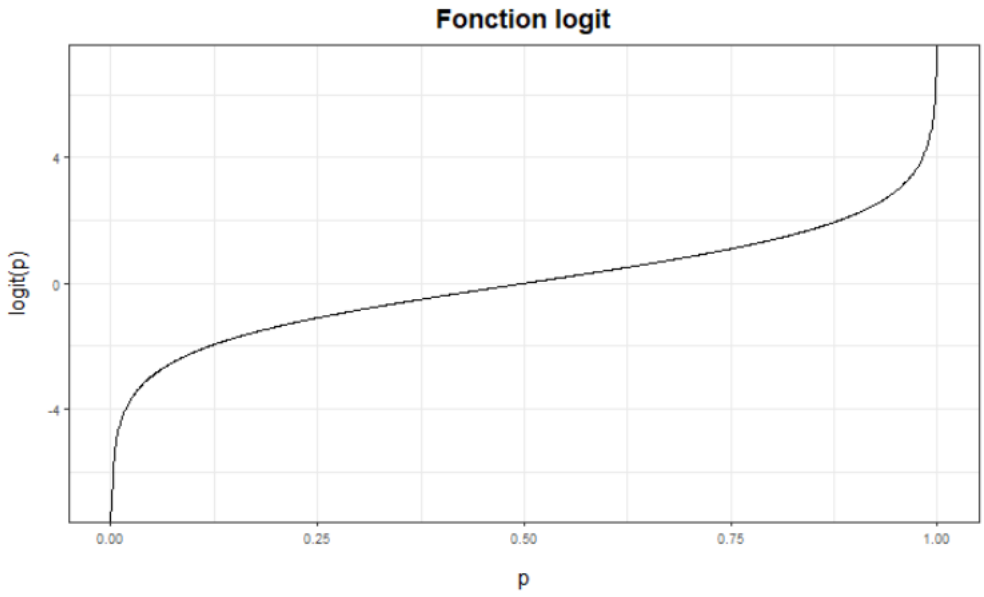

Figure 3.1 Représentation graphique de la fonction logit

3.1.3 Modèles linéaires mixtes généralisés

Les modèles à effets mixtes généralisés ou, plus simplement, les modèles mixtes généralisés sont des modèles statistiques qui intègrent à la fois des paramètres à effets fixes et des effets aléatoires (Bates, 2010). Il existe plusieurs situations pour lesquelles les modèles linéaires mixtes généralisés peuvent être un bon choix de modélisation. C'est le cas, par exemple, en présence de données longitudinales, où plusieurs observations d'un même sujet ont été recueillies à différents moments dans le temps.

Ce genre de modèles peut aussi être utilisé avec des données groupées. La structure d'un tel échantillon est formée de plusieurs groupes, ces derniers représentant les sujets de l'analyse. Chacun de ces sujets renferme un certain nombre varié d'observations, comme c'est le cas pour la base de données avec laquelle nous travaillons. Les différentes occupations du portefeuille de l'assureur représentent les sujets de l'analyse, et chacun de ces groupes renferme plusieurs différentes polices d'assurance. Ces dernières représentent tous les assurés d'une même occupation, et ces assurés forment un groupe de données au sein du portefeuille de l'assureur. Pour une base de données groupée comprenant j groupes, un modèle linéaire mixte généralisé peut être défini comme suit :

Définition 3.1

$$Y_j \mid \mathbf{u}_j \sim f_{Y_j|\mathbf{u}_j}(y_j \mid \mathbf{u}_j) \tag{3.2}$$

$$f_{Y_j|\mathbf{u}_j}(y_j \mid \mathbf{u}_j) = \exp\left\{[y_j\gamma_j - b(\gamma_j)]/\tau^2 - c(y_j, \tau)\right\} \tag{3.3}$$

$$E[Y_j \mid \mathbf{u}_j] = \mu_j \tag{3.4}$$

$$g(\mu_j) = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j \tag{3.5}$$

$$\mathbf{u}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \tag{3.6}$$

Le vecteur Y_j de l'équation 3.2 représente la variable d'intérêt du modèle. Ce vecteur, présenté ci-dessous, contient les n_j réponses de l'événement que nous modélisons pour le groupe j :

$$Y_j = \begin{pmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{n_j j} \end{pmatrix}$$

La base de données contient donc un vecteur Y_j pour chacun des j groupes de l'échantillon. Le premier élément du vecteur, Y_{1j}, représente la réponse pour la première observation du groupe. Il est à noter que chacun des groupes de la base de données peut contenir un nombre différent d'observations et que les n_j peuvent varier d'un groupe à l'autre.

La distribution de la variable réponse Y_j est conditionnelle aux effets aléatoires du modèle, contenu dans le vecteur $\mathbf{u}_j$. Comme il s'agit d'une généralisation des modèles linéaires mixtes, cette distribution conditionnelle n'est pas contrainte à la distribution gaussienne. Tel qu'illustré par l'équation 3.3, elle a comme seule contrainte de faire partie de la famille exponentielle. Les distributions Bernoulli, Poisson et Gamma sont des exemples de lois faisant partie de cette famille. Cependant, comme nous pouvons le voir sur les équations 3.4 et 3.5, cette généralisation implique une fonction lien $g(\mu_i)$ qui relie l'espérance conditionnelle μ_j de la variable d'intérêt Y_j et la prédiction linéaire du modèle.

Nous parlons de modèles à effets mixtes car cette prédiction linéaire, $\mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j$, contient à la fois des effets fixes et des effets aléatoires. Les effets fixes sont contenus dans le premier terme de l'équation. Le

premier élément de ce terme, $\mathbf{X}_j$, est une matrice de design de dimensions $n_j \times p$. Cette matrice contient les valeurs connues des p covariables associées aux effets fixes. Chacune des lignes de la matrice $\mathbf{X}_j$ ci-dessous représente le vecteur de covariables associé à chacune des n_j observation du groupe j, et chacune des colonnes est associée à une covariable. Il est à noter que si le modèle contient une ordonnée à l'origine, la première colonne de cette matrice ne sera composée que de 1.

$$\mathbf{X}_j = \begin{pmatrix} x_{1j}^{(1)} & x_{1j}^{(2)} & x_{1j}^{(3)} & \dots & x_{1j}^{(p)} \\ x_{2j}^{(1)} & x_{2j}^{(2)} & x_{2j}^{(3)} & \dots & x_{2j}^{(p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n_j j}^{(1)} & x_{n_j j}^{(2)} & x_{n_j j}^{(3)} & \dots & x_{n_j j}^{(p)} \end{pmatrix}$$

Le deuxième élément de ce terme, $\boldsymbol{\beta}$, représente le vecteur des p paramètres inconnus des effets fixes. Chacun de ces paramètres est associé à une covariable du modèle, à l'exception de l'ordonnée à l'origine. Ce dernier est plutôt associé à la première colonne de la matrice $\mathbf{X}_j$, cette dernière ne contenant que des 1. Ce vecteur $\boldsymbol{\beta}$, présenté ci-dessous, est estimé par l'ajustement du modèle.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Les effets aléatoires sont, quant à eux, contenus dans le deuxième terme de la prédiction linéaire du modèle. Le premier élément de ce terme, $\mathbf{Z}_j$, est une matrice de design de dimensions $n_j \times q$ très similaire à la matrice $\mathbf{X}_j$. Cette matrice contient les valeurs connues des q covariables associées aux effets aléatoires. Chaque ligne de la matrice $\mathbf{Z}_j$ ci-dessous représente le vecteur de covariables associé à chacune des n_j observation du groupe. De plus, chacune des colonnes est associée à une covariable q, cette dernière ayant un effet spécifique sur la variable réponse, et ce par rapport au groupe auquel il fait partie.

$$\mathbf{Z}_j = \begin{pmatrix} z_{1j}^{(1)} & z_{1j}^{(2)} & z_{1j}^{(3)} & \dots & z_{1j}^{(q)} \\ z_{2j}^{(1)} & z_{2j}^{(2)} & z_{2j}^{(3)} & \dots & z_{2j}^{(q)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n_j j}^{(1)} & z_{n_j j}^{(2)} & z_{n_j j}^{(3)} & \dots & z_{n_j j}^{(q)} \end{pmatrix}$$

Le deuxième élément de ce terme, $\mathbf{u}_j$, représente le vecteur des q paramètres inconnus des effets aléatoires, ces derniers étant associés aux q covariables de la matrice $\mathbf{Z}_j$. Les éléments de ce vecteur sont présentés ci-dessous :

$$\mathbf{u}_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ \vdots \\ u_{qj} \end{pmatrix}$$

Comme les paramètres $\mathbf{u}_j$ sont aléatoires, nous leur assignons une distribution *a priori*. Selon l'équation 3.6, nous émettons l'hypothèse que le vecteur aléatoire $\mathbf{u}_j$ suit une distribution gaussienne multivariée, ayant comme paramètres un vecteur de moyenne $\mathbf{0}$ et une matrice de variance-covariance $\mathbf{D}$, présentée ci-dessous. La diagonale de cette matrice $\mathbf{D}$ représente les variances de chacun des q effets aléatoires contenus dans le vecteur $\mathbf{u}_j$. Les éléments en dehors de la diagonale représentent les covariances entre les différents effets aléatoires du modèle.

$$\mathbf{D} = Var(\mathbf{u}_j) = \begin{pmatrix} Var(u_{1j}) & cov(u_{1j}, u_{2j}) & \dots & cov(u_{1j}, u_{qj}) \\ cov(u_{1j}, u_{2j}) & Var(u_{2j}) & \dots & cov(u_{2j}, u_{qj}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_{1j}, u_{qj}) & cov(u_{2j}, u_{qj}) & \dots & Var(u_{qj}) \end{pmatrix}$$

3.2 Choix du modèle

Selon le contexte de l'analyse, la sélection de chacun des différents types de modèles présentés ci-dessus peut être pertinente afin de modéliser les risques du domaine de l'assurance de dommage. Dans le cadre de notre recherche, nous avons choisi de modéliser les risques du portefeuille de l'assureur par des modèles linéaires à effets mixtes généralisés. Plusieurs facteurs ont influencé notre processus de sélection.

Tout d'abord, puisque nous considérons le portefeuille de l'assureur comme étant une base de données groupées, l'utilisation des effets mixtes devient intéressante. Lorsque les niveaux observés dans la base de données représentent un échantillon aléatoire de l'ensemble de tous les niveaux possibles, nous pouvons intégrer des effets aléatoires dans le modèle (Bates, 2010). C'est le cas pour le portefeuille de l'assureur, où les différents groupes représentent un sous-ensemble de tous les secteurs d'activités possibles. Nous avons donc associé cette variable à un effet aléatoire.

