

Mémoire présenté le 4 juillet 2023

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

Par : Marie DEROM

Titre : Modèles de prédiction de l'âge de retraite, dans le contexte de l'évaluation de l'engagement des indemnités de fin de carrière, selon l'IAS 19

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

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

Entreprise

Nom : Also Consulting

Signature :

Directeur de mémoire en entreprise

Nom : Jean-Baptiste CHALNOT

Signature :

Invité(e)

Nom : Candice POUJADE

Signature :

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

Signature du responsable entreprise

Signature du candidat

Membres présents du jury de l'Institut des Actuaires :

A. EYRAUD

.....
.....

Membres présents du jury de l'ISFA :

J. ARTHUR

.....
.....

Résumé

Mots clés : âge de retraite, indemnité de fin de carrière, évaluation d'engagements sociaux, IAS 19, modèle prédictif.

Ce mémoire a l'objectif d'améliorer le modèle d'âge de retraite d'une entreprise, qui est utilisé lors de l'évaluation de son engagement au titre des indemnités de fin de carrière (IFC). L'établissement de ce modèle représente un enjeu pour les entreprises car elles doivent représenter fidèlement leurs engagements (pour donner une image réaliste de leur situation financière) et elles ont intérêt à diminuer l'écart actuariel constaté à chaque évaluation (c'est-à-dire l'écart entre l'engagement réel et l'engagement prévu). Une procédure de construction et d'évaluation de modèles a été choisie pour s'adapter spécifiquement aux modèles prédictifs (et non explicatifs). Et l'une des métriques d'évaluation sélectionnées est spécifique au contexte : il s'agit de l'écart actuariel qu'il engendre. Après une présentation et une analyse des données, deux modèles ont été appliqués : la régression linéaire et l'arbre de régression. Ces modèles présentent l'avantage d'être facilement interprétables et applicables (particulièrement pour l'arbre de régression). En effet, ce mémoire est contraint de se restreindre à de tels modèles car ses résultats sont destinés à être présentés et expliqués à des responsables d'une entreprise, et qu'ils doivent être analysés et compris par des commissaires aux comptes. La prédiction des modèles construits s'est révélée être plus précise que le modèle qui est actuellement utilisé par l'entreprise. Toutefois, les résultats obtenus sur l'écart actuariel engendré ne sont pas satisfaisants, comparés à ceux du modèle actuel. Il faut toutefois souligner que ce dernier est plus sophistiqué que les modèles couramment utilisés lors d'évaluation d'engagement IFC. De plus, la prédiction de l'âge de retraite est en pratique difficilement réalisable, car il existe de nombreux facteurs décisifs au départ à la retraite d'un individu, alors que les caractéristiques des salariés mises à disposition dans ce mémoire sont restreintes.

Abstract

Keywords : retirement age, retirement indemnity, valuation of employee liabilities, IAS 19, predictive model.

The purpose of this thesis is to improve a company's retirement age model, which is used when valuing its liability for end-of-career benefits (IFC). Establishing this model is a challenge for companies because they need to accurately represent their liabilities (to give a realistic picture of their financial situation) and they have an interest in reducing the actuarial difference noted at each valuation (i.e. the difference between the actual liability and the expected liability). A model building and evaluation procedure was chosen to specifically accommodate predictive (not explanatory) models. And one of the evaluation metrics selected is specific to the context : it is the actuarial gap it creates. After a presentation and an analysis of the data, two models were applied : the linear regression and the regression tree. These models have the advantage of being easily interpretable and applicable (especially for the regression tree). Indeed, this thesis is constrained to such models because its results are intended to be presented and explained to managers of a company, and they must be analyzed and understood by auditors. The prediction of the constructed models proved to be more accurate than the model that is currently used by the company. However, the results obtained on the actuarial gap generated are not satisfactory compared to the current model. It should be noted, however, that the latter is more sophisticated than the models commonly used in IFC engagement evaluations. Moreover, the prediction of retirement age is in practice difficult to achieve, as there are many decisive factors for the retirement of an individual, while the characteristics of the employees made available in this thesis are limited.

Sommaire

1	Introduction	7
2	Contexte de l'étude	10
2.1	La retraite en France	10
2.1.1	Présentation des ancêtres du régime général de retraite en France	11
2.1.2	Présentation des fondements du régime général des retraites en France	12
2.1.3	Remises en question du régime général	13
2.1.4	Les chiffres clés de la retraite en France aujourd'hui	15
2.1.5	Observation du comportement des français face au départ en retraite	19
2.2	Présentation de l'IAS 19	23
2.2.1	Définition de l'IAS 19	23
2.2.2	Périmètre d'application de la norme IAS 19 dans l'UE	26
2.2.3	Méthodes de comptabilisation et de calculs selon l'IAS 19	27
2.2.4	Explication et conséquences d'une nouvelle interprétation de l'IAS 19	31
2.2.5	Présentation de l'évaluation de l'engagement au titre des IFC selon l'IAS 19.	33
2.3	L'importance de l'hypothèse de l'âge de retraite lors d'une évaluation de l'engagement au titre des IFC	36
2.3.1	La sensibilité de l'engagement face à cette hypothèse	36
2.3.2	L'impact de l'interprétation de l'IFRIC	37
3	Méthodologie adoptée pour répondre à la problématique	39
3.1	Présentation de la modélisation de l'âge de retraite actuellement utilisée dans l'entreprise	40

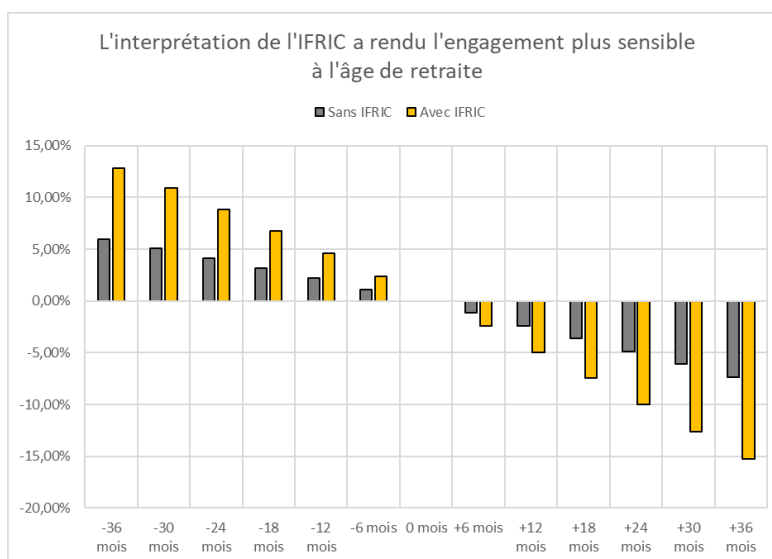
3.1.1	Définition du modèle	40
3.1.2	Le modèle suppose que les travailleurs partent entre ces deux âges.	42
3.2	Présentation des données fournies par l'entreprise	45
3.2.1	La base de données des retraités contient plus de 32 000 observations.	45
3.2.2	L'entreprise a fourni des données pour évaluer son engagement IFC à deux dates différentes.	47
3.3	Méthodologie adoptée lors du traitement des données	51
3.3.1	Deux variables dépendent de caractéristiques obsolètes ou représentatives d'une petite partie de la population.	52
3.3.2	La présence d'une caractéristique peu représentée est un risque de biaiser les modélisations.	55
3.3.3	Les écarts "âge retraite - âge minimal" éloignés de la moyenne doivent être étudiés de près.	56
3.3.4	La présence de variables qui prédisent le même comportement doit être évité.	59
3.3.5	Le traitement des données sera donc réalisé en 4 étapes.	61
3.4	Méthodologie adoptée lors de la modélisation	62
3.4.1	La base de données des retraités est divisée en deux échantillons.	62
3.4.2	Le phénomène de sur-apprentissage doit être évité pour bien prédire.	63
3.4.3	La stabilité du modèle doit être vérifiée.	66
3.4.4	La procédure de modélisation permet d'étudier la stabilité du modèle et d'optimiser son pouvoir prédictif	67
3.5	Métriques choisies d'évaluation du pouvoir prédictif d'un modèle.	69
3.5.1	L'écart actuariel d'expérience engendré a été choisi pour évaluer la performance des modèles.	69
3.5.2	La racine de l'erreur quadratique moyenne a également été choisie pour évaluer la performance d'un modèle.	72
4	Application de la méthodologie adoptée pour répondre à la problématique	74
4.1	Analyse et traitements de la base de données des retraités	75
4.1.1	Transformation des variables présentant des caractéristiques obsolètes ou représentatives d'une petite population	75
4.1.2	Identification et traitement des caractéristiques peu représentées	77
4.1.3	Identification et traitement des valeurs extrêmes de la variable à prédire	81

4.1.4	Identification et traitement des variables prédictives fortement liées	85
4.1.5	Présentation de la base de données des retraités après traitements	90
4.2	Modélisation de l'âge de retraite à partir d'une régression linéaire .	91
4.2.1	Présentation générale de la régression linéaire	91
4.2.2	Analyse du phénomène de sur-apprentissage et de la stabilité du modèle	93
4.2.3	Analyse du modèle final	95
4.3	Modélisation de l'âge de retraite à partir d'un arbre de régression .	99
4.3.1	Présentation générale de l'arbre de régression	100
4.3.2	Analyse du phénomène de sur-apprentissage et de la stabilité du modèle	102
4.3.3	Analyse du modèle final	104
4.4	Comparaison des modélisations avec le modèle de référence	106
4.4.1	Évaluation du pouvoir prédictif du modèle de référence . . .	106
4.4.2	Comparaison du modèle de référence avec les modèles construits	107
5	Conclusion	109

Chapitre 1

Introduction

L’objectif de ce mémoire est d’établir une hypothèse sur l’âge de retraite des salariés d’une entreprise, utilisée lors de l’évaluation d’un engagement social. En France, les entreprises sont engagées (par la loi) vis-à-vis de leurs salariés à verser une Indemnité de Fin de Carrière (IFC) au moment de leur départ en retraite. Cette indemnité représente une dette, qu’il faut savoir évaluer pour s’assurer d’être en capacité de la supporter, et pour donner une image fidèle de la situation financière de l’entreprise (lors de la clôture des comptes). L’évaluation de cet engagement nécessite de fixer une hypothèse sur l’âge de retraite des salariés, que ce mémoire propose d’établir pour une entreprise. La construction de cette hypothèse ne représente toutefois pas seulement une nécessité pour l’entreprise, mais également un enjeu. En effet, elle doit comptabiliser l’écart entre l’engagement réel et l’engagement prévu (susceptible d’être élevé lorsque l’âge de retraite prévu est éloigné de l’âge de retraite réel). Il est important pour elle que cette hypothèse soit bien calibrée, car cet écart vient impacter son compte de résultats ou ses capitaux propres (selon la norme comptable qu’elle applique). D’autre part, une interprétation de la norme comptable internationale (l’IAS 19) encadrant l’évaluation de cet engagement, a été validée récemment (en mai 2021) et a rendu l’évaluation de l’engagement plus sensible à l’hypothèse de l’âge de retraite. Un mauvais calibrage de cette hypothèse biaise alors la représentation de la situation financière de l’entreprise. Comme l’illustre le graphique ci-dessous, l’impact sur l’engagement d’un avancement ou d’un recul de l’âge de retraite est plus élevé après prise en compte de cette interprétation (réalisée par un comité nommé l’IFRIC, présenté dans 2.2.1).



Par conséquent, l'entreprise a intérêt à ce que cette hypothèse soit bien calibrée pour minimiser l'impact sur son compte de résultats ou ses capitaux propres, mais aussi pour représenter fidèlement sa situation financière. Ce mémoire s'intéressera donc à établir cette hypothèse de façon à minimiser l'écart entre l'engagement réel et l'engagement prévu, et à se rapprocher au mieux de la réalité.

L'entreprise étudiée utilise aujourd'hui une hypothèse, qui est déjà plus complexe que dans la majorité des cas. En général, l'âge de retraite des non-cadres est fixé à 62-63 ans et celui des cadres à 64-65 ans. En effet, il est supposé que les non-cadres commencent leur carrière professionnelle à 19-20 ans, alors que les cadres la commence à 21-22 ans. De plus, les salariés sont supposés partir à la retraite lorsqu'ils ont validé la durée minimale d'obtention de la pension retraite maximale (fixée à 43 ans, pour les plus jeunes générations). C'est donc pourquoi ces âges de retraite sont généralement fixés pour l'évaluation de l'engagement IFC. Toutefois, l'entreprise étudiée dans ce mémoire utilise aujourd'hui une hypothèse plus complexe : elle suppose qu'une certaine proportion des salariés parte à la retraite à l'âge d'ouverture des droits, et qu'une autre proportion parte à l'âge d'obtention de la pension maximale (taux plein). Les départs en retraite survenus dans l'entreprise permettent de déterminer ces proportions. Par conséquent, ce mémoire cherche à améliorer une hypothèse, qui est déjà plus sophistiquée que dans la majorité des cas. De plus, même si la norme comptable internationale (l'IAS 19) qui régit le calcul et la comptabilisation de cet engagement est vague sur l'établissement de cette hypothèse (elle impose simplement que l'hypothèse ne soit ni risquée, ni d'une prudence excessive, cf. article 74 de la référence [1]), les possibilités de modélisation de l'âge de retraite sont restreintes. En effet, ce mémoire a des contraintes à respecter puisqu'il s'inscrit dans un contexte corporatif :

il doit proposer une hypothèse facilement compréhensible et facilement applicable. Enfin, il devra se contenter des données salariales dont l'entreprise dispose, et qui sont souvent limitées. Ce dernier point représente une lourde contrainte puisqu'il a été démontré (cf. référence [2]) que les facteurs explicatifs du départ à la retraite en Europe sont très nombreux. Ils peuvent être regroupés en trois sphères : la sphère individuelle (âge, genre, état de santé, richesse, épargne, etc.), contextuelle (situation familiale, environnement de travail, etc.) et institutionnelle (systèmes de protection sociale). Ainsi, ce mémoire se place dans un contexte difficile : améliorer le calibrage d'une hypothèse qui est déjà plus sophistiquée que la moyenne, tout en proposant un modèle facilement interprétable et applicable, à partir de caractéristiques sur les salariés limitées.

La première partie de ce mémoire est consacrée à son contexte. Elle présente tout d'abord le système de retraite français, qui fait partie de l'un des facteurs explicatifs du départ à la retraite en Europe (dans la sphère institutionnelle). Elle définit et présente ensuite l'IAS 19, la norme comptable internationale qui encadre l'évaluation d'engagements sociaux, dont la méthode de calcul sera détaillée. Elle termine enfin par démontrer l'importance d'un bon calibrage de l'hypothèse pour cette évaluation.

La seconde partie de ce mémoire consiste à présenter la méthodologie suivie pour répondre à ses objectifs. Pour cela, la méthode et les métriques d'évaluation de modèle seront présentées dans un premier temps. Dans un deuxième temps, l'hypothèse de l'âge de retraite actuellement utilisée par l'entreprise sera présentée en détails. L'exploration, l'analyse et le traitement des données feront l'objet de la dernière section de cette deuxième partie.

La dernière partie de ce mémoire est consacrée à la présentation et l'évaluation des modèles sélectionnés. Avant cela, le modèle d'âge de retraite qui est actuellement en vigueur dans l'entreprise sera évalué : il constituera le modèle de référence de ce mémoire. Ensuite, les différents modèles sélectionnés pour satisfaire l'objectif de ce mémoire seront présentés, évalués et commentés au regard du modèle de référence.

Chapitre 2

Contexte de l'étude

L'âge de la retraite sera étudié au regard de l'évaluation de l'engagement au titre des IFC d'une entreprise.

Anticiper le départ en retraite d'un salarié est utile pour un employeur, notamment pour piloter son entreprise (choisir un futur remplaçant, réduire l'éventuelle perte de clientèle, etc.), mais également pour anticiper les pertes financières. En effet, le départ en retraite d'un salarié est nécessairement (en France) accompagné du versement d'une indemnité. En comptabilité, cette indemnité est un engagement (une dette) de l'entreprise vis-à-vis de ses salariés. Il faut savoir bien l'évaluer afin de représenter une image financière fidèle de l'entreprise. Étant publique et accessible à tous, cette image est souvent un point sur lequel l'employeur porte une grande attention (vis-à-vis de ses clients, ses prestataires, ses investisseurs, ses salariés, etc.). C'est dans le contexte de l'évaluation de cet engagement social, qui est l'indemnité de fin de carrière (IFC), que ce mémoire s'inscrit. Les règles de cette évaluation sont régies par l'IAS 19, une norme comptable internationale. Cette norme sera présentée au paragraphe 2.2 ci-après. La première partie de ce chapitre est consacrée au système de retraite français (cf. 2.1), car l'âge de départ en retraite est évidemment amené à varier selon celui-ci.

2.1 La retraite en France

Le système de retraite en France est un sujet intrinsèque de ce mémoire. En effet, il semble inévitable que celui-ci ait un impact sur l'âge de départ en retraite.

2.1.1 Le premier régime de retraite a été créé au XVIIIe siècle.

Le premier système de retraite a été fondé pour attirer des personnes à devenir des marins.

En 1784, l'ordonnance de Castries a créé le **premier véritable système de retraite** (cf. référence [3]). Il est fréquent que cette création soit rattachée à Jean-Baptiste Colbert car celui-ci a fondé en 1673 le premier régime social (par l'Edit de Nancy), qui a conduit à la création du premier régime de retraite. Contrôleur général des finances de Louis XIV, mais également secrétaire d'Etat de la marine, Jean-Baptiste Colbert se penche sur un nouveau système de recrutement des équipages de la marine royale (cf. référence [4]), afin de satisfaire la volonté du roi de devenir une grande puissance maritime. Dans ce contexte, il propose aux marins une certaine protection sociale, pour les pousser à s'engager et rester fidèle : la caisse des invalides de la marine royale fut alors créée en 1673. Elle était financée par tous les marins de la marine royale, et permettait de verser une pension aux marins blessés ou invalides. Cependant, cette pension ne fut considérée comme un droit (et non plus une faveur du roi) qu'au jour de l'adoption de l'ordonnance de Castries (à l'initiative du maréchal de cette ville) en 1784 (cf. référence [5]). Cette ordonnance avait pour but de refonder le système de recrutement des équipages de la marine royale, façonné par Colbert. Désormais, sous condition de durée effective de navigation, tous les marins âgés de plus de 60 ans bénéficiaient d'une pension (dont le montant variait selon leurs situations).

La retraite est alors perçue comme un bon moyen d'attirer des personnes à exercer un métier.

D'autres caisses de retraite ont ensuite vu le jour : il s'agit des systèmes de retraite spécifiques. Par exemple, un régime de retraite a été créé en 1768 pour certains fonctionnaires (les employés de la Ferme générale, qui est un service de collecte des impôts sous la royauté). Ce régime était un fonds de pension, qui fut alimenté par les employés versant une retenue annuelle. Ce système leur permettait de recevoir une somme égale à la totalité de leurs retenues. D'autre part, en 1850, les premières compagnies privées de chemin de fer ont bénéficié d'un système de retraite par capitalisation : les retenues sur salaire et les contributions patronales étaient versées sur un livret ouvert au nom de chacun des employés, qui leur permettait de verser leur pension de retraite. Le but de ces régimes est généralement d'attirer des salariés de qualité et de les fidéliser, car la perspective d'une pension de retraite est appréciée.

L'ancêtre du système général de retraite français fut ensuite créé durant l'entre-deux-guerres.

Après la Première Guerre Mondiale, la France constate qu'elle est en retard face aux autres pays européens, dans le domaine social :

elle met alors en place les assurances sociales en 1928. En effet, après la guerre 14-18, les habitants de l'Alsace et de la Lorraine disposaient d'une législation sociale bien plus large (issue des assurances sociales de Bismarck, cf. référence [6]) que les autres habitants français. La France a alors fondé des assurances sociales en 1928, et les a révisées deux ans plus tard. Parmi ces assurances sociales, le système de retraite élaboré est un système par capitalisation, qui fixe un minimum vieillesse à 40% du salaire moyen pour trente années de cotisations à 60 ans. Il présente beaucoup d'inconvénients :

- les cadres ne sont pas concernés, ce qui limite le degré de solidarité entre salariés ;
- il existe 80 caisses d'assurance vieillesse-décès et des caisses d'assurance invalidité, ce qui rend difficile leur gestion ;
- les ressources financières furent épuisées suite à la crise des années 30, à la guerre et aux besoins de l'Etat.

2.1.2 Le régime général des retraites en France fut fondé en 1945, lors de la création de la sécurité sociale.

Le régime général de retraite est un système par répartition.

En 1945, tous les salariés du secteur privé (hors agriculteurs) bénéficient d'un régime de retraite par répartition (cf. référence [7]). Ce type de régime est financé par les cotisations des actifs qui servent directement à payer les prestations des retraités (il s'oppose au régime par capitalisation, vu précédemment). L'âge minimal de départ à la retraite était fixé à 65 ans. Concernant la retraite des salariés agricoles, celle-ci s'est mise en place progressivement entre 1951 et 1958 (cf. référence [8]).

Les dispositions qui ont suivi viennent améliorer les conditions des retraités.

En 1947, un régime de retraite complémentaire pour les cadres fut créé. L'Association générale des institutions de retraite des cadres (AGIRC) coordonne un système de retraite qui diffère du système général par son mode d'acquisition des droits : le montant de la pension dépend du nombre de points acquis (via le versement des cotisations). L'âge minimal de départ à la retraite reste celui du système général.

En 1949, c'est au tour des travailleurs indépendants d'avoir leur propre caisse de retraite. Trois nouvelles caisses de retraite sont créées :

la CNAVPL (Caisse nationale d'assurance vieillesse des professions libérales), la Cancava (Caisse nationale de compensation d'assurances vieillesse des artisans) et l'Organic (Organisation autonome nationale d'assurance vieillesse de l'industrie et du commerce).

En 1956, le minimum vieillesse est créé. Financé par l'impôt, ce dispositif permet d'assurer une pension minimale aux personnes de plus 65 ans n'ayant pas suffisamment cotisé.

En 1961, l'Association pour le régime de retraite complémentaire des salariés (ARRCO) est créée. Elle élabore un régime de retraite pour les salariés du privé (cadres et non-cadres). Les cadres ont donc désormais deux systèmes de retraite complémentaires différents.

En 1971, les règles de calcul des pensions du régime général deviennent plus favorables aux retraités. La réforme Boulin augmente le taux de remplacement de la retraite à taux plein de 10 points : il passe de 40% à 50%. De plus, le salaire de référence ne correspond plus au salaire moyen des 10 dernières années mais au salaire moyen des 10 meilleures années.

En 1972, les retraites complémentaires AGIRC et ARRCO deviennent obligatoires. Par conséquent, la retraite obligatoire des salariés a 2 volets : la retraite de base et la retraite complémentaire.

En 1982, l'âge minimal de départ en retraite passe de 65 ans à 60 ans. François Mitterrand cherche, à travers cette mesure, à instaurer un « droit au repos que les travailleurs sont fondés à revendiquer en contrepartie des services rendus à la collectivité à l'issue d'une durée de carrière normale ».

2.1.3 A partir de 1991, l'équilibre financier du système de retraite est une réelle préoccupation.

En 1991, le Livre Blanc sur les retraites pointe le doigt sur les conséquences du vieillissement de la population sur l'équilibre du système de retraite. Plusieurs pistes de réforme sont proposées et ont été appliquées, en 1993, par la réforme Balladur qui :

- allonge progressivement la durée de cotisation de 37,5 ans à 40 ans,

Le bon fonctionnement du régime général de retraite fait partie petit à petit des préoccupations importantes de l'Etat français.

- allonge la durée du calcul du salaire de référence de 10 à 25 ans,
- modifie l'indexation des pensions de retraite sur les salaires par une indexation sur les prix,
- crée un Fonds de Solidarité Vieillesse (FSV) pour financer les dépenses de solidarité comme le minimum vieillesse, les trimestres de chômage, les majorations pour enfants.

De plus, en 1999, le Fonds de Réserve des Retraites (FRR) est créé pour anticiper le choc financier du "Papy-boom". Ce fonds est financé par une taxe sur les revenus du patrimoine et des placements, et sur les dotations exceptionnelles.

En 2000, la création du Conseil d'Orientation des Retraites (COR) illustre la volonté de suivre le bon fonctionnement du système de retraite. Ce conseil doit étudier ce système en réalisant des projections à plus ou moins long terme et des analyses thématiques. Il est placé sous la responsabilité du Premier Ministre et il est essentiellement composé de représentants d'organisations professionnelles et syndicales, de parlementaires et de hauts fonctionnaires (cf. référence [9]).

En 2003, la prise de conscience, initiée par le Livre Blanc, sur l'équilibre financier du régime de retraite continue par la réforme Fillon. En effet, cette réforme allonge encore la durée de cotisation de 40 ans à 41 ans et crée des produits d'épargne retraite par capitalisation (PERP, PERCO, PERE).

En 2007, une volonté d'harmoniser le système de retraite apparaît. Les régimes spéciaux, comme ceux de la Banque de France, la SNCF et la RATP, s'alignent effectivement progressivement sur le régime de la fonction publique (notamment par l'instauration d'un mécanisme de surcote / décote).

En 2010, le besoin de financement n'est toujours pas suffisant. Suite à la crise de 2008, le déficit s'aggrave beaucoup. La réforme Woerth :

- allonge alors la durée de cotisation de 41 ans à 41,5 ans,
- pousse l'âge minimal de départ en retraite de 60 ans à 62 ans,
- reporte l'âge automatique de la retraite à taux plein de 65 ans à 67 ans.

La solution beaucoup utilisée pour financer davantage le système est de rallonger la durée de cotisation.

- augmente le taux de cotisation des fonctionnaires de 7,85% à 10,55%, qui correspond au taux de cotisations des salariés du privé.

En 2014, la réforme Touraine allonge de nouveau la durée de cotisation. En effet, celle-ci passe de 41,5 ans à 43 ans. D'autre part, cette réforme instaure un compte pénibilité permettant de partir plus tôt dans certaines conditions, et abaisse la valeur d'un trimestre de 200 à 150 SMIC horaire brut.

En 2019, les deux systèmes de retraite complémentaires des salariés du privé fusionnent. Afin d'éviter la faillite du système, l'AGIRC et l'ARRCO fusionnent pour que leurs réserves financières soient en commun et pour harmoniser les règles : le système unique fonctionne par points.

2.1.4 Les paramètres du système de retraite français d'aujourd'hui

Les chiffres clés de la retraite dépendent du statut du salarié. Malgré la volonté d'harmoniser les systèmes de retraite, il existe encore aujourd'hui des différences, particulièrement entre les salariés du secteur privé et les agents du droit public. Au sein des fonctionnaires, deux catégories existent :

- La catégorie active qui réunit les emplois ayant un risque particulier ou procurant une fatigue exceptionnelle (ils sont classés par décret ou arrêté ministériel) ;
- La catégorie sédentaire qui réunit les autres emplois.

A noter que pour être considéré fonctionnaire de la catégorie active, il faut avoir effectué une durée minimale au sein d'un ou plusieurs emplois de cette catégorie. De plus, il n'est pas exigé que le salarié termine sa carrière en exerçant un emploi de la cette catégorie.

L'âge d'ouverture des droits à la retraite varie entre 55 ans et 62 ans. Afin de repousser progressivement cet âge minimal de départ à la retraite, celui-ci dépend de la génération du salarié et de son statut (cf. fiche 14 de la référence [10]).

L'âge minimal de départ en retraite, entre 55 ans et 62 ans, varie selon la génération et le statut du salarié.

Date de naissance	Age de retraite minimal	
	Salarié du privé et fonctionnaire sédentaire	Fonctionnaire actif
Avant juillet 1951	60 ans	55 ans
Juillet et décembre 1951	60 ans et 4 mois	
En 1952	60 ans et 9 mois	
En 1953	61 ans et 2 mois	
En 1954	61 ans et 7 mois	
Janvier 1955 et juin 1956	62 ans	
Juillet et décembre 1956		55 ans et 4 mois
En 1957		55 ans et 9 mois
En 1958		56 ans et 2 mois
En 1959		56 ans et 7 mois
1960 et après		57 ans

TABLE 2.1 – L'âge d'ouverture des droits à la retraite

Il existe toutefois des situations particulières permettant de partir avant l'âge d'ouverture des droits. Les salariés du droit privé dans l'une des situations suivantes peuvent partir plus tôt :

- Avoir effectué une carrière longue (cf. référence [11] pour plus de détails). Ce dispositif permettait de partir au plus tôt à 56 ans. Cependant, cet âge a augmenté par paliers de 4 mois à partir de la génération de naissance 1955, pour atteindre 58 ans pour les salariés nés en 1960 ou après ;
- Avoir un handicap (cf. référence [12] pour plus de détails), ce qui permet de partir au plus tôt à 55 ans ;
- Être atteint d'une incapacité permanente d'origine professionnelle (cf. référence [13] pour plus de détails), ce qui permet de partir au plus tôt à 60 ans ;
- Avoir été exposé à l'amiante dans le contexte professionnel (cf. référence [14] pour plus de détails), ce qui permet de partir au plus tôt à 50 ans.

Dans le secteur public, plusieurs situations exceptionnelles existent permettant un départ anticipé à la retraite :

Il existe des situations donnant droit à une retraite anticipée (avant l'âge d'ouverture des droits).

- Avoir effectué une carrière longue (cf. référence [15] pour plus de détails), ce qui permet de partir au plus tôt à 58 ans ;
- Être invalide (cf. référence [16] pour plus de détails), ce qui permet de partir dès que l'expiration des droits à congé maladie ;
- Être atteint d'une incapacité permanente d'au moins 50% ou être reconnu comme un travailleur avec un handicap (cf. référence [17] pour plus de détails), ce qui permet de partir au plus tôt à 55 ans ;
- Avoir un enfant atteint d'une invalidité d'au moins 80%, sous condition d'avoir au moins 15 ans de services (cf. référence [18] pour plus de détails), ce qui permet de partir dès que ces deux conditions sont remplies ;
- Être atteint (ou avoir un époux(se) atteint) d'une infirmité ou d'une maladie incurable rendant l'exercice de toute profession impossible, sous condition de faire partie des fonctionnaires d'Etat et d'avoir au moins 15 ans de services (cf. référence [19] pour plus de détails), ce qui permet de partir dès que ces conditions sont remplies.

Certains dispositifs de départ anticipé sont fermés mais appliqués encore aujourd'hui.

Les situations évoquées ci-dessus sont encore en vigueur, c'est-à-dire qu'elles s'appliquent pour les nouveaux travailleurs. Néanmoins, il peut exister d'autres dispositifs de retraite anticipée qui ne s'appliquent plus pour les nouveaux travailleurs mais qui s'appliquent pour une population fermée de travailleurs. Par exemple, les fonctionnaires remplissant les conditions ci-dessous avant 2012, peuvent demander leur retraite anticipée à tout moment (cf. référence [20]) :

- Avoir accompli au moins 15 ans de service civils ou militaires ;
- Être parent d'au moins 3 enfants ;
- Avoir cessé ou réduit son activité professionnelle pendant une durée minimum.

Pour obtenir la retraite à taux plein, il faut justifier d'une durée de cotisation ou atteindre l'âge de taux plein automatique. Le calcul de la pension de retraite du système général prend en compte le nombre de trimestres cotisé par le salarié : si celui-ci est supérieur ou égal à la durée de référence (cf. la fiche 14 de la référence [10]), alors la retraite est à taux plein (la pension est maximale). Si ce n'est pas le cas, il existe un âge auquel le taux plein est accordé quelque soit la durée de cotisation du salarié : il

s'agit de l'âge du taux plein automatique, aussi appelé l'âge d'annulation de la décote.

Année de naissance	Durée de cotisation		
	Salarié du privé	Fonctionnaire sédentaire	Fonctionnaire actif
1947	160 trimestres	158 trimestres	150 trimestres
1948		160 trimestres	
1949	161 trimestres	161 trimestres	152 trimestres
1950	162 trimestres	162 trimestres	154 trimestres
1951	163 trimestres	163 trimestres	156 trimestres
1952	164 trimestres	164 trimestres	158 trimestres
1953	165 trimestres	165 trimestres	160 trimestres
1954			161 trimestres
1955	166 trimestres	166 trimestres	162 trimestres
1956			163 trimestres
1957			165 trimestres
1958			
1959 et 1960	167 trimestres	167 trimestres	166 trimestres
1961	168 trimestres	168 trimestres	
1962 et 1963		167 trimestres	
1964 et 1965	169 trimestres		169 trimestres
1966		168 trimestres	
1967 et 1968	170 trimestres		170 trimestres
1969		169 trimestres	
1970 et 1971	171 trimestres		171 trimestres
1972		170 trimestres	
1973 et 1974	172 trimestres		172 trimestres
1975 à 1977		172 trimestres	171 trimestres
1978 et après		172 trimestres	172 trimestres

TABLE 2.2 – Durée de cotisation pour obtenir la retraite à taux plein

La durée de cotisation pour obtenir le taux plein se situe entre 37,5 ans et 43 ans.

Date de naissance	Age du taux plein automatique		
	Salarié du privé	Fonctionnaire sédentaire	Fonctionnaire actif
1946	65 ans	61 ans	sans objet
1947		61 ans et 6 mois	sans objet
1948		62 ans	sans objet
1949		62 ans et 3 mois	sans objet
1950		62 ans et 6 mois	sans objet
Janvier à juin 1951		62 ans et 9 mois	56 ans
Juillet à août 1951	65 ans et 4 mois	63 ans et 1 mois	56 ans
Septembre à décembre 1951		63 ans et 4 mois	56 ans
Janvier à mars 1952	65 ans et 9 mois	63 ans et 9 mois	56 ans et 6 mois
Avril à décembre 1952		64 ans	56 ans et 6 mois
Janvier à octobre 1953	66 ans et 2 mois	64 ans et 8 mois	57 ans
Novembre à décembre 1953		64 ans et 11 mois	57 ans
Janvier à mai 1954	66 ans et 7 mois	65 ans et 4 mois	57 ans et 3 mois
Juin à décembre 1954		65 ans et 7 mois	57 ans et 3 mois
1955		66 ans et 3 mois	57 ans et 6 mois
Janvier à juin 1956	67 ans	66 ans et 6 mois	57 ans et 9 mois
Juillet à août 1956			58 ans et 1 mois
Septembre à décembre 1956			58 ans et 4 mois
Janvier à mars 1957		66 ans et 9 mois	58 ans et 9 mois
Avril à décembre 1957			59 ans
Janvier à octobre 1958		67 ans	59 ans et 8 mois
Novembre à décembre 1958			59 ans et 11 mois
Janvier à mai 1959			60 ans et 4 mois
Juin à décembre 1959			60 ans et 7 mois
1960			61 ans et 3 mois
1961			61 ans et 6 mois
1962			61 ans et 9 mois
1963 et après			62 ans

TABLE 2.3 – Âge du taux plein automatique

L'âge du taux plein automatique est entre 56 ans et 67 ans.

Les travailleurs peuvent être mis à la retraite (sans leur consentement) à partir d'un certain âge : il s'agit de l'âge limite. Cet âge s'élève à :

A partir de l'âge limite (67 ans), l'employeur peut mettre son salarié à la retraite, sans son consentement.

- 70 ans pour les salariés du privé (cf. référence [21]) ;
- 67 ans pour les fonctionnaires de la catégorie sédentaire (cf. référence [22]) ;
- 62 ans pour les fonctionnaires de la catégorie active (cf. référence [22]).

Pour les fonctionnaires, la mise à la retraite est automatique, alors que dans le secteur privé, l'employeur doit en faire la demande. Néanmoins, certaines situations peuvent permettre aux fonctionnaires de travailler au-delà de l'âge limite, à condition qu'ils soient déclarés aptes à travailler (cf. référence [22]) :

Les fonctionnaires sont mis, quant à eux, automatiquement à la retraite dès qu'ils ont atteint leur âge limite (62 ans ou 67 ans, selon leur catégorie).

- Lorsque le départ à l'âge limite ne permet pas d'obtenir une retraite à taux plein, il est possible de travailler jusqu'à son obtention (sans dépasser 69 ans et 6 mois pour la catégorie sédentaire, et 64 ans et 6 mois pour la catégorie active). Cette situation est toutefois amenée à disparaître, puisque l'âge du taux plein automatique correspond, pour les nouvelles générations, à la limite d'âge (cf. tableau 2.3).
- Si le fonctionnaire a au moins un enfant à sa charge lorsqu'il atteint la limite d'âge, il peut alors continuer à travailler 1 an supplémentaire, par enfant à charge (sans dépasser 3 ans).
- Si le travailleur avait au moins 3 enfants vivants à charges lorsqu'il avait 50 ans (ou qu'il a un ou plusieurs enfants morts pour la France), il peut alors travailler 1 année supplémentaire.
- Si l'administration accepte sa demande, il peut alors travailler jusqu'à ses 67 ans.

2.1.5 L'étude sur l'âge de retraite en France démontre l'importance des paramètres du système.

Pour tous travailleurs confondus, l'âge conjoncturel moyen de départ en retraite est passé de 60,2 ans en 2010 à 62,3 ans en 2020. L'âge conjoncturel moyen de départ à la retraite mesure l'âge moyen de retraite d'une année donnée, en neutralisant les différences de taille de génération (cf. référence [23]). Son évolution depuis 2009 (cf. référence [24])

Le départ en retraite est repoussé depuis 2010.

illustre bien les réformes successives, qui ont pour objectif de repousser l'âge de retraite.

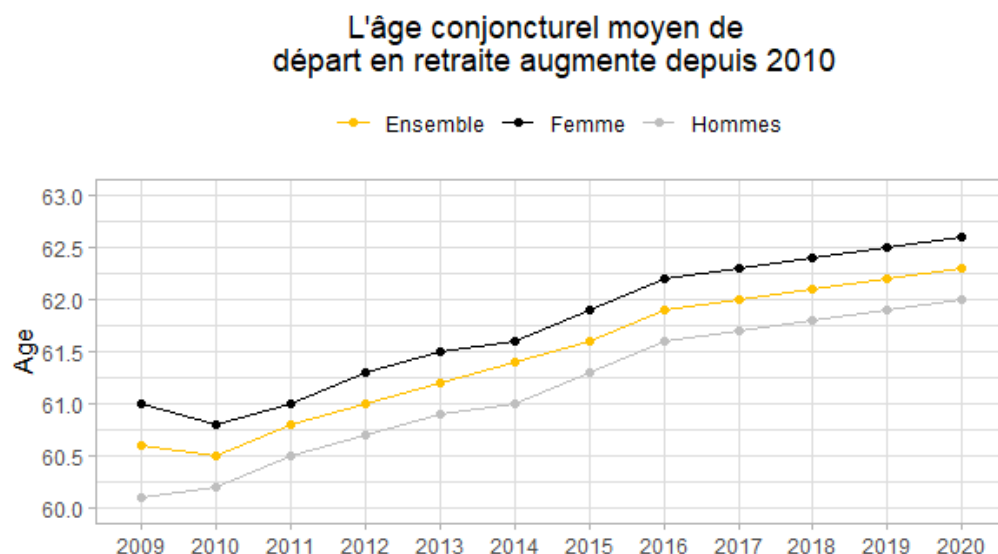


FIGURE 2.1 – Évolution de l'âge de retraite

Par ailleurs, il est possible d'observer que cet âge conjoncturel peut-être très différent selon la catégorie socioprofessionnelle des travailleurs (cf. référence [25]).

L'âge de retraite est différent selon la catégorie socio-professionnelle du travailleur.

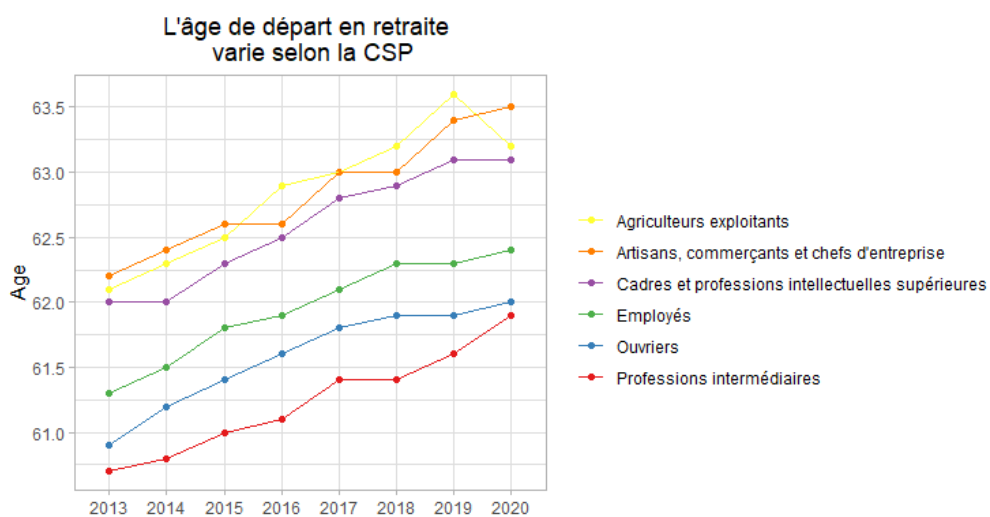


FIGURE 2.2 – Évolution de l'âge de retraite par CSP

Les français partent le plus souvent dès qu'ils ont atteint l'âge d'ouverture des droits. Selon les "Caractéristiques de tous les retraités résidant en France (EIR 2016)" (cf. référence [26]), 41,4% des retraités sont partis à l'âge minimum.

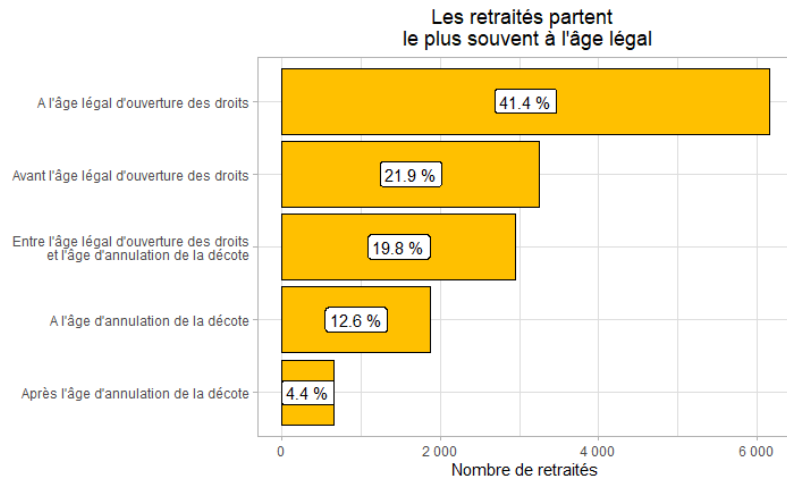


FIGURE 2.3 – Répartition des âges de liquidation

Le taux plein semble être une motivation pour travailler plus longtemps.

L'âge d'ouverture des droits semble donc important mais l'obtention du taux plein semble l'être également. En effet, 36,5% des retraités ont pris leur retraite lorsque leur durée de cotisation était suffisante pour obtenir le taux plein. Les Retraites Anticipées pour Carrière Longue (RACL) ainsi que les retraites avec surcote sont exclus de cette statistique. De plus, l'instauration de la surcote (mécanisme permettant d'augmenter la pension, en travaillant plus longtemps) ne semble pas inciter beaucoup de travailleur à repousser leur retraite : seulement 6 % des retraités sont partis avec une surcote.

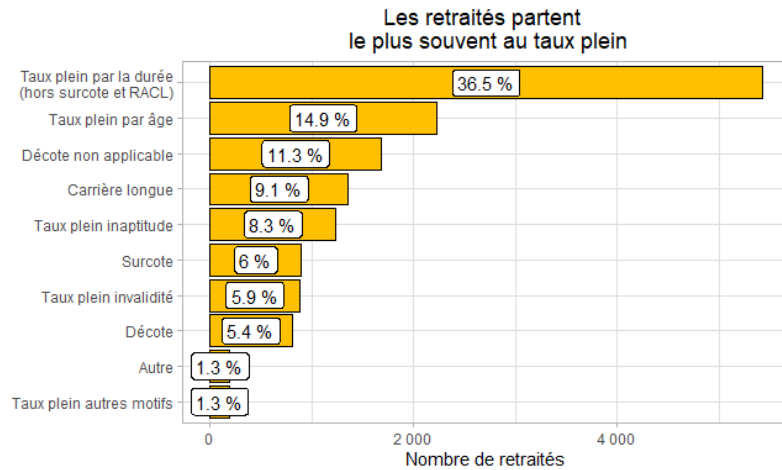


FIGURE 2.4 – Répartition des conditions de liquidation

Le sondage publié par DREES (cf. référence [27]) exposé ci-dessous, démontre bien que l'âge d'ouverture des droits et l'obtention du taux plein sont des paramètres décisifs.

La surcote de la pension ne joue pas un rôle décisif sur le report du départ en retraite.

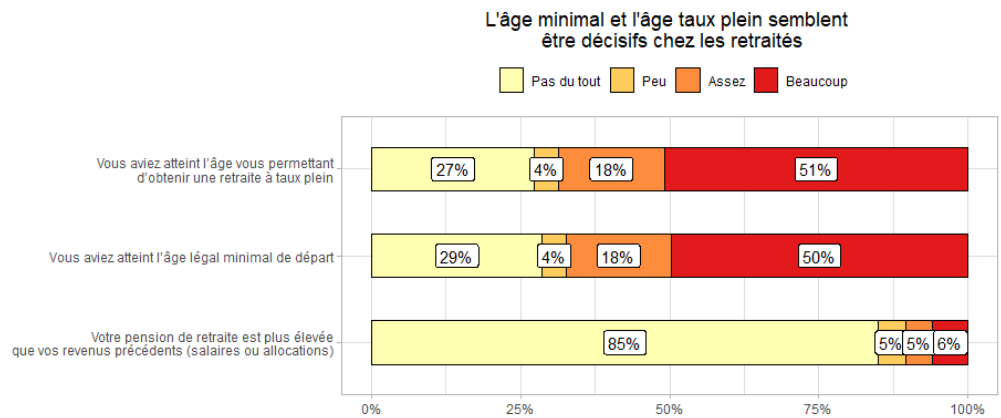


FIGURE 2.5 – Les motivations du départ à la retraite

Enfin, la quasi-totalité des français de plus de 66 ans sont à la retraite . En effet, selon DREES (cf. référence [28]), 80% des français de moins de 63 ans sont à la retraite, ainsi que la quasi-totalité des français de moins de 66 ans.

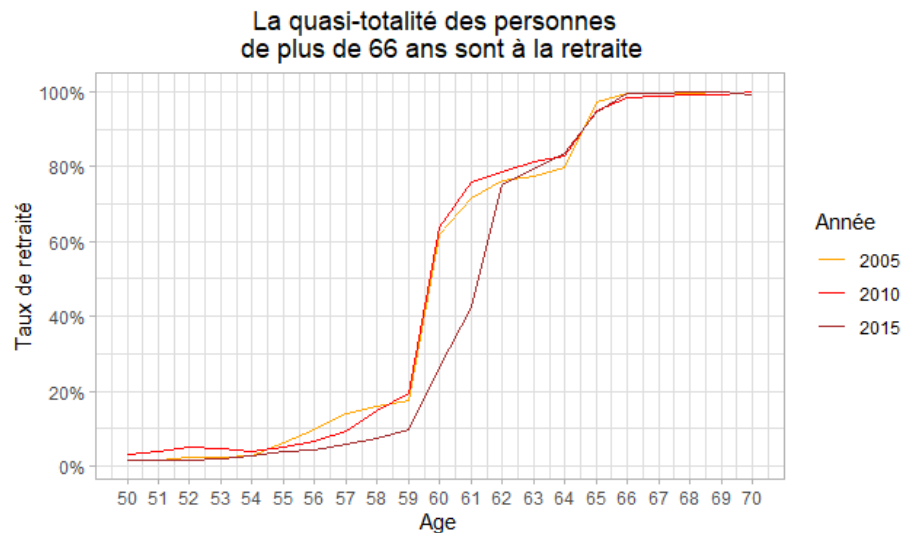


FIGURE 2.6 – Taux de retraités par âge

Deux pics d'âge de retraite sont révélés : entre 59 et 61 ans et entre 64 et 66 ans.

2.2 Présentation de l'IAS 19

2.2.1 L'IAS 19 est une norme qui impose des règles d'évaluation et de comptabilisation d'engagements sociaux.

L'IAS 19 est une norme comptable au coeur de l'évaluation des engagements sociaux.

Le terme "IAS" signifie *International Accounting Standard* (norme comptable internationale). Depuis 1973, il existe un conseil nommé l'IAS (*IAS Board*) qui développe des normes comptables pour établir des états financiers, les publie et facilite leur adoption (ref. [29]). Ce conseil a été créé par les instituts comptables de 9 pays (cf. figure 2.7).

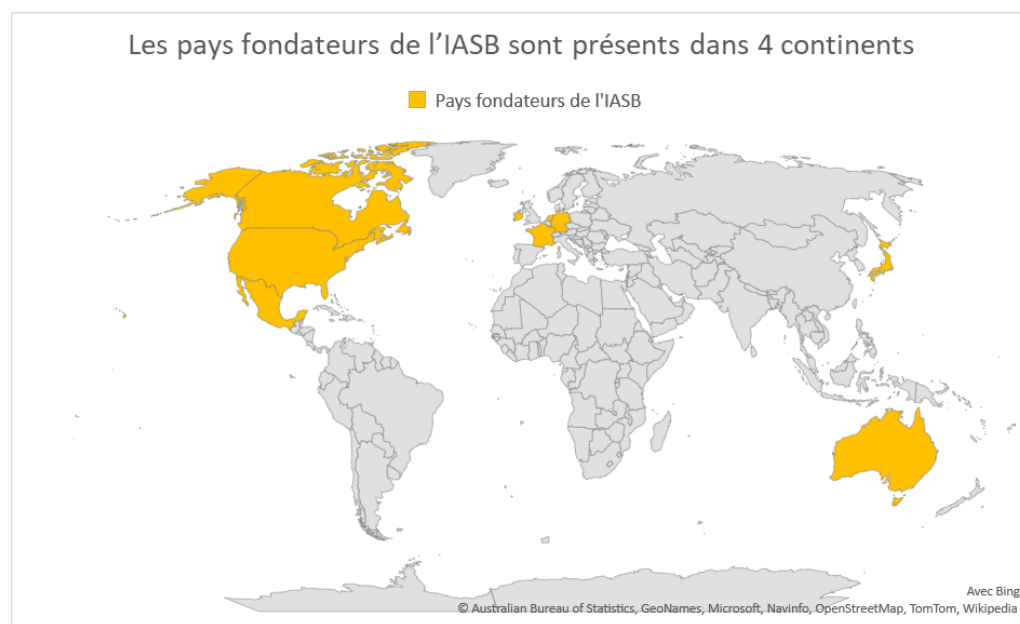


FIGURE 2.7 – Pays fondateurs de l'IASB

L'objectif de l'IASB est d'harmoniser mondialement les états financiers des entreprises. En effet, ce conseil veut que ces états soient comparables et compréhensibles afin de faciliter les transactions économiques internationales, de plus en plus nombreuses. De plus, l'IASB cherche à réduire les coûts engendrés et la complexité rencontrée lors de l'établissement d'un état financier.

L'IAS 19 a été publiée en 1998 par l'*IAS Committee*, devenue l'*IAS Board*.

Depuis le 01/04/2001, les nouvelles normes publiées par l'IASB sont intitulées **IFRS** pour *International Financial Reporting Standards* (normes internationales d'information financière), et non plus **IAS**. Par conséquent, nous parlons du référentiel IAS/IFRS pour prendre en compte l'ensemble des normes établies par l'IASB. D'ailleurs, l'IAS 19 a été publiée par l'IASB (appelé à cette époque, *International Accounting Standards Committee*) en Février 1998 (cf. référence [30]), ce qui justifie son appellation.

La Fondation IFRS assure la bonne organisation de l'IASB. Pour cela, cette instance a écrit un manuel détaillant les procédures que l'IASB doit suivre avant de publier une norme (ref. [31]). Ce dernier doit notamment respecter trois principes :

- Principe de transparence. Les réunions tenues par l'IASB sont alors

La procédure de publication d'une norme est encadrée par la fondation IFRS.

publiques, enregistrées et si possible, rediffusées en direct. De plus, toutes décisions importantes du conseil requièrent un nombre minimal de votes favorables. Par exemple, il est nécessaire d'obtenir la majorité qualifiée (obtenue par scrutin) pour publier une norme, fixée à plus de 60% (9 membres sur 14). D'ailleurs, les 14 membres de l'IASB sont nommés par les 22 membres de la fondation IFRS (appelés les *Trustees*), pour une période de 5 ans (pouvant être renouvelée pour une durée de 3 ans), selon l'article 24 de la constitution de la fondation IFRS (ref. [32]).

- Principe d'équité des traitements. Avant la publication d'une norme, l'IASB doit suivre des étapes permettant d'assurer la pertinence d'une proposition. Par exemple, il doit constituer un groupe de conseillers qui l'accompagnera tout au long du projet.
- Principe de responsabilité. L'IASB doit alors analyser les effets d'une norme, identifier les éventuels problèmes qu'elle pourrait engendrer et trouver des solutions.

L'interprétation des normes est à la charge de l'IFRIC.

Un comité, nommé l'IFRIC pour *IFRS Interpretations Committee* doit fournir des interprétations concrètes des IAS/IFRS. Ce comité, également sous la tutelle de la fondation IFRS, permet donc aux entreprises de concrétiser et d'harmoniser l'application de ces normes.

Chaque juridiction, ayant adopté ce référentiel comptable, doit retransmettre les normes et définir leur périmètre d'application.

Le référentiel comptable de l'IASB est désormais adopté dans plus de 160 juridictions. Selon le site de la fondation IFRS (ref. [33]), les IAS/IFRS sont appliquées (partiellement ou totalement) dans 166 juridictions à travers le monde (chiffre arrêté en février 2022), illustrées dans la figure 2.8 ci-dessous. Chaque juridiction détermine les sociétés concernées par le référentiel IAS/IFRS et retranscrit les normes (qu'elle a choisies) dans leur référentiel local. A titre d'exemple, l'UE a retranscrit la quasi-totalité des IAS/IFRS via le règlement de l'UE n°1126/2008 (ref. [1]).

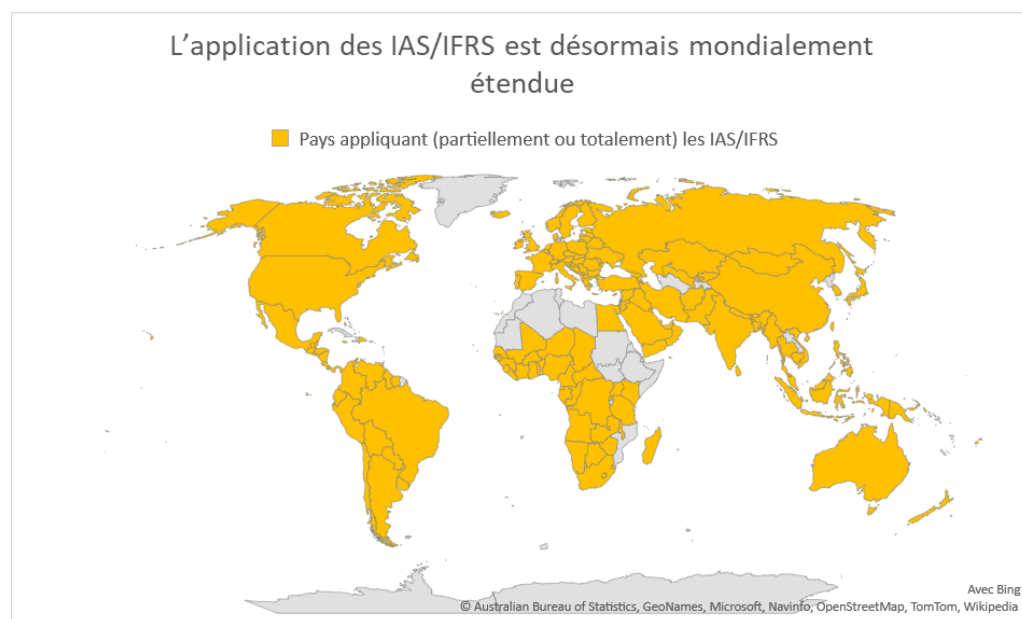


FIGURE 2.8 – Pays appliquant les IAS/IFRS

2.2.2 Au sein de l'UE, l'application des IAS/IFRS est obligatoire pour certaines sociétés.

Les comptes consolidés des sociétés cotées de l'UE doivent être établis selon le référentiel IAS/IFRS.

Depuis 2005, les sociétés cotées doivent appliquer ces normes pour leurs comptes consolidés. Selon l'article 4 du règlement des Communautés Européennes N°1606/2002 (ref. [34]), les entreprises sont obligatoirement concernées par le référentiel de l'IASB lorsqu'à la date de clôture du bilan, leurs titres sont admis à la négociation sur un marché réglementé d'un État membre de l'UE. A titre d'exemple, les marchés réglementés français, selon l'Autorité des Marchés Financiers (ref. [35]), sont : Euronext Paris (marché actions) et Euronext Liffe composé de Monep et MATIF (marchés de produits dérivés). Cette obligation est appliquée seulement pour les comptes consolidés. Selon l'article du cabinet L-Expert-Comptable (ref. [36]), ces comptes sont obligatoires pour les groupes composés d'une société mère détenant plusieurs filiales, dont les résultats dépassent deux des trois seuils suivants :

- Bilan supérieur à 15 millions d'euros,
- Chiffre d'affaire supérieur à 30 millions d'euros,
- Effectif supérieur à 250 personnes.

En France, les comptes consolidés des sociétés non cotées n'ont pas l'obligation de suivre le référentiel IAS/IFRS.

Concernant les sociétés non cotées et les comptes individuels, les Etats membres de l'UE peuvent leur autoriser ou leur imposer l'application de cette norme. L'article 5 du même règlement (ref. [34]) donne effectivement cette possibilité aux Etats membres. Par exemple, la législation française a seulement décidé d'autoriser les sociétés commerciales d'appliquer le référentiel de l'IASB pour établir leurs comptes consolidés (cf. article L. 233-24 du code de commerce). Par conséquent, les comptes individuels des sociétés françaises (cotées ou non cotées) doivent être préparés en conformité avec le référentiel comptable français. Néanmoins, l'Autorité des Normes Comptables a pour objectif de rapprocher le référentiel comptable français du référentiel IAS (ref. [37]). C'est pourquoi, en pratique, les sociétés françaises, qui provisionnent (de manière facultative) leurs engagements de retraite et avantages similaires, appliquent des règles de comptabilisation et d'évaluation conformes à celles de l'IAS 19.

2.2.3 L'IAS 19 impose des méthodes de comptabilisation et de calculs.

L'entité doit comptabiliser un passif lorsqu'elle est engagée à verser, dans le futur, des avantages au personnel. Selon la retranscription de l'IAS 19 par l'UE (ref. [1]), il existe différentes dispositions selon la catégorie dans laquelle se trouve l'avantage du personnel en question. Les catégories possibles et concernées par l'IASB 19, sont listées ci-dessous.

- Les avantages à court terme (ex : salaire, cotisations de sécurité sociale), définis comme des avantages dus intégralement dans les douze mois suivant les services rendus par le personnel qui ont permis d'acquérir ces avantages.
- Les avantages postérieurs à l'emploi (ex : indemnité de départ en retraite) ;
- Les avantages à long terme, n'entrant pas dans la catégorie précédente (ex : médailles de travail), définis comme des avantages qui ne sont pas à court terme.
- Les indemnités de fin de contrat de travail (ex : indemnité de licenciement), définis comme des indemnités versées après la résiliation du contrat de travail (avant l'âge normal de départ à la retraite).

L'indemnité de fin de carrière est un avantage postérieur à l'emploi.

Les dispositions qui nous intéressent sont donc celles des avantages postérieurs à l'emploi. Dans cette catégorie, il est nécessaire de distinguer deux types de régime d'avantages postérieurs à l'emploi :

L'indemnité de fin de carrière suit un régime à prestation définie.

- Les régimes à cotisations définies : l'entité est engagée sur le montant des cotisations. Par conséquent, le risque actuariel (risque que les prestations soient plus coûteuses que prévu) incombe aux membres du personnel.
- Les régimes à prestations définies : l'entité est engagée sur le montant reçu par les membres du personnel. Ainsi, le risque actuariel incombe à l'entité.

L'indemnité de fin de carrière est une prestation définie. Par conséquent, l'étude de l'IAS 19 sera orientée autour de ce type de régime.

L'IAS 19 impose de calculer la valeur actuelle probable des prestations définies pour évaluer l'engagement au titre de cet avantage.

Selon l'article 50 du règlement (ref. [1]), le montant des prestations doit intégrer des hypothèses démographiques (âge de retraite, mortalité, rotation du personnel) et financières (augmentations futures des salaires, taux d'actualisation). Ces hypothèses sont nécessaires pour estimer le montant de la prestation elle-même, ainsi que sa valeur actuelle probable (imposée par la norme). Le montant de la prestation est effectivement inconnue dans la plupart des cas, puisqu'elle dépend généralement des caractéristiques du salarié et de la date de versement. A titre d'exemple, le montant d'une IFC dépend très souvent du salaire et de l'ancienneté dans l'entreprise du salarié lors de son départ en retraite. . Une fois que le montant de la prestation est estimé, la norme impose d'estimer sa valeur actuelle probable. Cette valeur résulte de la multiplication du montant (estimé) de la prestation, par un facteur d'actualisation et la probabilité de versement. La probabilité de versement passe par la probabilité que le salarié soit en vie et que les conditions de versement soit remplies lors du versement. Pour les IFC, seule la condition de présence dans l'effectif de l'entreprise est requise. Le facteur d'actualisation permet de prendre en compte la valeur temporelle de l'argent : celui-ci prend de la valeur simplement parce que le temps passe. Ce phénomène est pertinent dans le contexte de l'évaluation d'engagement au titre d'avantages postérieurs à l'emploi, puisque les prestations risquent d'être versées dans quelques années. Ainsi, les prestations ont aujourd'hui moins de valeur que lors de leur versement : leur valeur actuelle est donc plus faible que leur valeur lors du versement. Par exemple, en supposant que l'argent gagne 2% par an, une prestation de 100 € à verser dans 10 ans, vaut environ 82 € aujourd'hui. En effet, $100 \times (1 + 2\%)^{-10} = 82,03$. Le taux fixé à 2% dans cet exemple est appelé le taux d'actualisation.

L'engagement doit être évalué à partir de la valeur actuelle probable (VAP) des prestations

Le taux d'actualisation est réglementé. Selon l'article 78 du règlement qui retranscrit les IAS/IFRS dans le référentiel comptable européen (ref. [1]), le taux d'actualisation doit faire référence à la valeur d'un taux de marché, prise à la date d'évaluation de l'engagement. Ce taux doit être fondé :

Le taux utilisé pour actualiser la prestation doit respecter certaines conditions.

- sur les taux d'intérêt des obligations¹ d'entités de première catégorie (notées au moins AA),
- dont la monnaie et la maturité sont cohérentes avec la monnaie et la durée de cet engagement².

Puisque la monnaie française est l'euro et que la durée de l'engagement d'une IFC est souvent supérieure à 10 ans, le taux d'actualisation généralement sélectionné est le taux *iBoxx Euro Corporates AA 10+* (taux iBoxx³ des obligations des sociétés cotées AA, du secteur privé de la zone euro, de maturité 10 ans et plus).

L'entité doit utiliser une certaine méthode pour évaluer son engagement à une date précise. Selon l'article 64 du règlement (ref. [1]), la méthode des unités de crédit projetées doit être appliquée pour calculer le montant de la provision à comptabiliser à la date de clôture de l'exercice comptable. Conformément au paragraphe 68 de la norme IAS 19 (ref. [30]), cette méthode consiste à découper en sous-périodes, la période totale de services qui sera effectuée par le salarié en contrepartie de la prestation. Ces sous-périodes donneront lieu ensuite à l'attribution d'une "quantité d'engagement" à la date d'évaluation : plus le nombre de sous-périodes accomplies sera élevé à la date d'évaluation, plus l'entreprise est engagée financièrement (et plus la provision sera élevée). La valeur de l'engagement à une date dépend donc :

L'évaluation de l'engagement à une date, doit être définie de façon à ce qu'il soit proportionnel à la période de services effectués en contrepartie de la prestation, déjà entamée à la date.

- de la définition de la période de services effectués en contrepartie de la prestation ;

1. Une obligation est un morceau de dette émis par une entreprise, une collectivité territoriale ou un État (cf. référence [38]), qui se caractérise notamment par une durée (la durée de l'emprunt) appelée "maturité" et un taux d'intérêts (l'entité s'engage à verser des intérêts auprès de l'acheteur de l'obligation).

2. La durée d'un engagement correspond à la durée de vie moyenne des prestations que l'entité doit verser (autrement dit la durée moyenne des durées entre la date actuelle et les dates de versement), pondérée par la VAP de cette prestation (cf. référence [39]).

3. iBoxx est une branche du service financier de l'entreprise "IHS Markit", chargée notamment de calculer et fournir des indices du marché obligataire.

- de la méthode de découpage de cette période ;
- de la "quantité d'engagement" attribuée à chaque sous-période accomplie à la date d'évaluation.

Cette méthode est indispensable car la norme IAS 19 impose d'évaluer :

- l'engagement à la date d'évaluation (elle correspond à la provision) ;
- l'engagement qui viendra s'ajouter lors de la prochaine évaluation, du fait de l'écoulement de la durée entre les deux dates (elle constitue une charge).

En pratique, cette formule de calcul consiste simplement à proratiser la valeur actuelle probable des prestations selon la période de services déjà écoulée à la date d'évaluation de l'engagement. Par conséquent :

- l'engagement à la date d'évaluation correspond à la proportion de la VAP qui coïncide avec la proportion de la période de services déjà accomplie à la date d'évaluation ;
- l'engagement qui sera ajouté lors de la prochaine évaluation, correspond à la proportion de la VAP qui coïncide avec proportion de la période de services qui sera accomplie entre la date d'évaluation actuelle et celle de la prochaine évaluation.

Par exemple, si le salarié a déjà effectué 75% de la période de services (en contrepartie de l'avantage) à la date d'évaluation, l'engagement s'élève à 75% de la VAP de la prestation qui lui sera versée. Et si la période de services qu'il aura accomplie à la prochaine évaluation augmente de 3%, alors 3% de la VAP de la prestations qui lui sera versée, seront ajoutés à l'engagement.

Les hypothèses actuarielles doivent respecter des critères peu contraignants. Selon l'article 72 du règlement (ref. [1]), les hypothèses doivent être :

- Objectives : c'est-à-dire ni risquées, ni d'une prudence excessive (cf. article 74 du règlement).
- Mutuellement compatibles : c'est-à-dire traduisant les rapports économiques qui existent entre certains facteurs tels que l'inflation, les taux d'augmentation des salaires, les taux d'actualisation.

Puisque le second critère concerne les hypothèses financières, l'hypothèse démographique de l'âge de départ en retraite n'a donc qu'un seul critère à respecter : celui de l'objectivité.

En pratique, la VAP de la prestation est proratisée selon la durée des services déjà effectués par le salarié.

L'hypothèse de l'âge de départ en retraite doit être établie objectivement.

2.2.4 En mai 2021, l'IASB a validé une nouvelle interprétation de l'IAS 19 par l'IFRIC.

Une nouvelle interprétation de l'IAS 19 a conduit à modifier l'évaluation de l'engagement des IFC.

L'IFRIC a rédigé une note en avril 2021 concernant certains avantages postérieurs à l'emploi. Cet addendum (ref. [40]), concerne les régimes à prestations définies :

- Attribuant une prestation lors du départ à la retraite, à condition que l'employé fasse partie de l'effectif de l'entreprise à cette date ;
- Attribuant une prestation dont le montant dépend de la durée de service de l'employé et est plafonné.

L'indemnité de fin de carrière fait donc partie de ces régimes : elle est versée lors du départ en retraite d'un salarié de l'entreprise, elle est proportionnelle à son ancienneté et elle est souvent plafonnée.

La période des services effectués en contrepartie de la prestation est modifiée.

L'IFRIC précise que la période de services à définir pour évaluer l'engagement à une date, doit participer totalement à l'acquisition de l'avantage. Cette note additionnelle concerne donc la méthode pour définir la période de services effectués en contrepartie de l'avantage (paragraphe 70 à 74 de la norme). Selon l'IFRIC, cette période doit donner droits, pour chaque durée de services accomplie dans celle-ci, à plus d'avantages. Prenons l'exemple d'une entreprise qui est engagée à verser à ses salariés, lors de leur départ en retraite (s'ils sont toujours présents dans l'entreprise), une IFC égale à :

- un demi mois de salaire, par année de services accomplis dans l'entreprise,
- sans pouvoir excéder 8 mois de salaire.

En supposant qu'un salarié de cette entreprise y soit entré le 31/12/2000 (à ses 40 ans), puis qu'il soit parti à la retraite le 31/12/2022 (soit 22 ans plus tard, à ses 62 ans), l'IFC qui lui a été versée s'élève alors à 8 mois de son salaire. Néanmoins, si le salarié était entré le 31/12/2006 dans l'entreprise, il aurait perçu la même indemnité (en effet, celle-ci ne peut pas dépasser 8 mois de salaires). Par conséquent, l'IFRIC suppose que la période de services effectuée en contrepartie de l'IFC est celle entre le 31/12/2006 et le 31/12/2022. Avant son interprétation, la période était supposée être celle entre le 31/12/2000 et le 31/12/2022.

En pratique, la prise en compte de cette interprétation peut diminuer la valeur de l'engagement à une date. En reprenant l'exemple au paragraphe précédent, il est possible de constater que la période de services effectués en contrepartie de l'avantage peut être bien plus courte que celle prise en compte avant l'interprétation (elle est passée de 22 ans à 16 ans). Le coefficient de prorata appliqué à la VAP de la prestation est alors plus faible (cf. figure 2.9).

La période des services à prendre en compte peut donc être plus courte, ce qui diminue l'engagement évalué à une date.

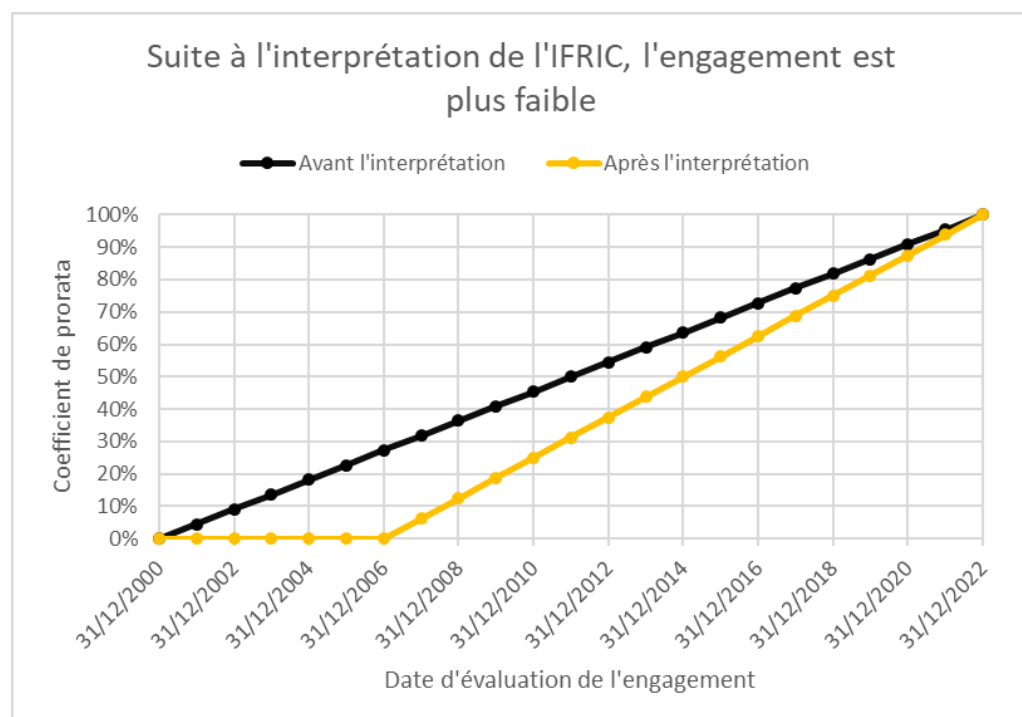


FIGURE 2.9 – Exemple d'un coefficient de prorata appliqué à la VAP de l'IFC pour déterminer l'engagement, en fonction de la date d'évaluation

Toutefois, l'impact de cette interprétation dépend beaucoup de la démographie. En effet, une entreprise ayant une majorité de salariés proches de leur retraite, aura un impact très limité. Alors qu'une entreprise ayant une majorité de jeunes salariés verra son engagement considérablement diminuer. Il est possible de constater ceci à partir du jeu de données mis à disposition pour ce mémoire (cf. graphique ci-dessous).