L'avantage des effets aléatoires est la possibilité de calculer une prédiction pour un nouvel assuré ne faisant pas partie des groupes déjà observés par l'assureur. L'utilisation des effets fixes limite la portée de l'inférence (on ne peut pas extrapoler les estimations des effets fixes à de nouveaux groupes) (Bolker *et al.*, 2009). Comme l'effet fixe génère un paramètre pour chaque groupe de la variable, nous devons connaître cette information afin d'utiliser le bon paramètre dans le calcul de notre prédiction. À l'opposé, avec un effet aléatoire, si un nouvel assuré œuvre dans un secteur d'activité jamais observé, les modèles linéaires mixtes permettent tout de même de faire une prédiction par l'utilisation de la distribution a priori de l'effet aléatoire.

Un autre avantage des effets aléatoires est la possibilité de modéliser une variable catégorielle ayant plusieurs niveaux à partir d'un seul paramètre. Lorsque nous utilisons plutôt un effet fixe, nous obtenons un paramètre fixe par catégorie, ce qui peut mener à un nombre excessif de paramètres à estimer. L'occupation des entreprises du portefeuille que nous modéliserons est assez diversifiée et contient plusieurs niveaux. Comme nous avons associé cette variable à un effet aléatoire plutôt qu'à un effet fixe, nous n'aurons qu'à estimer les paramètres associés à la distribution de l'effet aléatoire.

Lorsque nous nous retrouvons à travailler avec une variable présentant des données clairsemées, comme c'est le cas pour la variable OCCUPATION, certains types de modèles auront de la difficulté à estimer certains paramètres. C'est le cas notamment pour les modèles linéaires incluant seulement des effets fixes.

Lorsque certains niveaux d'une variable de classification sont peu peuplés, et qu'il n'existe pas beaucoup de données sur lesquelles fonder l'estimation du coefficient de régression, ces modèles estimerons toujours un coefficient pour ces niveaux, mais ceux-ci auront une erreur type d'estimation élevée [31].

Afin de régler ce problème, une solution est d'utiliser un modèle linéaire mixte permettant de modéliser la base de données selon des niveaux. Lorsque nous faisont ce genre d'ajustement, 2 types d'information sont considérés. D'un côté, il y a l'information qui est spécifique à chacun des groupes, et d'un autre côté, l'information que tous les groupes partagent. Losqu'un groupe est clairsemé, il est plus difficile d'obtenir de l'information spécifique à ce groupe. Alors, l'estimation de sa ditribution *a posteriori* sera basée en plus grande partie sur l'information que tous les groupes partagent.

Le modèle que nous avons choisi est un cas spécifique du modèle présenté par la définition 3.1, où la distribution conditionnelle de la variable d'intérêt est spécifiée. Il peut être défini comme suit :

Définition 3.2

$$Y_{ij} \mid \mathbf{u}_j \sim Bernoulli(p_{ij}) \tag{3.7}$$

$$logit(p_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}_j \tag{3.8}$$

$$\mathbf{u}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

Y_{ij}, la variable d'intérêt du modèle, modélise la réalisation d'une réclamation de feu pour la structure des entreprises clientes faisant partie du portefeuille de l'assureur. Comme nous considérons cette variable dichotomique, nous avons opté pour une régression logistique. Sa distribution conditiennelle, illustrée par l'équation 3.7, suit donc une loi Bernoulli. En ce sens, Y_{ij} peut prendre seulement deux valeurs. Si la police de l'assuré compte une réclamation, cette variable prend une valeur de 1. Dans le cas contraire où la police de l'assuré ne compte aucune réclamation, elle prend alors la valeur de 0.

Dans la base de données d'entraînement, qui sert à ajuster ce modèle, les entreprises sont regroupées en 13 différents groupes, selon la covariable OCCUPATION. Chacun de ces groupes sont désigné par l'indice j. À l'intérieur de chacun de ces groupes, nous retrouvons n_j assurés. Ces derniers sont désignés par l'indice i. Ci-dessous, le vecteur réponse Y_3 regroupe donc les réclamations des n_3 assurés faisant partie du groupe d'occupation $j = 3$.

$$Y_3 = \begin{pmatrix} Y_{13} \\ Y_{23} \\ Y_{33} \\ \vdots \\ Y_{n_3 3} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Selon l'équation 3.7, la variable Y_{ij}, conditionnellement à l'effet aléatoire $\mathbf{u}_j$, suit une distribution Bernoulli de paramètre p_{ij}. Ce dernier représente le risque spécifique d'une observation du portefeuille de l'assureur. Nous pouvons interpréter ce paramètre comme étant la probabilité que l'entreprise i du groupe j fasse une réclamation à son assureur. De plus, comme nous pouvons le voir par l'équation 3.8, dans le cas d'une régression logistique, la fonction $g(\mu_i)$ qui relie le risque p_{ij} et la prédiction linéaire est une fonction logit.

Les éléments du premier terme de cette prédiction linéaire sont associés aux effets fixes du modèle. $\mathbf{X}_{ij}$ représente le vecteur des valeurs des covariables spécifique à l'assuré i du groupe j. Pour l'assuré $i = 1$ de l'occupation $j = 3$, ce vecteur correspond à la première ligne de la matrice de design $\mathbf{X}_3$ ci-dessous :

$$\mathbf{X}_3 = \begin{pmatrix} x_{13}^{(1)} & x_{13}^{(2)} \\ x_{23}^{(1)} & x_{23}^{(2)} \\ \vdots & \vdots \\ x_{n_3 3}^{(1)} & x_{n_3 3}^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 1,04 \\ 1 & 1,62 \\ \vdots & \vdots \\ 1 & 0,75 \end{pmatrix}$$

La première colonne de cette matrice ne contient que des 1 et est associée à l'ordonnée à l'origine du modèle. La deuxième colonne contient les valeurs de la covariable X.STD. Comme nous l'avons standardisée, cette variable a une moyenne de 0 et un écart-type de 1.

Le deuxième élément du premier terme de la prédiction linéaire, le vecteur $\boldsymbol{\beta}$, représente les paramètres fixes inconnus du modèle. Comme nous avons une seule covariable associée aux effets fixes, ce vecteur, présenté ci-dessous, comprend alors seulement deux paramètres. β_0 représente l'ordonnée à l'origine du modèle, et β_1 est associé à la covariable X.STD.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Comme ces paramètres sont inconnus, nous devons les estimer. Dans le cadre de notre recherche, nous avons ajusté notre modèle de référence à partir de la fonction *glmer* du paquet *lme4*, ce dernier représentant un standard pour l'ajustement de modèles à effets mixtes (Broström et Holmberg, 2011). Cette méthode estime les paramètres par maximum de vraisemblance, et les résultats de cette estimation sont présentés par le vecteur $\widehat{\boldsymbol{\beta}}$ ci-dessous :

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -3,0068 \\ -0,2092 \end{pmatrix}$$

Les éléments du deuxième terme de cette prédiction linéaire sont associés aux effets aléatoires du modèle. Le vecteur $\mathbf{Z}_{ij}$, de la même manière que $\mathbf{X}_{ij}$, représente les valeurs des covariables spécifiques à la police i du groupe j, mais pour les effets aléatoires. Pour la police $i = 1$ de l'occupation $j = 3$, ce vecteur correspond à la première ligne de la matrice de design $\mathbf{Z}_3$ ci-dessous :

$$\mathbf{Z}_3 = \begin{pmatrix} z_{13}^{(1)} \\ z_{23}^{(1)} \\ \vdots \\ z_{n_3 3}^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Comme le modèle ne comprend qu'un seul effet aléatoire, la matrice de design $\mathbf{Z}_3$ se trouve à avoir une seule colonne. De plus, parce que cet effet aléatoire est l'ordonnée à l'origine, cette colonne ne comprend que des 1. De ce fait, chacune des n_j observations i comprise dans le même groupe j ont le même effet aléatoire u_j. Nous pouvons donc réécrire la prédiction linéaire comme suit, en omettant la matrice de design $\mathbf{Z}_j$:

$$logit(p_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{u}_j \tag{3.9}$$

Le deuxième élément du deuxième terme de la prédiction linéaire représente le vecteur $\mathbf{u}_j$. Ce sont les paramètres aléatoires associés au groupe j. Comme les effets aléatoires de notre modèle de base n'incluent qu'une ordonnée à l'origine, le nombre d'effets aléatoires est donc $q = 1$. De ce fait, le vecteur des effets aléatoires du groupe j, présenté ci-dessous, ne contient qu'un élément :

$$\mathbf{u}_j = (u_{1j})$$

Comme il n'y a qu'un effet aléatoire dans le modèle, la matrice de variance-covariance $\mathbf{D}$ est, comme nous pouvons le voir ci-dessous, elle aussi réduite à un seul élément :

$$\mathbf{D} = Var(\mathbf{u}_j) = Var(u_{1j}) = \sigma^2_{u1}$$

Comme nous assumons que l'effet aléatoire u_{1j} suit une loi normale de moyenne 0 et de variance constante σ^2_u, nous pouvons réécrire l'équation 3.9 comme suit :

$$u_j \sim \mathcal{N}(0, \sigma^2_u)$$

Cette équation représente alors la distribution *a priori* de l'effet aléatoire u_j. Nous avons retiré l'indice 1 de la notation car, ayant un seul effet aléatoire, il n'est plus nécessaire de le différencier des autres.

3.3 Calcul des prédictions

Une fois que le modèle avec lequel nous allons travailler est sélectionné, nous pouvons l'ajuster à la base de données d'entraînement. Cet ajustement nous permettra d'obtenir 14 distributions *a posteriori* reliées à l'effet aléatoire. 13 d'entre elles sont spécifiques à une des 13 occupations de la base de données d'entraînement. La 14^{e}, plus générale, modélise la variabilité entre les différents groupes. Ces distributions permettent d'échantillonner l'effet aléatoire, et ainsi calculer des prédictions. À partir d'un modèle à effets mixtes, nous pouvons obtenir 2 types de prédictions : celles provenant des groupes déjà observés, et celles provenant de groupes jamais observés.