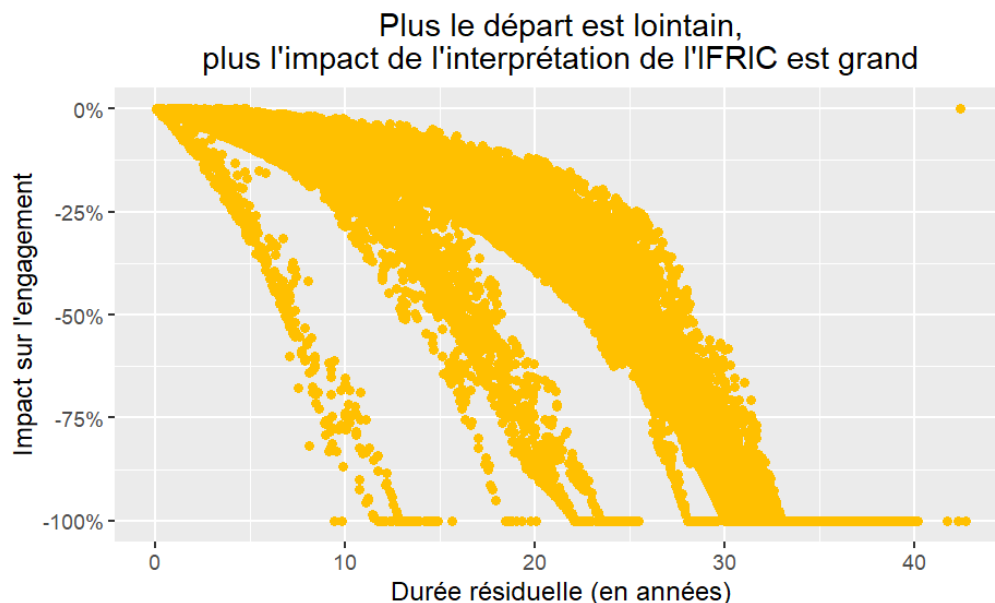


FIGURE 2.10 – Impact IFRIC en fonction de la durée résiduelle.

L'impact de l'interprétation IFRIC évolue directement selon la durée résiduelle de l'engagement.

Ce graphique met en relation l'impact de l'interprétation de l'IFRIC sur l'engagement d'un salarié (en pourcentage), en fonction de sa durée résiduelle (autrement dit, le nombre d'année restant avant son départ théorique à la retraite). Il est ainsi possible de constater une relation proportionnelle : plus le départ du salarié est prévu dans longtemps (plus la durée résiduelle est grande), plus l'impact est grand, jusqu'à annuler totalement celui-ci. De plus, le graphique révèle trois tendances distinctes : elles apparaissent car l'impact est plus important pour les montants d'indemnité élevés (et celle-ci augmente par palier).

2.2.5 Présentation de l'évaluation de l'engagement au titre des IFC selon l'IAS 19.

En France, le code du travail impose aux entreprises de verser une IFC à leurs salariés partant à la retraite : elles ont donc un engagement au titre de cette indemnité. Selon l'article D1237-1 du code du travail, le montant de cette indemnité varie entre 0,5 mois et 2 mois de salaire, selon l'ancienneté du salarié dans l'entreprise (néanmoins, les accords d'entreprise et les conventions collectives nationales prévoient fréquemment des montants supérieurs). Selon l'IAS 19, cette indemnité est un avantage postérieur à l'emploi (car elle est versée au départ à la retraite), qui s'inscrit dans un régime à prestations définies (puisque son montant est

Le versement d'une IFC est obligatoire en France.

garanti). L'évaluation de l'engagement au titre des IFC est donc déterminé selon la méthode des unités de crédits projetées, pour certaines sociétés. En effet, l'application des IAS/IFRS est obligatoire pour les comptes consolidés des sociétés françaises cotées (cf. 2.2.2). Par conséquent, ces sociétés sont tenues de calculer leur provision au titre des IFC selon la norme IAS 19 qui impose d'appliquer la méthode de crédits projetés à la section 2.2.3.

L'évaluation de l'engagement d'une société passe par l'évaluation de son engagement envers chacun de ses salariés. Tout d'abord, l'engagement d'une société au titre des IFC est appelé PBO pour *Projected Benefit Obligation*. Cet engagement à la date d'évaluation E correspond à la somme des engagements envers chacun de ses salariés, évalués à la même date E . Autrement dit, pour S , le nombre de salariés dans l'entreprise, nous avons :

$$\text{PBO}(E) = \sum_{s=1}^S \text{PBO}_s(E).$$

L'engagement vis-à-vis d'un salarié dépend de la VAP de la prestation qui lui sera versée et de la période qu'il a déjà accomplie en contrepartie de cette prestation. En reprenant les notations ci-dessus, l'engagement vis-à-vis d'un salarié vérifie la formule ci-dessous :

$$\text{PBO}_s(E) = \text{VAP}_s(E) \times \frac{\text{durée entre } D_s \text{ et } E}{\text{durée entre } D_s \text{ et } R_s},$$

où :

- $\text{VAP}_s(E)$ est la VAP à la date d'évaluation E de la prestation prévue d'être versée au salarié s ,
- D_s est la date de début de la période de services effectués en contrepartie de l'IFC du salarié s ,
- R_s est la date de retraite prévue pour le salarié s .

L'hypothèse de l'âge de retraite intervient donc de manière évidente dans le coefficient de prorata de la VAP de l'IFC (à travers le terme R_s), mais également dans le calcul de la VAP elle-même. En effet, la VAP de la prestation est obtenue à partir de la formule suivante :

$$\text{VAP}_s(E) = \text{IFC}_s(R_s) \times p_s(R_s) \times (1 + r(E))^{-(\text{durée entre } E \text{ et } R_s)},$$

où :

L'évaluation de l'engagement résulte de calculs individuels.

L'engagement vis-à-vis d'un salarié correspond à la VAP de l'IFC prévue, proratisée selon la durée de services effectuée.

La VAP s'obtient à partir d'hypothèses démographiques et financières.

- $\text{IFC}_s(R_s)$ est le montant de l'IFC qu'il est prévu de verser au salarié s lors de son départ en retraite à la date R_s ,
- $p_s(R_s)$ correspond à la probabilité de versement de l'IFC au salarié s à la date de retraite R_s ;
- $r(E)$ correspond au taux annuel d'actualisation à la date d'évaluation E .

Pour rappel, le montant de l'IFC est une estimation puisqu'elle dépend de l'ancienneté dans l'entreprise du salarié et de son salaire à la date de sa retraite R_s . De plus, la probabilité de versement doit nécessairement prendre en compte deux phénomènes (selon l'article 73 du règlement qui retranscrit les IAS/IFRS dans le référentiel comptable européen, ref. [1]) :

- la mortalité (le salarié doit être vivant à la date R_s),
- la rotation du personnel (le salarié doit être présent dans l'entreprise à la date R_s).

Comme évoqué précédemment, l'IAS 19 n'impose pas de critères contraignants sur les hypothèses démographiques permettant de calculer ces probabilités, mais la pratique en France est très uniformisée.

La probabilité de versement doit prendre en compte la probabilité de survie du salarié à sa retraite.

La probabilité de survie est généralement déterminée à partir de tables exogènes. L'IAS 19 précise que cette probabilité doit faire référence à une table de mortalité qui représente au mieux la population concernée (cf. paragraphe 81), et suivre l'évolution de la mortalité au fil du temps (cf. paragraphe 82). Ces contraintes supposent donc de préférer les tables endogènes aux tables exogènes. Cependant, l'établissement de telles tables requièrent un grand nombre d'individus étudiés et un grand nombre d'années d'observation, pour qu'elles soient robustes. Par conséquent, les tables de mortalité utilisées pour l'évaluation des IFC sont généralement des tables exogènes. Comparé à des engagements viagers, l'engagement au titre des IFC est court (généralement autour de 15 ans), ce qui justifie l'utilisation de tables de mortalité instantanées (au lieu des tables générationnelles). Ainsi, il est très commun d'appliquer les tables de l'INSEE (une table pour les hommes, et une autre pour les femmes) pour les entreprises françaises, puisque :

- elles sont basées sur la population française,
- elles sont statiques,

- l'INSEE valide régulièrement de nouvelles tables, pour prendre en compte l'évolution de l'espérance de vie.

Ces tables associent un nombre de survivants pour chaque âge, en partant d'une population de 100 000 individus. Pour obtenir la probabilité de survie d'un individu à son départ à la retraite, il suffit alors de diviser le nombre de survivants à son âge de retraite par le nombre de survivants à son âge actuel.

La probabilité de versement doit aussi prendre en compte la probabilité que le salarié soit présent dans l'entreprise à sa retraite.

La probabilité de rotation du personnel (turnover) se repose, quant à elle, sur les démissions observées dans l'entreprise. Aucune précision n'est ajoutée dans la version de l'IAS 19, sur cette probabilité. Néanmoins, la CNCC (Compagnie Nationale des Commissaires aux Comptes) a publié un avis (EC 2018-17, cf. référence [41]) qui souligne le fait que la probabilité de turnover doit correspondre à la probabilité de démissions : les licenciements, les ruptures conventionnelles et tout autre type de sorties ne doivent pas être pris en compte pour établir l'hypothèse sur le turnover de l'entreprise. En pratique, cette hypothèse se traduit par une table qui associe une probabilité de démissionner au cours de l'année suivante, à un âge. Pour obtenir la probabilité que le salarié n'ait pas démissionné à la date de retraite, il suffit de multiplier successivement les probabilités de ne pas démissionner jusqu'à l'âge de départ en retraite.

2.3 L'importance de l'hypothèse de l'âge de retraite lors d'une évaluation de l'engagement au titre des IFC

2.3.1 L'engagement est sensible à cette hypothèse

L'hypothèse de l'âge de départ est une hypothèse centrale dans l'évaluation. En effet, si l'âge de départ en retraite est modifié, alors :

- Le coefficient de prorata de la VAP de l'IFC sera différent ;
- Le montant de l'IFC sera également impacté car il dépend de l'ancienneté et du salaire du salarié à sa retraite ;
- La probabilité de versement sera modifiée puisque la date de versement correspond à la date de retraite ;
- Le coefficient d'actualisation sera différent puisque la durée entre E et R_s sera modifiée.

L'hypothèse de l'âge de départ en retraite est omniprésente dans le calcul de l'engagement.

L'engagement IFC est donc sensible à cette hypothèse

Par conséquent, l'hypothèse de l'âge de départ en retraite peut considérablement modifier le résultat d'une évaluation. Un mauvais calibrage de celle-ci entraînerait donc une mauvaise représentation de la dette de l'entreprise face à cet engagement. A partir des données présentées plus loin, il est possible de constater (en appliquant l'interprétation de l'IFRIC) nettement l'impact d'un changement de date de retraite sur l'engagement (cf. graphique ci-dessous).

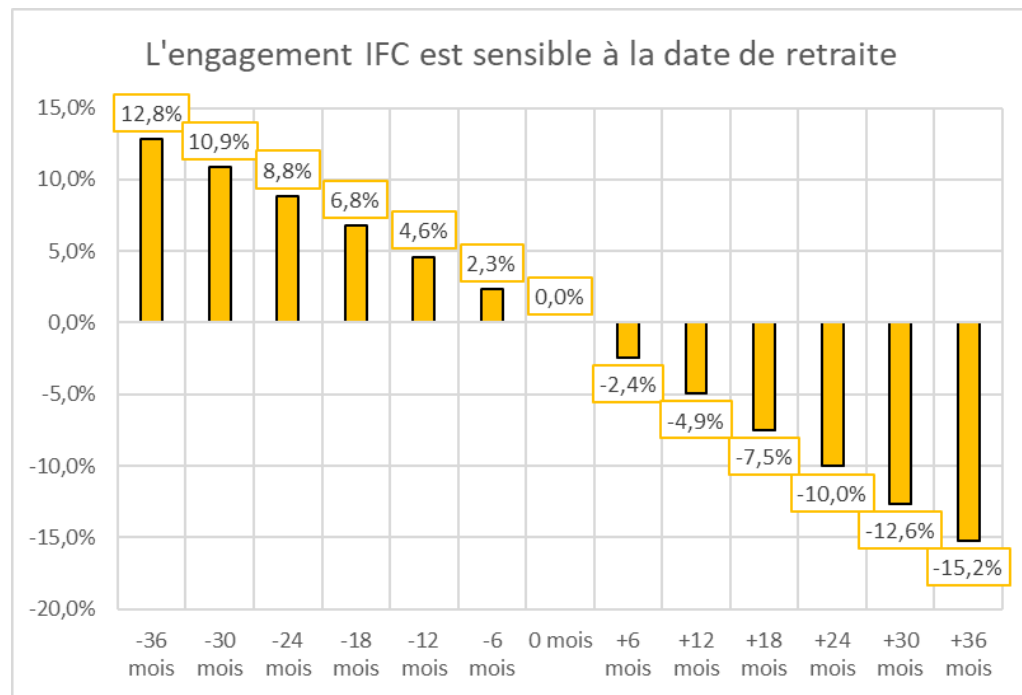


FIGURE 2.11 – Sensibilité de l'engagement IFC face à l'âge de retraite

2.3.2 L'interprétation de l'IFRIC a accru l'importance de bien calibrer l'âge de retraite.

L'interprétation de l'IFRIC a intensifié cette sensibilité.

La sensibilité de l'engagement face à l'âge de retraite s'est intensifiée. En effet, le graphique ci-dessous compare l'impact du changement de l'âge de retraite, entre un engagement calculé avec l'interprétation IFRIC et un engagement calculé sans l'interprétation IFRIC.

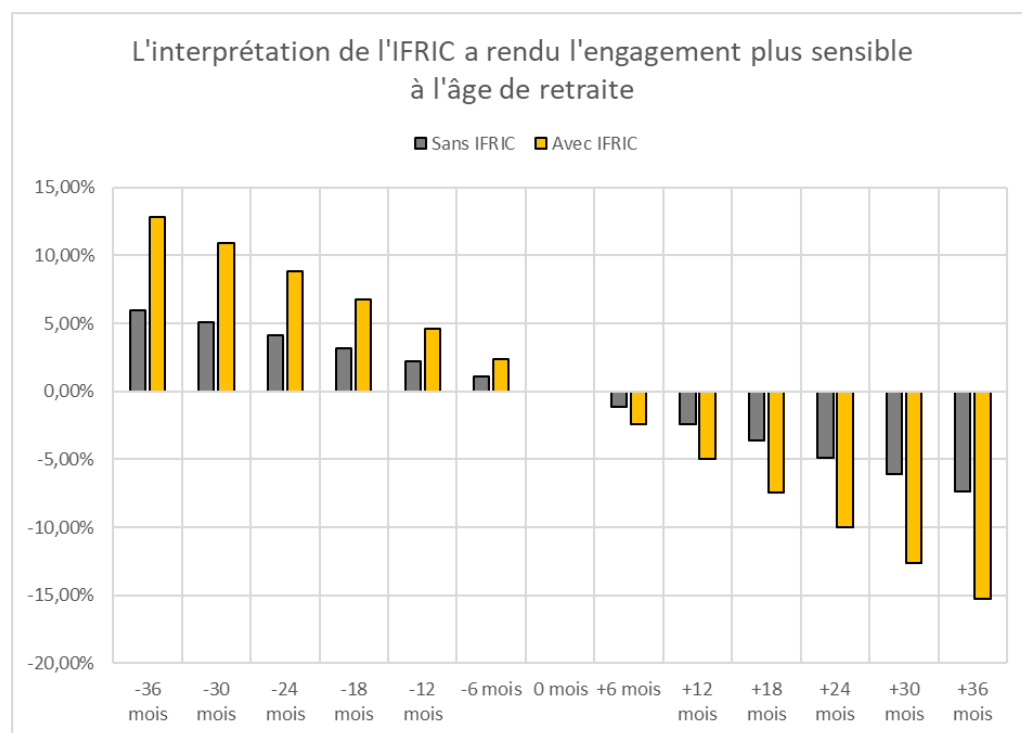


FIGURE 2.12 – Comparaison de l'impact d'un changement d'âge de retraite avec/sans la prise en compte de l'interprétation de l'IFRIC.

Chapitre 3

Méthodologie adoptée pour répondre à la problématique

Ce chapitre présente le modèle de référence, les données et la méthodologie adoptée.

La problématique de ce mémoire est la suivante : existe-il des modélisations de l'âge de départ en retraite qui sont plus performantes que celle utilisée actuellement dans une entreprise, pour évaluer son engagement IFC ? Le modèle d'âge de retraite utilisé actuellement dans l'entreprise étudiée, constituera donc le modèle de référence de ce mémoire, et sera présenté dans la première partie de ce chapitre. De plus, l'entreprise étudiée fait partie des plus grandes entreprises françaises : elle dispose donc d'un grand nombre de données. Il a alors été décidé de modéliser l'âge de retraite en se basant exclusivement sur les nombreuses observations sur les ex-salariés partis en retraite de l'entreprise. Ces données, au coeur de l'étude, sont présentées dans la seconde partie de ce chapitre. L'entreprise a également fourni des données liées à l'évaluation de l'engagement IFC (bases de données d'actifs et hypothèses d'évaluation), pour que l'étude s'inscrive au mieux dans ce contexte d'engagement social. Ces données seront donc également présentées dans cette seconde partie. D'autre part, la problématique de ce mémoire sous-entend de construire des modèles destinés à être appliqués sur de futures bases de données d'actifs (puisque le périmètre d'évaluation de l'engagement IFC correspond aux salariés actifs à date). Ainsi, les modèles doivent avoir un meilleur pouvoir prédictif que celui du modèle actuel. Or, le pouvoir prédictif s'oppose au pouvoir explicatif dans la modélisation statistique (d'après Leo Breiman, cf. référence [42]) : l'accent est mis sur la qualité des prédictions obtenues à partir de nouvelles données, et non sur la qualité des prédictions obtenues à partir des données observées. Ceci a donc

des conséquences sur les méthodologies adoptées lors des 3 grandes étapes de modélisation : l'analyse et le traitement des données sur les retraités, la construction des modèles et leur évaluation. Ces points sont explicités et justifiés (avant d'être appliqués dans le chapitre suivant) dans les trois dernières parties de ce chapitre.

3.1 Présentation de la modélisation de l'âge de retraite actuellement utilisée dans l'entreprise

Les modèles généralement rencontrés dans ce contexte sont moins individualisés et moins sophistiqués que le modèle de référence.

En général, les modèles utilisés pour évaluer l'engagement IFC d'une entreprise sont construits indépendamment des départs en retraite intervenus dans l'entreprise : ils consistent simplement à fixer un âge de retraite pour les non-cadres à 62-63 ans et pour les cadres à 64-65 ans. Ces modèles sont ainsi construits pour deux raisons : les non-cadres commencent généralement leur carrière professionnelle plus tôt (19-20 ans) que les cadres (21-22 ans) ; les salariés partent à la retraite lorsqu'ils ont validé la durée de cotisation minimale pour obtenir la retraite à taux plein (43 ans pour les générations 1978 et après, cf. tableau 2.2). Ainsi, dans la majorité des entreprises, seule la catégorie socio-professionnelle (cadre/non-cadre) des salariés est prise en compte. Toutefois, l'entreprise étudiée dans ce mémoire ne fait pas partie de cette majorité : elle applique un modèle plus individualisé et sophistiqué. Ce modèle est donc présenté dans cette partie car il est rarement rencontré lors d'une évaluation d'engagement IFC, et il représente le modèle de référence de ce mémoire.

3.1.1 Le modèle se base sur trois caractéristiques des travailleurs pour calculer deux âges de retraite.

Le modèle de référence calcule l'âge d'ouverture des droits de chaque salarié, ainsi que leur âge taux plein.

La génération de naissance et le statut des salariés sont utilisés pour déterminer leur âge d'ouverture des droits à la retraite. En effet, d'après le tableau 2.1, l'âge minimal auquel un salarié peut prétendre partir à la retraite dépend de sa date de naissance, et de son statut, c'est-à-dire s'il est un salarié du secteur privé, un fonctionnaire de la catégorie active (emplois ayant un risque) ou un fonctionnaire de la catégorie sédentaire (autres emplois). Le modèle utilisé par l'entreprise étudiée dans ce mémoire calcule donc de manière individuelle, l'âge d'ouverture des droits des travailleurs.

La catégorie socio-professionnelle du salarié est également prise en compte pour estimer son âge d'obtention de la retraite au taux plein. Cet âge correspond au minimum entre :

- l'âge d'acquisition de la durée de cotisation requise pour le taux plein (cf. tableau 2.2),
- l'âge du taux plein automatique (cf. tableau 2.3).

L'âge taux plein correspond au minimum entre deux âges : l'âge d'acquisition d'une certaine durée de cotisation et l'âge taux plein automatique.

A l'instar de l'âge d'ouverture des droits, l'âge du taux plein automatique est entièrement défini par la génération de naissance et le statut du salarié (cf. tableau 2.3). Ce n'est toutefois pas le cas pour l'âge d'acquisition de la durée de cotisation requise pour le taux plein : la durée à acquérir est connue (également grâce à la génération de naissance et le statut du salarié, cf. tableau 2.2), mais l'âge auquel le salarié aura acquis cette durée est bien plus difficile à connaître. En théorie, pour savoir précisément quand le salarié aura acquis cette durée, il est nécessaire de posséder ses rémunérations annuelles perçues tout au long de sa carrière. En effet, un trimestre de cotisation est acquis lorsque la rémunération du salarié est au moins équivalente à 150 heures de travail payé au Salaire Minimum Interprofessionnel de Croissance (SMIC), sans pouvoir valider plus de 4 trimestres par an. Par conséquent, l'âge auquel le salarié aura acquis une certaine durée de cotisation est en pratique inconnue pour l'entreprise car :

- la durée de cotisation qu'il a acquise avant d'entrer dans l'entreprise est inconnue : l'entreprise ne possède pas ses anciennes fiches de salaire ;
- la durée de cotisation qu'il acquerra jusqu'à son départ à la retraite est inconnue : le salarié peut passer à temps partiel, demander un congé sabbatique, etc. (ce qui modifierait le nombre de trimestres de cotisation qu'il acquiert habituellement jusqu'ici).

C'est donc pour ces raisons que l'entreprise ne peut qu'estimer l'âge d'acquisition de la durée de cotisation requise pour le taux plein, en fixant les hypothèses ci-dessous :

Pour déterminer l'âge d'acquisition d'une certaine durée de cotisation, des hypothèses doivent être fixées.

- les travailleurs ont validé 4 trimestres de cotisation par an, tout au long de leur carrière effectuée jusque-là.
- les travailleurs non-cadres commencent leur carrière professionnelle à 22 ans, et les cadres la commencent à 24 ans. A noter que si l'âge d'entrée dans l'entreprise d'un salarié est plus faible que ces âges théoriques, le modèle retient évidemment l'âge d'entrée dans l'entreprise.

- les travailleurs valideront 4 trimestres de cotisation par an, jusqu'à leur départ en retraite.

Cet âge est donc estimé en ajoutant la durée de cotisation requise (cf. tableau 2.2) en année, à l'âge de début de carrière théorique (ou l'âge d'entrée dans l'entreprise, si celui-ci est plus faible). Finalement, le modèle sélectionne l'âge le plus faible entre l'âge ainsi obtenu et l'âge du taux plein automatique. C'est ainsi que l'âge du taux plein du salarié est obtenu (c'est-à-dire à l'âge de retraite lui permettant d'obtenir une pension à taux plein).

3.1.2 Le modèle suppose que les travailleurs partent entre ces deux âges.

Le modèle attribut un poids à ces deux âges, qui dépend de la génération de naissance du salarié.

Finalement, les travailleurs sont supposés partir entre l'âge d'ouverture des droits à la retraite et l'âge du taux plein. Un poids est alors associé à chacun de ces âges pour réaliser une moyenne pondérée de ces âges et obtenir ainsi l'âge de retraite final estimé par le modèle. Ce poids, qui dépend de la génération de naissance du salarié, est fixé selon une analyse des départs en retraite survenus dans l'entreprise. L'objectif de cette analyse est d'obtenir la proportion des départs survenus à la date d'ouverture des droits, par génération de naissance. Cette proportion permet ensuite de fixer le poids associé à l'âge d'ouverture des droits. La proportion restante permet, quant à elle, d'ajuster le poids associé à l'âge du taux plein. Toutefois, ce processus de fixation des poids rencontre plusieurs problèmes : comment les proportions de départs survenus à la date d'ouverture des droits sont-elles déterminées, alors qu'en pratique il est rare qu'un départ survienne le jour où le salarié atteint l'âge d'ouverture des droits ? Quel poids est fixé aux générations de naissance dont aucun départ en retraite n'est encore survenu ?

Ces poids sont déterminés selon une étude des départs en retraite survenus dans l'entreprise.

Lors de l'analyse, les départs observés sont classés en deux groupes : "Départs à l'âge d'ouverture des droits" et "Départs à l'âge d'obtention du taux plein". Néanmoins, les départs en retraite ne surviennent généralement pas le jour d'anniversaire du salarié, et encore moins le jour d'anniversaire de son âge d'ouverture des droits ou de son âge d'obtention du taux plein. Par conséquent, une règle a dû être fixée pour classer les départs en retraite :

- Lorsque le départ en retraite est survenu avant ou au cours du mois pendant lequel l'âge d'ouverture des droits est atteint, alors ce départ est classé dans le groupe "Départ à l'âge d'ouverture des droits".

- Sinon, il est classé dans le groupe "Départ à l'âge du taux plein".

Une fois que les départs sont classés selon cette règle, il est possible d'obtenir la proportion de départs classés "Départs à l'âge d'ouverture des droits" (ainsi que la proportion de départs classés "Départs à l'âge taux plein"), par génération de naissance observée. Toutefois, les générations de naissances observées sont nécessairement plus anciennes que celles des salariés actifs actuellement dans l'entreprise. Ceci pose un problème puisque l'objectif est d'estimer l'âge de retraite des salariés actuellement actifs dans l'entreprise.

Les étapes de l'étude sont partiellement connues.

Les proportions observées sont ensuite attribuées à des plus jeunes générations. Une hypothèse a donc été fixée pour réaliser cette attribution : la dernière génération de naissance observée dont la proportion de "Départs à l'âge d'ouverture des droits" est nulle, doit coïncider avec la génération de naissance qui a actuellement 62 ans. En effet, cet âge correspond au maximum des âges d'ouverture des droits (cf. tableau 2.1). Par conséquent, les salariés encore actifs à 62 ans, partiront nécessairement à la retraite après leur date d'ouverture des droits. C'est ainsi que les proportions observées sont "décalées" et associées à des générations plus jeunes.

Toutefois, plusieurs étapes réalisées par l'entreprise restent inconnues. Par exemple, à partir de quelle génération observée, il est considéré que les proportions sont faussées ? En effet, il y a nécessairement des générations observées (les plus jeunes) dont les départs en retraite sont tous classés "Départs à l'âge d'ouverture des droits". Ceci s'explique par la troncature à droite des observations : le départ en retraite des générations les plus jeunes ne sont pas encore survenus à la date d'observation, ce qui biaise la proportion de "Départs à l'âge taux plein". L'entreprise n'a pas précisé comment ce phénomène a été traité pour déterminer le poids de l'âge d'ouverture des droits pour 25 générations de naissance.

Finalement, les poids de l'âge d'ouverture des droits utilisés par l'entreprise sont inférieurs à 40%. En effet, le graphique ci-dessous représente ces poids pour chaque génération de naissance (utilisée lors de l'évaluation de l'engagement IFC au 31/12/2019).

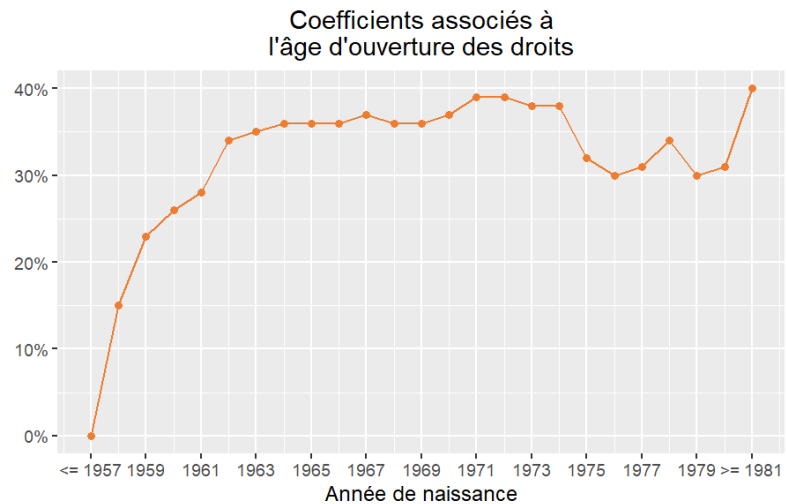


FIGURE 3.1 – Poids de l'âge d'ouverture des droits utilisés lors de l'évaluation de l'engagement IFC de l'entreprise

Pour chaque salarié, un premier engagement IFC est évalué en supposant un départ à la retraite à l'âge d'ouverture des droits, et un second en supposant un départ à l'âge taux plein.

A partir des poids déterminés précédemment, l'engagement final d'un salarié est la moyenne pondérée des deux engagements évalués.

Il est donc considéré, par exemple, que l'âge d'ouverture des droits des salariés nés en 1963 pèse 35% et que leur âge taux plein pèse 65%. C'est ainsi que le modèle estime l'âge de retraite des salariés, mais cet âge moyen n'est en réalité pas utilisé directement pour évaluer l'engagement IFC.

A partir des poids et des âges évoqués précédemment, deux engagements IFC sont évalués par salarié. Le premier engagement est évalué en supposant que le salarié partira à la retraite à l'âge d'ouverture des droits. Le second en engagement est, quant à lui, évalué en supposant que le salarié partira à l'âge taux plein. Enfin, l'engagement final correspond à la moyenne de ces engagements, pondérés par les poids déterminés auparavant.

Finalement, le modèle d'âge de retraite utilisé par l'entreprise est assez complexe. Comparé aux modèles couramment rencontrés lors d'une évaluation d'engagement IFC (présentés en introduction), le modèle de l'entreprise est effectivement plus individualisé et bien plus sophistiqué. Il est plus individualisé car il détermine l'âge d'ouverture des droits et l'âge taux plein des salariés, à partir de leurs caractéristiques individuelles (date de naissance, statut et catégorie socio-professionnelle). Alors qu'en général, les paramètres sont identiques pour tous les salariés (la durée de cotisation nécessaire au taux plein est fixée à 43 ans, pour correspondre à celle des plus jeunes générations). D'autre part, il est plus sophistiqué car il prend finalement en compte 3 âges de retraite : l'âge d'ouverture des droits, l'âge

d'obtention du taux plein automatique, l'âge d'acquisition de la durée de cotisation nécessaire à l'obtention du taux plein. De plus, il prend en compte les départs en retraite survenus dans l'entreprise (pour donner un poids à l'âge d'ouverture des droits), alors qu'en général cette analyse n'est pas effectuée.

3.2 Présentation des données fournies par l'entreprise

Du fait de sa grandeur, l'entreprise étudiée ici dispose de beaucoup d'observations sur les départs en retraite. L'exploitation de cette richesse d'information représente donc un bon moyen de modéliser l'âge de retraite des salariés de l'entreprise. La base de données des retraités est donc présentée en détails dans cette partie. D'autre part, cette modélisation doit s'inscrire dans un contexte particulier : celui de l'évaluation de l'engagement IFC de l'entreprise. Ainsi, l'objectif n'est pas seulement de bien prédire l'âge de retraite, mais également de bien évaluer l'engagement IFC. Pour savoir cela, l'entreprise a fourni les données nécessaires pour évaluer l'engagement à deux dates : au 30/06/2019 et au 31/12/2019. Ces données sont donc également présentées dans cette partie.

La base de données des départs en retraite est au coeur de l'étude.

La base de données des actifs permet de s'inscrire dans le contexte de l'évaluation d'engagement IFC.

3.2.1 La base de données des retraités contient plus de 32 000 observations.

Les 32 851 départs en retraite observés sont survenus entre 2010 et 2019. L'entreprise a donc fourni 10 ans d'historique pour l'étude, ainsi que les dates exactes de départ en retraite. Pour chaque départ en retraite, deux informations personnelles sont fournies :

- La date de naissance : les retraités observés sont nés entre 1942 et 1974 ;
- Le genre : parmi les retraités observés, 70% sont des hommes.

A partir de la date de retraite et de la date de naissance, l'âge de retraite a pu être calculé : celui-ci vacille entre 36,9 ans et 70,8 ans, et il est en moyenne égal à 61,20 ans. Une analyse approfondie sera réalisée, afin d'identifier les âges de retraite aberrants.

Les autres informations fournies pour chaque départs en retraite concernent la situation professionnelle. En effet, la base de données contient :

- **L'âge d'entrée** : la moyenne des âges d'entrée observés s'élève à 23,2 ans, avec un écart-type de 5,3 ans.
- **Le statut** : parmi les retraités observés, la majorité est constituée de fonctionnaires de la catégorie sédentaire (78,7%), puis une grande partie correspond à des fonctionnaires de la catégorie active (15,9%), enfin seulement 5,4% des salariés sont dans le secteur privé.
- **La catégorie définie par la Convention Collective Nationale (CCN)** : 8 catégories sont possibles et définissent le poste du salarié selon sa complexité, son niveau d'autonomie, l'impact des décisions prises par le salarié sur d'autres postes de travail et les connaissances requises. Pour conserver l'anonymat de l'entreprise, ces catégories ont été renommées par les premières lettres de l'alphabet : A à H. Cette variable est ordonnée puisque plus la lettre de la catégorie est éloignée du 'A', plus le poste du salarié est complexe et à hautes responsabilités. Les catégories C et D sont les plus représentées dans la base puisqu'elles contiennent à elles deux 65,2% des observations.
- **Le domaine du métier** : le métier exercé lors du départ en retraite est classé en 5 groupes. Ces groupes sont : "Accompagnement", "Client", "Fonctions support", "Innovation et technologie" et "Domaine non défini". Les domaines les plus présents dans l'entreprise sont "Innovation et technologie" (40,7% des observations) et "Client" (30,1% des observations).
- **Le bassin géographique du lieu de travail** : l'entreprise a de nombreux établissements situés en France (métropolitaine et d'outre-mer). Ces différents lieux ont été regroupés en 7 groupes :
 - L'île de France ;
 - Le Grand Nord Ouest (Normandie, Bretagne, Pays de la Loire, Centre-Val de Loire), qui est la plus représentée dans la base (21,3% des retraités travaillaient dans cette zone avant leur départ) ;
 - Le Grand Nord Est (Hauts-de-France, Grand Est, Bourgogne-Franche-Comté) ;
 - Le Grand Sud Ouest (Nouvelle-Aquitaine et Occitanie) ;

Plusieurs caractéristiques sur les retraités sont fournies : elles constitueront les variables prédictives des modèles.

- Le Grand Sud Est (Auvergne-Rhône-Alpes, Provence-Alpes-Côte d’Azur, Corse).
 - La Réunion ;
 - Les Caraïbes (Guadeloupe, Martinique, Saint-Barthélemy, Saint-Martin, mais également la Guyane et Saint-Pierre-et-Miquelon).
- **Le taux d’activité** : celui-ci vacille entre 20% et 100%.
 - **Le salaire annuel (y compris part variable)** : celui-ci se trouve entre 20 871 € et 168 146 €.

Finalement, beaucoup d’observations sont fournies mais seulement 8 caractéristiques sur les retraités sont connues : le genre, l’âge d’entrée, le statut, la catégorie définie par la CCN, le domaine du métier, le bassin géographique du lieu de travail, le taux d’activité et le salaire. Dans un contexte comme celui des évaluations IFC, il faut noter qu’il est rare d’avoir autant d’informations. Cependant, il a été démontré que de nombreux facteurs jouent un rôle dans la décision du départ à la retraite d’un salarié (cf. référence [2]), comme ceux liés à la situation familiale, l’état de santé, etc. Par conséquent, de multiples facteurs décisifs ne sont pas observés ce qui éloignera nécessairement les prédictions de la réalité.

3.2.2 L’entreprise a fourni des données pour évaluer son engagement IFC à deux dates différentes.

Les hypothèses d’évaluation au 31/12/2018, la base des actifs au 31/12/2018 et celle au 31/12/2019 ont été fournies par l’entreprise. A l’aide de ces données, il est possible de savoir si les modèles construits évaluent bien l’engagement IFC de l’entreprise. En effet, ces données permettent d’évaluer les engagements suivants :

- L’engagement IFC projeté au 31/12/2019 (évalué au 31/12/2018) avec les hypothèses d’évaluation au 31/12/2018 ;
- L’engagement IFC réel au 31/12/2019 avec les mêmes hypothèses d’évaluation.

L’écart entre ces deux engagements correspond finalement à l’écart entre "le prévu" et "le réel" (il est appelé *écart actuariel d’expérience* et il sera explicité dans la dernière partie de ce chapitre). Ainsi, il présente un bon moyen de savoir si le modèle d’âge de retraite utilisé évalue bien l’engagement IFC. C’est pourquoi l’entreprise a fourni ces données, en complément de celles concernant les retraités.

Deux bases d’actifs, ainsi que les hypothèses d’évaluation ont été fournies pour savoir si les modèles d’âge de retraite permettent de bien évaluer l’engagement IFC de l’entreprise.

Les hypothèses d'évaluation (hors âge de départ en retraite) utilisées sont classiques. Elles sont regroupées en deux groupes : les hypothèses démographiques et les hypothèses financières. Le premier groupe réunit les deux hypothèses suivantes :

- **L'hypothèse de survie**, qui permet de calculer la probabilité que le travailleur soit en vie lors de son départ en retraite. Cette hypothèse se traduit par une table associant un nombre restant de survivants (d'une population initiale de 10 000 personnes) à chaque âge. En pratique, une table est établie pour les hommes et une autre pour les femmes (car il existe une différence significative entre l'espérance de vie des hommes et celle des femmes). Deux types de table de survie existent : les tables générationnelles (proposant une évolution du nombre de survivants pour chaque année de naissance) et les tables non générationnelles. Celles utilisées par l'entreprise correspondent aux tables non générationnelles de l'INSEE, qui sont très souvent utilisées pour l'évaluation de cet engagement. L'INSEE valide des nouvelles tables chaque année. Les dernières tables validées par l'INSEE au 31/12/2018 correspondaient aux tables 2012-2014.
- **L'hypothèse de *turnover***, qui permet de calculer la probabilité que le travailleur n'ait pas démissionné avant son départ en retraite. Cette hypothèse est également traduite par une table, associant cette fois-ci un taux annuel de démissions à chaque âge. Contrairement aux tables de survie, elles ne sont pas exogènes : elles sont construites à partir d'une analyse des démissions survenues dans l'entreprise. Plusieurs pratiques sont toutefois d'usage dans le contexte d'évaluation d'engagements sociaux : les taux de démission sont décroissants avec l'âge et les taux s'annulent à partir de 60 ans (voire 55 ans). Il existe deux types de table de turnover : les tables par paliers (constantes par tranche d'âge) et les tables lissées. La table de turnover utilisée par l'entreprise est par palier et les probabilités vacillent entre 0% et 3,5%.

Les tables de survie utilisées sont celles validées par l'INSEE

L'entreprise utilise une unique table de turnover par paliers.

Tranche d'âge	Probabilité de démission au cours d'année
< 30 ans	3,5%
30 - 40 ans	2,4%
40 - 50 ans	1,1%
50 - 60 ans	0,3%
>= 60 ans	0,0%

FIGURE 3.2 – Table de turnover appliquée lors des évaluations d’engagement IFC

Les hypothèses financières font référence, quant à elles, aux hypothèses suivantes :

- **L’hypothèse d’évolution des salaires**, qui permet d’estimer le salaire que le travailleur percevra avant de partir en retraite. A l’instar de l’hypothèse de turnover, l’hypothèse d’évolution des salaires est établie à partir d’une analyse des évolutions de salaire observées dans l’entreprise. Elle doit représenter également la politique salariale adoptée par l’entreprise. Cette hypothèse se traduit généralement par un taux d’évolution annuelle par catégorie socio-professionnelle. Il arrive toutefois qu’une table par âge et par catégorie socio-professionnelle soit établie pour cette hypothèse : c’est le cas pour l’entreprise étudiée ici.

L’entreprise utilise deux tables d’évolution annuelle des salaires par âge : une pour les non-cadres et une pour les cadres.

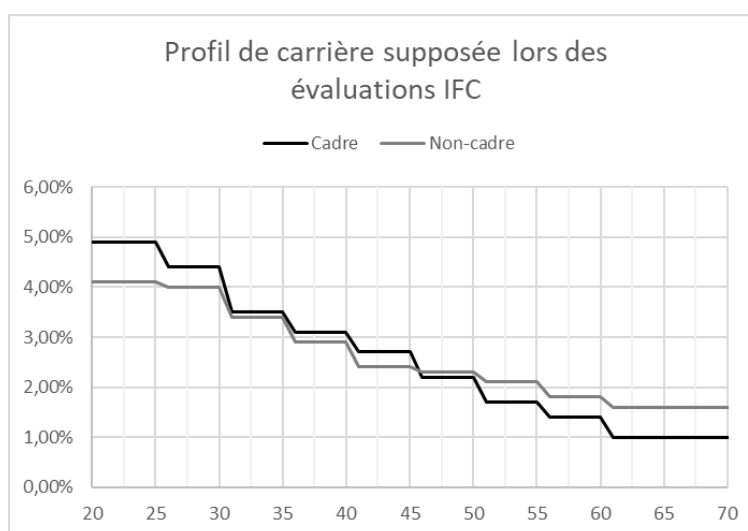


FIGURE 3.3 – Tables des taux d’évolution annuelle des salaire appliqués lors des évaluations d’engagement IFC

- **L’hypothèse du taux de charges sociales patronales**, qui permet d’estimer le coût total de l’IFC pour l’entreprise. Cette hypothèse est généralement donnée par l’entreprise, qui retient la moyenne des taux appliqués. L’entreprise étudiée dans ce mémoire a indiqué qu’elle appliquait en moyenne 46% de charges patronales sur les revenus de ses salariés.

- **L'hypothèse du taux d'actualisation**, qui permet d'estimer la valeur de l'IFC à la date d'évaluation (cf. partie 2.2.3 pour plus de détails). Le taux d'actualisation est réglementé, ce qui explique qu'en pratique, le même taux est utilisé entre les entreprises françaises : il s'agit du taux iBoxx Corporate Bonds AA 10+. Ce taux représente les taux d'obligation du maturité supérieure à 10 ans, du marché des entreprises cotées et évaluée AA. Au 31/12/2018, ce taux s'élevait approximativement à 1,55%.

Finalement, toutes les hypothèses d'évaluation utilisées par l'entreprise sont classiques, contrairement à l'hypothèse d'âge de retraite (présentée en détails dans la première partie de ce chapitre).

La table des droits IFC est également classique. Elle attribue un nombre de mois de salaire pour plusieurs tranches d'ancienneté. Elle est croissante et par paliers, ce qui est très répandu pour le calcul des IFC. Le fait qu'elle soit par paliers justifie d'ailleurs que l'entreprise soit concernée par l'interprétation de l'IFRIC validée en mai 2021 (cf. partie 2.2.4 pour plus de détails). A noter que même si les dates d'évaluation sont antérieures à la validation de cette interprétation, celle-ci a tout de même été appliquée pour représenter au mieux les futurs écarts entre "le prévu" et "le réel" rencontrés. La table des droits IFC est définie par la Convention Collective Nationale (CCN) dont dépendent les salariés de l'entreprise. Afin de respecter l'anonymat de l'entreprise, elle ne sera pas explicitée.

La table des droits IFC est unique et par paliers.

Les bases des actifs sont constituées des travailleurs titulaires d'un contrat à durée indéterminée. En effet, les IFC ne sont pas versées aux autres travailleurs, ce qui explique qu'ils soient exclus de l'évaluation. D'autre part, ces bases contiennent les caractéristiques des salariés, qui sont nécessaires pour l'évaluation :

Plusieurs caractéristiques des actifs sont fournies pour évaluer l'engagement IFC.

- Le genre (utile pour l'hypothèse de survie) ;
- La date de naissance (utile pour déterminer la date de retraite, la probabilité de turnover, le coefficient de revalorisation du salaire, etc.) ;
- La date d'ancienneté (utile pour calculer le montant de l'IFC, puisqu'elle dépend de l'ancienneté acquise) ;
- La catégorie socio-professionnelle (utile pour l'hypothèse de turnover) ;
- Le salaire (utile pour déterminer le montant de l'IFC, puisqu'elle est exprimée en mois de salaire).

Finalement, l'entreprise a fourni tous les éléments nécessaires pour évaluer l'écart entre l'engagement prévu et l'engagement réel au 31/12/2019. Cet écart, appelé *écart actuariel*, peut ainsi faire partie des critères d'évaluation des modèles (cf. dernière partie de ce chapitre).

3.3 Méthodologie adoptée lors du traitement des données

L'analyse des données permet d'améliorer la modélisation.

Avant toutes modélisations, les données utilisées pour construire le modèle (à savoir, les données sur les retraités) doivent être analysées en détails. Cette analyse amène généralement à réaliser des traitements sur les données pour améliorer la modélisation. Elle peut se scinder en deux étapes : l'analyse univariée (étude des variables une à une) et l'analyse bivariée (étude de la liaison entre deux variables). La première étape permet de faire plusieurs "découvertes" : les erreurs de saisie, les modalités peu représentées (ayant un faible effectif) et les valeurs extrêmes (valeurs éloignées de la moyenne) de la variable à prédire. La seconde étape permet, quant à elle, de mettre en évidence les couples de variables liées. Lors de chacune de ces "découvertes", les questions suivantes sont posées : est-ce que cela risque de détériorer la modélisation ? Si oui, quel traitement faut-il appliquer pour éviter cela ? Dans le contexte de ce mémoire, il sera considéré qu'il y a un risque de détériorer la modélisation, lorsque le pouvoir prédictif du modèle risque de diminuer. En effet, l'objectif de ce mémoire est de construire un modèle qui sera utilisé lors de l'évaluation de l'engagement IFC d'une entreprise. Le modèle construit est donc destiné à être appliqué sur de futures bases de données de salariés actifs : il doit ainsi savoir prédire l'âge de retraite d'un individu qu'il n'a pas observé auparavant. A titre d'exemple, les erreurs de saisie risquent de détériorer la modélisation car le modèle construit aura considéré que ces observations sont vraies et risquera de répercuter cette fausse information sur ses prédictions. S'il n'est évidemment pas possible de les corriger, le traitement pour éviter ce risque consiste à retirer ces observations de la base. Concernant les autres "découvertes" citées, les réponses à ces questions méritent d'être développées : l'objectif de cette partie est donc d'expliquer les problèmes causés par ces "découvertes", ainsi que les traitements choisis pour les pallier (l'application de ces traitements sera explicitée dans le prochain chapitre). Avant d'évoquer les "découvertes" citées, cette partie développera un problème (et les traitements choisis pour le pallier) spécifique à la construction de modèle prédictif, à savoir : la présence de variables qui dépendent d'une caractéristique qui se retrouvera très peu

Dans cette partie, les traitements de données pour optimiser sont justifiés et explicités.

chez les actifs.

3.3.1 Deux variables dépendent de caractéristiques obso- lètes ou représentatives d'une petite partie de la population.

La variable à prédire correspondra à l'écart entre l'âge de retraite et l'âge d'ouverture des droits.

L'âge de retraite dépend de la génération de naissance. En effet, il est très probable que la date de départ en retraite dépende de la date à laquelle le salarié peut théoriquement partir, c'est-à-dire la date d'ouverture des droits. Or cette date, définie par l'âge d'ouverture des droits, dépend à son tour de la génération de naissance du salarié (cf. tableau 2.1). Ceci pose problème puisque la génération de naissance des salariés actifs est logiquement plus jeune que celle des individus observés : le modèle construit aura observé des générations très peu (voire pas du tout) présentes dans les futures bases des actifs et sa prédiction sur de telles bases risquent ainsi d'être mauvaises. La solution choisie pour pallier ce problème est d'étudier l'évènement "partir en retraite X ans avant/après l'âge d'ouverture des droits", plutôt que l'évènement "partir en retraite à l'âge de X ans". Cet évènement est effectivement considéré être indépendant de la génération de naissance des salariés. Il faudra donc traiter la variable à prédire, de façon à obtenir l'écart entre l'âge de retraite et l'âge d'ouverture des droits. Toutefois, la base de données des retraités est constituée de fonctionnaires, dont l'âge d'ouverture des droits dépend du nombre d'années de services effectués au sein d'un emploi de la catégorie active. Il est effectivement nécessaire d'effectuer une durée de services minimale dans la catégorie active, pour dépendre de son régime de retraite (cette durée varie selon les emplois effectués). Or, cette information n'a pas été fournie par l'entreprise. Ainsi, l'attribution de l'âge d'ouverture des droits a été réalisée en fonction du statut du salarié, lors de son départ en retraite.

Le salaire dépend de l'année de versement. En effet, il existe en France un salaire horaire minimal depuis 1950, appelé "SMIG" (Salaire Minimal Interprofessionnel Garanti), puis renommé "SMIC" (Salaire Minimal Interprofessionnel de Croissance) en 1970. Les employeurs ne peuvent donc pas verser une rémunération plus faible que ce montant. De plus, selon le site officiel de l'administration française (cf. référence [43]), ce montant est automatiquement revalorisé au 1er janvier de chaque année. Le graphique ci-dessous¹ illustre l'évolution du SMIG / SMIC depuis 1950.

1. La France a connu trois monnaies différentes depuis 1950 : l'ancien franc (avant 1960), le franc (de 1960 à 2002) et l'euro (à partir de 2002). Pour obtenir le graphique 3.4

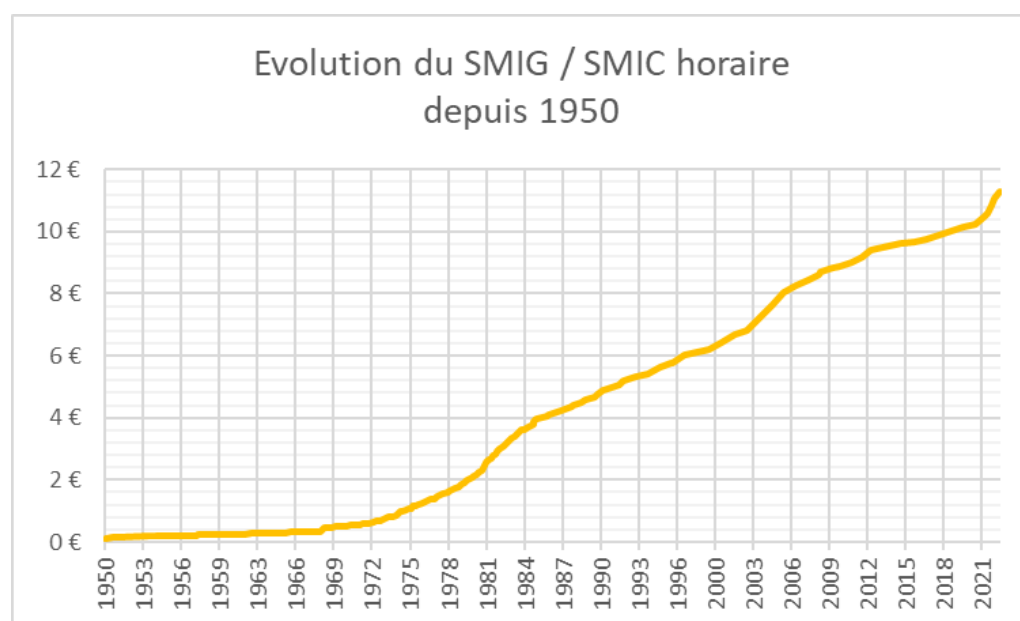


FIGURE 3.4 – Evolution du salaire minimal interprofessionnel depuis 1950

Ainsi, un salaire au SMIC augmente nécessairement avec le temps. De plus, il est fréquent de considérer que ce phénomène impacte indirectement les salaires supérieurs, qui *in fine* augmentent alors également avec le temps. Ceci pose problème lors de la modélisation puisque la présence de la variable "salaire" permet avant tout de représenter le niveau de vie de l'individu. Et un salaire peut être considéré plus faible qu'un autre, alors qu'en réalité le niveau de vie de l'individu est peut-être plus élevé (du fait qu'il ait été versé bien plus tôt). Pour pallier ce problème, il a été décidé de représenter le niveau de vie d'un individu par le rapport entre son salaire (horaire) et le SMIC en vigueur lors du versement (plutôt que le salaire lui-même).

Le niveau de vie d'un individu sera représenté par le rapport entre son salaire et le SMIC, plutôt que son salaire lui-même.

Le salaire dépend également de l'âge de l'individu à la date de versement. Comme le démontre l'hypothèse d'évolution des salaires fixée par l'entreprise (cf. graphique 3.3), l'entreprise adopte généralement une politique salariale qui consiste à augmenter le revenu de leurs salariés selon leur âge. Toutefois, le rapport entre le salaire et le SMIC n'est, quant à lui, pas nécessairement croissant avec l'âge. En effet, le tableau 3.3 a été établi pour l'évaluation de l'engagement IFC au 31/12/2019, alors que l'évolution annuelle du SMIC constatée à cette date s'élevait à 1,52% (puisque le SMIC est passé de 9,88 € au 31/12/2018 à 10,03 € au 31/12/2019). Ainsi, la poli-

en euros, les montant avant 2002 ont été déduits de façon à obtenir les mêmes évolutions du SMIG / SMIC.

tique salariale de l'entreprise sur le rapport entre le salaire et le SMIC peut être illustrée par le tableau ci-dessous.

Tranche d'âge	Cadre		Non-cadre		Moyenne	
	Evolution du SMIC comprise	Au-delà de l'évolution du SMIC	Evolution du SMIC comprise	Au-delà de l'évolution du SMIC	Evolution du SMIC comprise	Au-delà de l'évolution du SMIC
- de 26	4,90%	3,38%	4,10%	2,58%	4,50%	2,98%
26 - 30	4,40%	2,88%	4,00%	2,48%	4,20%	2,68%
31 - 35	3,50%	1,98%	3,40%	1,88%	3,45%	1,93%
36 - 40	3,10%	1,58%	2,90%	1,38%	3,00%	1,48%
41 - 45	2,70%	1,18%	2,40%	0,88%	2,55%	1,03%
46 - 50	2,20%	0,68%	2,30%	0,78%	2,25%	0,73%
51 - 55	1,70%	0,18%	2,10%	0,58%	1,90%	0,38%
56 - 60	1,40%	-0,12%	1,80%	0,28%	1,60%	0,08%
61 et +	1,00%	-0,52%	1,60%	0,08%	1,30%	-0,22%

TABLE 3.1 – Politique salariale de l'entreprise étudiée

Ainsi, le niveau de vie d'un salarié ayant accompli toute sa carrière dans cette entreprise n'a pas été constant au cours de sa vie. Ceci pose problème dans la modélisation, puisque les salaires observés ont été perçus juste avant de partir en retraite, lorsque les salariés avaient des âges élevés. Par conséquent, le modèle construit n'aura observé que les niveaux de vie des individus âgés, alors qu'il est destiné à être appliqué à des bases de données d'actifs contenant également des jeunes individus. Ses prédictions sur ces individus risquent donc d'être mauvaises. Pour pallier ce problème, il a été décidé d'observer le niveau de vie d'un individu à un âge fixé, à partir de son niveau de vie connu. L'âge choisi a été fixé à 15 ans afin de s'assurer qu'il soit possible de réaliser ce traitement pour tous les individus. En effet, il aurait été impossible d'estimer le niveau de vie à 30 ans à partir du niveau de vie d'un individu de 20 ans, puisque le montant du SMIC en vigueur à la date de son 30ème anniversaire est encore inconnu. De plus, pour réaliser cette estimation, il a été décidé d'utiliser la politique salariale moyenne au-delà de l'évolution du SMIC, de l'entreprise (cf. dernière colonne du tableau 3.1). Il est donc supposé que cette politique est représentative des anciennes politiques salariales adoptées par l'entreprise. La politique salariale moyenne a été retenue, plutôt que celle spécifique aux non-cadres et cadres, puisque la caractéristique "non-cadre / cadre" est rarement constante au cours d'une carrière et qu'elle n'est pas observée dans la base de données des retraités.

Le niveau de vie à 15 ans sera une variable prédictive du modèle.

Finalement, plusieurs traitements doivent être appliqués à la variable "salaire" :

- **Etape n°1 : calculer les salaires horaires.** Les salaires communiqués sont annuels, alors que les SMIC sont horaires. Pour les comparer,

il faut donc les exprimer à la même échelle. Pour cela, les salaires communiqués seront divisés par 1 607, qui correspond à la durée légale de travail depuis 2000.

- **Etape n°2 : estimer les salaires horaires perçus lors du 15ème anniversaire des salariés.** Pour cela, la politique salariale moyenne au-delà de l'évolution du SMIC, de l'entreprise est supposée être constante.
- **Etape n°3 : en déduire les niveaux de vie à 15 ans.** Pour cela, chaque salaire estimé est divisé par le SMIC en vigueur lors du 15ème anniversaire des salariés.

La variable prédictive "salaire" du modèle sera donc transformée par la variable "niveau de vie à 15 ans", afin que l'année et l'âge de versement ne jouent pas un rôle dans la modélisation.

3.3.2 La présence d'une caractéristique peu représentée est un risque de biaiser les modélisations.

Une caractéristique ayant un faible effectif a des chances d'être mal représentée. En effet, il y a un risque élevé que les quelques individus ayant cette caractéristique ne soient pas représentatifs de la population. Par conséquent, le modèle risque de mal interpréter le lien entre cette caractéristique et la variable à prédire (et de donner de mauvaises prédictions). C'est pourquoi, il a été décidé d'étudier ces caractéristiques de près. A noter que ces caractéristiques peuvent être une modalité d'une variable qualitative (par exemple, "Grand Nord Est" de la variable "Bassin géographique") mais aussi une valeur considérée "extrême" d'une variable quantitative.

Les modalités ayant moins de 100 individus sont sous-représentées.

L'effectif est considéré faible lorsqu'il ne dépasse pas 100. Ainsi, une modalité d'une variable qualitative est considérée être mal représentée, lorsque son effectif est inférieur à 100. Concernant les variables quantitative, les 50 valeurs les plus faibles et les 50 valeurs les plus fortes sont considérées être des valeurs extrêmes. Toutefois, cette dernière règle n'a de sens que si la répartition des valeurs est en "cloche" (c'est-à-dire lorsqu'il y a peu de valeurs proches du minimum et peu de valeurs proches du maximum), et si les valeurs ne font pas partie de la "cloche". Pour vérifier ces deux points, il sera pertinent de tracer l'histogramme des valeurs (si la répartition n'est pas "en cloche", la variable sera alors considérée comme une variable qualitative) ainsi que le nuage de points sur une dimension, appelé "stripchart" (si les

valeurs considérées extrêmes font partie de la "masse principale" alors il conviendra d'ajuster les seuils de valeurs extrêmes).

Pour les variables quantitatives, les 100 valeurs à la marge sont retirées.

Les modalités mal représentées ont été réunies et les valeurs extrêmes ont été retirées. Le retrait d'observations est plus prudent pour éviter de biaiser la modélisation (et il est envisageable car la base de données des retraités contient beaucoup d'observations). Toutefois, cette pratique n'est pas adaptée lorsque cela implique la suppression d'une modalité. En effet, il est nécessaire que toutes les modalités possibles fassent partie de la base de données des retraités, afin que le modèle puisse prédire une valeur pour tous les individus (dont ceux ayant cette modalité). C'est pourquoi il est plus adapté de réunir les individus d'une modalité avec les individus d'une autre modalité (plutôt que de supprimer ces individus). La réunion de deux modalités doit toutefois être réalisée de façon à ne pas modifier le sens de la variable, ni la modélisation. Ainsi, les règles suivantes ont été définies pour sélectionner la modalité à réunir avec celle qui est mal représentée :

Une modalité sous-représentée est réunie avec une autre.

- Lorsque la variable qualitative est ordonnée, les modalités candidates sont celles qui "entourent" la modalité mal représentée. Si plusieurs modalités sont candidates, la règle appliquée aux variables qualitatives non ordonnées est suivie.
- Lorsque la variable qualitative n'est pas ordonnée, la modalité la plus "proche" de celle mal représentée est sélectionnée. En pratique, il est considéré que deux modalités sont "proches" lorsque leur moyenne sur la variable à prédire est proche.

3.3.3 Les écarts "âge retraite - âge minimal" éloignés de la moyenne doivent être étudiés de près.

Lors d'une modélisation, les valeurs extrêmes de la variable à prédire représentent un risque de biaiser les prédictions. En effet, la présence de ces valeurs peut conduire le modèle à prédire des écarts qui sont plus proches de ces valeurs extrêmes, au détriment des bonnes projections sur d'autres valeurs (non extrêmes). Ce phénomène risque de survenir lorsque le modèle n'a pas pu "justifier" ces valeurs extrêmes, à partir des caractéristiques observées des individus. Par conséquent, il convient d'analyser les caractéristiques des individus ayant des valeurs extrêmes : si elles ne sont pas similaires, ces valeurs extrêmes ont de grande chance d'éloigner les prédictions des valeurs non-extrêmes.

Les valeurs extrêmes de la variable à prédire doivent être analysées.

Avant l'analyse, il a fallu fixer une méthode pour considéré qu'un écart "âge retraite - âge minimal" soit extrême. La méthode choisie consiste à considérer qu'un écart est extrême lorsque la situation de liquidation des droits à la retraite est rare. A première vue, les départs anticipés (avant l'âge d'ouverture des droits) et les départs tardifs (après l'âge limite) semblent être une bonne méthode pour déterminer les situations "extrêmes". Toutefois, ces situations ne sont en réalité pas si rares :

- Selon le rapport "Les retraités et les retraites" (édition 2021) de la DREES (cf. référence 2.1.4), les départs anticipés pour carrière longue représentent 21% des départs en 2019 pour la CNAV et 14% pour la fonction publique civile de l'état.
- Les départs en retraite après l'âge limite ne sont pas rares pour les fonctionnaires de la catégorie active. Ceci s'explique par l'affectation de l'âge limite. A défaut d'information, l'âge limite des fonctionnaires de la catégorie active, à savoir 62 ans, est affectés à tous les salariés détenant le statut "Public actif" à leur départ en retraite. Cependant, il faut avoir effectué au moins 17 ans de services dans un emploi de la catégorie active pour être considéré comme un fonctionnaire de la catégorie active (cf. référence 2.1.4), et donc pour dépendre du système de retraite de ce statut. Par conséquent, les départs en retraite après 62 ans pour les salariés ayant le statut "Public actif" n'est pas rare, puisqu'ils peuvent dépendre du système de retraite des fonctionnaires de la catégorie sédentaire (dont l'âge limite s'élève à 67 ans).

La première situation évoquée ci-dessus suppose de retirer les départs anticipés pour carrière longue des situations considérées "extrêmes". Toutefois, l'entreprise n'a pas fourni ce renseignement. Il a donc été décidé de considérer que tous les départs en retraite anticipés entre l'âge minimal possible dans le cadre d'une carrière longue (qui vacille entre 56 et 58 ans, selon la génération de naissance) et l'âge d'ouverture des droits, sont des départs anticipés pour carrière longue. Afin de s'assurer de cela, une étude a été réalisée sur la proportion ces départs. Cette proportion a été comparée (cf. graphique ci-dessous) à celle des départs anticipés pour carrière longue observés à l'échelle nationale (cf. graphique 6 de la fiche 17 du rapport "Les retraités et les retraites" (édition 2021) de la DREES, référence [10]).

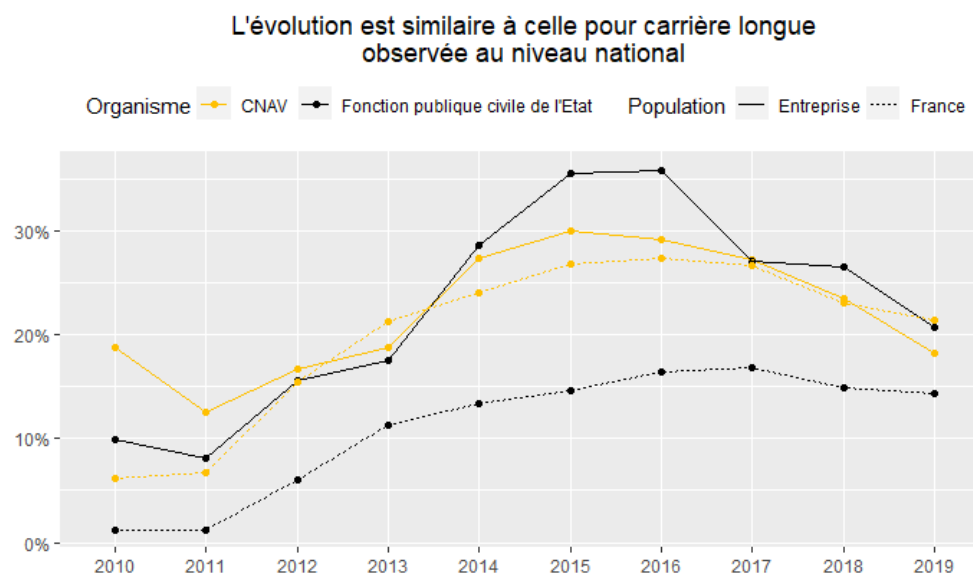


FIGURE 3.5 – Évolution de la proportion des départs anticipés à l'âge minimal de carrière longue ou après, comparée à celle des départs anticipés pour carrière longue observés à l'échelle nationale

Par exemple, l'entreprise a constaté que 27,3% de ses salariés du privé partis en retraite en 2014, sont partis entre l'âge minimal de carrière longue et l'âge d'ouverture des droits. Au niveau national, il a été constaté que, parmi les départs en retraite survenus en 2014 du secteur privé, 24,1% sont des départs en retraite pour carrière longue. Ce graphique permet ainsi de s'assurer que la considération suivante a du sens : les départs anticipés après l'âge minimal de carrière longue (observés au sein de l'entreprise) sont des départs anticipés pour carrière longue. Par conséquent, les deux situations suivantes ont été considérées comme "rares" :

Les valeurs extrêmes ont été définies d'après le comportement national vis-à-vis de la retraite.

- Les départs anticipés avant l'âge minimal possible dans le cadre de carrières longues ;
- Les départs après 70 ans pour les salariés du privé (ce qui correspond à l'âge limite) et les départs après 67 ans pour les fonctionnaires de la catégorie sédentaire et active (ce qui correspond à l'âge limite des fonctionnaires de la catégorie sédentaire).

Finalement, les individus observés qui sont dans l'une de ces deux situations ont des écarts "âge de retraite - âge minimal" considérés "extrêmes".

Pour analyser ces valeurs extrêmes, une AFDM a été réalisée.

L'analyse factorielle des données mixtes (AFDM) peut être réalisée pour déterminer si les individus avec des écarts "extrêmes" ont des caractéristiques similaires. En effet, cette analyse permet de visualiser les individus dans un espace réduit, comme l'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC). Elle se distingue de l'ACP et de l'AFC, car elle est applicable en présence de variables qualitatives et quantitatives (l'algorithme agit comme celui de l'ACP pour les variables qualitatives et comme celui de l'AFC pour les variables quantitatives). Par conséquent, cette analyse met en évidence les individus aux caractéristiques similaires (grâce à leur projection sur les axes principales), quelle que soit la nature de ces caractéristiques. Il a ainsi été décidé d'appliquer cette analyse pour représenter les individus avec des écarts "extrêmes", puisqu'ils sont décrits par 8 variables de nature différente et que l'objectif est de déterminer si leurs caractéristiques sont similaires. Il sera alors considéré que ces individus ont des caractéristiques similaires, si leur représentation sur les premiers axes principales forme un groupe (éloigné de l'origine).

3.3.4 La présence de variables qui prédisent le même comportement doit être évité.

Un modèle est davantage compréhensible et stable, lorsque le nombre de variables prédictives est faible. Or la compréhension et la stabilité du modèle sont des caractéristiques recherchées dans le contexte de l'évaluation d'engagement IFC : le modèle doit être compréhensible par les dirigeants de l'entreprise pour qu'ils veuillent l'intégrer dans l'évaluation de leur engagement ; le modèle doit être stable pour garantir que les résultats obtenus par cette étude sont représentatifs des résultats qui seront obtenus prochainement. Par conséquent, il est préférable de réduire le nombre de variables prédictives du modèle, sans pour autant supprimer des variables qui jouent un rôle important dans la prédiction de l'âge de retraite.

L'identification de variables fortement corrélées permet de réduire la complexité du modèle.

Lors de l'analyse de données, il convient alors de détecter les variables prédictives qui prédisent le même comportement, afin de n'en conserver qu'une seule. Une solution pour détecter ces variables est d'étudier leur corrélation. Deux variables fortement corrélées joueront nécessairement le même rôle dans la prédiction de l'âge de retraite. Par exemple, deux variables qualitatives 'A' et 'B' (de modalités a_1, a_2 et b_1, b_2) sont fortement corrélées : un individu ayant la modalité a_1 a très souvent la modalité b_1 également (idem avec les modalités a_2 et b_2). Par conséquent,

l'âge de retraite associé généralement aux individus ayant la modalité a_1 est très souvent associé également aux individus ayant la modalité b_1 . Les variables 'A' et 'B' prédisent ainsi le même comportement vis-à-vis de l'âge de retraite. La présence de ces deux variables dans le modèle n'a donc pas grand intérêt. De plus, comme évoqué précédemment, il est préférable de réduire le nombre de variables d'un modèle. Par conséquent, il a été décidé de retirer l'une des variables : celle dont le niveau de corrélation avec la variable à prédire est la plus faible.

Il a été décidé de considérer qu'une corrélation est forte lorsqu'elle est évaluée à 75% ou plus.

Pour étudier la corrélation entre deux variables, il a été convenu de suivre 3 étapes.

- Etape 1 : réaliser un test d'indépendance. Le seuil de confiance a été fixé à 5% : si la p-value du test est inférieure à 5%, alors l'hypothèse nulle d'indépendance des variables est rejetée.
- Etape 2 : calculer un coefficient de corrélation. Ceci permet d'évaluer le niveau de corrélation entre les variables. Il a été convenu que deux variables sont fortement liées lorsque le coefficient de corrélation est supérieur à 75%.
- Etape 3 : illustrer la corrélation à partir d'un graphique. Cette étape permet de mettre en évidence la liaison (ou l'indépendance) entre les variables.

Toutefois, il existe de nombreux tests d'indépendance, de coefficients de corrélation et de graphiques. Ces derniers dépendent de la nature (qualitative ou quantitative) des variables étudiées :

Selon la nature du couple de variables étudiées, trois outils analytiques ont été choisis.