3.3.1 Groupes d'occupation déjà observés

Comme mentionné plus haut, en ajustant le modèle, nous obtenons 13 distributions *a posteriori* spécifiques à chacun des groupes de la base de données d'entraînement. De ce fait, pour un nouvel assuré faisant partie d'un groupe déjà observé par l'assureur, nous pouvons échantillonner des effets aléatoires propres à son groupe. À partir de ces effets aléatoires, nous pouvons alors calculer des prédictions spécifiques à son profil de risque. De cette façon, nous nous retrouvons ainsi à segmenter les risques par rapport à l'occupation des entreprises. Cependant, le portefeuille de l'assureur ne contient seulement qu'un sous-ensemble des différentes occupations possibles pour un nouvel assuré. Il est donc possible que nous voulions estimer les risques d'un assuré faisant partie d'un nouveau groupe jamais observé par l'asureur.

3.3.2 Groupes d'occupation jamais observés

Dans le cadre de notre recherche, nous nous intéressons spécifiquement aux assurés oeuvrant dans un secteur d'activité jamais observé par l'assureur. C'est une des raisons pour lesquelles nous avons choisi de travailler avec un modèle à effets mixtes. Ce type de modèles nous permet de calculer des prédictions pour tous les nouveaux assurés du portefeuille, même s'ils ne font pas partie d'un groupe d'occupation déjà observé. Dans ce cas, les effets aléatoires sont échantillonnés à partir de la 14^{e} distributions *a posteriori*, celle modélisant la variabilité entre les différents groupes. Cependant, cette distribution ne permet pas d'échantillonner un effet aléatoire spécifique à l'occupation. Alors, toutes les obervations venants des occupations jamais obervées auront une prédiction contenant le même effet aléatoire.

Dans la base de données initiale, nous avions 25 différentes occupations. 13 d'entre elles ont été sélectionnées pour former la base de données d'entraînement, et elles sont considérées comme des groupes étant déjà observés par l'assureur. Les 12 autres occupations constituent la base de données de test, et sont considérées comme des groupes jamais observés. Une fois que le modèle est ajusté à la base de données d'entraînement, nous calculons des prédictions à partir de la base de données de test, pour chacune des observations des 12 occupations jamais observés. La procédure du calcul de ces prédictions est présentée ci-dessous.

Procédure de calcul

Les occupations présentes dans la base de données de test sont considérées comme oeuvrant dans des

secteurs d'activités ne faisant pas partie de la part de marché de l'assureur. Ces risques sont alors modélisés par la variable Y_{ag}, pour un nouvel assuré a $(a = 1, ..., n_g)$, faisant partie d'un nouveau groupe d'occupation g $(g = 1, ..., G)$. Afin d'évaluer la performance du modèle, nous comparons la variable réponse y_{ag} avec l'estimation de son espérance conditionnelle, $\widehat{E[Y_{ag} \mid u_g]}$, aussi appelée prédiction. Dans le cas d'une distribution Bernoulli, l'espérance de la variable aléatoire est équivalente à son paramètre.

Comme la variable d'intérêt Y_{ag}, conditionnellement à l'effet aléatoire u_g, suit une distribution Bernoulli, son espérance conditionnelle $E[Y_{ag} \mid u_g]$ est donc équivalente à son paramètre p_{ag}. Afin d'obtenir une estimation de ce paramètre, nous devons passer à travers certaines étapes :

- **Étape 1 :** La première étape consiste à échantillonner l'effet aléatoire $\hat{u}_g$. Nous avons comme hypothèse que sa distribution suit une loi normale centrée à zéro de variance σ_u^2. La distribution *a priori* de l'estimation de l'effet aléatoire est donc la suivante :

$$\hat{u}_g \sim \mathcal{N}(0; \widehat{\sigma}_u^2)$$

  En utilisant la base de données d'entrainement, la valeur du paramètre $\widehat{\sigma}_u^2 = 6.68$. Nous échantillonnons donc les k itérations de l'effet aléatoire $\widehat{u}_g$ à partir de la distribution ci-dessous :

$$\hat{u}_g \sim \mathcal{N}(0; 6,68) \tag{3.10}$$

  Pour chacune des observations a de la base de données de test, nous échantillonnons $k = 1000$ itérations $\hat{u}_g^{(k)}$ de l'effet aléatoire à partir de la ditribution 3.10. Nous obtenons alors un total de k prédictions par observation. Cet échantillon nous permettra d'estimer la prédiction de réclamation moyenne d'un assuré faisant partie d'un nouveau groupe d'occupation g jamais observé par l'assureur.

- **Étape 2 :** Une fois que nous avons obtenu un échantillon de l'effet aléatoire, nous pouvons passer à la deuxième étape. Elle consiste à calculer une prédiction linéaire $\widehat{PL}_{ag}^{(k)}$ à partir de chacun des k effets aléatoires obtenus à l'étape 1, selon l'équation ci-dessous :

$$\widehat{PL}_{ag}^{(k)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{ag} + \widehat{u}_g^{(k)}$$

De cette façon, nous obtenons 1000 prédictions linéaires $\widehat{PL}_{ag}^{(k)}$ par observation a de la base de données de test.

- **Étape 3 :** Dans un modèle linéaire mixte généralisé, la variable réponse et la prédiction linéaire sont liées par une fonction lien $g(\cdot)$. Nous devons donc transformer chacune des prédictions linéaires $\widehat{PL}_{ag}^{(k)}$ afin d'obtenir la prédiction de réclamation $\widehat{p}_{ag}^{(k)}$. Pour ce faire, nous utilisons la fonction sigmoïde, l'inverse de la fonction logit, comme suit :

$$\widehat{p}_{ag}^{(k)} = \left(\frac{exp^{\widehat{PL}_{ag}}}{1 + exp^{\widehat{PL}_{ag}}} \right)$$

Nous nous retrouvons donc avec un échantillon de k prédictions $\widehat{p}_{ag}^{(k)}$ par observation a de la base de données de test.

- **Étape 4 :** À la quatrième étape, pour chaque observation a, nous calculons une estimation $\widehat{p}_{ag}$ du paramètre de la distribution de la variable aléatoire Y_{ag}. Cette estimation est calculée à l'aide de l'estimateur ci-dessous :

$$\widehat{p}_{ag} = \widehat{E[Y_{ag} \mid u_g]} = \frac{\sum_{k=1}^{K} \widehat{p}_{ag}^{(k)}}{K}$$

Pour deux nouveaux assurés ayant les mêmes caractéristiques de risque, mais oeuvrant dans de nouveaux secteurs d'activité différents, leurs prédictions $\widehat{p}_{ag}$ calculées à l'aide de la procédure ci-dessus seront les mêmes. Il est possible d'oberver des légères différences, et cela s'explique par le caractère aléatoire de l'échantillonnage de l'effet aléatoire. Bien qu'ils soient tous générés par la même distribution, tous les $\widehat{u}_g^{(k)}$ obtenus peuvent être différents les uns des autres, ce qui mène à des prédictions $\widehat{p}_{ag}^{(k)}$ différentes. Cependant, plus le nombre d'itérations k est grand, moins ces différences seront importantes. Il est donc important de choisir un nombre suffisant d'itérations pour que les résultats soient stables.

Cette procédure permet de calculer les prédictions d'un modèle à effets mixtes. Cependant, elle ne permet pas d'inclure l'avis des experts dans le modèle. Dans la prochaine section, nous présentons comment nous avons modifié cette méthode afin d'intégrer la connaissance des souscripteurs dans nos prédictions. Cette procédure nous permettra de segmenter le risque des nouveaux groupes d'occupations, et ainsi affiner la modélisation des réclamations.

3.4 Échantillonnage de l'effet aléatoire

Comme expliqué plus haut, l'échantillonnage de l'effet aléatoire à partir de la distribution générale de l'effet aléatoire, l'équation 3.10, ne permet pas de segmenter le risque par rapport à l'occupation de l'entreprise assurée. Cependant, nous croyons que les souscripteurs de la ligne d'affaire commerciale possèdent une connaissance par rapport à ces nouveaux groupes d'occupation, de par leur expérience, qui serait intéressante d'inclure dans les modèles de tarification. De ce fait, nous proposons ici une méthode pour inclure l'élicitation de ces experts, de façon à améliorer les prédictions de réclamation pour les nouveaux assurés faisant partie de groupes d'occupation jamais observés par l'assureur.

3.4.1 Modèle de référence

Tout d'abord, nous devons sélectionner un modèle de référence avec lequel nous pourons comparer la performance des modèles incluant l'avis d'experts, afin d'évaluer si cette inclusion améliore vraiment les prédictions. Ce modèle de référence est défini à la section 3.1.3, et nous désignons ce modèle par l'étiquette MODELE.REF. La procédure de calcul des prédictions pour ce modèle a été décrite dans la section 3.3.2. Comme nous l'avons mentionné, l'effet aléatoire des nouveaux groupes d'occupation est échantillonné à partir de la distribution de l'équation 3.10, et cette distribution est illustrée par la figure 3.2.

La figure 3.3 illustre la fréquence de réclamation prédite en fonction de X.STD. Nous pouvons observer que la covariable X.STD est négativement corrélée à la fréquence des réclamations. Pour le modèle de référence, aucune segmentation n'est faite par rapport à l'occupation de l'entreprise assurée. Cela veut dire que, pour deux entreprises ayant deux occupations différentes jamais observées, pour une même valeur X.STD, la prédiction de fréquence sera la même.