- Lorsque les deux variables sont qualitatives, il a été décidé d'utiliser :
 - Le test d'indépendance du khi-deux ;
 - Le coefficient de corrélation "V de Cramer" ;
 - Le graphique à barres empilées en pourcent.
- Lorsque les deux variables sont quantitatives, il a été décidé d'utiliser :
 - Les tests de corrélation de Pearson et de Kendall ;
 - Les coefficients de corrélation de Pearson et de Kendall ;
 - Le nuage de points des valeurs et des rangs des valeurs.
- Lorsqu'une variable est qualitative et l'autre variable est quantitative, il a été décidé d'utiliser :

- Le test de Wilcoxon-Mann-Whitney, lorsque la variable qualitative n'a que deux modalités. Sinon, le test de Kruskal-Wallis ;
- Le rapport de corrélation ;
- Le graphique "boîte à moustache" qui met en évidence la répartition de la variable quantitative au sein des modalités.

En cas de forte corrélation, la variable prédictive la moins corrélée avec la variable à prédire sera retirée.

Lorsqu'un couple de variables prédictives est jugé comme étant fortement corrélé, il convient de ne conserver que l'une des deux variables : celle étant la plus corrélée avec la variable à prédire. En effet, cette variable prédictive aura ainsi plus de chance de contribuer à la bonne prédiction de l'écart "âge retraite - âge minimal". Les 3 étapes citées précédemment devront donc être suivies pour déterminer cette variable (et pour retirer la seconde variable du modèle). Toutefois, si les variables prédictives fortement corrélées ne sont pas de la même nature (l'une est qualitative et l'autre est quantitative), alors la comparaison de leur liaison avec la variable prédictive est difficile. En effet, le coefficient de corrélation utilisé sera différent : l'un correspondra au rapport de corrélation et l'autre au coefficient de Pearson et de Kendall. Afin d'être prudent, il a donc été décidé de seulement se fier au test d'indépendance : si celui-ci conclut que l'une des variables prédictives est indépendante avec la variable à prédire, alors cette variable sera retirée. Sinon, il conviendra de créer deux bases de données de retraités : l'une sans la variable prédictive qualitative et l'autre sans la variable prédictive quantitative.

3.3.5 Le traitement des données sera donc réalisé en 4 étapes.

Pour conclure, les étapes qui seront suivies lors de l'analyse et du traitement des données sont citées ci-dessous.

- La première étape consiste à traiter les variables qui dépendent de caractéristiques obsolètes (la variable à prédire et le salaire) ou qui représentent une petite partie de la population (le salaire).
- La deuxième étape consiste, quant à elle, à étudier les variables prédictives une à une, dans le but d'identifier les valeurs extrêmes (pour les retirer) ou les modalités sous-représentées (pour les réunir avec d'autres).
- La troisième étape sera consacrée à l'analyse détaillée de la variable à prédire, pour identifier les valeurs extrêmes non "justifiées" par les

L'étape de traitement des données est ainsi divisée en 4 sous-étapes.

caractéristiques observées.

- La quatrième et dernière étape est consacrée à l'analyse des corrélations entre les variables prédictives. Cette analyse a pour but d'identifier les couples de variables expliquant les mêmes comportements face à la retraite. Il est donc pertinent de ne conserver qu'une seule de ces variables, afin de réduire la complexité du modèle.

3.4 Méthodologie adoptée lors de la modélisation

L'objectif de ce mémoire est de construire un modèle avec un bon pouvoir prédictif. Cet objectif n'a pas que des conséquences sur l'analyse et le traitement des données (comme évoqué dans la partie précédente), mais également sur la construction du modèle. En effet, lors de cette étape, il convient d'optimiser le pouvoir prédictif et d'assurer que celui-ci est stable. L'optimisation du pouvoir prédictif passe par la division des observations en deux échantillons et l'analyse du phénomène de sur-apprentissage. D'autre part, pour assurer que le modèle prédictif construit est stable, il est possible d'utiliser une technique utilisée en apprentissage automatique : la validation croisée. Cette partie explicite ainsi ces 3 pratiques qui ont un objectif commun : construire un modèle avec un bon pouvoir prédictif. A noter que le choix des modèles n'est pas évoqué dans cette partie, et que l'application des pratiques sera présentée dans le chapitre suivant.

3.4.1 La base de données des retraités est divisée en deux échantillons.

Pour évaluer le pouvoir prédictif d'un modèle, une partie des observations est exclue de la phase de construction : il s'agit de l'échantillon test.

Pour savoir si les prédictions d'un modèle sont bonnes, il faut le tester sur des données qui sont nouvelles pour lui. Pour cela, une partie des données sur les retraités sera retirée de la phase de construction du modèle : il s'agit de l'*échantillon test*. Le modèle, une fois construit, n'a donc pas connaissance de cet échantillon, ce qui représente un très bon moyen d'évaluer son pouvoir prédictif. En effet, il sera possible d'appliquer le modèle aux données de l'échantillon test, et de comparer les prédictions obtenues avec les valeurs réelles de la variable à prédire : le modèle se retrouve donc face à de nouvelles données. Les données utilisées pour construire le modèle (celle qui ne font pas partie de l'échantillon test) constituent l'*échantillon d'apprentissage*. Par conséquent, la base de données des retri-

tés n'est pas intégralement utilisée pour construire le modèle : une partie est utilisée pour évaluer son pouvoir prédictif. Cette évaluation permettra, en pratique, de déterminer si le modèle est plus performant que le modèle actuellement utilisé par l'entreprise.

Il n'existe pas de règle d'or sur la manière de diviser la base de données en deux. Toutefois, il est d'usage de consacrer 80% des observations à l'échantillon d'apprentissage (et donc, 20% des observations à l'échantillon test), lorsque le nombre total d'observations est grand. D'autre part, il existe plusieurs méthodes pour affecter les observations à un des deux échantillons, comme la méthode aléatoire et la méthode stratifiée (qui permet de conserver les proportions de la base de données complète). La méthode stratifiée assure que l'échantillon d'apprentissage est représentatif de la base complète. Toutefois, en présence d'un grand nombre de données, la méthode aléatoire est suffisante pour faire de même.

La taille et la constitution de l'échantillon test est imposée par l'une des métriques d'évaluation choisies (cf. partie suivante).

Dans ce mémoire, la manière de diviser la base de données en deux échantillons (test et apprentissage) est imposée. En effet, ce mémoire doit s'inscrire dans le contexte de l'évaluation d'un engagement social, ce qui conduit à utiliser une métrique spécifique pour évaluer le pouvoir prédictif d'un modèle : l'écart actuariel d'expérience 2019 (cf. partie suivante). Ainsi, les départs en retraite survenus après le 31/12/2018 vont constituer l'échantillon test. Après l'analyse et le traitement des données (cf. chapitre suivant), la base de données des retraités contiendra 32 127 observations et ces départs en retraite correspondront 4 900 observations, soit 15,25% de la base.

3.4.2 Le phénomène de sur-apprentissage doit être évité pour bien prédire.

Le phénomène de sur-apprentissage survient lorsque le modèle est excessivement ajusté aux données de l'échantillon d'apprentissage. Ceci se détecte par une variance élevée du modèle. La variance d'un modèle correspond à l'écart moyen entre la moyenne des prédictions et les prédictions elles-mêmes : elle mesure donc à quel point le modèle s'éloigne de sa moyenne pour coller aux données observées (cf. référence [44]). Lorsque la variance d'un modèle est élevée, cela signifie donc que les prédictions sont variées et, *a fortiori*, proches des données observées. A première vue, la maximisation de la variance semble être un bon objectif lors d'une modélisation. Ce n'est pourtant pas le cas dans une démarche prédictive, car

une variance trop élevée suppose que le modèle est trop ajusté aux données observées et qu'il s'adaptera alors mal à la présence de nouvelles données (alors qu'il s'agit de l'objectif principal des modèles prédictifs). Un tel modèle risque donc de capter des "bruits" (c'est-à-dire des faibles liaisons entre des variables prédictives et la variable à prédire qui n'ont pas de sens). Les valeurs prédites par le modèle qui ont été obtenues à partir de l'échantillon d'apprentissage seront alors proches des données de l'échantillon, mais celles obtenues à partir de l'échantillon test risquent, quant à elles, d'être éloignées de la réalité. Ce phénomène, appelé *sur-apprentissage*, est donc à éviter. À noter toutefois que le phénomène inverse, le *sous-apprentissage*, est également à éviter puisque de fortes liaisons risquent de ne pas être détectées. En conclusion, la construction des modèles doit s'assurer que le phénomène de sur-apprentissage n'apparaisse pas, tout en évitant le phénomène de sous-apprentissage.

Pour éviter le sur-apprentissage, plusieurs constructions du modèle seront réalisées, avec un niveau de complexité différent.

Pour se rendre compte que le phénomène de sur-apprentissage apparaît, il est possible de construire des modèles de plus en plus complexes et d'évaluer, pour chaque niveau de complexité, le pouvoir prédictif du modèle (à partir de l'échantillon test). Lorsque le pouvoir prédictif est moins bon que le pouvoir prédictif évalué pour un niveau de complexité inférieur, le phénomène de sur-apprentissage apparaît alors. En pratique, un modèle est considéré plus complexe lorsque l'échantillon d'apprentissage contient plus de variables prédictives. Ainsi, plusieurs échantillons d'apprentissage seront établis (avec une combinaison de variables prédictives différente) et, pour chacun de ces échantillons d'apprentissage, un modèle sera construit puis évalué. À noter que le nombre "réel" de variables prédictives prend en compte les modalités des variables qualitatives : si une variable qualitative possède n modalités, alors elle cache en réalité $n - 1$ variables "effectives". Par conséquent, le nombre d'échantillons d'apprentissage pouvant être établis peut très vite être gigantesque. Par exemple, il est supposé que seulement 2 variables prédictives qualitatives composent la base : l'une (nommée "A") avec deux modalités ("a1" et "a2") et l'autre (nommée "B") avec trois modalités ("b1", "b2" et "b3"). Dans ce cas (5 modalités au total), les échantillons d'apprentissage possibles sont en nombre de 10 :

- L'échantillon ne contenant aucune variable prédictive (0 variable prédictive "effective") ;
- L'échantillon contenant la variable "A" (1 variable prédictive "effective") ;

- L'échantillon contenant partiellement la variable "B" avec les modalités "b1" et "b2+b3" (1 variable prédictive "effective");
- L'échantillon contenant partiellement la variable "B" avec les modalités "b2" et "b1+b3" (1 variable prédictive "effective");
- L'échantillon contenant partiellement la variable "B" avec les modalités "b3" et "b1+b2" (1 variable prédictive "effective");
- L'échantillon contenant la variable "A" et partiellement la variable "B" avec les modalités "b1" et "b2+b3" (2 variables prédictives "effectives");
- L'échantillon contenant la variable "A" et partiellement la variable "B" avec les modalités "b2" et "b1+b3" (2 variables prédictives "effectives");
- L'échantillon contenant la variable "A" et partiellement la variable "B" avec les modalités "b3" et "b1+b2" (2 variables prédictives "effectives");
- L'échantillon contenant la variable "B" (2 variables prédictives "effectives").
- L'échantillon contenant la variable "A" et la variable "B" (3 variables prédictives "effectives").

Tous les niveaux de complexité ne pourront pas être évalués.

Malgré le petit nombre de variables prédictives de la base de données des retraités (8), le nombre total de modalités est important (21). Il n'est donc pas envisageable d'établir tous les échantillons d'apprentissage possibles. Il a donc été décidé d'établir les échantillons sans prendre en compte les modalités, mais seulement les variables prédictives. En reprenant l'exemple précédent, seulement 3 échantillons seraient établis :

- Celui sans variable prédictive (0 variable prédictive "effective");
- Celui avec la variable "A" (1 variable prédictive "effective");
- Celui avec la variable "B" (2 variables prédictives "effectives");
- Celui avec les deux variables (3 variables prédictives "effectives").

Cette pratique permettra ainsi de faire apparaître (éventuellement) le phénomène de sur-apprentissage. Le modèle ayant le meilleur pouvoir prédictif (selon l'échantillon d'apprentissage établi) sera ainsi conservé. A noter

qu'après l'analyse et le traitement des données (cf. chapitre suivant), 128 échantillons différents seront établis (ce nombre correspond au nombre de combinaisons possibles de variables prédictives, de tailles différentes).

3.4.3 La stabilité du modèle doit être vérifiée.

Dans une démarche prédictive, il est important de s'assurer que le modèle est stable. Un modèle est considéré stable lorsque ses paramètres et ses prédictions ne sont pas extrêmement sensibles aux observations. Dans une démarche prédictive, il est important de s'assurer que le modèle est stable car il est possible que :

- la présence de certaines observations dans l'échantillon d'apprentissage modifie abusivement les paramètres du modèle. Ceci démontrerait alors que le modèle choisi est instable.
- les prédictions soient hasardeusement proches des valeurs réelles. En effet, les données de l'échantillon test peuvent être excessivement similaires à celles de l'échantillon d'apprentissage, ce qui a peu de chance d'être le cas pour les données futures. L'évaluation du pouvoir prédictif du modèle ne serait donc pas représentative de la réalité.

Pour s'assurer de la stabilité du modèle, la technique de validation croisée en k blocs sera retenue.

Une technique existe pour vérifier la stabilité du modèle : la validation croisée. Elle consiste à répéter le processus de construction avec un couple "échantillon d'apprentissage / échantillon test" différent. Toutefois, le couple "échantillon d'apprentissage / échantillon test" de cette étude est imposée (afin d'appliquer une métrique d'évaluation du pouvoir prédictif des modèles, cf. partie suivante), ce qui conduit à appliquer la validation croisée sur l'échantillon d'apprentissage (et non pas sur la base de données complète) : l'échantillon d'apprentissage sera donc divisé en deux sous-échantillons (ce qui reste acceptable, puisque le nombre d'observations est grand). Ces deux sous-échantillons seront appelés "échantillon d'apprentissage" (qui servira à construire le modèle) et "échantillon de validation" (qui servira à évaluer le pouvoir prédictif du modèle construit). Plusieurs types de validation croisée existent, comme :

- La *k-fold cross-validation* qui consiste à diviser l'échantillon d'apprentissage initial en k groupes égaux (de manière aléatoire ou stratifiée) et d'assigner, à chaque construction, un bloc comme étant l'échantillon de validation. Cette application entraîne ainsi k constructions de modèle.

- La *leave-one-out cross-validation* qui consiste à placer, à chaque construction, une seule observation dans l'échantillon de validation. Pour une base contenant n observations, cela implique que n modèles seront construits.

Le nombre de blocs a été fixé pour que le modèle soit construit sur 80% des données.

La première technique a été retenue, car elle est plus adaptée à la sélection du "meilleur modèle", en évitant le phénomène de sur-apprentissage (cf. point précédent). En effet, l'échantillon de validation contient plusieurs données (que celui de la *leave-one-out cross-validation*), ce qui permet de mieux évaluer le pouvoir prédictif des modèles de complexité différente. Il reste toutefois à déterminer le nombre de blocs (noté k), ainsi que la méthode d'affectation des observations à chaque bloc. Comme indiqué précédemment, la méthode d'affectation aléatoire est adaptée en présence d'un grand nombre d'observations : il a donc été décidé d'appliquer cette méthode pour constituer les k blocs. D'autre part, il est fréquent que l'échantillon test représente 20% de la base de données (lorsque celle-ci est grande). Par conséquent, il a été décidé de déterminer k de façon à ce que la somme du nombre d'observations dans un bloc et du nombre d'observations dans l'échantillon test imposé corresponde approximativement à 20% de la base de données complète. Après l'analyse et le traitement des données (cf. chapitre suivant), la base de données des retraités contiendra 32 127 observations et l'échantillon test sera composé de 4 900 observations. Il a donc été décidé d'affecter 1 601 observations à 7 blocs et 1602 observations à 10 autres blocs, car 6 501 (la somme de 4 900 et 1 601) correspond à 20,2% de la base de données des retraités et que cela revient à créer 17 blocs de taille quasi-identique (soit $k = 17$).

3.4.4 La procédure de modélisation permet d'étudier la stabilité du modèle et d'optimiser son pouvoir prédictif

La procédure commence par la construction de l'échantillon de test. Cet échantillon est constitué des départs en retraite survenus après le 31/12/2018 (15,25% de la base). Cette division est justifiée dans la prochaine partie, qui présente les métriques choisies d'évaluation du pouvoir prédictif d'un modèle. Cet échantillon permet de déterminer si le modèle est plus performant que celui utilisé actuellement par l'entreprise.

Le reste de la base de données est ensuite divisé en 17 blocs de taille quasi-identique. La taille est fixée de manière à ce que la somme

du nombre d'observations dans un bloc et celui dans l'échantillon test représente approximativement 20% de la base de données des retraités. L'affectation d'une observation dans un bloc est réalisée de manière aléatoire. Ces 17 blocs permettront :

- d'établir 17 échantillons d'apprentissage différents (constitués des données de 16 blocs) et de réaliser ainsi 17 modélisations.
- d'évaluer le pouvoir prédictif des 17 modèles construits grâce à l'échantillon de validation (constitué du bloc non utilisé pour établir l'échantillon d'apprentissage).

Cette pratique permet d'analyser la stabilité du modèle "au sens large" à travers la comparaison des différents jeux de paramètres obtenus et des différentes évaluations du pouvoir prédictif obtenues.

En parallèle, le phénomène de sur-apprentissage sera surveillé. Pour chacune des 17 modélisations, il a été décidé de construire des modèles à des niveaux de complexité différents. Pour cela, 128 échantillons d'apprentissage seront établis à partir de l'échantillon d'apprentissage sélectionné (un échantillon d'apprentissage pour chaque combinaison possible de variables prédictives à des niveaux de complexité différents). Le nombre total de modèles construits (et évalués) est donc égal à $17 \times 128 = 2176$. Cette pratique permettra, en parallèle de l'étude de la stabilité du modèle "au sens large", d'étudier l'évolution du pouvoir prédictif en fonction de la complexité du modèle construit (c'est-à-dire du nombre de variables prédictives "effectives" de l'échantillon d'apprentissage).

Enfin, plus de 2 000 modèles seront construits.

Finalement, le modèle ayant le meilleur pouvoir prédictif (parmi les 2 176 construits) sera retenu. Puis, celui-ci sera de nouveau évalué à partir, cette fois-ci, de l'échantillon test. Le schéma ci-dessous résume les pratiques évoquées dans cette partie.

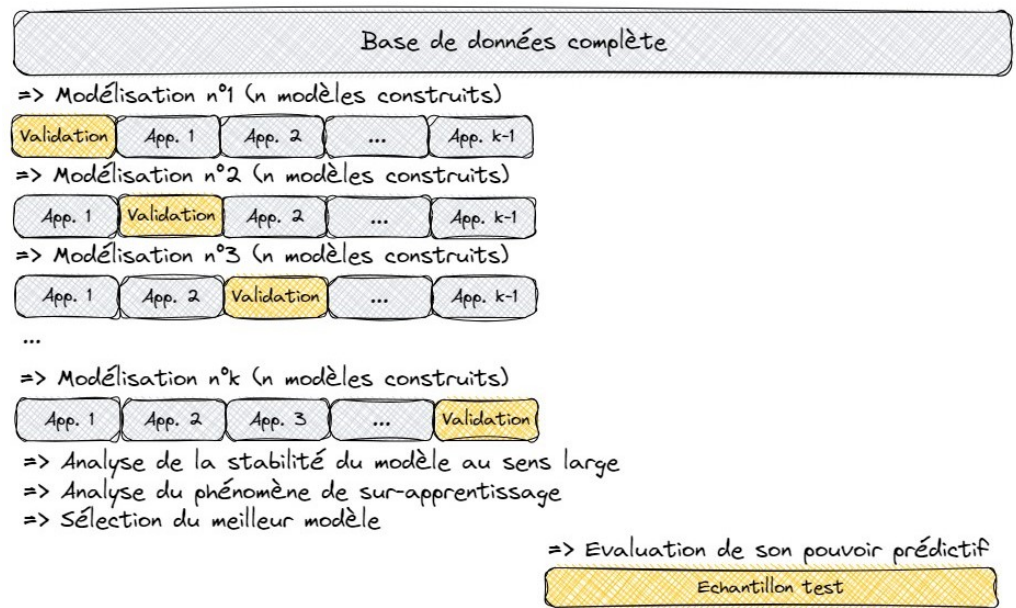


FIGURE 3.6 – Illustration de la procédure de construction d'un modèle adoptée

3.5 Métriques choisies d'évaluation du pouvoir prédictif d'un modèle.

3.5.1 L'écart actuariel d'expérience engendré a été choisi pour évaluer la performance des modèles.

L'écart actuariel correspond à la différence entre l'engagement réel et l'engagement prévu

L'écart actuariel d'expérience évalue la précision du modèle d'évaluation. Plus précisément, il est rattaché à un exercice défini entre deux dates d'évaluation E_0 et E_1 et il correspond à la différence entre :

- L'engagement réel : l'engagement évalué à E_1 , avec les données démographiques à E_1 et les hypothèses d'évaluation à E_0 ;
- L'engagement prévu : l'engagement projeté à E_1 qui a été calculé lors de l'évaluation à E_0 .

L'engagement projeté à E_1 correspond à la somme des éléments suivants (calculés lors de l'évaluation précédente, à E_0) :

- L'engagement évalué à E_0 ;

- Le *service cost* (ou charges normales) de l'exercice entre E_0 et E_1 ;
- L'*interest cost* (ou charges d'intérêt) du même exercice ;
- L'opposé du total des prestations prévues lors du même exercice (venant donc diminuer la somme).

A noter qu'il arrive de parler de l'*engagement prévu (hors prestations)*. Ce terme fait référence à l'engagement projeté, auquel les prestations prévues ne sont pas retraitées.

Le *service cost* calculé à E_0 d'un salarié s (noté $SC_s(E_0)$) correspond à la part d'engagement à ajouter suite à l'augmentation de son ancienneté entre E_0 et E_1 . Plus précisément :

$$SC_s(E_0) = VAP_s(E_0) \times \frac{\text{durée entre } \tilde{E}_0 \text{ et } \tilde{E}_1}{\text{durée entre } D_s \text{ et } R_s},$$

avec :

- $\tilde{E}_0$, la plus récente date entre E_0 et D_s (cette dernière désigne la date de début de période de services effectué par s) ;
- $\tilde{E}_1$, la plus ancienne date entre E_1 et R_s (cette dernière désigne la date de départ en retraite de s).

Si D_s est postérieure à E_1 , alors le *service cost* est nul.

Concernant l'*interest cost* calculé à E_0 d'un salarié s (noté $IC_s(E_0)$), cette charge correspond à la part d'engagement à ajouter suite à la capitalisation de la valeur actuelle probable de l'IFC. Plus précisément :

$$IC_s(E_0) = (PBO_s(E_0) + SC_s(E_0)) \times r(E_0).$$

A noter que si la date de retraite est prévue entre E_0 et E_1 , alors il est considéré que l'indemnité sera versée en milieu d'exercice prochain. Par conséquent, l'*interest cost* correspond plus exactement à :

$$IC_s(E_0) = \left(PBO_s(E_0) + SC_s(E_0) - \frac{IFC_s(R_s)}{2} \right) \times r(E_0).$$

Il est fréquent de parler "des" écarts actuariels qui, une fois additionnés, donne l'écart actuariel global. En effet, il est possible de justifier un écart actuariel par un événement spécifique : ce qui explique l'intérêt de parler de plusieurs écarts actuariels. Par exemple, l'entrée d'un salarié dans l'effectif. Cet événement entraîne un écart actuariel (positif,

L'engagement projeté correspond à l'engagement précédent, auquel plusieurs éléments sont ajoutés.

puisque l'entreprise est engagée à verser une IFC supplémentaire) qui est égal à l'engagement réel du salarié. Il existe d'autres événements engendrant un écart actuariel : la sortie d'un salarié de l'effectif (retraite, décès, démission, licenciement, etc.), la modification de l'une des informations sur un salarié (catégorie, date d'ancienneté, salaire, etc.). Pour chacun de ces événements, il est possible de connaître précisément le montant de l'écart actuariel qu'il a engendré.

L'écart actuariel engendré par le modèle d'âge de retraite sera utilisé pour évaluer le modèle.

Il a donc été décidé d'évaluer le pouvoir prédictif d'un modèle d'âge de retraite, selon l'écart actuariel qu'il engendre. Ceci permet ainsi d'inscrire au mieux l'étude dans le contexte d'évaluation d'engagement IFC. Pour rappel, l'entreprise a fourni la base de données des actifs au 31/12/2018, celle au 31/12/2019 et les hypothèses d'évaluation au 31/12/2018 (cf. deuxième partie du présent chapitre). Par conséquent, l'écart actuariel (engendré par le modèle d'âge de retraite) qu'il est possible d'obtenir correspond à celui de l'exercice 2019. D'autre part, l'écart actuariel engendré par le modèle d'âge de retraite correspond à la somme de trois écarts actuariels ci-dessous :

- *L'écart dû aux IFC* : la différence entre le montant des IFC prévues de verser au cours de l'exercice (entre E_0 et E_1) et le montant des IFC réellement versées au cours de l'exercice. Le montant des IFC prévues de verser correspond finalement à l'impact sur l'engagement que le modèle d'âge de retraite a prévu : plus cet écart est proche de 0, plus le modèle a été proche de la réalité.
- *L'écart dû aux retraites réalisées* : la différence entre le montant des IFC versées et l'engagement projeté (hors prestations) des salariés partis en retraite. Pour un modèle d'âge de retraite "individuel" (c'est-à-dire qu'une unique date de retraite est fixée pour un salarié), cet écart est nul lorsque la date de retraite prévue (et unique) correspond à la date de retraite réelle. En effet, l'engagement projeté (hors prestations) est, dans ce cas, égal à l'IFC prévue qui correspond elle-même à l'IFC versée. Pour un modèle d'âge de retraite "global" (c'est-à-dire que plusieurs dates de retraite sont fixées pour un même salarié, à l'instar du modèle utilisé par l'entreprise), cet écart n'est pas nul, même lorsque l'une des dates de retraite prévues correspond à la date de retraite réelle. En effet, l'engagement projeté (hors prestations) est, dans ce cas, la moyenne pondérée des engagements projetés (hors prestations) obtenus avec les différentes dates de retraite fixées (cf. première partie de ce chapitre).

- *L'écart dû aux retraites non réalisées* : la différence entre l'engagement réel et l'engagement projeté (hors prestations) des salariés dont la date de retraite prévue (ou l'une des dates de retraite prévues) est entre E_0 et E_1 et qui font encore partie de l'effectif à E_1 . Le modèle d'âge de retraite a donc prévu leur départ en retraite, alors qu'ils ne sont finalement pas partis.

Ainsi, la somme de ces 3 écarts actuariels constitue l'écart actuariel (de l'exercice 2019) engendré par le modèle d'âge de retraite. Cet écart permettra d'évaluer le pouvoir prédictif du modèle : plus il est proche de 0, plus le modèle a prédit des départs en retraite à des dates proches de la réalité et donc, plus le pouvoir prédictif de ce modèle sera considéré bon. De plus, cet écart permettra également de déterminer si le modèle d'âge de retraite est plus performant que celui utilisé actuellement par l'entreprise : s'il est plus proche de 0 que celui engendré par le modèle de l'entreprise, alors il sera considéré plus performant.

L'échantillon test doit être composé des départs survenus en 2019.

Cette métrique impose d'affecter les départs en retraite réalisés entre le 01/01/2019 et le 31/12/2019 à l'échantillon test. Il a été effectivement évoqué (cf. quatrième partie du présent chapitre) que l'évaluation du pouvoir prédictif du modèle doit se réaliser à partir de l'échantillon test, contenant des données "nouvelles" pour le modèle. Or, la métrique choisie correspond à l'écart actuariel de l'exercice 2019 engendré par le modèle : elle évalue donc si le modèle a prédit des dates de départ en retraite en 2019 qui se sont réellement réalisées en 2019. Par conséquent, les départs en retraite réalisés en 2019 doivent faire partie de l'échantillon test : l'échantillon d'apprentissage ne contiendra donc pas ces observations, ce qui ne biaisera pas l'évaluation du pouvoir prédictif du modèle (à partir de cette métrique). A noter que cette métrique ne peut pas être obtenue sur l'échantillon de validation (cf. schéma en fin de partie précédente), puisque celui-ci contient des départs en retraite réalisés entre 2010 et 2018 inclus. L'écart actuariel de l'exercice 2019 engendré par le modèle ne peut donc pas comparer les dates de retraite prévues et réelles pour l'échantillon de validation, puisque les départs en retraite sont survenus avant 2019.

3.5.2 La racine de l'erreur quadratique moyenne a également été choisie pour évaluer la performance d'un modèle.

L'erreur quadratique moyenne est un indicateur largement utilisé pour évaluer la qualité prédictive d'un modèle. Pour un échantillon

test de taille $n \in \mathbb{N}^*$, cette erreur (plus connue sous l'appellation anglaise *Mean Squared Error*, notée MSE) est égale à :

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2,$$

où $(\hat{y}_i)_{i \in \mathbb{N}^*}$ et $(y_i)_{i \in \mathbb{N}^*}$ correspondent respectivement aux prédictions du modèle et aux valeurs réelles de l'échantillon test.

La *RMSE* sera utilisée pour évaluer le modèle.

Pour faciliter son interprétation, la racine de cette erreur sera calculée : soit la *Root Mean Squared Error* (RMSE). En effet, la RMSE permet d'affirmer que le modèle se trompe en moyenne de x années (où x correspond à la RMSE).

Cette métrique sera utilisée sur l'échantillon de validation et l'échantillon test. Pour rappel (cf. partie précédente), 2 176 modèles seront construits puis leur pouvoir prédictif sera évalué (autrement dit, la RMSE sera calculée à partir des données de l'échantillon de validation). Cette pratique permettra :

- d'étudier la stabilité de la RMSE (et donc la stabilité du modèle "au sens large") ;
- de surveiller l'apparition du phénomène de sur-apprentissage à travers l'évolution de la RMSE en fonction de la complexité du modèle (le phénomène de sur-apprentissage apparaît dès que la RMSE d'un modèle plus complexe est plus petite).

D'autre part, une fois que les 2 176 modèles seront construits et évalués, le modèle ayant le meilleur pouvoir prédictif sera retenu : autrement dit, le modèle dont la RMSE est la plus petite sera retenue.

Enfin, la RMSE du modèle ainsi retenu sera calculée sur l'échantillon test. Finalement, cette métrique est très utile car elle permet d'assurer que le modèle est stable, d'assurer que le phénomène de sur-apprentissage n'apparaisse pas, de sélectionner le meilleur modèle et de comparer son pouvoir prédictif avec le modèle de l'entreprise (si sa RMSE est plus petite que celle du modèle de l'entreprise, alors il sera considéré plus performant).

Chapitre 4

Application de la méthodologie adoptée pour répondre à la problématique

Ce chapitre présente les résultats des modélisations.

L'objectif de ce chapitre est de présenter et commenter les résultats obtenus, après avoir appliqué les différentes méthodologies évoquées dans le chapitre précédent. La première partie sera ainsi consacrée à l'analyse et aux traitements appliqués sur les données des retraités, qui est une étape préliminaire importante avant toutes modélisations. A ce stade, la base de données des retraités peut ensuite être divisée en échantillon test et en 17 blocs pour commencer la modélisation de l'âge de retraite (c'est-à-dire la construction des 2 176 modèles). Après la première partie de ce chapitre, les parties seront donc consacrées à la présentation des modélisations de l'âge de retraite selon la méthodologie présentée dans le chapitre précédent : une première modélisation selon une régression linéaire et une seconde modélisation selon un arbre de régression. La présentation de ces modélisations passera donc par l'étude du phénomène de sur-apprentissage, l'étude de la stabilité du modèle, la sélection du meilleur modèle et l'évaluation de son pouvoir prédictif sur l'échantillon test (grâce aux deux métriques présentées au chapitre précédent). Enfin, la dernière partie de ce chapitre sera consacrée à l'évaluation du pouvoir prédictif (sur l'échantillon test) du modèle de référence (c'est-à-dire le modèle actuellement utilisé par l'entreprise, présentée dans la première partie du chapitre précédent), et à la comparaison de celui-ci avec les deux modèles construits.

4.1 Analyse et traitements de la base de données des retraités

Dans le chapitre précédent, 4 étapes d'analyse et de traitements des données (sur les retraités) ont été présentées : la transformation de variables présentant des caractéristiques obsolètes ou représentatives d'une petite population ; l'identification et le traitement des caractéristiques peu représentées ; l'identification et le traitement des valeurs extrêmes de la variable à prédire ; l'identification et le traitement des variables prédictives fortement liées. Ces étapes permettent aux données d'être adaptées à une modélisation prédictive et ainsi, d'optimiser le pouvoir prédictif des modèles. Leur application sont présentées dans cette partie.

4.1.1 Transformation des variables présentant des caractéristiques obsolètes ou représentatives d'une petite population

La variable à prédire est l'écart entre l'âge de retraite et l'âge d'ouverture des droits.

La variable à prédire "Age de retraite" a été transformée en "Écart". Comme évoqué lors du chapitre précédent, l'âge de retraite dépend de la génération de naissance, ce qui est une caractéristique obsolète dans le contexte d'une modélisation prédictive : le modèle est effectivement destiné à être appliqué sur des générations futures. Par conséquent, il a été décidé de prédire l'écart entre l'âge de retraite et l'âge d'ouverture des droits à la retraite (appelé aussi l'âge minimal). L'écart "âge retraite - âge minimal" varie entre -25 ans et 12 ans, avec une médiane de 0,07 ans et une moyenne de 0,58 ans. La répartition des valeurs de cette variable (cf. graphique ci-dessous) démontre que certains départs en retraite sont survenus bien avant l'ouverture des droits, mais que la majorité (72%) survient après l'ouverture des droits.

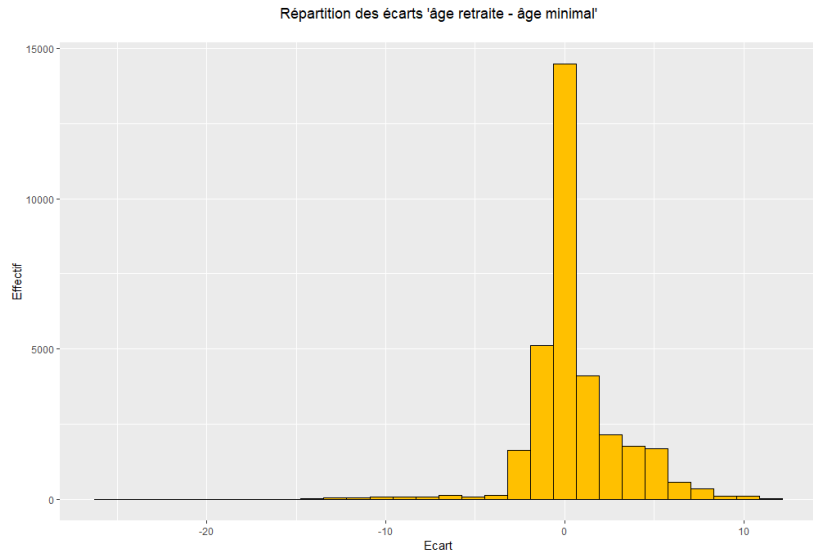


FIGURE 4.1 – Répartition de l'écart "âge retraite - âge minimal"

Grâce à la politique salariale de l'entreprise, le salaire a pu être traité pour s'adapter à une modélisation prédictive.

La variable prédictive "Salaire" a été transformée en "Niveau de vie à 15 ans". Le salaire dépend de l'âge du salarié, alors que les observations concernent des salariés partis en retraite. Ainsi, les salaires observés sont des salaires versés à des âges élevés : il représente donc une petite population, puisque le modèle est destiné à être appliqué sur une population de salariés actifs (dont les âges peuvent varier entre 20 et 70 ans). Pour pallier cela, il a été décidé d'estimer le salaire perçu à 15 ans (selon la politique salariale moyenne de l'entreprise). De plus, le salaire dépend de l'année de son versement à travers le phénomène d'inflation et de la revalorisation annuelle obligatoire du SMIC. A l'instar de l'âge de retraite, le salaire dépend donc d'une caractéristique obsolète dans le contexte d'une modélisation prédictive : le modèle sera appliqué sur des populations futures. Par conséquent, il a été décidé de calculer le rapport entre le salaire et le SMIC en vigueur lors de son versement : ce rapport est appelé ici "Niveau de vie". C'est ainsi que la variable prédictive "Salaire" a été transformée en "Niveau de vie à 15 ans". La nouvelle variable ainsi créée varie entre 71% et 533%, avec une médiane de 113% et une moyenne de 126% (du SMIC en vigueur lors du quinzième anniversaire du salaire). Le graphique ci-dessous illustre la répartition des niveaux de vie à 15 ans observés.

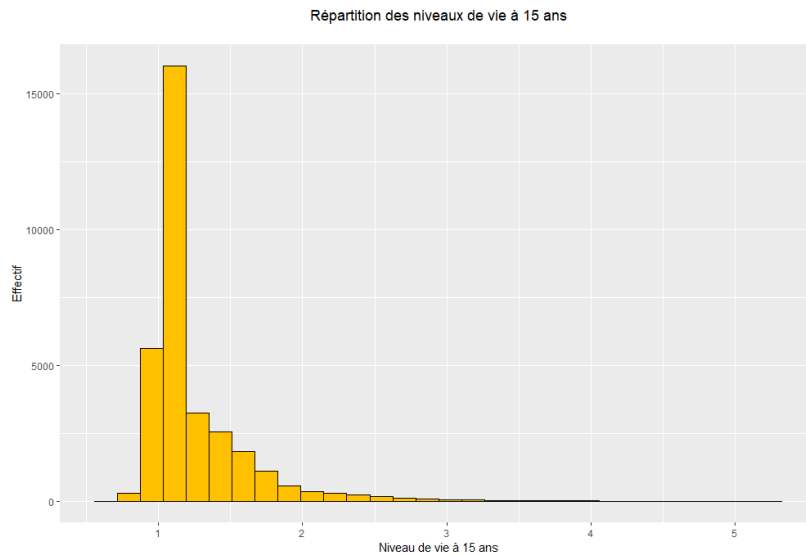


FIGURE 4.2 – Répartition des niveaux de vie à 15 ans

4.1.2 Identification et traitement des caractéristiques peu représentées

Les individus de plus petite catégorie ont été réunis avec ceux de la catégorie supérieure.

La variable "Catégorie" a une modalité sous-représentée. En effet, la première catégorie (nommée "A") n'est représentée que par 2 individus (cf. graphique ci-dessous). Puisque cette variable est ordonnée et que la seule modalité qui l'entoure est la deuxième catégorie ("B"), la modalité "A" a été réunie avec la modalité "B". Afin de se rappeler de ce traitement, la variable a été renommée par "Catégorie v2".

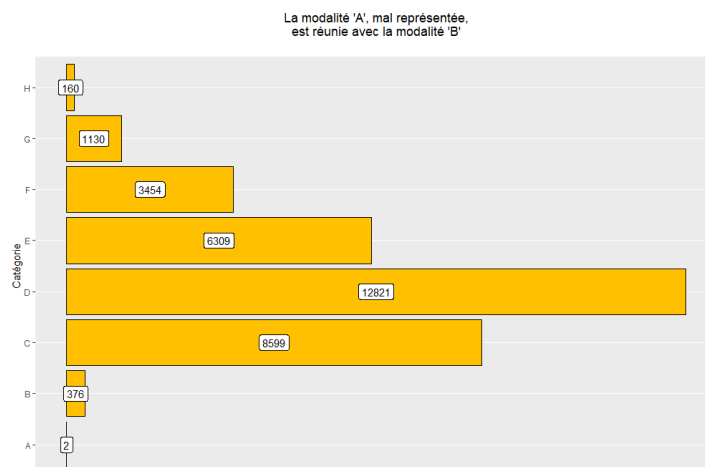


FIGURE 4.3 – Répartition des catégories

Les métiers classés dans un domaine non défini ont été réunis avec les métiers de l'innovation et de la technologie.

La variable "Domaine métier" a une modalité sous-représentée. En effet, la modalité "Domaine non défini" n'est représentée que par 5 individus (cf. graphique ci-dessous).

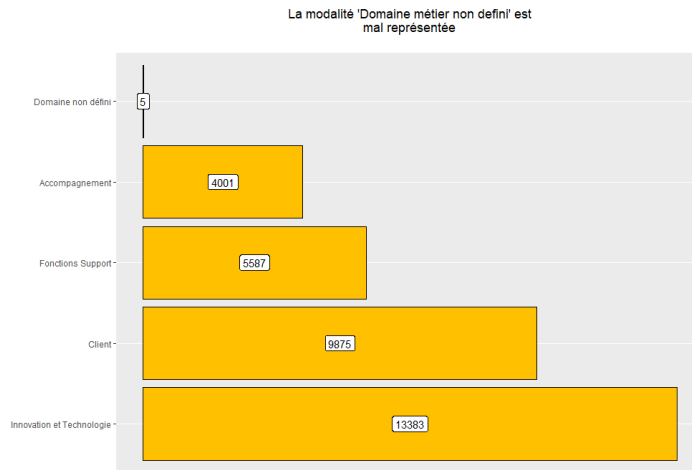


FIGURE 4.4 – Répartition des domaines de métier

Cette variable n'est pas ordonnée : il est donc nécessaire de comparer les écarts "âge de retraite - âge minimal" moyens de chaque modalité (cf. graphique ci-dessous).

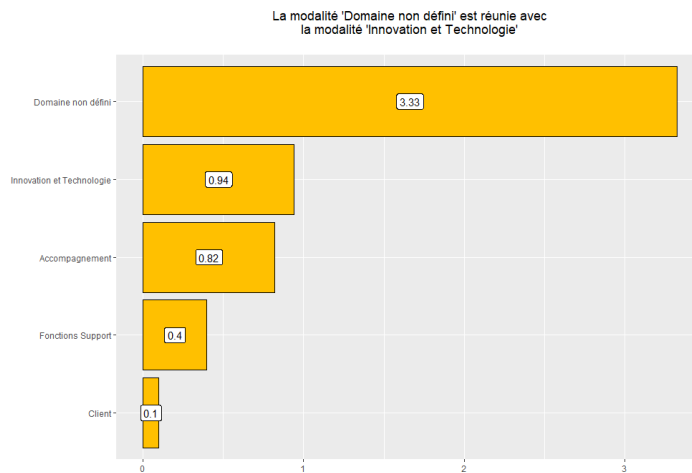


FIGURE 4.5 – Moyenne des écarts "âge retraite - âge minimal" par domaine de métier

L'écart moyen observé au sein des salariés ayant un métier dans un domaine non défini se rapproche le plus de celui observé au sein des salariés ayant un métier dans l'innovation et la technologie. Par conséquent, la modalité "Domaine non défini" a été réunie avec la modalité "Innovation et technologie". Afin de se rappeler de ce traitement, la variable a été renommée par "Domaine métier v2".

Le taux d'activité a été transformé en variable qualitative à 4 modalités.

La variable "Taux d'activité" a été transformée en variable qualitative. Il a été décidé de réaliser cette transformation, car la répartition des taux d'activité n'est pas en "cloche" (cf. graphique ci-dessous).

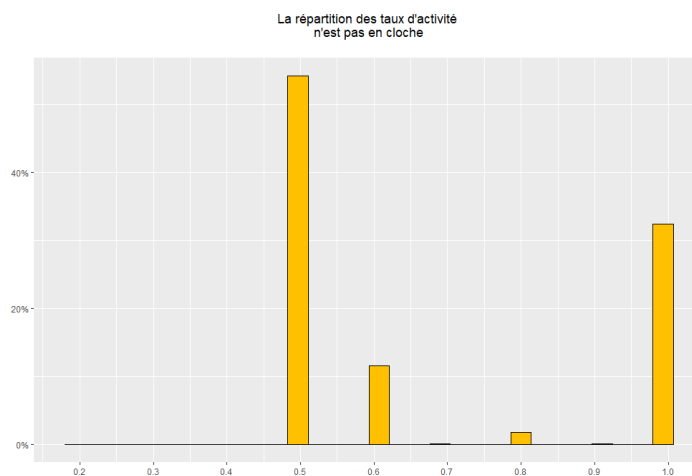


FIGURE 4.6 – Répartition des taux d'activité

La règle de réunion des modalités sous-représentées d'une variable qualitative ordonnée (cf. chapitre précédent) a été ensuite appliquée, pour obtenir finalement 4 modalités : $[0,50]$, $]50,80[$, $[80,90]$, $]90,100]$ (cf. graphique ci-dessous). Afin de se rappeler de ce traitement, la variable a été renommée par "Taux activité v2".

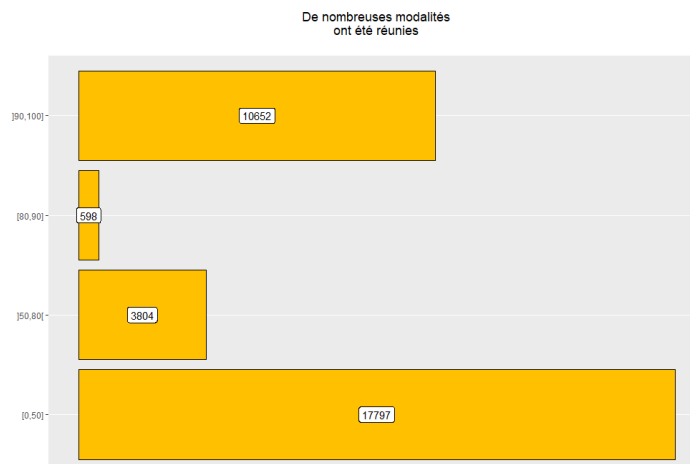


FIGURE 4.7 – Répartition des taux d'activité v2

Les 100 âges d'entrée les plus extrêmes ont été retirés. Après avoir vérifié que la répartition des âges d'entrée est en "cloche", les 50 âges d'entrée les plus petits (inférieurs à 17,39 ans) et les 50 âges d'entrée les plus grands (supérieurs à 56,61 ans) ont été retirés. Ces deux seuils ont été tout d'abord validés graphiquement (à partir d'un stripchart) puisque ces âges ne font pas partie de la "masse principale" (cf. graphique ci-dessous). Afin de se rappeler de ce traitement, la variable a été renommée par "Age entrée v2".

100 individus ont été retirés à cause de leur âge d'entrée "extrême".

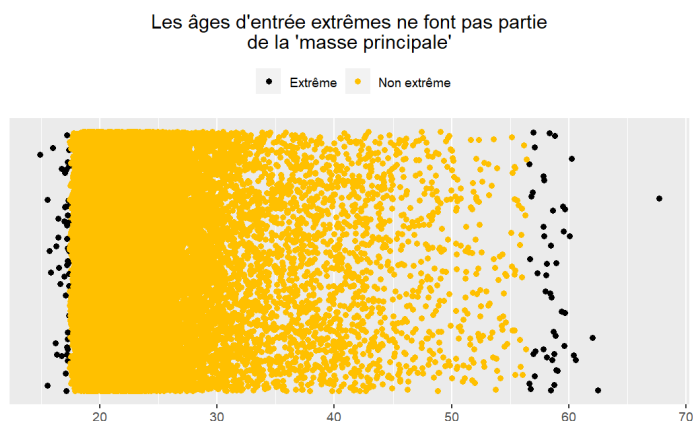


FIGURE 4.8 – Répartition des âges d'entrée

Les 100 niveaux de vie à 15 ans les plus extrêmes ont été retirés. Après avoir vérifié que la répartition des niveaux de vie à 15 ans est en

95 individus ont été également retirés à cause de leur niveau de vie à 15 ans "extrême".

"cloche", les 50 niveaux les plus petits (inférieurs à 81,6%) et les 50 niveaux les plus grands (supérieurs à 391,5%) ont été retirés. Ces deux seuils ont été tout d'abord validés graphiquement (à partir d'un stripchart) puisque ces niveaux ne font pas partie de la "masse principale" (cf. graphique ci-dessous). Afin de se rappeler de ce traitement, la variable a été renommée par "Niveau de vie 15 ans v2".

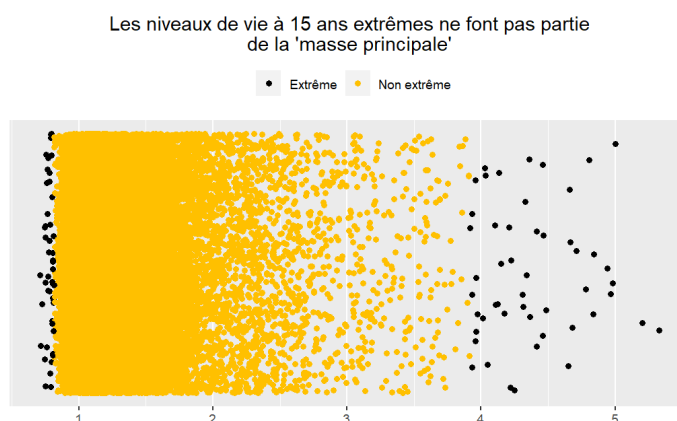


FIGURE 4.9 – Répartition des niveaux de vie

A noter que la taille de la base de données est passée de 32 851 à 32 656 (soit une baisse de 195) car 95 individus ont été retirés pour avoir un âge d'entrée jugé extrême, 95 autres individus ont été retirés pour avoir un niveau de vie à 15 ans jugé extrême, et 5 autres individus ont été retirés pour avoir un âge d'entrée et un niveau de vie à 15 ans jugés extrêmes.

Les 3 variables prédictives restantes ont été conservées comme telles, à savoir : le genre, le statut et le bassin géographique. Ces variables, qualitatives, n'ont effectivement aucune modalité représentée par moins de 100 individus.

4.1.3 Identification et traitement des valeurs extrêmes de la variable à prédire

Certains individus ont été identifiés comme ayant un écart "âge retraite - âge minimal" extrême. Pour rappel, ces individus sont dans l'une des situations suivantes :

- Le départ en retraite est survenu avant l'âge minimal possible dans le cadre de carrières longues ;

622 écarts "âge retraite - âge minimal" sont considérés "extrêmes".

- Le salarié travaille dans le secteur privé et le départ en retraite est survenu après 70 ans (ce qui correspond à l'âge limite) ;
- Le salarié travaille dans le secteur public et le départ en retraite est survenu après 67 ans (ce qui correspond à l'âge limite des fonctionnaires de la catégorie sédentaire).

C'est ainsi que 622 individus ont été identifiés avec un écart "âge retraite - âge minimal" extrêmes (soit 1,90% des observations) : la première situation concerne 573 individus (soit 1,75% des observations) et, la deuxième et troisième situations concernent, à elles deux, 49 individus (soit 0,15% des observations).

Une Analyse Factorielle des Données Mixtes (AFDM) a été réalisée. Cette analyse a l'objectif de savoir si les individus identifiés forment un groupe aux caractéristiques homogènes. Si ce n'est pas le cas, ces individus seront retirés de la base car ils risquent de biaiser les prédictions. Tout d'abord, les trois premiers axes principaux ont été retenus puisqu'ils expliquent une part importante de la variance totale des données (26,5%), et que le critère du coude suppose de retenir l'axe 1 ou les axes 1 à 3 (un changement abrupt de la pente est formé par le graphique des valeurs propres ci-dessous).

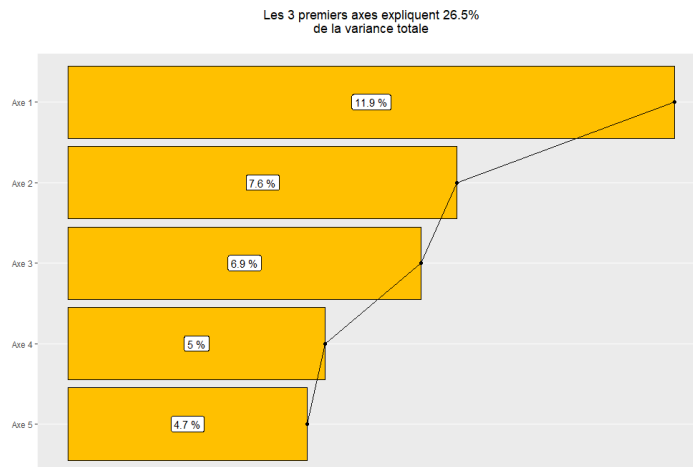


FIGURE 4.10 – Représentation des valeurs propres de l'AFDM

Chaque axe est caractérisé par certaines variables : celles qui contribuent le plus à la variance expliquée par l'axe. Puisqu'il y a 29 variables prédictives "effectives" (c'est-à-dire, 29 variables prédictives en prenant compte les modalités des variables qualitatives), la contribution des variables s'élèverait à

Les 3 premiers axes de l'AFDM ont été sélectionnés.

3,4%, si les variables contribuaient de manière uniforme. Ainsi, toutes les variables ayant une contribution supérieure à 3,4% sont considérées comme les variables qui caractérisent l'axe. Par exemple (cf. graphique ci-dessous), le premier axe est caractérisé par 8 variables : le niveau de vie à 15 ans, le statut privé, les catégories F, G et H, l'âge d'entrée, le domaine métier "Accompagnement" et le bassin géographique "Ile de France". L'axe 2 et 3 sont respectivement caractérisés par 7 et 10 variables prédictives "effectives".

La contribution des variables aux axes a été analysée.

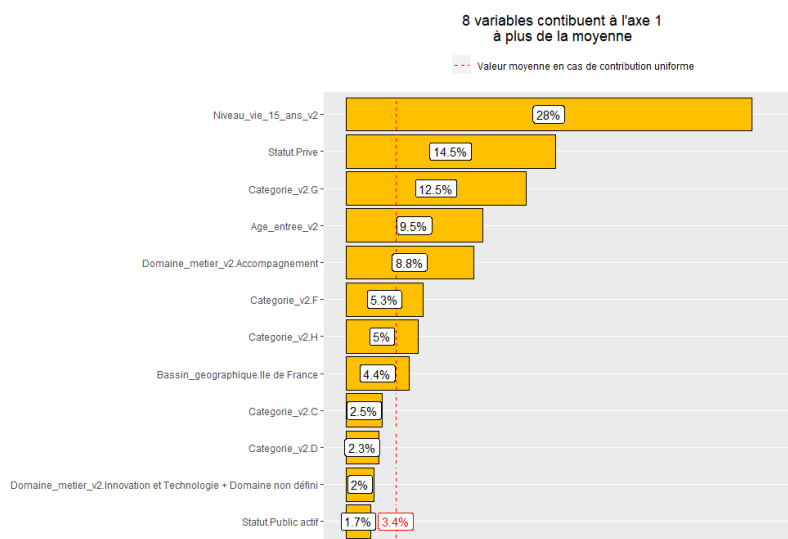


FIGURE 4.11 – Contribution des variables au premier axe

Ceci permet de sélectionner les variables à afficher lors de l'analyse des individus. Puisque 3 axes ont été retenus, il y a 3 plans possibles pour représenter les individus : le plan "axe 1 - axe 2", le plan "axe 1 - axe 3" et le plan "axe 2 - axe 3". Sur le plan "axe 1 - axe 2" (cf. graphique ci-dessous), il est constaté que les individus partis en retraite "extrêmement" tôt (c'est-à-dire avant l'âge minimal possible dans le cadre de carrières longues) sont proches du 0 de l'axe 1 et s'étalent sur toute la longueur de l'axe 2. Ainsi, l'axe 1 ne permet pas de faire de conclusion et l'axe 2 ne permet pas d'affirmer que ces individus forment un groupe d'individus ayant des caractéristiques similaires.

Concernant les individus partis en retraite "extrêmement tard" (c'est-à-dire après 67 ou 70 ans, selon leur statut), ils sont étalés sur la partie droite du plan (supérieure et inférieure), ce qui indique également qu'ils ne forment pas un groupe aux caractéristiques similaires.



FIGURE 4.12 – Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 1 - axe 2"

Sur le plan "axe 1 - axe3" (cf. graphique ci-dessous), les individus partis "extrêmement tôt" sont trop proches de l'origine pour les analyser. Quant aux individus partis "extrêmement" tard, ils sont trop éloignés les uns des autres pour déduire qu'ils forment un groupe.

Les 622 individus ont été visualisés sur les 3 plans : ils ne forment aucun groupe.

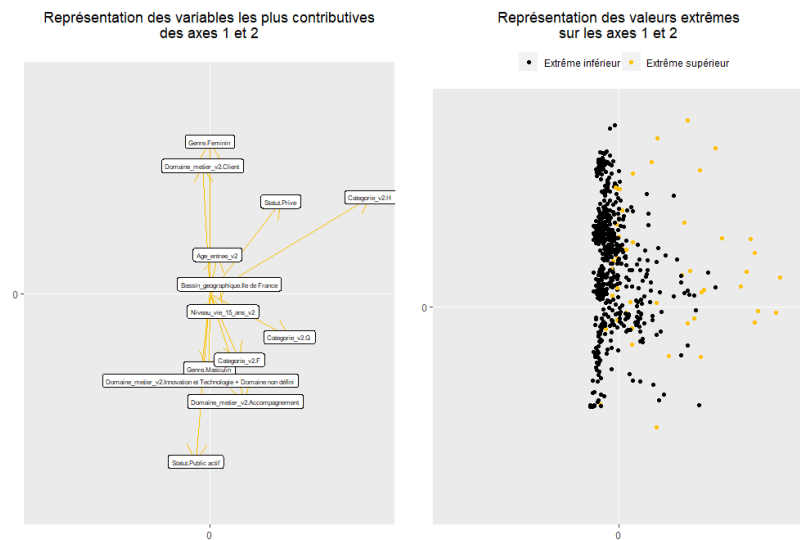


FIGURE 4.13 – Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 1 - axe 3"

Enfin, sur le plan "axe 2 - axe 3" (cf. graphique ci-dessous), les individus ne forment pas de groupe : ils sont étalés sur tout le plan et sont éloignés les uns des autres.

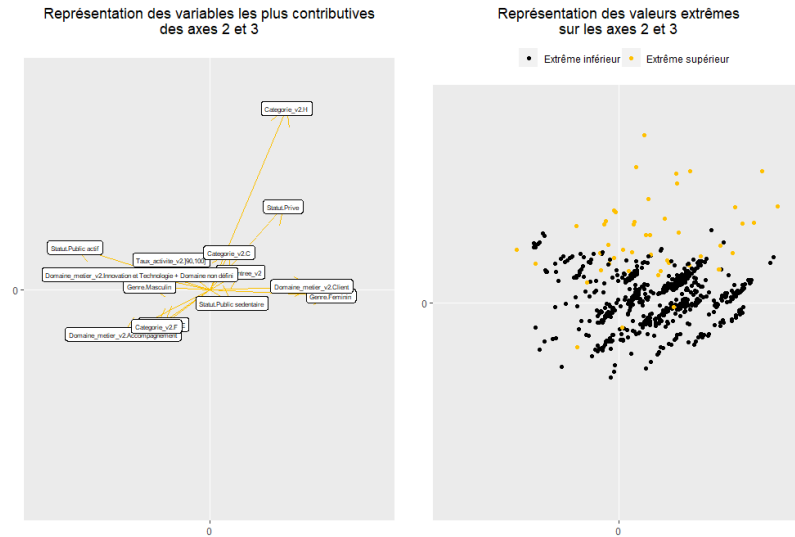


FIGURE 4.14 – Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 2 - axe 3"

Par conséquent, l'AFDM a permis de s'assurer que les individus identifiés ne forment pas un groupe d'individus aux caractéristiques similaires. Ils ont été ainsi retirés de la base de données, dont la taille est passée de 32 656 observations à 32 034 observations (soit une baisse de 622 observations).

4.1.4 Identification et traitement des variables prédictives fortement liées

Aucun couple de variables prédictives qualitatives n'a été identifié comme un couple de variables fortement liées. Dans un premier temps, l'indépendance des variables a été testée à partir du test du khi-deux. Celui-ci a systématiquement rejeté l'hypothèse d'indépendance (au seuil de 5%) : il existe ainsi un lien entre chacune des variables prédictives qualitatives. Toutefois, ces liens (mesurés grâce au V de Cramer) ne sont pas supérieurs à 75% (cf. graphique ci-dessous).

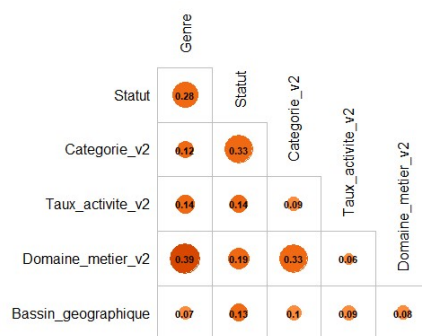


FIGURE 4.15 – V de Cramer entre chaque variables prédictives qualitatives

Les V de Cramer n'excèdent pas 75%.

La plus forte liaison est effectivement évaluée à 39% et s'établit entre "Domaine métier v2" et "Genre". Le graphique à barres empilées en pourcent (cf. ci-dessous) indique que les hommes sont majoritaires au sein des métiers des domaines "Innovation et Technologie et Domaine non défini" (87%) et "Accompagnement" (76,6%). Néanmoins, cette corrélation n'est pas assez forte pour supposer que la variable "Domaine métier" va prédire les mêmes comportements vis-à-vis de la retraite que la variable "Genre".

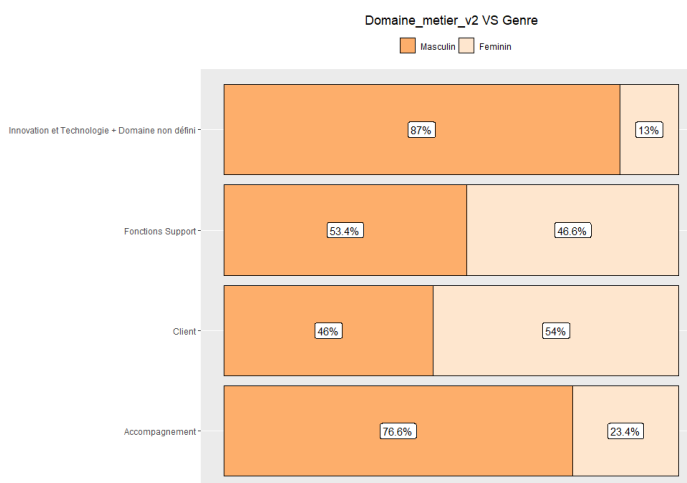


FIGURE 4.16 – Répartition des genres selon le domaine du métier

Les variables "Age d'entrée v2" et "Niveau de vie à 15 ans v2" ne sont pas fortement liées. Pour rappel, il s'agit des seules variables quantitatives de la base de données. Le test de Pearson et de Kendall ont rejeté l'hypothèse d'indépendance (au seuil de 5%) entre ces deux variables, mais leur liaison n'est pas évaluée à plus de 75% : 21% pour le coefficient de Pearson (qui mesure une corrélation linéaire) et -7% pour le coefficient

Les variables qualitatives ne sont pas fortement liées.

de Kendall (qui mesure une corrélation linéaire des rangs). Ces deux types de corrélations sont illustrés par les graphiques ci-dessous, qui confirment que la liaison entre les variables est faible.

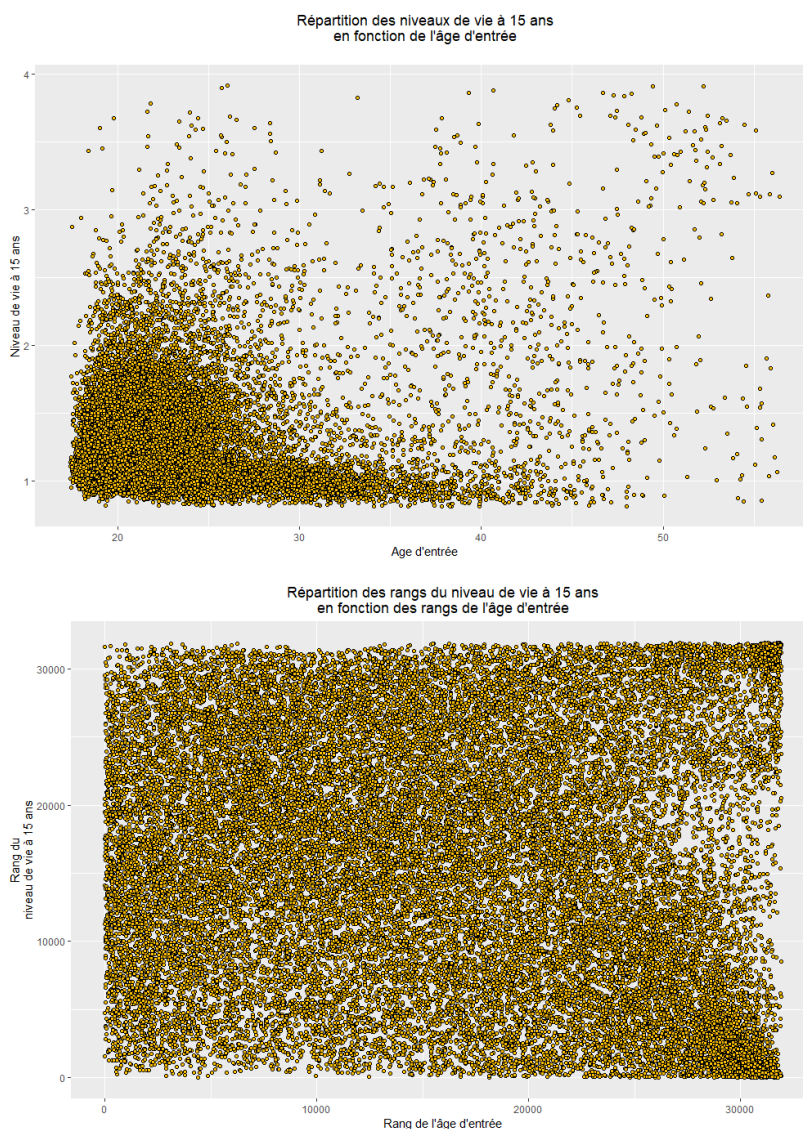


FIGURE 4.17 – Représentation des niveaux de vie à 15 ans en fonction de l'âge d'entrée dans l'entreprise

Une variable qualitative et une variable quantitative sont fortement liées : il s'agit de "Catégorie v2" et "Niveau de vie à 15 ans v2". Tout d'abord, des tests d'indépendance ont été réalisés pour identifier les couples de variables (de natures différentes) indépendantes : le test de

Wilcoxon-Mann-Whitney (lorsque la variable qualitative n'a que deux modalités) ou Kruskal-Wallis (lorsqu'elle en possède plus). Ces tests ont rejeté l'hypothèse d'indépendance pour chacun des couples de variables. Ensuite, le rapport de corrélation a alors été calculé pour évaluer à quel point les variables d'un couple sont dépendantes entre elles (cf. graphique ci-dessous).

La catégorie et le niveau de vie à 15 ans des individus sont fortement liés.

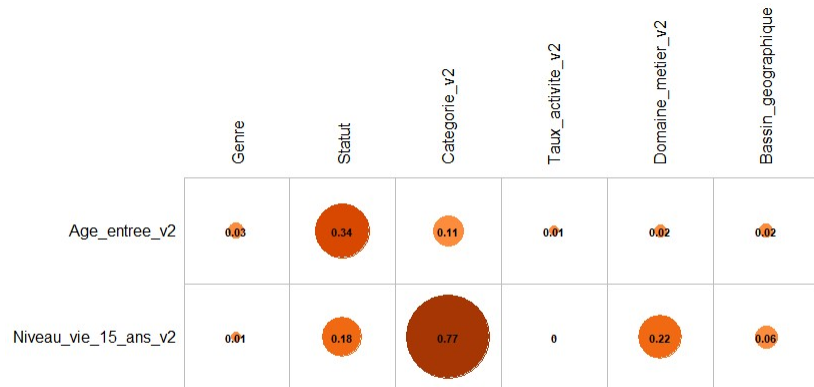


FIGURE 4.18 – Rapports de corrélation des couples de variables de natures différentes

Un seul couple de variables a un rapport de corrélation supérieur à 75% : il s'agit du couple "Catégorie v2" et "Niveau de vie à 15 ans v2". Cette corrélation a ensuite été vérifiée à l'aide du graphique "boîte à moustache" qui met en évidence la répartition des niveaux de vie à 15 ans au sein des différentes catégories (cf. ci-dessous) : plus la catégorie de l'individu est élevée, plus son niveau de vie à 15 ans l'est également.

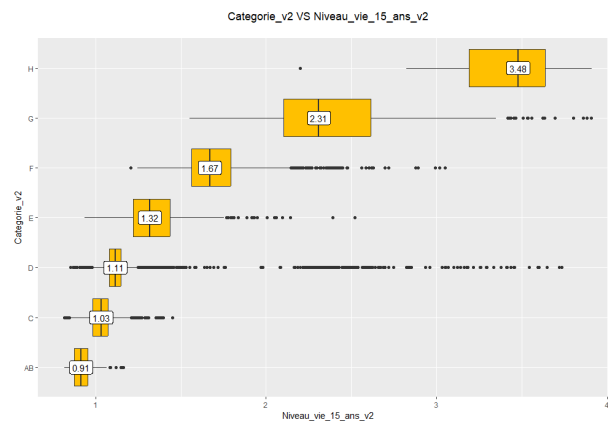


FIGURE 4.19 – Répartition des niveaux de vie à 15 ans selon la catégorie

Par conséquent, il est considéré que ces deux variables vont prédire les mêmes comportement vis-à-vis de la retraite : leur présence dans la base de données n'est donc pas recommandée. Toutefois, pour déterminer la variable à retirer, il a été décidé de réaliser un test d'indépendance entre elle et la variable à prédire (à savoir, l'écart entre "âge retraite - âge minimal") :

Le niveau de vie a été retiré car sa liaison avec la variable à prédire est très faible.

- le test d'indépendance de Kruskal-Wallis a rejeté l'hypothèse d'indépendance (au seuil de 5%) entre "Catégorie v2" et "Ecart v2" ;
- le test d'indépendance de Pearson a rejeté l'hypothèse d'indépendance (au seuil de 5%) entre "Niveau de vie à 15 ans v2" et "Ecart v2" ;
- le test d'indépendance de Kendall n'a pas rejeté l'hypothèse d'indépendance (au seuil de 5%) entre "Niveau de vie à 15 ans v2" et "Ecart v2".

Ainsi, une incertitude sur la dépendance entre "Niveau de vie à 15 ans v2" et "Ecart v2" existe puisque le test de Pearson rejette l'indépendance, tandis que le test de Kendall ne la rejette pas. Puisque le coefficient de Pearson s'élève à seulement 5% (et que le nuage de point ci-dessous ne révèle pas une forte corrélation), il a été décidé de considérer que la liaison linéaire entre ces deux variables était trop faible.

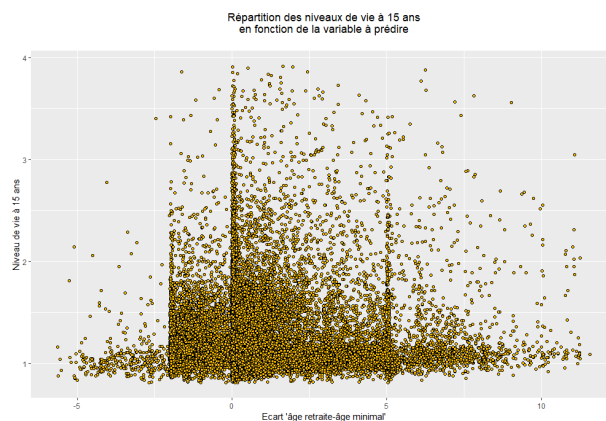


FIGURE 4.20 – Répartition des niveaux de vie à 15 ans en fonction de l'écart "âge retraite - âge minimal"

Ainsi, le niveau de vie à 15 ans a été retiré de la base de données des retraités. Or, lors de l'identification et du traitement des caractéristiques

peu représentées, 100 individus ont été retirés de la base pour avoir un niveau de vie à 15 ans jugé extrême. Parmi eux, 5 ont un âge d'entrée jugé extrême et 2 ont un écart "âge retraite - âge minimal" jugé extrême. Par conséquent, 93 de ces individus ont été réintégrés dans la base, dont la taille est passée de 32 034 observations à 32 127 observations.

4.1.5 Présentation de la base de données des retraités après traitements

Après l'analyse et le traitement des données, la base contient finalement 32 127 observations et 7 variables prédictives :

La base de données des retraités contient finalement 32 127 observations et 7 variables prédictives.