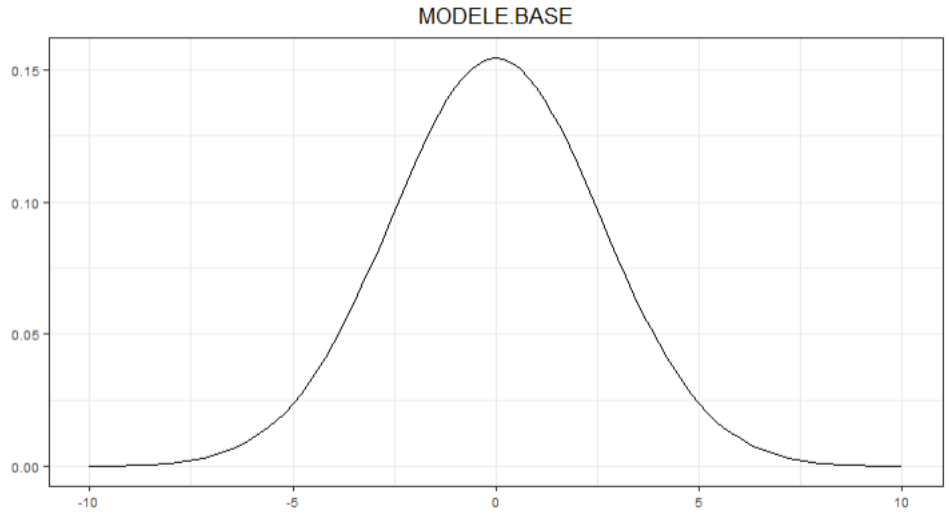

Figure 3.2 Distribution d'échantillonnage de l'effet aléatoire $\widehat{u}_g$: modèle de référence

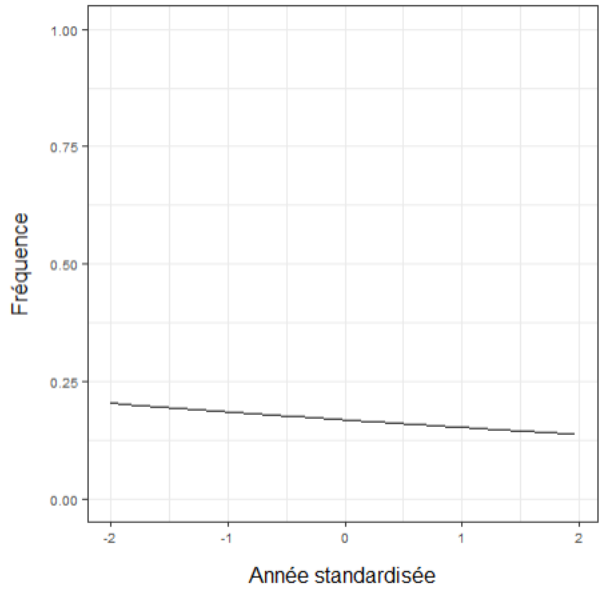

Figure 3.3 Courbe de régression du modèle de référence

3.4.2 Modèles incluant l'avis des experts

Dans (Ni *et al.*, 2018), les auteurs proposent une méthode permettant d'inclure l'avis des experts du domaine de la santé dans des modèles de prédictions de survie de maladies. Selon les auteurs, il y aurait des hôpitaux qui sont en mesure de mieux traiter certaines maladies que d'autres, et ce par rapport à différentes caractéristiques spécifiques de l'hôpital. De ce fait, ils croient que l'inclusion de l'avis d'un expert ayant une connaissance de ces différentes caractéristiques peut améliorer la performance de prédiction de guérison des modèles.

Dans le domaine de l'assurance commerciale, les souscripteurs possèdent une connaissance des risques qui peut être utile à la modélisation d'un portefeuille. De la même façon que les auteurs de (Ni *et al.*, 2018), nous croyons que l'élicitation de leur connaissance et son intégration dans les algorithmes de tarification peut permettre de segmenter les nouveaux assurés faisant partie des secteurs d'activité jamais observés par l'assureur. Ces derniers, de la même façon que les hôpitaux, possèdent certaines caractéristiques spécifiques que nous allons intégrer dans notre algorithme de tarification.

3.4.2.1 Méthode d'élicitation

Afin d'inclure l'avis des experts dans un modèle de tarification, nous devons faire une élicitation de leur connaissance. Dans le cadre de notre recherche, nous utilisons la méthode d'élicitation proposée dans (Ni *et al.*, 2018). Cette méthode consiste à tronquer la distribution à partir de laquelle nous échantillons l'effet aléatoire. Tel que présenté dans une section précédente, afin de calculer les prédictions de réclamation d'un assuré, nous devons tout d'abord échantillonner les effets aléatoires à partir de leur distribution.

Dans le cas du modèle de référence, cet échantillonnage se fait à partir de leur distribution complète (figure 3.2). Dans le cas des modèles incluant l'avis des experts, tel que présenté dans (Ni *et al.*, 2018), cet échantillonnage se fait plutôt à partir d'une certaine zone de la distribution. L'élicitation de l'avis des souscripteurs se fera alors par la sélection de cette zone. Cette sélection se fera à partir de leur connaissance du risque et des caractéristiques spécifiques des nouveaux secteurs d'activité jamais observés.

Prenons par exemple la figure 3.4. Cette courbe représente la distribution de l'effet aléatoire à partir de laquelle il sera échantillonné. Il s'agit de la même distribution que la figure 3.2. Cependant, nous pouvons voir que la courbe est divisée en deux différentes zones par une ligne pointillée ; une première en-dessous

de zéro, et une deuxième au-dessus de zéro. L'élicitation de l'avis des souscripteurs se fera par la sélection d'une de ces deux zones. Si le souscripteur croit que cette nouvelle occupation représente un risque de réclamation plus faible que la moyenne des occupations du portefeuille, les effets aléatoires seront générés à partir de la zone inférieure de la distribution. Cependant, si au contraire, l'expert croit que cette classe d'assurés représente un risque de réclamation plus élevé que la moyenne des occupations du portefeuille, l'effet aléatoire sera plutôt échantillonné à partir de la zone supérieure.

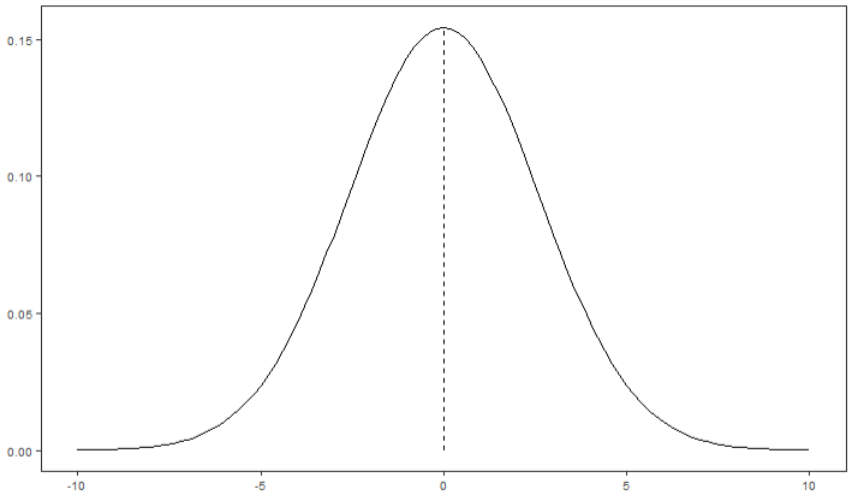

Figure 3.4 Exemple de distribution tronquée

Dans l'exemple ci-dessus, la distribution a été divisée en seulement deux zones. Cela implique une segmentation des assurés selon deux niveaux de risque. Si nous voulons obtenir une segmentation plus précise, nous pouvons diviser la distribution en plusieurs zones afin d'obtenir une segmentation des risques plus précise.

3.4.2.2 Précision de la segmentation

Le niveau de précision de l'avis des experts dépend de la taille de la zone à partir de laquelle les effets aléatoires sont échantillonnés. Plus la taille de cette zone sera petite, plus la précision sera grande. Sur la figure 3.5, les trois graphiques représentent les différents niveaux de précision que nous utilisons pour l'élicitation des souscripteurs. Chacun de ses niveaux de segmentation est associé à un modèle : MODELE.BAS, MODELE.MOYEN et MODELE.HAUT. Ces distributions sont divisées en un certain nombre de zones égales, représentant la région à partir de laquelle l'effet aléatoire sera échantillonné.

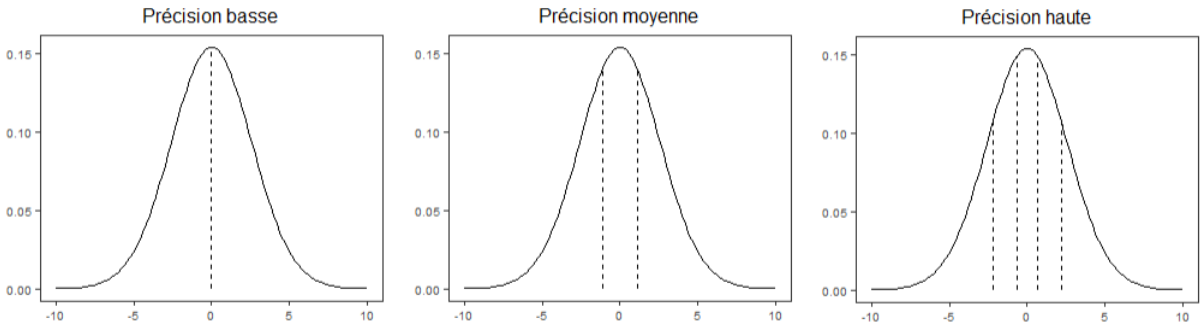

Figure 3.5 Niveaux de précision de la segmentation

Le graphique de gauche ne comprend que deux zones. Les régions d'échantillonnage sont donc plus grandes que sur les deux autres graphiques. De ce fait, il est associé au modèle ayant la segmentation des risques la moins précise : MODELE.BAS. Au centre, la distribution est divisée en 3 zones, représentant une précision plus grande que celle de gauche, mais plus petite que celle de droite. Il est donc associé au modèle ayant un niveau de précision de segmentation moyen : MODELE.MOYEN. Finalement, la courbe de droite est divisée en cinq différentes régions. Les zones sont donc plus petites que sur les deux autres graphiques. De ce fait, elle est associée au modèle incluant la segmentation de risque la plus précise des trois modèles : MODELE.HAUT. Mise à part le niveau de segmentation des risques, les spécifications des trois modèles sont identiques au modèle de référence. Autrement dit, la seule chose qui différencie ces quatre modèles est la zone à partir de laquelle nous échantillonnons l'effet aléatoire.

3.4.2.3 Simulation de l'avis des souscripteurs

Le but de notre recherche est de présenter une méthode afin d'inclure l'avis des souscripteurs dans un modèle de tarification. Cependant, nous ne disposons pas concrètement de cet avis. Nous avons donc développé une méthode afin de le simuler. Cette simulation se fait à partir de la base de données de test, et elle consiste à calculer une estimation du mode de la distribution *a posteriori* du paramètre aléatoire associé à chacune des différentes classes d'occupation.

Pour ce faire, nous utilisons la fonction abc du paquet R du même nom. Le nom de cette fonction provient de l'expression *Approximate Bayesian Computation*, et elle consiste à estimer la distribution *a posteriori* d'un paramètre (Turner et Van Zandt, 2012). Le principal avantage de l'utilisation de cette forme d'approximation est qu'elle ne requiert pas le calcul d'une fonction de vraisemblance (Csilléry *et al.*, 2012). Pour de plus amples informations sur cet algorithme, les auteurs de (Csilléry *et al.*, 2012) présentent une description

détaillée du paquet R abc.

Sur la table 3.1, les colonnes « Occupation » représentent les différentes occupations contenues dans la base de données de test. Les colonnes « Mode » affichent quant à elles les résultats d'estimation du mode de la distribution *a posteriori* du paramètre aléatoire associé à chacune des différentes classes d'occupation. Les différentes occupations sont classées en ordre croissant selon cette estimation.

Occupation	Mode	Occupation	Mode
L	-3,9402	I	0,3825
T	-3,4885	G	1,6493
V	-3,2983	B	2,5045
R	-3,2217	D	2,7215
U	-2,6668	K	3,1254
X	-1,3541	A1	3,4603

Table 3.1 Estimation du mode *a posteriori* des effets aléatoires

C'est à partir de ces estimations que nous formulons l'avis des experts. Selon la valeur de cette estimation, nous définissons à partir de quelle zone l'effet aléatoire sera échantillonné. Par exemple, considérons le graphique de droite de la figure 3.5. Comme mentionné précédemment, ce graphique présente deux zones d'échantillonnage séparées par la ligne verticale pointillée. Prenons alors l'occupation L. Sur la table 3.1, nous pouvons voir que l'estimation du mode pour cette occupation est de -3,9402. Comme cette valeur se situe dans la zone de gauche du graphique de la figure 3.5, c'est à partir de cette zone que les effets aléatoires seront échantillonnés. À l'opposé, pour l'occupation I présentant une estimation au-dessus de zéro, l'effet aléatoire sera plutôt échantillonné dans la zone de droite du graphique.

Il est important de mentionner que, dans le cadre de notre recherche, nous avons comme hypothèse que l'expert ne se trompe jamais. Cependant, nous pouvons croire que, dans certains cas, l'expert peut sélectionner une mauvaise zone à échantillonner, réduisant ainsi la performance du modèle. En ce sens, dans le cadre de futures recherches, il pourrait être intéressant d'évaluer l'impact d'une certaine marge d'erreur sur la performance des modèles. Cependant, ici, nous considérons seulement le cas où l'expert sélectionne la bonne région à échantillonner.

3.4.2.4 Présentation des modèles

Tel que présenté précédemment, notre recherche comprend trois différents niveaux de segmentation des risques, chacun étant associé à un modèle : MODELE.BAS, MODEL.MOYEN et MODELE.HAUT. Dans cette section, nous présentons ces trois modèles de façon détaillée.

- **MODELE.BAS**

  Sur la figure 3.6, nous pouvons voir les 2 différentes régions à partir desquelles nous échantillonnons l'effet aléatoire pour chacune des occupations de la base de données de test. La région à droite, en bleu, représente la zone A. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant une estimation de leur mode *a posteriori* inférieure à zéro. La région à gauche, en vert, représente la zone B. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant une estimation de leur mode *a posteriori* supérieure à zéro. Sur la table 3.2, nous pouvons voir, dans la colonne « Zone », à partir de quelle région l'effet aléatoire de chacune des occupations a été échantillonné.

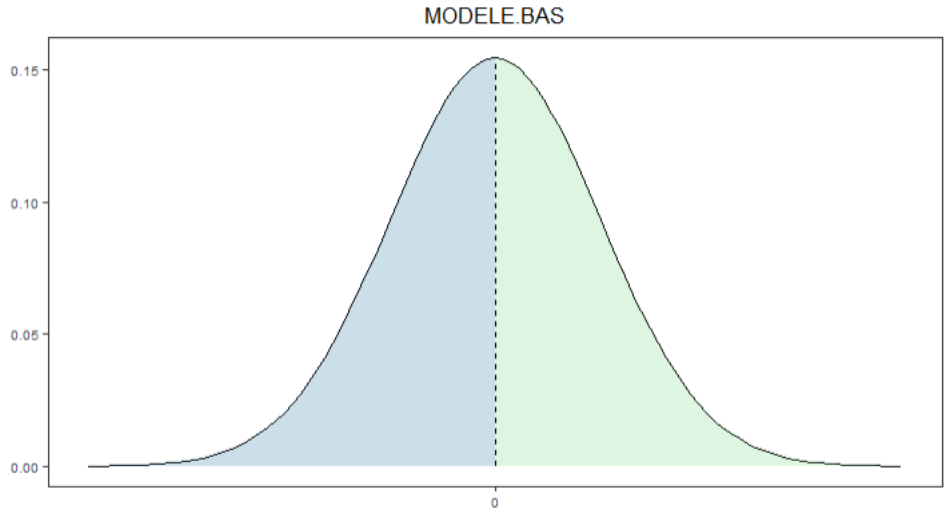

Figure 3.6 Régions d'échantillonnage de l'effet aléatoire pour MODELE.BAS

Les régions d'échantillonnage de MODELE.BAS sont les mêmes que celles présentées sur le graphique de gauche de la figure 3.5. Le niveau de précision est le plus bas des trois modèles. Dans l'éventualité d'obtenir un vrai avis d'expert, le souscripteur doit définir, pour chacun des nouveaux groupes d'occupation, s'il croit que ce groupe représente un risque plus faible que la moyenne des

MODELE.BAS					
Occupation	Mode	Zone	Occupation	Mode	Zone
L	-3,9402	A	I	0,3825	B
T	-3,4885	A	G	1,6493	B
V	-3,2983	A	B	2,5045	B
R	-3,2217	A	D	2,7215	B
U	-2,6668	A	K	3,1254	B
X	-1,3541	A	A1	3,4603	B

Table 3.2 Distribution des zones d'échantillonnage par occupation : MODELE.BAS

occupations déjà observées par l'assureur. Si c'est le cas, l'effet aléatoire pour ce secteur d'activités sera échantillonné à partir de la zone A. Si, au contraire, l'expert croit plutôt que ce groupe est un plus grand risque que la moyenne du portefeuille, l'effet aléatoire sera plutôt généré à partir de la zone B.

La figure 3.7 représente les courbes de régression pour les 2 niveaux de segmentation de MODELE.BAS. L'abcisse représente les différentes valeurs que peut prendre la covariable X.STD, cette dernière étant associée à l'effet fixe du modèle. La courbe bleue représente la fréquence de réclamation pour les occupations faisant partie de la zone A, et la courbe verte représente celle pour les occupations dont nous avons échantillonné l'effet aléatoire dans la zone B. Nous pouvons voir, pour les deux niveaux de risque, une corrélation négative entre la fréquence de réclamation et X.STD. Cependant, pour la zone B, la pente de la courbe est plus prononcée que pour la zone A. De plus, nous pouvons voir que la segmentation des risques par l'élicitation des experts a un important impact sur les prédictions. La courbe de régression pour la zone B se situe au-dessus de 25 %, alors que celle pour la zone A reste près de zéro.

- **MODELE.MOYEN**

  Dans le cas de MODELE.MOYEN, la distribution de l'effet aléatoire est divisée en trois zones. Sur la figure 3.8, nous pouvons distinguer ces trois différentes régions à partir desquelles nous échantillonnons l'effet aléatoire pour chacune des occupations de la base de données de test. La région à droite, en bleu, représente la zone A. Les itérations de l'effet aléatoire sont générées à partir de cette

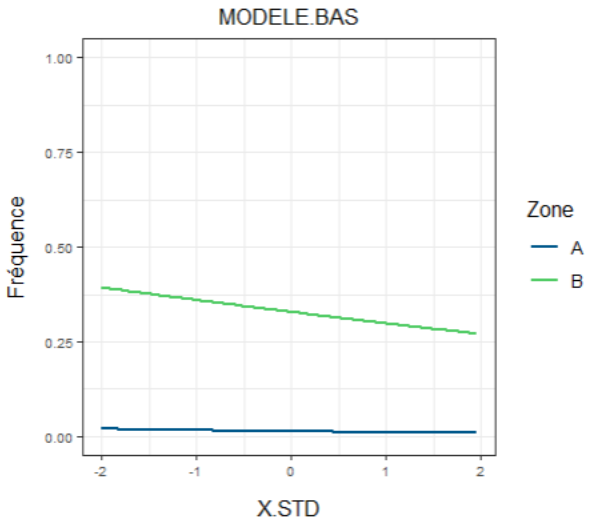

Figure 3.7 Courbes de régression pour MODELE.BAS

région pour les occupations ayant une estimation de leur mode *a posteriori* inférieur à environ -1,11. La région au centre, en turquoise, représente la zone B. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant une estimation de leur mode *a posteriori* se situant entre -1,11 et 1,11. Finalement, la région à gauche, en vert, représente la zone C. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant une estimation de leur mode *a posteriori* supérieur à environ 1,11. De la même façon que pour MODELE.BAS, sur la table 3.3, les différentes occupations de la base de données de test ont été distribuées dans chacune de ces zones, selon l'estimation du mode de leur distribution *a posteriori*.

MODELE.MOYEN					
Occupation	Mode	Zone	Occupation	Mode	Zone
L	-3,9402	A	I	0,3825	B
T	-3,4885	A	G	1,6493	C
V	-3,2983	A	B	2,5045	C
R	-3,2217	A	D	2,7215	C
U	-2,6668	A	K	3,1254	C
X	-1,3541	A	A1	3,4603	C

Table 3.3 Distribution des zones d'échantillonnage par occupation : MODELE.MOYEN

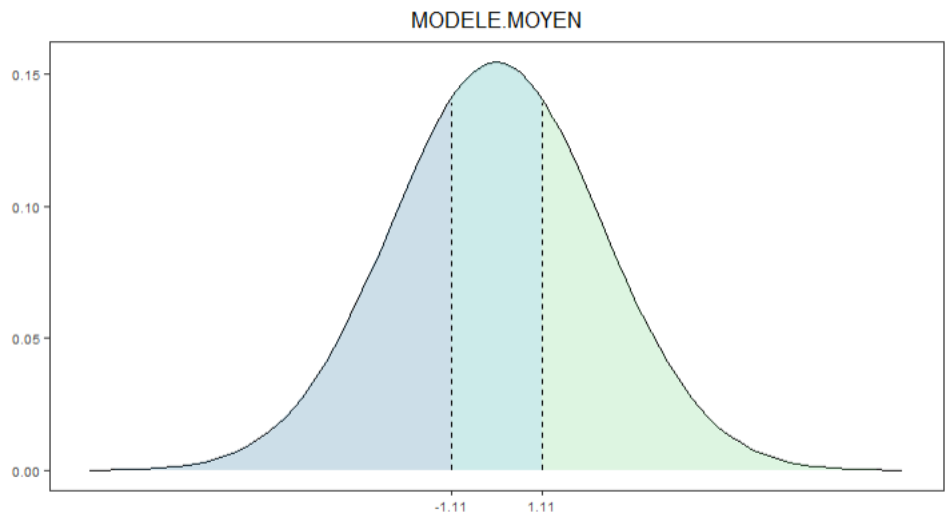

Figure 3.8 Région d'échantillonnage de l'effet aléatoire pour MODELE.MOYEN

Les régions d'échantillonnage de MODELE.MOYEN sont les mêmes que celles représentées par le graphique du centre de la figure 3.5. Ayant trois zones d'échantillonnage au lieu de deux, le niveau de précision de ce modèle est plus élevé que celui de MODELE.BAS. Dans l'éventualité d'obtenir un vrai avis d'expert, le souscripteur doit définir, pour chacun des nouveaux groupes d'occupation, s'il croit que ce groupe représente un niveau de risque se situant dans le premier tiers des différents secteurs d'activités des entreprises possiblement clientes de l'assureur. Si c'est le cas, l'effet aléatoire pour ce secteur d'activités sera échantillonné à partir de la zone A. Il en va de même pour les deux autres zones. Si l'expert croit plutôt que ce groupe représente un niveau de risque se situant dans le deuxième ou troisième tiers du portefeuille, l'effet aléatoire sera plutôt généré à partir, respectivement, de la zone B ou de la zone C.