- le genre (conservé comme tel),
- le statut (conservé comme tel),
- la catégorie (dont la modalité "A" a été réunie avec la modalité "B"),
- le domaine du métier (dont la modalité "Domaine non défini" a été réuni avec la modalité "Innovation et Technologie"),
- le bassin géographique (conservé comme tel),
- le taux d'activité (qui a été transformé en variable qualitative pour obtenir 4 modalités : [0,50], [80,90],]50,80[,]90,100]),
- l'âge d'entrée (sachant que les âges les plus extrêmes ont été retirés).

L'échantillon test est composé de 4 900 observations. Le reste de la base est divisé en 17 blocs de tailles égales.

De plus, l'échantillon test est composé de 4 900 observations (soit 15,25% de la base), ce qui laisse 27 227 observations pour la modélisation. Pour rappel, à chaque modélisation, il faut s'assurer de la stabilité du modèle. C'est pourquoi plusieurs modèles seront construits (avec un échantillon d'apprentissage différent) et évalués (avec un échantillon de validation différent). Pour établir les données de l'échantillon d'apprentissage et celles de l'échantillon de validation lors d'une construction d'un modèle, les 27 227 observations ont été divisés en 17 blocs de tailles similaires (7 blocs de 1 601 observations et 10 blocs de 1 602 observations) : 16 blocs seront sélectionnés pour établir l'échantillon d'apprentissage et le bloc restant sera destiné à l'échantillon de validation (d'où la construction de 17 modèles avec un échantillon d'apprentissage et de validation différent). Toutefois, plusieurs niveaux de complexité du modèle doivent être construits afin de surveiller que le phénomène de sur-apprentissage n'apparaisse pas. Pour cela, lors de chacune des 17 constructions, seule une partie des variables prédictives

seront retenues. Puisqu'il existe 7 variables prédictives, il est possible de réaliser $2^7 = 128$ combinaisons de variables prédictives (à des niveaux de complexité différents). Par exemple, une construction consistera à retenir seulement les variables "Genre" et "Statut", alors qu'une autre retiendra les variables "Genre", "Catégorie" et "Bassin géographique". Par conséquent, $17 \times 128 = 2176$ modèles seront construits, pour chaque modélisation.

4.2 Modélisation de l'âge de retraite à partir d'une régression linéaire

La première modélisation proposée est la régression linéaire. Elle a été sélectionnée car il s'agit d'un modèle connu et simple à interpréter. Son application peut représenter toutefois une difficulté, qui peut être effacée par la création d'un "outil" destiné au client. Dans un premier temps, la modélisation de la régression linéaire sera présentée. Ensuite, les résultats de l'application de la méthodologie seront présentés, à savoir : la stabilité du modèle et le phénomène de sur-apprentissage. La troisième partie sera consacrée à la présentation du modèle final retenu (parmi les 2 176 modèles construits) ainsi que l'évaluation de son pouvoir prédictif sur l'échantillon test (à l'aide des deux métriques présentée au chapitre précédent).

4.2.1 Présentation générale de la régression linéaire

La régression linéaire a été choisie car elle est facilement interprétable.

La variable à prédire est supposée être égale à une combinaison linéaire des variables prédictives. La méthode des moindres carrés ordinaires (MCO) permet de calculer les coefficients de la combinaison linéaire. Elle suppose que les variables prédictives sont indépendantes entre elles. En pratique, cette hypothèse est difficile à satisfaire : il est donc toléré lorsqu'aucune liaison entre variable n'est forte. Lors de l'analyse et du traitement des données, ceci a dû être vérifié pour éviter la présence de variables expliquant le même comportement. Par conséquent, il est supposé que la base de données (après traitements) est adaptée pour appliquer une régression linéaire. D'autre part, la significativité de chaque coefficient de la combinaison linéaire est testée (à l'aide du test de Student). Lorsqu'un coefficient est jugé non significatif (au seuil de 5%), son terme est alors retiré : il s'agit de l'ajustement du modèle. En pratique, si le terme fait référence à une variable quantitative, cela revient à retirer cette variable de l'échantillon d'apprentissage. Si le terme fait référence à une modalité d'une variable qualitative, cela revient à réunir cette modalité avec la modalité

de référence. Une fois cette correction faite sur l'échantillon d'apprentissage, une nouvelle régression linéaire est calculée. L'opération est répétée jusqu'à ce que tous les coefficients soient significatifs (au seuil de 5%). Il convient ensuite de vérifier les hypothèses de validité d'une régression linéaire, à savoir : le modèle est linéaire, les résidus (c'est-à-dire la différence entre les écarts réels et les écarts prédits, sur l'échantillon d'apprentissage) suivent une loi normale centrée, sont indépendants et homoscedastiques (la variance est uniforme). Ces hypothèses peuvent être vérifiées à l'aide d'un test statistique, accompagné d'un ou plusieurs graphiques :

Cette modélisation est paramétrique : des hypothèses doivent être vérifiées.

- Linéarité du modèle : le test de Rainbow dont l'hypothèse nulle est "Le modèle est linéaire" est adapté. Ce test est généralement accompagné d'un graphique en nuage de points, représentant l'évolution des résidus en fonction des prédictions. Pour vérifier la linéarité du modèle, les points doivent être symétriquement répartis de part et d'autre de la droite $y = 0$ (sans être éloignés de cette droite) et aucune forme (rond, entonnoir, etc.) n'est formée.
- Normalité des résidus : le test de Kolmogorov-Smirnov peut être réalisé, dont l'hypothèse nulle est "l'échantillon suit une loi gaussienne". Il est accompagné de deux graphiques : l'histogramme et le diagramme quantile-quantile. L'histogramme doit former une cloche dont la hauteur et l'aplatissement doivent être similaires à ceux de la densité de la loi normale de moyenne nulle et d'écart-type identique à celui des résidus. Le diagramme quantile-quantile consiste à représenter l'évolution de différents quantiles des résidus en fonction des quantiles de la loi normale centrée. La normalité des résidus est donc vérifiée si les points de ce graphique sont situés sur la première diagonale.
- Indépendance des résidus : il est possible de réaliser le test de Durbin-Watson dont l'hypothèse nulle est "Les résidus sont non-autocorrélés". Ce test est accompagné du graphique des résidus en fonction des prédictions, à l'instar du test de Rainbow (il doit d'ailleurs être interprété de la même façon).
- Homoscedasticité des résidus : le test de Breusch-Pagan, dont l'hypothèse nulle est "Les résidus sont homoscedastiques". Ce test est accompagné également du graphique des résidus en fonction des prédictions, à l'instar du test de Rainbow et du test de Durbin-Watson (avec la même interprétation).

L'analyse du phénomène de sur-apprentissage sera réalisée grâce au nombre de termes présents dans la régression.

Les coefficients des termes de la régression seront analysés pour déterminer la stabilité du modèle.

Le phénomène de sur-apprentissage sera vérifié au regard du nombre de termes dans la combinaison linéaire. Pour rappel, 128 bases de données des retraités ont été construits : chacune d'elle ayant une combinaison différente de variables prédictives. Ceci permettra donc de varier le nombre de termes dans la combinaison linéaire et d'étudier l'évolution du pouvoir prédictif (la RMSE) en fonction de ce nombre. A noter qu'il existera plusieurs modèles avec le même nombre de termes, donc l'étude portera plus précisément sur l'évolution de la RMSE moyenne. De plus, le nombre de termes ne découlera pas nécessairement du nombre de variables (et de modalités) de l'échantillon d'apprentissage. En effet, l'ajustement du modèle peut conduire à retirer des variables : le nombre de termes est donc plus faible.

La stabilité du modèle sera étudiée au regard des coefficients obtenus et de la RMSE calculée. Pour rappel, les observations de la base de données qui ne sont pas destinées à l'échantillon test ont été réparties dans 17 blocs distincts. Cette répartition permet de construire 17 régressions linéaires (ayant le même nombre de termes), à partir d'un échantillon d'apprentissage composé de 16 blocs. Le bloc restant (qui n'a pas fait partie de l'échantillon d'apprentissage) permet d'évaluer le pouvoir prédictif de la régression linéaire construite. Ainsi, la stabilité du modèle pourra être analysée à travers la dispersion des 17 coefficients de chaque variable prédictive et des 17 RMSE calculée.

La régression linéaire retenue sera celle avec le niveau de complexité "optimal" et la plus petite RMSE. Finalement $128 \times 17 = 2176$ modèles seront construits. Dans un premier temps, le niveau de complexité ayant une RMSE moyenne plus faible sera retenu. Ceci permet de s'assurer que le phénomène de sur-apprentissage n'apparaisse pas. Ensuite, parmi les modèles construits de même complexité, le modèle retenu sera celui avec la plus petite RMSE (calculée sur l'échantillon de validation)

4.2.2 Analyse du phénomène de sur-apprentissage et de la stabilité du modèle

Le phénomène de sur-apprentissage apparaît. Selon le graphique ci-dessous, il est effectivement constaté que la RMSE est parfois supérieure à une autre dont le niveau de complexité est moindre. Par exemple, lorsque le nombre de termes s'élève à 19, les modèles ont en moyenne une RMSE

La RMSE est parfois moins faible pour un nombre de termes plus élevé : le phénomène de sur-apprentissage semble apparaître.

de 1,59 ans. Tandis que lorsque ce nombre s'élève à 20, les modèles ont en moyenne une RMSE de 1,61 ans. Le nombre de termes retenu est donc 19.

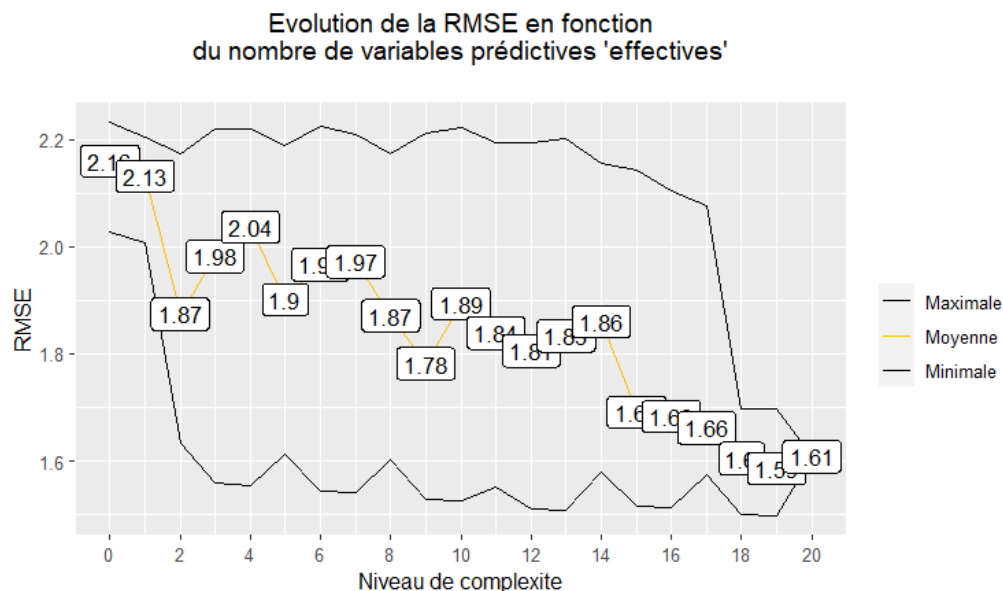


FIGURE 4.21 – Evolution de la RMSE en fonction du niveau de complexité

47 régressions linéaires ont 19 termes. L'analyse de la stabilité du modèle se réalise donc sur ces 47 modèles. Tout d'abord, il est constaté que les variables "Genre - Masculin", "Age d'entrée v2", "Domaine métier v2 - Fonction support" et "Domaine métier v2 - Innovation et Technologie + Domaine non défini" apparaisse respectivement dans 46, 34, 13 et 1 modèle(s). Sinon, toutes les autres variables font partie des 47 combinaisons linéaires. Afin d'analyser la stabilité du modèle, l'écart-type des coefficients des variable a été calculé (cf. graphique ci-dessous). Il est constaté que les coefficients de l'interception varient de 2,1 ans en moyenne. La variation des autres coefficients est plus modérée mais reste toutefois assez élevée. Les coefficients ne sont pas stables, ce qui est un inconvénient dans une démarche prédictive : les données observées ne semblent donc pas adaptée pour prédire l'âge de retraite.

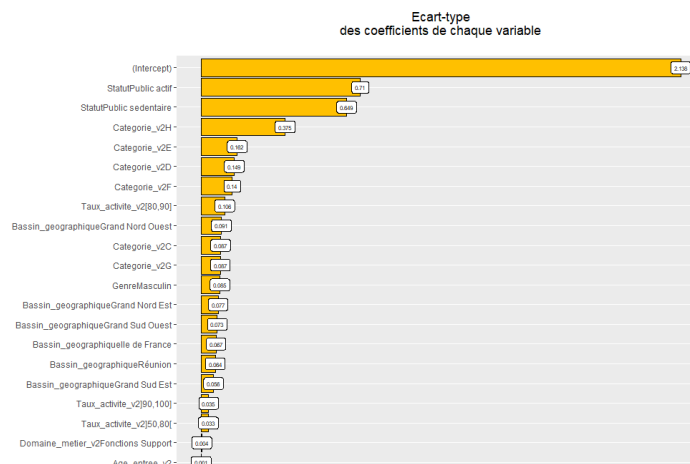


FIGURE 4.22 – Dispersion des coefficients des variables

D'autre part, la RMSE des 47 variables varie entre 1,5 ans et 1,7 ans (cf. graphique ci-dessous), avec un écart type de 0,05 ans. Contrairement à la stabilité des coefficients, la stabilité de la RMSE est acceptable puisque 0,05 ans représente moins d'un mois. Par conséquent, même si les modèles ont des coefficients différents, la RMSE reste sensiblement identique. Ceci confirme alors que les données observées ne sont pas adaptées pour prédire l'âge de retraite.

Les coefficients ne sont pas stables, contrairement aux RMSE obtenues.

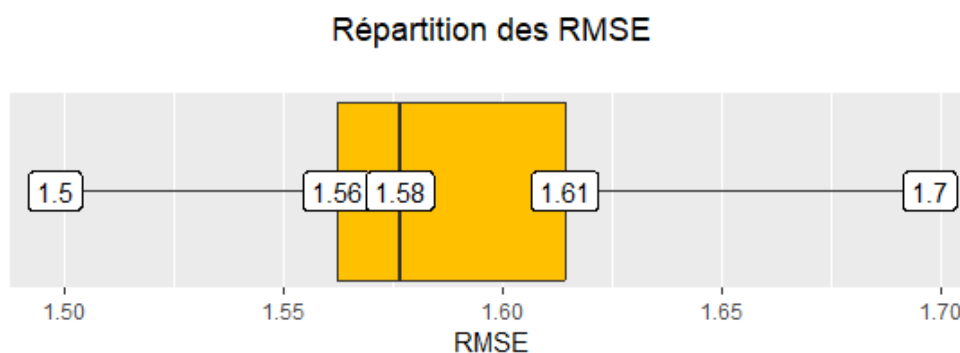


FIGURE 4.23 – Dispersion des RMSE

4.2.3 Analyse du modèle final

Le modèle final correspond à celui ayant la plus petite RMSE (calculée sur l'échantillon de validation). La valeur de ces coefficients sont présentées dans le graphique ci-dessous.

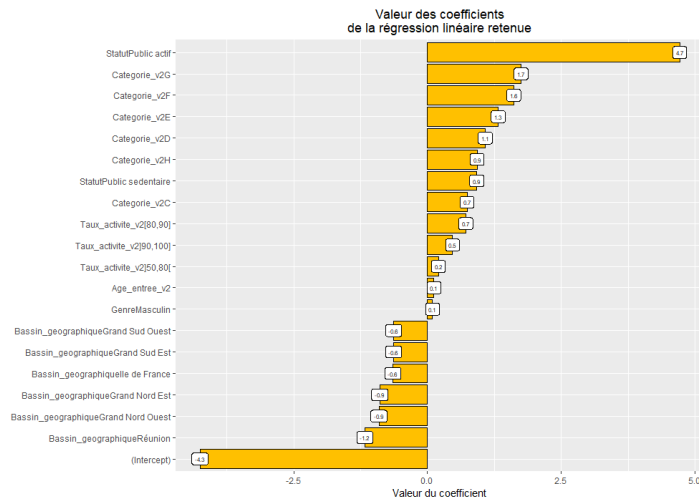


FIGURE 4.24 – Valeurs des coefficients de la régression linéaire retenue

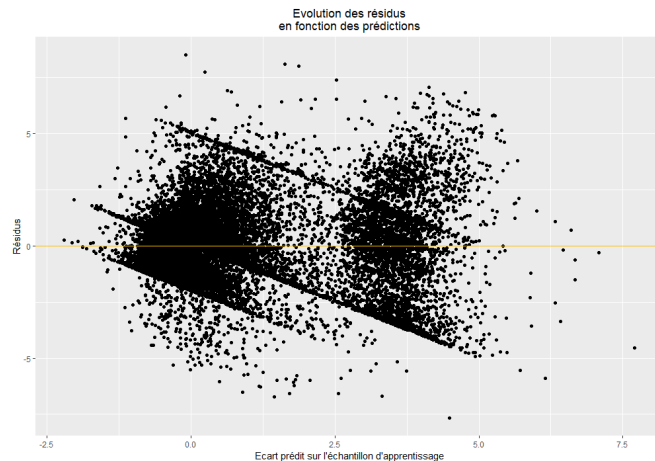
L'avantage principal de la régression vient de sa facilité d'interprétation. En effet, à partir des coefficients du modèle, il est possible d'affirmer les points suivants (par exemple) :

Les coefficients d'une régression linéaire permettent d'interpréter clairement l'impact des variables prédictives sur la variable à prédire.

- Toutes choses égales par ailleurs, le départ en retraite des fonctionnaires de la catégorie active survient en moyenne 4,7 ans plus tard que les salariés du privés, vis-à-vis de la date d'atteinte de l'âge d'ouverture des droits.
- Toutes choses égales par ailleurs, un salarié entré dans l'entreprise une année plus tard, part en moyenne 0,1 an (soit 1,2 mois) plus tard à la retraite, vis-à-vis de sa date d'ouverture des droits.

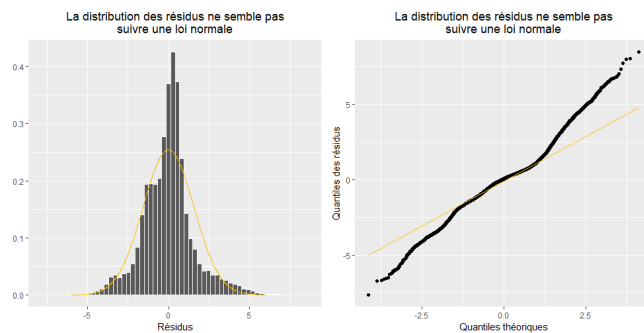
Toutefois, ce modèle ne vérifie aucune hypothèses de validité de la régression linéaire :

- La linéarité n'est pas vérifiée par le test de Rainbow (car la p-value est inférieure à 5%), ni par le graphique (cf. ci-dessous) qui met en évidence deux groupes.



- La normalité des résidus n'est pas vérifiée par le test de Kolmogorov-Smirnov, ni par les graphiques (cf. ci-dessous) qui démontrent que les résidus ne s'ajustent pas à une loi normale : l'histogramme ne forme pas la cloche de la densité de la loi normale centrée de même écart-type, et le graphique quantile-quantile ne forme pas la première diagonale.

Les hypothèses de validité de la régression linéaire ne sont pas vérifiées.



- L'indépendance et l'homoscédasticité des résidus ne sont pas vérifiées par le test de Durbin-Watson et par le test de Breusch-Pagan, ni par le graphique (comme vu précédemment).

Le pouvoir prédictif du modèle a été évalué sur l'échantillon test, par le calcul de la RMSE. Tout d'abord, la RMSE obtenue s'élève à 1,61 ans, autrement dit : le modèle prédit un écart "âge retraite - âge minimal" qui est éloigné en moyenne de 1,61 ans de l'écart réel. Dans le contexte d'évaluation de l'engagement IFC, cette erreur moyenne pose problème. En effet, l'engagement IFC est (au moins) évalué une fois par an (à la date de clôture de l'exercice comptable) et projeté à la prochaine date de

clôture (soit 1 an plus tard). Ainsi, l'erreur moyenne du modèle ne devrait pas dépasser quelques mois, pour que celui-ci prédise au mieux les départs qui surviendront au cours de l'année prochaine. Le graphique ci-dessous, qui compare les écarts prédits aux écarts réels, permet de bien illustrer la RMSE.

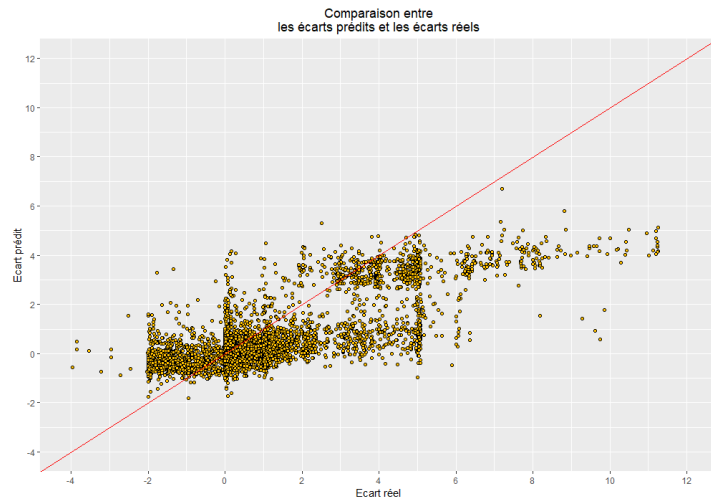


FIGURE 4.25 – Comparaison des écarts prédits et des écarts réels, sur l'échantillon test

Les prédictions s'éloignent en moyenne de 1,61 ans de la réalité.

Il est constaté que beaucoup de points sont éloignés de la première diagonale (autrement dit, les prédictions sont éloignées de la réalité). D'ailleurs, le résidus moyen (la différence entre les écarts réels et les écart prédits) s'élève à 0,58. Ceci signifie donc que le modèle a tendance à anticiper les départs en retraite. D'autre part, les écarts prédits sont répartis sur un intervalle bien plus restreint (de -1,8 ans à 6,7 ans) que celui sur lequel se répartissent les écarts réels (de -4,0 ans à 11,3 ans). Le modèle n'a donc pas réussi à capter les différences de "comportements" existants. Enfin, il est possible de constater 2 groupes : le premier groupe (le plus abondant) dont les écarts prédits sont situés autour de 0, et le second groupe dont les écarts prédits sont plutôt entre 3 et 4 ans. Après analyse, il s'avère que ces groupes sont caractérisés par le statut des individus : le premier groupe est principalement composé de salariés du privé et des fonctionnaires de la catégorie sédentaire. Le second groupe est, quant à lui, composé essentiellement de fonctionnaires de la catégorie active. Puisque l'âge d'ouverture des droits de ces derniers est faible (entre 55 et 57 ans, selon la génération de naissance), cela suppose que l'attribution de l'âge d'ouverture des droits est faussée. Pour rappel, celui-ci a été effectué selon le statut du salarié actuel. Or, un fonctionnaire peut bénéficier d'un départ au titre de la catégorie active, seulement s'il a

effectué une durée minimale de services actifs au cours de sa carrière. Par conséquent, cette attribution semble finalement être trop approximative.

Le pouvoir prédictif du modèle a été ensuite évalué par le calcul de l'écart actuariel 2019 qu'il engendre. L'écart actuariel sur l'exercice 2019 engendré par ce modèle s'élève à 10,1% de l'engagement projeté. Il est essentiellement expliqué par l'écart dû aux IFC puisque celui-ci s'élève à 9,8%. L'écart dû aux retraites réalisées et celui dû aux retraites non réalisées s'élèvent, quant à eux, à 0,1% et 0,2% respectivement. Comme introduit par la valeur de la RMSE, le pouvoir prédictif de ce modèle est très mauvais. Le modèle prédit une unique date de retraite par individu, ce qui permet d'étudier individuellement les écarts actuariels :

L'écart actuariel engendré par la régression linéaire est très conséquent : 10,1% de l'engagement projeté.

- Parmi les départs en retraite prévus au cours de l'année 2019, seulement 55,5% ne se sont pas réellement réalisés ;
- Parmi les individus pour qui le départ en retraite était prévu après l'année 2019, 1,5% sont partis en retraite.

L'inconvénient principal du modèle est qu'il anticipe trop les départs en retraite (ce qui a d'ailleurs été confirmé par le résidu moyen positif). Ceci peut s'expliquer par le fait qu'il ne considère pas l'âge actuel du salarié : il peut ainsi prévoir un âge de retraite qui a déjà été dépassé. Toutefois, cela n'a été le cas que pour 5,9% des individus dont la moitié sont finalement bien partis en retraite en 2019.

Pour conclure, la régression linéaire n'est pas adaptée. Tout d'abord, les hypothèses de validation ne sont pas respectées, ce qui suppose que aucune liaison linéaire n'existe entre les variables prédictives et la variable à prédire. De plus, les prédictions sur l'échantillon test sont éloignées en moyenne de 1,61 ans de la réalité et engendrent 10,1% d'écart actuariel sur l'exercice 2019. Cet écart est très conséquent et inacceptable.

4.3 Modélisation de l'âge de retraite à partir d'un arbre de régression

Suite à la régression linéaire (et les hypothèses de validité non respectées), l'arbre de régression a été choisi pour modéliser l'âge de retraite, car il est non-paramétrique. De plus, il est facilement interprétable et applicable grâce à sa représentation graphique. A l'instar de la partie précédente, cette

partie se structure de la façon suivante : présentation du modèle, analyse du phénomène de sur-apprentissage et de la stabilité du modèle et analyse du modèle final retenu.

4.3.1 Présentation générale de l'arbre de régression

L'arbre de régression est une méthode supervisée non paramétrique. Autrement dit, le modèle se construit à partir d'observations de départs en retraite, et il n'impose aucune hypothèse sur la loi de la variable à prédire. L'arbre de régression est appelé ainsi car la variable à prédire est continue (sinon, on parle d'arbre de classification) : le terme général pour évoquer ce modèle est la méthode CART pour "Classification And Regression Trees".

L'arbre de régression a été choisi car il est non paramétrique, facilement interprétable et applicable.

Ce modèle consiste à créer des groupes d'individus similaires au regard de la variable à prédire. Pour ce faire, il commence par diviser la base de données (appelée "racine") en deux groupes selon une règle de division se basant sur une variable prédictive (par exemple, les travailleurs sont des salariés du secteur privé ou non). Puis, il divise de nouveau les deux groupes obtenus (appelés "noeuds") en deux groupes (et ainsi obtenir 4 groupes), selon une autre règle de division. En régression, la règle de division est choisie de sorte à maximiser l'écart entre :

- La variance (de la variable à prédire) au sein du noeud à diviser ;
- La somme des variances au sein des noeuds descendants (qui sont nécessairement en nombre de deux).

Autrement dit, la division est choisie de sorte à minimiser la variance au sein des noeuds descendants. En théorie, il faudrait toutes les règles de division possible et sélectionner la plus optimale. Mais en pratique, il est impossible de tester toutes les règles de division existantes (le calcul serait très coûteux en temps). Dans ce cas, certaines règles sont testées aléatoirement jusqu'à obtenir une règle optimale "locale". Lorsque les derniers groupes sont obtenus (appelés "feuilles"), la moyenne des valeurs de la variable à prédire des individus composant un groupe correspond à la prédiction de l'arbre.

Un critère d'arrêt de division doit être fixé. En effet, l'arbre pourrait continuer à diviser les noeuds, jusqu'à obtenir des feuilles constituées d'un seul individu (l'arbre ainsi obtenu est appelé l'"arbre maximal"). Mais ceci est évidemment inintéressant : le phénomène de sur-apprentissage apparaît nécessairement. Par conséquent, il faut fixer un critère d'arrêt, appelé

Pour construire un arbre de régression, le paramètre de complexité a été fixé selon la règle du pouce et la validation croisée à 10 blocs.

"paramètre de complexité", pour que le processus de division s'arrête plus tôt. Ce critère est un seuil qui porte sur la qualité de la division : si aucune division testée ne permet d'obtenir un écart "variance noeud à diviser - somme variance noeuds descendants" supérieur à ce seuil, alors le noeud ne sera pas divisé. Toutefois, il reste à savoir comment fixer ce seuil. Pour cela, l'arbre maximal est construit grâce à un échantillon d'observations. A chaque division, deux éléments sont stockés :

- la RMSE de l'arbre obtenu (au stade des divisions réalisées), calculée sur un autre échantillon d'observations ;
- l'écart maximisé "variance noeud à diviser - somme variances noeuds descendants" de la division.

Le phénomène de sur-apprentissage apparaîtra nécessairement, c'est-à-dire qu'il existe une division $n^{\circ}d$ associée à une RMSE qui est supérieure à la RMSE de la division $n^{\circ}d - 1$. Ainsi, le paramètre de complexité optimal correspondrait à l'écart maximisé "variance noeud à diviser - somme variances noeuds descendants", qui est associé à la RMSE la plus faible. Toutefois cette méthode n'est pas stable, car elle dépend beaucoup de l'échantillon qui a servi à construire l'arbre maximal et de celui qui a servi à calculer les RMSE. Par conséquent, la validation croisée en $k = 10$ blocs est généralement utilisée pour associer, à chaque numéro de division, 10 RMSE (plutôt qu'une seule). Grâce à cette pratique, une méthode de fixation du paramètre de complexité (appelée "règle du pouce") peut être appliquée. Elle consiste à :

- Calculer la moyenne des 10 RMSE de chaque nombre de divisions, ainsi que l'écart-type ;
- Sommer la RMSE moyenne la plus faible et son écart-type ;
- Identifier les nombres de divisions dont la RMSE moyenne est inférieure à cette somme ;
- Parmi eux, sélectionner le nombre de divisions le plus faible ;
- Retenir le paramètre de complexité associé à ce nombre de divisions.

Cette méthode ne cherche donc pas seulement à fixer le paramètre de complexité pour minimiser la RMSE, mais également à minimiser le nombre de divisions réalisées. C'est méthode a été retenue pour modéliser l'âge de retraite.

Une fois le paramètre de complexité fixé, l'arbre est élagué. À partir de ce paramètre, il est possible de récupérer un "sous-arbre" de l'arbre maximal, tel que toutes les divisions réalisées ont permis d'obtenir un écart "variance noeud à diviser - somme variances noeuds descendants" supérieur au paramètre de complexité.

4.3.2 Analyse du phénomène de sur-apprentissage et de la stabilité du modèle

Le phénomène de sur-apprentissage apparaît localement. Il est possible de le constater à partir de l'évolution de la moyenne des RMSE pour chaque nombre de variables prédictives effectives du modèle (cf. graphique ci-dessous).

La RMSE évolue en dents de scie, en fonction du nombre de variables prédictives "effectives".

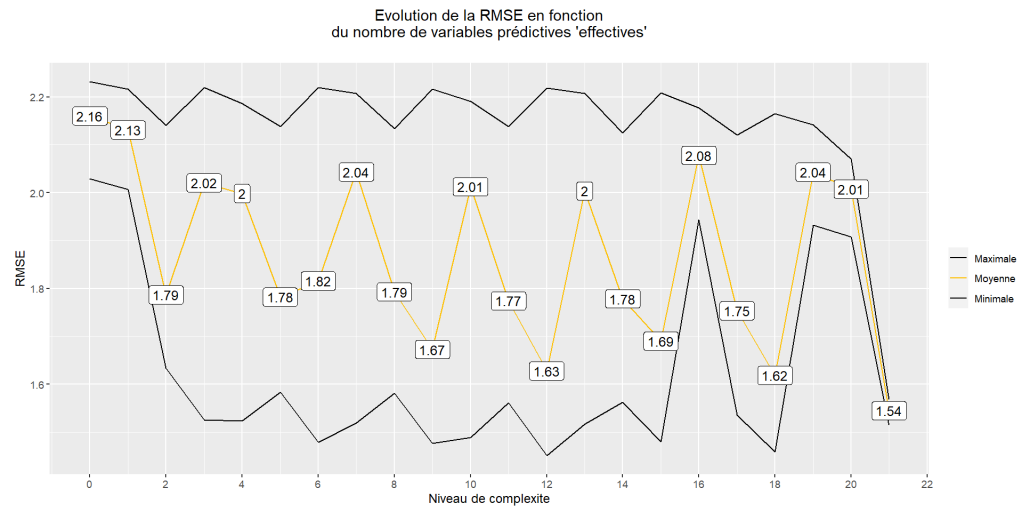


FIGURE 4.26 – Évolution de la RMSE en fonction du nombre de variables prédictives effectives

Toutefois, le phénomène n'apparaît que localement car le nombre maximal de variables prédictives effectives (21) est associé à la RMSE moyenne la plus petite. Par conséquent, il a été tout de même décidé de conserver le niveau de complexité maximal. Celui-ci a été atteint 4 fois : 1 fois à partir d'une base de données ayant 21 variables effectives (ce qui suppose une situation exceptionnelle) et 3 autres fois par la base de données complète. La base de données complète a donc été retenue.

La stabilité des paramètres a été vérifiée à travers l'importance des variables. Cette dernière correspond à la somme des écarts "variance

noeud à diviser - somme variances noeuds descendants" calculés lorsque la règle de division fait intervenir la variable en question. Ainsi, ce critère permet de mettre en avant les variables qui ont le plus été "utiles" pour regrouper les individus. D'après le graphique ci-dessous, l'importance des variables est assez stable puisque l'écart-type ne représente pas plus de 13,2% de la moyenne.

L'importance des variables et la RMSE sont stables.

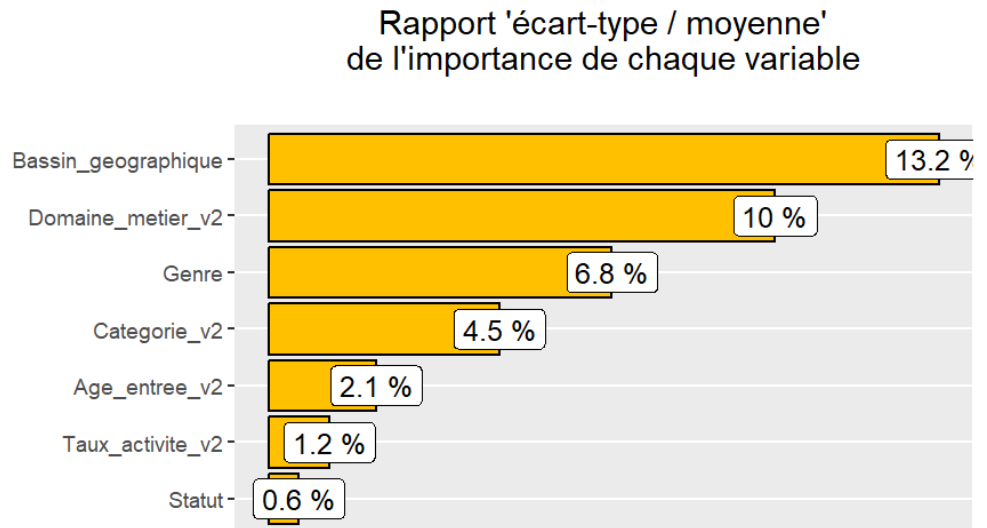


FIGURE 4.27 – Écart-type de l'importance des variables, selon sa moyenne

D'autre part, la RMSE (calculée sur l'échantillon de validation) des 17 modèles se situe entre 1,46 ans et 1,57 ans (cf. graphique ci-dessous).

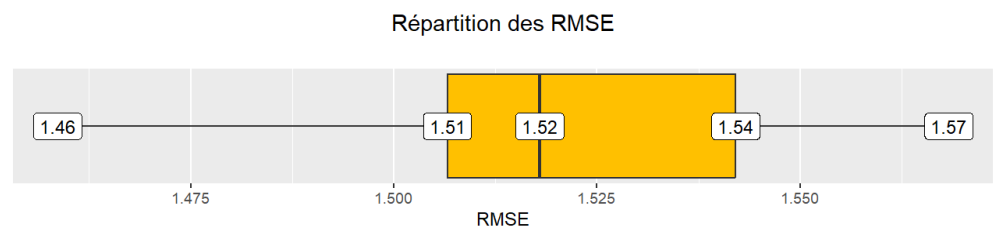


FIGURE 4.28 – Ecart-type

Finalement, il est possible de conclure que l'arbre de régression est stable pour construire un modèle sur l'âge de retraite. En effet, les paramètres et le pouvoir prédictif du modèle ne vacillent pas beaucoup selon l'échantillon d'apprentissage et l'échantillon de validation sélectionnés.