Sur la figure 3.9, nous pouvons comparer les trois courbes de régression pour les différents niveaux de risque de MODELE.MOYEN. La courbe bleue représente la fréquence de réclamation pour les occupations faisant partie de la zone A, la courbe turquoise trace les différentes fréquences pour la zone B, et la courbe verte représente celle pour les occupations dont nous avons échantillonné l'effet aléatoire dans la zone C. Nous pouvons voir que les trois courbes ne sont pas distribuées uniformément sur le graphique. En effet, Les régressions des zones A et B sont beaucoup plus rapprochées l'une de l'autre que la régression de la zone C. La courbe de la zone A se situe très près de zéro, présentant des fréquences très basses et une pente presque nulle. Pour la zone B, la pente est un peu plus prononcée, et les fréquences se trouvent à être un peu plus élevées que pour la zone A.

Finalement, la zone C présente une courbe ayant une pente encore plus prononcée que les deux autres, et une fréquence beaucoup plus élevée.

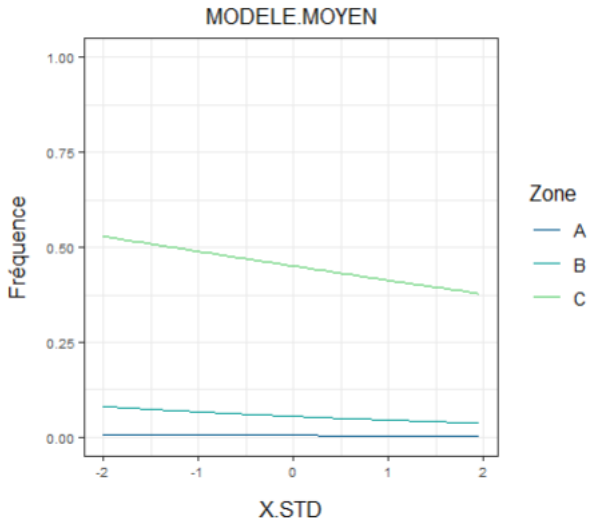

Figure 3.9 Courbe de régression pour le groupe de modèles MODELE.BAS

- **MODELE.HAUT**

  Sur la figure 3.10, nous pouvons maintenant distinguer 5 différentes régions à partir desquelles nous échantillonnons l'effet aléatoire pour chacune des occupations de la base de données de test. La région à droite, en mauve, représente la zone A. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant un mode *a posteriori* inférieur à environ -2,18. La région juste à droite, en bleu, représente la zone B. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant un mode *a posteriori* se situant entre -2,18 et -0,65. La région du centre, en turquoise, représente la zone C. Elle regroupe les occupations ayant un mode *a posteriori* se situant entre -0,65 et 0,65. Juste à droite, la zone D, en vert, permet d'échantillonner l'effet aléatoire des occupations ayant un mode *a posteriori* se situant entre 0,65 et 2,18. Finalement, la région complètement à gauche, en jaune, représente la zone E. Les itérations de l'effet aléatoire sont générées à partir de cette région pour les occupations ayant un mode *a posteriori* supérieur à environ 2,18.

  Sur la table 3.4, les différentes occupations de la base de données de test ont été distribuées dans chacune de ces zones, selon l'estimation du mode de leur distribution *a posteriori*. Nous pouvons voir qu'il y a seulement une occupation dans les zones B, C et D. Le reste des occupations est distribué

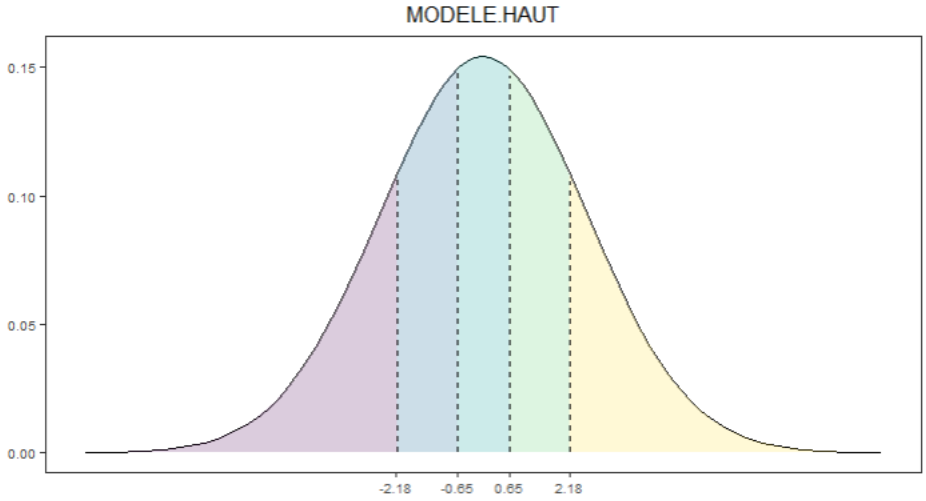

Figure 3.10 Région d'échantillonnage de l'effet aléatoire $\widehat{u}_g$ pour MODELE.HAUT

MODELE.HAUT					
Occupation	Mode	Zone	Occupation	Mode	Zone
L	-3,9402	A	I	0,3825	C
T	-3,4885	A	G	1,6493	D
V	-3,2983	A	B	2,5045	E
R	-3,2217	A	D	2,7215	E
U	-2,6668	A	K	3,1254	E
X	-1,3541	B	A1	3,4603	E

Table 3.4 Distribution des zones d'échantillonnage par occupation : MODELE.HAUT

dans les zones A et E. Les régions d'échantillonnage de MODELE.HAUT sont les mêmes que celles représentées par le graphique de droite de la figure 3.5. Ayant 5 zones d'échantillonnage, le niveau de précision de ce modèle est le plus élevé des trois modèles. Afin d'obtenir un avis d'expert, le souscripteur doit être en mesure d'évaluer dans quelle portion de la distribution l'effet aléatoire doit être échantillonné.

Sur la figure 3.11, nous pouvons comparer les cinq courbes de régression pour les différents niveaux de risque de MODELE.HAUT. Nous pouvons constater que, de la même manière que pour MODELE.MOYEN, la distribution des régressions n'est pas uniforme entre les différentes zones. Plus nous échantillonnons l'effet aléatoire dans une région supérieure de la distribution, plus l'impact sur la fréquence est élevé. Effectivement, l'échantillonnage dans les zones A ou B n'aura pas beaucoup d'impact sur les prédictions. Les courbes mauve et bleue sont très rapprochées l'une de l'autre. À l'opposé, l'échantillonnage dans la zone D ou E changera énormément la fréquence de réclamation. Les courbes de régression verte et jaune sont très loin l'une de l'autre.

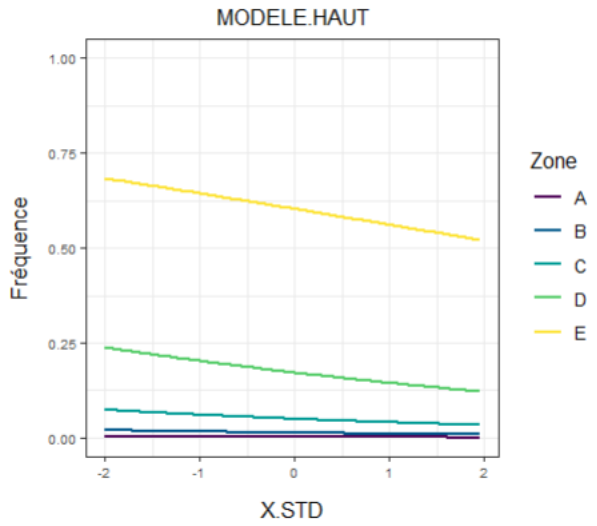

Figure 3.11 Courbe de régression pour le groupe de modèles MODELE.HAUT

3.5 Performance des modèles

Une fois que nous avons calculé les prédictions des observations de la base de données de test, nous pouvons maintenant calculer la performance des différents modèles, et les comparer avec le modèle de référence. Nous quantifions cette performance à l'aide des métriques ROC-AUC et PR-AUC que nous avons

présentées dans le premier chapitre.

Sur la table 3.6, nous pouvons voir les résultats obtenus. La première colonne nommée « Modèle » représente le modèle à partir duquel nous avons calculé les prédictions utilisées pour le calcul des métriques de performance. La deuxième colonne représente l'aire sous la fonction d'efficacité du récepteur (ROC-AUC), plus connue sous le nom de courbe ROC (*Receiver Operating Characteristic*). La troisième colonne représente l'aire sous la courbe de précision-rappel (PR-AUC). Pour chacune de ces deux métriques, avec un modèle complètement aléatoire, nous pouvons nous attendre à une aire sous la courbe de 0,50.

Modèle	ROC-AUC	PR-AUC
MODELE.BASE	0,5267	0,5445
MODELE.BAS	0,5784	0,5541
MODELE.MOYEN	0,6126	0,5728
MODELE.HAUT	0,6183	0,5777

Table 3.5 Résultats des métriques de performance

Sur la figure 3.12, nous pouvons voir une représentation graphique de la courbe ROC pour les différents modèles. La courbe pointillée, sur chacun des graphiques, représente la courbe ROC moyenne d'un modèle complètement aléatoire. Cette courbe correspond à une ROC-AUC de 0,5. Sur le premier graphique en haut à gauche, la courbe noire représente la courbe ROC obtenue par le modèle de référence. L'aire sous cette courbe est de 0,5267. Comme nous pouvons le constater, le modèle de référence a un pouvoir prédictif plutôt faible. Effectivement, la courbe ROC obtenue est très près de la courbe moyenne d'une modélisation aléatoire.

En regardant les trois autres graphiques, nous pouvons voir que les régions entre les courbes ROC et pointillée sont plus grandes. En ce sens, l'avis des experts semble améliorer, dans les trois cas, le pouvoir prédictif du modèle. Respectivement, l'aire sous les courbes ROC de MODELE.BAS, MODELE.MOYEN et MODELE.HAUT est augmentée respectivement de 0,0517, 0,0858 et 0,0915 comparativement au MODEL.REF.

Maintenant, sur la figure 3.13, nous pouvons voir une représentation graphique de la courbe de précision-rappel (PR) des différents modèles. La courbe pointillée, sur chacun des graphiques, représente la courbe PR provenant d'un modèle complètement aléatoire. Cette courbe correspond à une PR-AUC de 0,5. Le premier

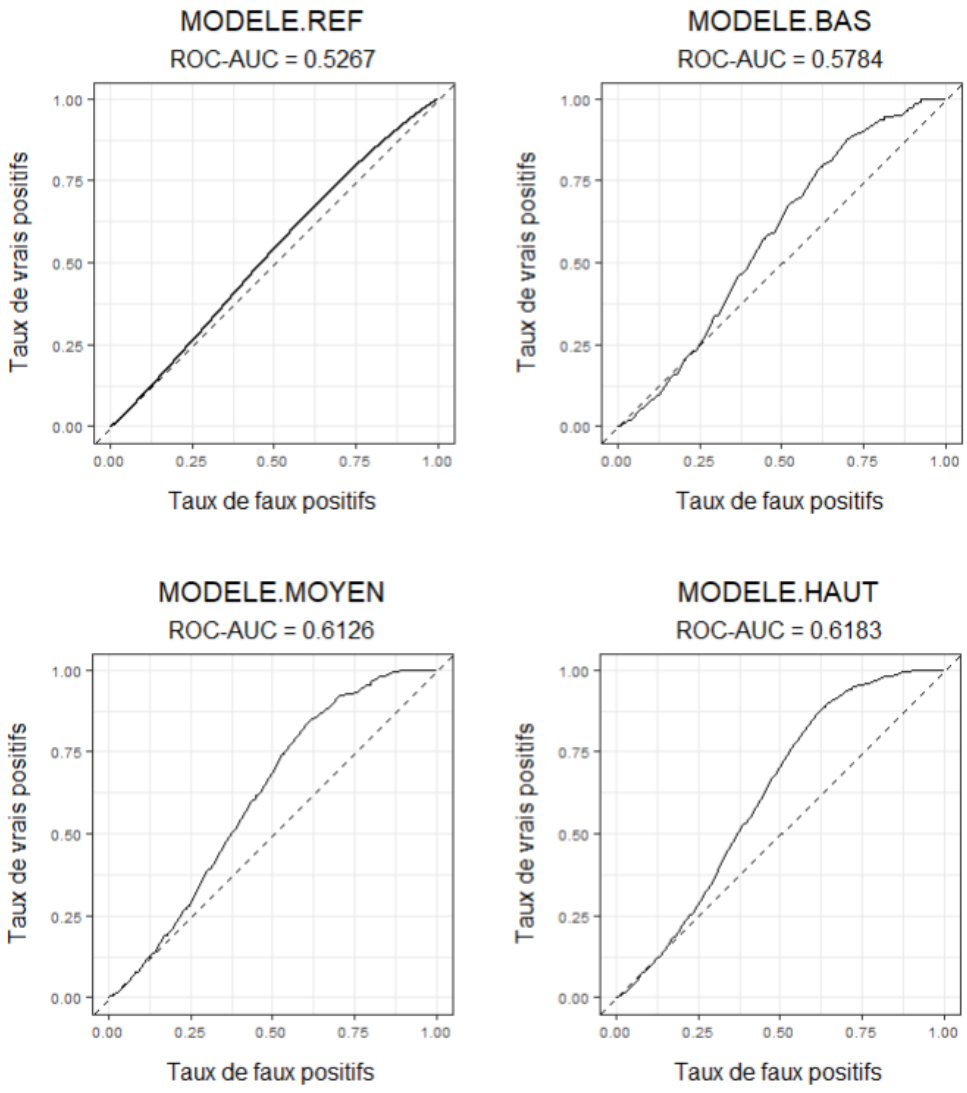

Figure 3.12 Graphique des courbes ROC

graphique en haut à gauche illustre la courbe PR obtenue par le modèle de référence. L'aire sous cette courbe est de 0,5445. Comme nous pouvons le constater, de la même manière que sur le graphique de la courbe ROC, le modèle de référence semble avoir un pouvoir prédictif plutôt faible. Effectivement, la courbe PR obtenue est assez près de celle d'une modélisation aléatoire.

En regardant les trois autres graphiques, nous pouvons voir que la région entre les courbes PR et pointillée est plus grande. En ce sens, l'avis des experts semble améliorer, dans les trois cas, le pouvoir prédictif du modèle. Respectivement, l'aire sous les courbes PR de MODELE.BAS, MODELE.MOYEN et MODELE.HAUT est augmentée de 0,0096, 0,0283 et 0,0332 comparativement au MODELE.REF.

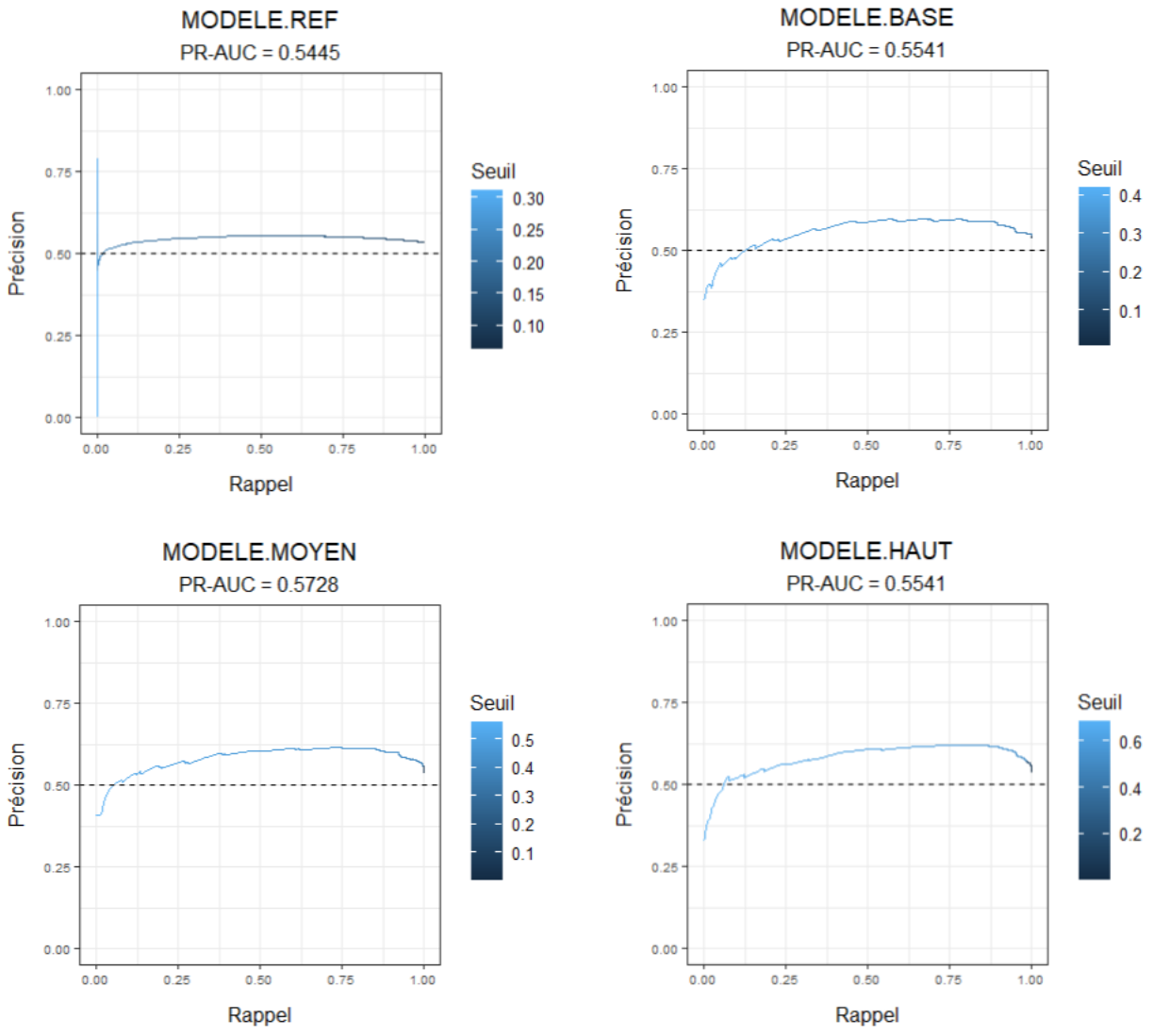

Figure 3.13 Graphique des courbes PR

3.6 Procédure d'élicitation d'un vrai avis d'expert

Dans les modèles présentés ci-dessus, nous utilisons une simulation d'avis d'experts afin de proposer une méthode d'élicitation de connaissance dans le cadre de la tarification en assurance. En ce sens, nous n'utilisons pas un vrai avis d'expert. L'objectif de cette présente section est de proposer une méthode qui permettra d'utiliser un vrai avis dans le cadre du travail quotidien des souscripteurs. Les modèles proposés plus haut visent à inclure l'avis d'experts en influençant la simulation de l'effet aléatoire dans un modèle GLMM. Plus précisément, l'avis de l'expert influence la simulation de l'effet aléatoire en restreignant la distribution utilisée pour l'échantillonnage. Cette idée repose toutefois sur une hypothèse fondamentale : la capacité de l'expert à évaluer correctement la qualité d'un risque.

3.6.1 Définition d'un expert

Dans un contexte opérationnel, l'application de cette méthode doit être encadrée, car elle repose sur un jugement subjectif. Il ne s'agit pas simplement de « croire » que le risque est bon, mais d'avoir une expérience suffisante pour reconnaître les bons profils de manière fiable. Dans le cadre de ce modèle, un expert n'est pas simplement un employé sénior. C'est une personne qui :

- Connaît bien le portefeuille de l'entreprise et ses spécificités.
- A une compréhension fine des types de risques assurés.
- Est capable de reconnaître les signaux faibles d'un bon (ou mauvais) risque.
- A un historique de décisions cohérent avec les résultats observés.

Ce niveau d'expertise s'acquiert avec le temps, l'exposition à divers cas, et le retour d'expérience sur les décisions passées. Il serait donc pertinent que l'organisation mette en place des critères internes (nombre d'années, validation par un mentor, performance passée) pour autoriser aux souscripteurs l'utilisation de leur avis.

3.6.2 Quatre niveaux de modèles selon l'avis de l'expert

L'un des apports de cette méthode est de formaliser ce que plusieurs souscripteurs expérimentés font déjà de manière implicite : ajuster leur jugement sur un risque en fonction de leur intuition professionnelle. Ici, cette intuition est intégrée dans le modèle via la sélection d'une portion de la distribution. Cela permet de

donner un cadre méthodologique rigoureux à un processus jusqu'ici informel. La méthode que nous avons développée propose différents niveaux de confiance exprimés par le souscripteur, selon la portion de la distribution utilisée pour simuler l'effet aléatoire :

- **Modèle de référence** : Utilisation de la distribution complète. Aucune influence de l'expert. C'est l'approche par défaut, entièrement guidée par les données.
- **Modèle BAS** : Utilisation de la moitié de la distribution. L'expert émet une opinion prudente : le risque se situe au-dessus ou au-dessous de la moyenne, sans précision.
- **Modèle MOYEN** : Utilisation du tiers de la distribution. L'expert est confiant que le risque se situe autour de la moyenne, ou plus loin en-dessous ou au-dessus.
- **Modèle HAUT** : Utilisation du cinquième de la distribution. L'expert a une forte certitude de la qualité du risque.

Ces différents niveaux de modèles permettent de faire varier le degré d'intervention du jugement humain dans la tarification. En résumé, cette approche reconnaît la valeur du jugement expert, tout en s'assurant qu'il est appliqué de manière prudente, progressive et encadrée — en particulier en fonction de l'expérience du souscripteur.

3.6.3 Niveaux d'autorité

Dans les pratiques actuelles de souscription, les compagnies d'assurance encadrent les décisions individuelles à l'aide de directives de souscription. Ces dernières définissent les règles, procédures et marges de manœuvre autorisées pour les souscripteurs, et elles sont conçues pour assurer la cohérence, la conformité réglementaire et la rentabilité des décisions de souscription.

Un élément central de ces directives est la notion de niveaux d'autorité, qui déterminent jusqu'à quel point un souscripteur peut engager l'entreprise (par exemple : accepter un risque complexe, appliquer une dérogation, ajuster une prime, etc.). Ces niveaux sont généralement structurés selon l'expérience, la formation, les résultats passés et parfois la spécialisation du souscripteur. Dans le tableau ci-dessous, nous proposons quatre différents niveaux d'autorité dans l'élaboration de notre méthode. Plus l'impact du jugement sur la tarification est important, plus il est nécessaire que le souscripteur dispose d'un niveau d'autorité élevé.

Niveau	Profil	Autorité	Accès aux modèles
1	Nouveau souscripteur	Risques simples	Référence
2	Souscripteur intermédiaire	Risques standards	Référence/BAS
3	Souscripteur sénior	Risques complexes	Référence/BAS/MOYEN
4	Expert ou chef de souscription	Risques hors normes	Référence/BAS/MOYEN/HAUT

Table 3.6 Niveaux d'autorité proposés

Afin d'implémenter notre méthode, nous proposons d'ajouter une section dans les directives de souscription définissant les conditions d'utilisation des modèles BAS, MOYEN et HAUT. De plus nous proposons d'associer à chaque modèle une description du niveau de confiance requis, et un niveau d'autorité minimal pour y accéder. Il serait aussi intéressant de mettre en place un processus de validation ou de revue pour les cas où un souscripteur souhaite utiliser un modèle auquel il n'a pas encore accès (par exemple, en justifiant son choix auprès d'un responsable). Nous croyons que cette structure permettrait de favoriser l'utilisation du jugement expert tout en garantissant un contrôle de qualité adapté à l'expérience du souscripteur.

CONCLUSION

Dans un environnement en constante évolution, où l'incertitude, la réglementation et la pression concurrentielle exigent des analyses rigoureuses, l'actuariat représente une discipline pour laquelle la recherche et l'innovation sont constament mise de l'avant par les entreprises. Dans ce mémoire, nous nous sommes penchés sur deux importantes problématiques en actuariat : le débalancement de données et la segmentation des risques. En ce sens, nous avons proposé une approche permettant une meilleure estimation des pertes en assurance commerciale.

Cette approche se divise en deux volets. Le premier se résume au développement d'une méthode d'échantillonnage pour rebalancer la base de données. Le second consiste à inclure l'avis des experts dans un modèle linéaire à effets mixtes généralisé (GLMM) par l'utilisation de distributions tronquées pour les effets aléatoires. Notre méthode démontre que l'élicitation des souscripteurs peut améliorer le pouvoir prédictif d'un modèle de tarification traditionnel. Cependant, cette étude a également mis en évidence certaines limites.

D'une part, nous avons travaillé avec une base de données réelles pour laquelle la distribution de la variable d'intérêt était largement débalancée. De ce fait, nous avons exploré plusieurs algorithmes afin de choisir celui qui se prête le mieux à notre échantillon. Cependant, dans le cadre de cette recherche, nous nous sommes contraints à l'analyse des méthodes d'échantillonnage. Cependant, à ce jour, il existe un vaste éventail de méthodes qui pourraient être explorées, et une méthode mieux adaptée au contexte de notre étude pourrait être identifiée.

D'autre part, dans ce travail, nous ne disposons pas concrètement de l'avis des experts. Alors, afin de proposer une méthode permettant d'inclure l'élicitation des souscripteurs dans un modèle, une simulation de cet avis a été obtenue. Pour de futures analyses, il serait intéressant d'amener cette recherche plus loin en utilisant un vrai avis d'expert. Cela permettra de confirmer la pertinence de cette méthode, et d'évaluer la robustesse des modèles développés face au biais des souscripteurs.

BIBLIOGRAPHIE

Baesens, B., Höppner, S., Ortner, I. et Verdonck, T. (2021). robrose : A robust approach for dealing with imbalanced data in fraud detection. *Statistical Methods Applications*, *30*, 841—-861.

Bates, D. M. (2010). lme4 : Mixed-effects modeling with r. *Humaine Association Conference on Affective Computing and Intelligent Interaction (Geneva : IEEE)*, 245—-251.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. et White, J.-S. S. (2009). Generalized linear mixed models : a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127–135.

Broström, G. et Holmberg, H. (2011). Generalized linear models with clustered data : Fixed and random effects models. *Computational Statistics and Data Analysis*, *55*(12), 3123–3134.

Chawla, N. V., Bowyer, K. W., Hall, L. O. et Kegelmeyer, W. P. (2002). Smote : Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–-357.

Chen, M.-H., Ibrahim, J. G., Shao, Q.-M. et Weiss, R. E. (2003). Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, *3*, 57–76.

Csilléry, K., François, O. et Blum, M. G. B. (2012). abc : an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, *3*, 475–479.

David, M. (2015). A review of theoretical concepts and empirical literature of non-life insurance pricing. *Procedia Economics and Finance*, *20*, 157–162.

El Kassimi, F. et Zahi, J. (2021). Non-life insurance ratemaking techniques : A literature review of the classic methods. *International Journal of Accounting*, *Finance*, *Auditing*, *Management and Economics*, *2*(1), 344–361.

Ferri, C., Hernández-Orallo, J. et Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*, 27–38.

Grize, Y. L. (2015). Applications of statistics in the field of general insurance : An overview. *International Statistical Review*, *83*(1), 135–159.

Harris, G. T. et Rice, M. E. (2013). Bayes and base rates : What is an informative prior for actuarial violence risk assessment ? *Behavioral Sciences and the Law*, *31*(1), 103–124.

He, H., Bai, Y., Garcia, E. A. et Li, S. (2008). Adasyn : Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 1322–1328.

He, H. et Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Ibrahim, J. G. et Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*,

15(1), 46–60.

Jeni, L. A., Cohn, J. F. et De La Torre, F. (2013). Facing imbalanced data – recommendations for the use of performance metrics. *HUMAINE Association Conference on Affective Computing and Intelligent Interaction*, 245–251.

Kassimi, F. E. et Zahi, J. (2021). Non-life insurance ratemaking techniques : A literature review of the classic methods. *International Journal of Accounting, Finance, Auditing, Management and Economics*, *2*(1), 344–361.

Krestenitis, M., Orfanidis, G., Ioannidis, K., Avgerinakis, K., Vrochidis, S. et Kompatsiaris, I. (2019). Oil spill identification from satellite images using deep neural networks. *Remote Sens.*, *11*, 1762.

Kuhnert, P. M. (2011). Four case studies in using expert opinion to inform priors. *Environmetrics*, *22*(5), 662–674.

Kynn, M. (2008). The 'heuristics and biases' bias in expert elicitation. *Journal of the Royal Statistical Society, Series A*, *171*(1), 239–264.

Low Choy, S., O'Leary, R. et Mengersen, K. (2009). Elicitation by design in ecology : using expert opinion to inform priors for bayesian statistical models. *Ecology*, *90*(1), 265–277.

Lunardon, N., Menardi, G. et Torelli, N. (2014). Rose : A package for binary imbalanced learning. *The R Journal*, *6*(1), 79–89.

Macedo, L. (2009). The role of the underwriter in insurance. *The Primer Series on Insurance*, (8).

Ni, H., Groenwold, R. H. H., Nielen, M. et Klugkist, I. (2021). Expert opinion as priors for random effects in bayesian prediction models : Subclinical ketosis in dairy cows as an example. *Plos One*, *16*(1).

Ni, H., van der Drift, S., Klugkist, I., Jorritsma, R., Hooijer, G. et Nielen, M. (2018). Prediction models for clustered data with informative priors for the random effects : a simulation study. *BMC Medical Research Methodology*, *83*(83).

O'Leary, R. A., Low Choy, S., Murray, J. V., Kynn, M., Denham, R., Martin, T. G. et Mengersen, K. (2008). Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby petrogale penicillata. *Environmetrics*, *20*, 379–398.

Raina, R., Ng, A. Y. et Koller, D. (2006). Constructing informative priors using transfer learning. *Proceedings of the 23rd international conference on Machine learning*, 713–720.

Turner, B. M. et Van Zandt, T. (2012). A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85.

Zhang, J. et Miljkovic, T. (2018). Ratemaking for a new territory : Enhancing glm pricing model with a bayesian analysis. *Casualty Actuarial Society E-Forum*, *2*.