4.3.3 Analyse du modèle final

L'arbre de régression retenu correspond à celui ayant la plus faible RMSE sur son échantillon de validation. Celui-ci contient toutes les variables, sauf le genre et le domaine métier : il a donc 18 variables prédictives effectives. D'autre part, il a regroupé les individus en 23 groupes distincts. Le schéma ci-dessous illustre les 10 premières divisions de l'arbre (la représentation complète de celui-ci serait illisible).

La représentation graphique de l'arbre de régression est très explicite.

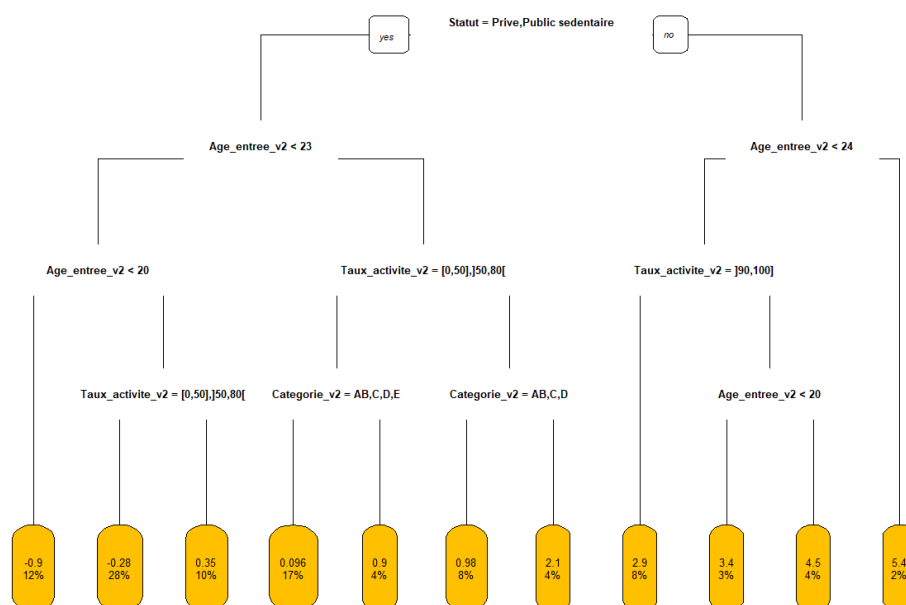


FIGURE 4.29 – Représentation simplifiée de l'arbre de régression retenu pour modéliser l'âge de retraite

A partir de cette représentation, il est possible de constater facilement que, par exemple, l'écart moyen "âge retraite - âge minimal" s'élève à -0,9 ans, au sein des salariés du secteur privé ou de la catégorie sédentaire du secteur public et dont l'âge d'entrée dans l'entreprise est inférieur à 20 ans. L'avantage principal de cette modélisation réside sur sa facilité d'interprétation et d'application, comme illustré ici.

La RMSE du modèle a été calculée sur l'échantillon test. Elle s'élève à 1,59 ans, ce qui suppose que le pouvoir prédictif du modèle est légèrement meilleur que celui de la régression linéaire (dont la RMSE s'élève à 1,61 ans). Toutefois, cela reste un mauvais pouvoir prédictif dans

L'arbre de régression prédit un écart "âge retraite - âge minimal" qui est en moyenne à 1,59 ans de l'écart réel.

le contexte des évaluations d'engagement IFC, puisque le modèle ne permet pas de bien prédire sur l'horizon d'une année. Cette RMSE peut être illustrée par le graphique ci-dessous qui compare les prédictions avec les valeurs réelles.

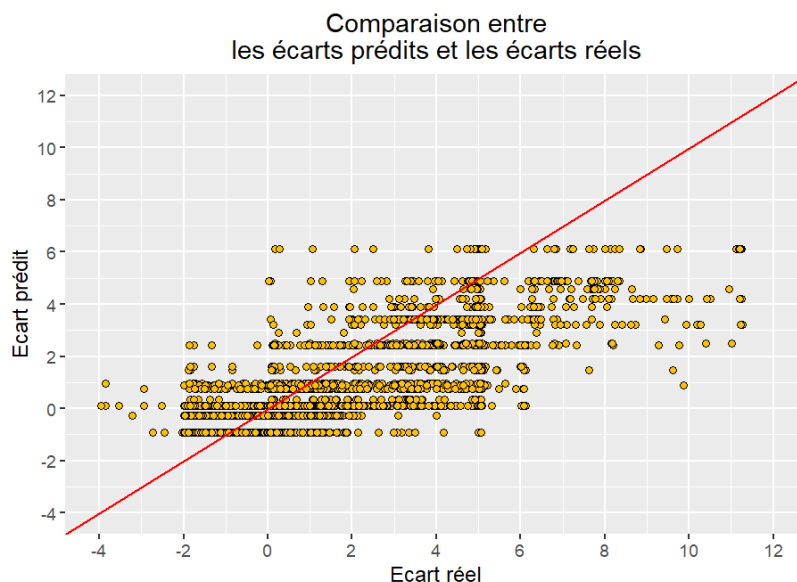


FIGURE 4.30 – Comparaison entre les écarts prédits et les écarts réels.

Il est constaté que des "lignes" se forment : cela provient du fait que le modèle ne prédit que 23 valeurs différentes. Tout comme la régression linéaire, les points sont éloignés de la première diagonale, ce qui signifie que les prédictions sont mauvaises. Il est aussi possible de remarquer que les 23 prédictions différentes s'étalent sur un intervalle allant de -1,0 à 6,0 ans, alors que les écarts "âge retraite - âge minimale" réels se situent dans un intervalle bien plus large (de -4,0 à 11,3 ans).

L'écart actuariel engendré par l'arbre est très conséquent : 11,7% de l'engagement projeté.

L'écart actuariel engendré par le modèle a ensuite été calculé. Il s'élève à 11,7% de l'engagement projeté. Il est essentiellement expliqué par l'écart dû aux IFC puisque celui-ci s'élève à 11,4%. L'écart dû aux retraites réalisées et celui dû aux retraites non réalisées s'élèvent, quant à eux, à 0,1% et 0,2% respectivement. Comme introduit par la valeur de la RMSE, le pouvoir prédictif de ce modèle est très mauvais. Le modèle prédit une unique date de retraite par individu, ce qui permet d'étudier individuellement les écarts actuariels :

- Parmi les départs en retraite prévus au cours de l'année 2019, seulement 55,3% ne se sont pas réellement réalisés ;

- Parmi les individus pour qui le départ en retraite était prévu après l'année 2019, 1,3% sont partis en retraite.

L'inconvénient principal du modèle est qu'il anticipe trop les départs en retraite, ce qui peut être confirmé par l'écart moyen entre les observations et les prédictions : 0,67 ans. De plus, il faut noter que 6,1% des âges de retraite prédits ont été dépassés, et 51,8% de ces individus sont effectivement partis en retraite en 2019.

Pour conclure, l'arbre de régression ne donne pas de résultats satisfaisants. Les prédictions sur l'échantillon test sont éloignées en moyenne de 1,59 ans de la réalité et engendrent 11,7% d'écart actuariel sur l'exercice 2019. Cet écart est très conséquent et inacceptable.

4.4 Comparaison des modélisations avec le modèle de référence

4.4.1 Évaluation du pouvoir prédictif du modèle de référence

Le modèle de référence prédit un âge de retraite qui est en moyenne à 1,75 ans de l'âge réel.

Le modèle actuellement utilisé par l'entreprise a été évalué sur l'échantillon test. Comparé aux modèles construits, celui ne prédit pas un âge de retraite par individu. Il prédit effectivement deux âges de retraite (l'âge d'ouverture des droits et l'âge du taux plein) dont chacun est associé à une probabilité (fixée selon les observations). Ainsi, pour évaluer la RMSE de ce modèle, il a été considéré que l'âge de retraite prédit correspond à la moyenne pondérée de ces âges. La RMSE obtenue s'élève à 1,75 ans. Le graphique ci-dessous indique que l'éventail des prédictions est restreint (entre 0 et 5 ans).

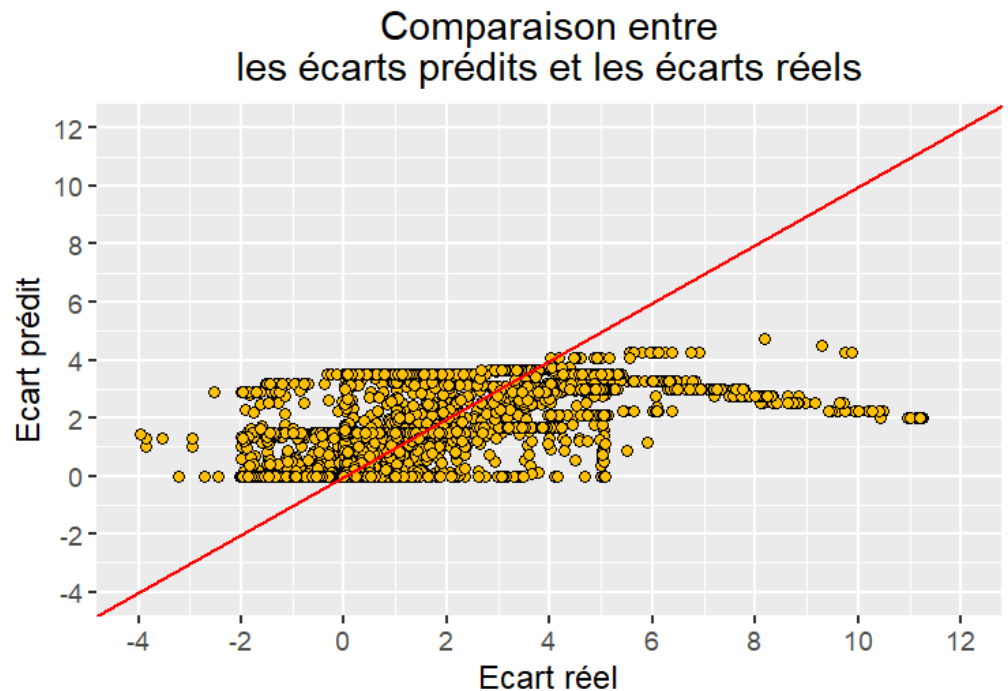


FIGURE 4.31 – Comparaison des écarts 'âge retraite - âge minimal' du modèle de référence

A noter que l'écart moyen entre les observations et les prédictions s'élève à -0,09 ans. Ce modèle a donc une légère tendance à retarder les départs en retraite.

L'écart actuariel engendré par le modèle de référence est acceptable : 3,6% de l'engagement projeté.

L'écart actuariel 2019 engendré par le modèle de référence est très acceptable. Il s'élève effectivement à 3,6% : l'impact dû aux prestations est de 3,3% ; celui dû aux retraites réalisées est de 0,2% ; et celui dû aux retraites non réalisées est de 0,1%. Toutefois, le modèle de référence a prédit deux dates de retraite pour 76% des individus (sinon il n'a prédit qu'une seule date).

4.4.2 Comparaison du modèle de référence avec les modèles construits

Les modèles construits ont une RMSE plus petite que celle du modèle de référence. La RMSE du modèle de référence est effectivement la plus élevée puisqu'elle est de 1,75 ans, alors que celles de la régression linéaire et de l'arbre de régression sont autour de 1,60 ans. Cela signifie

Les modèles construits captent plus de comportements différents que le modèle de référence.

que les modèles construits captent plus de comportements différents que le modèle de référence. De plus, cela peut également se remarquer grâce à la largeur des intervalles contenant les prédictions :

- Les prédictions du modèle de référence vont de 0 à 5 ans ;
- Les prédictions de la régression linéaire vont de -1,8 à 6,7 ans ;
- Les prédictions de l'arbre de régression vont de -1,0 à 6,0 ans.

Le modèle de référence engendre toutefois beaucoup moins d'écart actuariel d'expérience. En effet, il engendre 3,6% d'écart actuariel sur l'exercice 2019, alors que la régression linéaire et l'arbre de régression engendrent respectivement 10,1% et 11,7% d'écart actuariel. L'explication à cela vient du fait que le modèle de référence a tendance à retarder les départs en retraite, alors que les autres modèles ont tendance à les anticiper. En effet :

- Le résidu moyen (c'est-à-dire l'écart entre les observations et les prédictions) s'élève à -0,09 ans pour le modèle de référence, à 0,58 ans pour la régression linéaire et à 0,67 ans pour l'arbre de régression.
- Le modèle de référence a prévu (avec une probabilité de 100%) 6 549 départs en retraite en 2019, alors que la régression linéaire en a prévu 8 687 et l'arbre de régression en a prévu 9 114.

De plus, les données fournies par l'entreprise ne semblent pas assez complètes pour prédire tous les comportements. Selon un article publié en 2009 (cf. référence [2]), les facteurs sont nombreux et peuvent être répartis en 3 sphères :

Les modèles construits sont nettement moins performants pour évaluer l'engagement IFC que le modèle de référence.

- La sphère individuelle : âge, genre, état de santé, richesse, épargne.
- La sphère contextuelle : situation familiale, environnement de travail.
- La sphère institutionnelle : système de protection sociale, système de retraite.

Ainsi, les données fournies sont loin de regrouper tous ces facteurs : l'état de santé, l'épargne, la situation familiale, l'environnement de travail (mis à part le bassin géographique du lieu de travail) et le système de protection sociale sont inconnus, alors qu'ils jouent un rôle clé dans la décision du départ en retraite.

Chapitre 5

Conclusion

Les entreprises françaises sont engagées auprès de leurs salariés partant à la retraite, à verser une indemnité de fin de carrière (IFC). Cette indemnité représente donc une dette, qu'il est obligatoire de provisionner pour les comptes consolidés des sociétés cotées de l'Union Européenne. Toutefois, l'évaluation de cet engagement demande de fixer plusieurs hypothèses, dont celle de l'âge de départ en retraite des salariés de l'entreprise. En général, cette hypothèse est fixée selon la durée de cotisation nécessaire à l'obtention de la pension de retraite maximale (le taux plein) des générations les plus jeunes (de 43 ans aujourd'hui), et selon un âge de début de carrière (souvent fixé à 19-20 ans pour les non-cadres, et 21-22 ans pour les cadres). Néanmoins, plus de facteurs pourraient être pris en compte : l'âge d'entrée du salarié, son statut (salarié du privé, fonctionnaire de la catégorie active ou sédentaire), son lieu de travail, son métier, son salaire, etc. De plus, malgré le contexte comptable dans lequel s'inscrit l'évaluation de cet engagement, l'établissement d'une telle hypothèse n'est pas très encadrée : elle ne doit être ni risquée, ni d'une prudence excessive. Par conséquent, les modèles d'âge de retraite actuellement utilisés dans ce contexte pourraient être améliorés.

Dans le cas d'une entreprise spécifique, l'objectif a été de proposer des modèles adaptés au contexte qui améliorent le modèle actuellement utilisé. Premièrement, le contexte corporatif impose de proposer des modèles facilement compréhensibles et applicables, puisque les résultats sont destinés à des responsables de l'entreprise et aux commissaires aux comptes. De plus, il faut souligner que le modèle utilisé actuellement par l'entreprise (représentant le modèle de référence) est plus complexe que le modèle généralement utilisé (présenté ci-dessus) : il suppose que le salarié a une certaine probabilité de partir à son âge

d'ouverture des droits, et autre probabilité de partir à son âge d'obtention du taux plein. Le modèle de référence est donc plus sophistiqué que la moyenne, car il prend en compte la génération de naissance des salariés et leur statut. De plus, il se base sur les départs en retraite survenus au sein de l'entreprise pour déterminer les probabilités de départ. Par conséquent, l'objectif de ce mémoire était difficile à atteindre.

Les modèles ont été construits pour prédire l'écart entre l'âge de retraite et l'âge d'ouverture des droits du salarié. En effet, ce choix a été très vite imposé puisqu'il a été démontré de nombreuses fois que l'âge de retraite des individus dépendait du système de retraite en vigueur. Par conséquent, la construction d'un modèle sur l'âge de retraite serait biaisée par les réformes passées. De plus, il est pertinent de prédire cet écart (plutôt que l'âge de retraite) pour anticiper les éventuelles réformes à venir. La construction de chaque modèle a suivi un processus spécifique, afin de donner une grande importance sur le pouvoir prédictif du modèle. En effet, il a été tout d'abord décidé de construire un échantillon test qui permet d'évaluer les prédictions du modèle selon deux métriques sélectionnées : la racine carrée de l'erreur moyenne et l'écart actuariel. La première métrique est très répandue lorsque la variable prédictive est numérique. Concernant la seconde métrique, celle-ci est spécifique au contexte de l'évaluation d'un engagement social selon l'IAS 19 (norme comptable internationale). Elle correspond à l'écart entre l'engagement réel et l'engagement prévu. Elle est pertinente puisque les entreprises ont intérêt à minimiser cet écart, car celui-ci vient impacter les comptes de résultats ou les capitaux propres, selon le référentiel comptable en vigueur. Après fixation de l'échantillon test et des métriques, le modèle est construit et validé un grand nombre de fois. A chaque fois, seule une partie (différente à chaque fois) des données est utilisée pour construire le modèle et la partie restante est utilisée pour le tester. Le processus est donc axé sur le pouvoir prédictif du modèle puisque celui-ci est évalué de nombreuses fois sur des données qui lui sont inconnues.

Les départs en retraite survenus au sein de l'entreprise ont été analysés. L'analyse s'est articulée autour de l'objectif suivant : adapter les données au contexte prédictif. Ainsi, elle s'est réalisée en 4 étapes : transformation de variables "obsolètes", regroupement ou retrait des caractéristiques peu représentées, analyse et retrait des écarts "âge retraite - âge minimal" extrêmes, analyse et retrait de variables fortement corrélées. La première étape a du sens car les caractéristiques observées représentent nécessairement des salariés en fin de carrière (puisque la base de données est constituée de retraités), ce qui pose problème puisque le modèle est destiné à être appliqué (également) sur des salariés en début de carrière. Une solution pour pallier ce problème est d'estimer la caractéristique à un âge fixé,

pour la rendre comparable. Toutefois, ceci demande de connaître l'évolution de la caractéristique au fil de la carrière, ce qui n'est le cas que pour une caractéristique : le salaire (à partir de la politique salariale de l'entreprise). Ce dernier a donc été transformé et les autres caractéristiques sont restées comme telles. La deuxième étape de l'analyse (regroupement ou retrait des caractéristiques peu représentées) permet d'éviter au modèle de capter de "faux comportements" spécifiques aux observations. La troisième étape (analyse et retrait des écarts "âge retraite - âge minimal" extrêmes) permet d'éviter de capter des situations exceptionnelles. Enfin, la dernière étape (analyse et retrait des variables fortement corrélées) évite d'avoir un nombre de variables prédictives inutilement trop élevé.

A l'issue de l'analyse et du traitement des données, des modèles facilement interprétables et applicables ont été testés. La régression linéaire étant connue par une large population, présente un bon candidat, même s'il s'avère parfois fastidieux de l'appliquer. Sa construction s'est tout d'abord révélée être instable : elle est assez sensible à la composition de l'échantillon d'apprentissage. Toutefois, les résultats de prédiction sur les différents échantillons de validations sont stables, ce qui permet de s'assurer de la pertinence de l'évaluation du pouvoir prédictif de ce modèle. Après analyse de celui-ci, il s'avère que les âges prédits sont en moyenne plus proches des âges réels que le modèle de référence. Toutefois, l'écart actuariel engendré par ce modèle a considérablement augmenté. L'origine de ce phénomène s'explique par le fait que la régression linéaire a tendance à anticiper tous les départs en retraite et qu'elle ne prend pas en compte l'âge actuel du salarié. De plus, la régression linéaire est un modèle paramétrique : elle demande alors de fixer une loi sur la variable à prédire. Après analyse des résidus, les hypothèses de validation du modèle ne sont pas vérifiées (homoscédasticité des résidus, normalité, indépendances, etc.). C'est pourquoi un modèle non paramétrique a donc été sélectionné ensuite : l'arbre de régression. Il s'agit d'un bon candidat puisqu'il est très visuel, ce qui rend son interprétation et son application faciles. Le pouvoir prédictif de ce modèle est comparable avec celui de la régression linéaire. Mais l'écart actuariel qu'il engendre est encore le plus élevé, car il a encore plus tendance à anticiper les départs en retraite. Il a donc été démontré à quel point la prédiction de l'âge de retraite dans ce contexte pouvait être difficile. L'utilisation d'un modèle tel que le modèle de référence, qui se base seulement un peu sur les observations, n'est donc pas injustifiée.

Sans contrainte, d'autres modèles auraient pu être envisagés. Les contraintes de ce mémoire portaient à la fois sur les données (l'étude a dû se contenter des données fournies par l'entreprise) et sur les modèles (seuls les modèles facilement interprétables et applicables ont été envisagés). Si ces contraintes

n'existaient pas, la première piste d'exploration serait celle de l'enrichissement des données. En effet, un article publié en 2009 nommé "Les facteurs explicatifs du départ à la retraite en Europe" (cf. référence [2]) précise que de nombreux facteurs jouent un rôle sur la décision du départ en retraite d'un travailleur. Ces facteurs ont été regroupés dans 3 sphères : la sphère individuelle (âge, genre, état de santé, richesse, épargne, etc.), la sphère contextuelle (situation familiale, environnement de travail, etc.) et la sphère institutionnelle (systèmes de protection sociale, de retraite, etc.). L'enrichissement des données par celles citées dans cet article a donc de grande de chance de permettre à la modélisation de capter plus de comportements différents. Une deuxième piste d'exploration serait celle d'explorer davantage les techniques de modélisation de *Machine Learning* (forêt aléatoire, gradient boosting, réseau de neurones, etc.). Ces techniques présentent l'inconvénient d'être perçues comme des "boîtes noires" : les données en entrée et en sortie sont connues, mais la modélisation elle-même est un mystère (du fait de sa complexité). Une dernière piste d'exploration serait celle de prédire l'évènement "partir en retraite au i -ème anniversaire" (plutôt que simplement l'évènement "partir en retraite"). Cela conduirait à exploiter les modèles de durée pour obtenir des tables de "survie dans la vie professionnelle" (à noter que ce type de modèles requiert davantage de données sur les salariés encore en activité). Toutefois, le coût en temps de l'application de ces modèles pour évaluer l'engagement IFC serait très conséquent, puisque cela demanderait d'évaluer beaucoup d'engagements par individu (une évaluation pour chaque départ en retraite possible au i -ème anniversaire de l'individu).

Bibliographie

- [1] Journal officiel de l'Union EUROPÉENNE. *Règlement (CE) n°1126/2008 de la commission du 3 novembre 2008 portant adoption de certaines normes comptables internationales conformément au règlement (CE) n°1606/2002 du Parlement européen et du Conseil.*
- [2] Debrand THIERRY et Sirven NICOLAS. « Les facteurs explicatifs du départ à la retraite en Europe ». In : *Retraite et société* 2009/1 (n° 57) (2009), p. 35.
- [3] Confédération Française des RETRAITÉS (CFR). *Les grandes lignes de l'histoire de la retraite.* URL : <https://www.retraite-cfr.fr/wp-content/uploads/2016/03/R-11-Histoire-retraite.pdf> (visité le 09/04/2022).
- [4] Consultation de l'inscription maritime du COTENTIN SUR INTERNET (CIMARCONET). *Le système des classes de 1665 à 1784.* URL : https://www.unicaen.fr/ufr/histoire/cimarconet/inscription_maritime/historique2.php (visité le 09/04/2022).
- [5] Consultation de l'inscription maritime du COTENTIN SUR INTERNET (CIMARCONET). *La réorganisation des années 1784-1815.* URL : https://www.unicaen.fr/ufr/histoire/cimarconet/inscription_maritime/historique3.php (visité le 09/04/2022).
- [6] CAIRN INFO. *Des premiers systèmes obligatoires de protection sociale aux assurances sociales.* URL : <https://www.cairn.info/revue-vie-sociale-2015-2-page-31.htm> (visité le 26/05/2022).
- [7] La retraite en CLAIR. *Les grandes dates de la retraite en France.* URL : <https://www.la-retraite-en-clair.fr/retraite-france-monde/retraite-france-evenements-marquants/grandes-dates-retraite-france#:~:text=C'est%5C%20en%5C%201945%5C%20avec,compte%5C%20les%5C%20cons%5C%C3%5C%A9quences%5C%20du%5C%20vieillissement.> (visité le 18/05/2022).

- [8] Statistiques MSA. *Qui sommes-nous ?* URL : <https://statistiques.msa.fr/qui-sommes-nous/> (visité le 26/05/2022).
- [9] La retraite en CLAIR. *Le Conseil d'orientation des retraites : organisation et projections*. URL : <https://www.la-retraite-en-clair.fr/retraite-france-monde/systeme-retraite-france-enjeux/conseil-orientation-retraites-organisation-projections> (visité le 26/05/2022).
- [10] DREES. *Les retraités et les retraites - édition 2021*. URL : <https://drees.solidarites-sante.gouv.fr/publications-documents-de-reference/panoramas-de-la-drees/les-retraites-et-les-retraites-edition-0> (visité le 01/03/2023).
- [11] Denayrolles EMILIE et Guilain MARIE. « Retraite anticipée pour carrière longue : 10 années d'évolutions réglementaires ». In : *Retraite et société* 70 (1) (2015), p. 151-166.
- [12] SERVICE-PUBLIC.FR. *Retraite anticipée pour handicap du salarié*. URL : <https://www.service-public.fr/particuliers/vosdroits/F16337> (visité le 11/03/2023).
- [13] SERVICE-PUBLIC.FR. *Retraite anticipée du salarié pour incapacité permanente (ou pénibilité)*. URL : <https://www.service-public.fr/particuliers/vosdroits/F14101> (visité le 11/03/2023).
- [14] SERVICE-PUBLIC.FR. *Préretraite amiante du salarié du secteur privé*. URL : <https://www.service-public.fr/particuliers/vosdroits/F2110> (visité le 11/03/2023).
- [15] SERVICE-PUBLIC.FR. *Retraite anticipée pour carrière longue du fonctionnaire*. URL : <https://www.service-public.fr/particuliers/vosdroits/F13946> (visité le 11/03/2023).
- [16] SERVICE-PUBLIC.FR. *Retraite pour invalidité dans la fonction publique*. URL : <https://www.service-public.fr/particuliers/vosdroits/F550> (visité le 11/03/2023).
- [17] SERVICE-PUBLIC.FR. *Retraite anticipée pour handicap dans la fonction publique*. URL : <https://www.service-public.fr/particuliers/vosdroits/F14060> (visité le 11/03/2023).
- [18] SERVICE-PUBLIC.FR. *Retraite anticipée du fonctionnaire parent d'un enfant handicapé*. URL : <https://www.service-public.fr/particuliers/vosdroits/F20659> (visité le 11/03/2023).

- [19] SERVICE-PUBLIC.FR. *Un ancien fonctionnaire qui devient invalide a-t-il droit à une retraite pour invalidité ?* URL : <https://www.service-public.fr/particuliers/vosdroits/F18217> (visité le 11/03/2023).
- [20] SERVICE-PUBLIC.FR. *Un fonctionnaire ayant 3 enfants peut-il encore partir en retraite anticipée ?* URL : <https://www.service-public.fr/particuliers/vosdroits/F2766#:~:text=succ%C3%A8s%20aux%20destinataires.-,Un%20fonctionnaire%20ayant%203%20enfants%20peut,encore%20partir%20en%20retraite%20anticip%C3%A9e%20%5C%3F&text=Non%2C%20le%20dispositif%20de%20retraite,retraite%20anticip%C3%A9e%20%5C%3%A0%20tout%20moment.> (visité le 11/03/2023).
- [21] AG2R La MONDIALE. *Retraite : quel âge maximum pour travailler ?* URL : <https://www.ag2rlamondiale.fr/retraite/prendre-sa-retraite/conseil-retraite-quel-age-maximum-pour-travailler> (visité le 11/03/2023).
- [22] Ministère de L'INTÉRIEUR. *Jusqu'à quel âge peut-on travailler dans la fonction publique ?* URL : <https://www.demarches.interieur.gouv.fr/particuliers/retraite-fonction-publique-limite-age#:~:text=Contractuel-,Quel%20est%20l'âge%20limite%20d'activité%20plus%20tard%20%5C%3%A0%2067%20ans.> (visité le 11/03/2023).
- [23] INSEE. *Âge conjoncturel de départ à la retraite.* URL : <https://www.insee.fr/fr/metadonnees/definition/c2180> (visité le 26/05/2022).
- [24] INSEE. *Âge conjoncturel moyen de départ à la retraite selon le sexe.* URL : <https://www.insee.fr/fr/statistiques/5209769#graphique-figure1> (visité le 26/05/2022).
- [25] DREES. *Âge de départ à la retraite et conditions de fin de carrière selon la catégorie socioprofessionnelle.* URL : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/departretraite_parccsp/information/ (visité le 31/05/2022).
- [26] DREES. *Données statistiques sur les caractéristiques des retraités.* URL : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/3013_donnees-statistiques-sur-les-caracteristiques-des-retraites/information/ (visité le 31/05/2022).

- [27] DREES. *Les motivations de départ à la retraite*. URL : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/2586_les-motivations-de-depart-a-la-retraite/information/ (visité le 31/05/2022).
- [28] DREES. *Les âges de départ à la retraite et les transitions entre emploi et retraite*. URL : https://data.drees.solidarites-sante.gouv.fr/explore/dataset/2932_les-ages-de-depart-a-la-retraite-et-les-transitions-entre-emploi-et-retrait/information/ (visité le 31/05/2022).
- [29] IFRS FOUNDATION. *Who we are*. URL : <https://www.ifrs.org/about-us/who-we-are/#about-us> (visité le 23/02/2022).
- [30] International Accounting Standard Board (IASB). *IAS 19 Employee Benefits*.
- [31] IFRS FOUNDATION. *Due Process Handbook*.
- [32] IFRS FOUNDATION. *Constitution effective from 1 December 2016*.
- [33] IFRS FOUNDATION. *Who uses IFRS Standards?* URL : <https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/> (visité le 20/02/2022).
- [34] Journal officiel des COMMUNAUTÉS EUROPÉENNES. *Règlement (CE) N°1606/2002 du parlement européen et du conseil du 19 juillet 2002 sur l'application des normes comptables internationales*.
- [35] AMF (Autorité des MARCHÉS FINANCIERS). *Quels sont les différents marchés français de cotation?* URL : [https://www.amf-france.org/fr/quels-sont-les-differents-marches-francais-de-cotation-1#:~:text=Euronext%5C%20Paris%5C%20g%5C%C3%5C%A8re%5C%203%5C%20march%5C%C3%5C%A9s,Euronext%5C%20Growth%5C%20\(anciennement%5C%20Alternext\)](https://www.amf-france.org/fr/quels-sont-les-differents-marches-francais-de-cotation-1#:~:text=Euronext%5C%20Paris%5C%20g%5C%C3%5C%A8re%5C%203%5C%20march%5C%C3%5C%A9s,Euronext%5C%20Growth%5C%20(anciennement%5C%20Alternext)) (visité le 20/02/2022).
- [36] L-EXPERT-COMPTABLE. *Les comptes consolidés : définition et obligations*. URL : <https://www.l-expert-comptable.com/a/52548-les-comptes-consolides.html> (visité le 20/02/2022).
- [37] FOCUSIFRS. *IAS/IFRS et PME : Travaux de l'ANC (anciennement CNC)*. URL : http://www.focusifrs.com/menu_gauche/normes_et_interpretations/qui_est_concerne/ias_ifrs_et_pme_travaux_de_l_anc_anciennement_cnc (visité le 20/02/2022).
- [38] ECONOMIE.GOUV.FR. *Placements : qu'est-ce qu'une obligation?* URL : <https://www.economie.gouv.fr/particuliers/placement-obligation> (visité le 16/03/2022).

- [39] SPAC-ACTUAIRES.FR. *Duration définition*. URL : <https://www.spac-actuaires.fr/lexique/duration/#:~:text=La%5C%20duration%5C%20est%5C%20la%5C%20dur%5C%C3%5C%A9e,est%5C%20%5C%C3%5C%A9gale%5C%20%5C%C3%5C%A0%5C%20la%5C%20maturit%5C%C3%5C%A9>. (visité le 16/03/2022).
- [40] IFRS FOUNDATION. *IFRIC Update April 2021*. URL : <https://www.ifrs.org/news-and-events/updates/ifric/2021/ifric-update-april-2021/> (visité le 09/03/2022).
- [41] Compagnie Nationale des COMMISSAIRES AUX COMPTES. *ENGAGEMENTS DE RETRAITE ET AVANTAGES SIMILAIRES – Application de la recommandation n° 2013-02 de l’Autorité des normes comptables – IFRS – Détermination du taux de rotation du personnel pour les besoins de l’évaluation de l’engagement ou de la provision*.
- [42] Breiman LEO. « Statistical Modeling : The Two Cultures ». In : *Statistical Science* 16 (3) (2001), p. 199-231.
- [43] SERVICE PUBLIC. *Smic (salaire minimum de croissance)*. URL : <https://www.service-public.fr/particuliers/vosdroits/F2300> (visité le 01/03/2023).
- [44] Sireyjol ANTOINE. *Le dilemme biais variance dans la modélisation de données*. URL : <https://blog.statoscop.fr/le-dilemme-biais-variance-dans-la-modelisation-de-donnees.html#liens-avec-loverfitting-et-lunderfitting-dans-les-modeles-de-machine-learning> (visité le 11/03/2023).

Table des figures

2.1	Évolution de l'âge de retraite	20
2.2	Évolution de l'âge de retraite par CSP	20
2.3	Répartition des âges de liquidation	21
2.4	Répartition des conditions de liquidation	22
2.5	Les motivations du départ à la retraite	22
2.6	Taux de retraités par âge	23
2.7	Pays fondateurs de l'IASB	24
2.8	Pays appliquant les IAS/IFRS	26
2.9	Exemple d'un coefficient de prorata appliqué à la VAP de l'IFC pour déterminer l'engagement, en fonction de la date d'évaluation .	32
2.10	Impact IFRIC en fonction de la durée résiduelle.	33
2.11	Sensibilité de l'engagement IFC face à l'âge de retraite	37
2.12	Comparaison de l'impact d'un changement d'âge de retraite avec/sans la prise en compte de l'interprétation de l'IFRIC.	38
3.1	Poids de l'âge d'ouverture des droits utilisés lors de l'évaluation de l'engagement IFC de l'entreprise	44
3.2	Table de turnover appliquée lors des évaluations d'engagement IFC	48
3.3	Tables des taux d'évolution annuelle des salaire appliqués lors des évaluations d'engagement IFC	49
3.4	Evolution du salaire minimal interprofessionnel depuis 1950	53
3.5	Évolution de la proportion des départs anticipés à l'"âge minimal de carrière longue" ou après, comparée à celle des départs anticipés pour carrière longue observés à l'échelle nationale	58
3.6	Illustration de la procédure de construction d'un modèle adoptée . .	69
4.1	Répartition de l'écart "âge retraite - âge minimal"	76
4.2	Répartition des niveaux de vie à 15 ans	77
4.3	Répartition des catégories	77

4.4	Répartition des domaines de métier	78
4.5	Moyenne des écarts "âge retraite - âge minimal" par domaine de métier	78
4.6	Répartition des taux d'activité	79
4.7	Répartition des taux d'activité v2	80
4.8	Répartition des âges d'entrée	80
4.9	Répartition des niveaux de vie	81
4.10	Représentation des valeurs propres de l'AFDM	82
4.11	Contribution des variables au premier axe	83
4.12	Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 1 - axe 2"	84
4.13	Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 1 - axe 3"	84
4.14	Représentation des individus partis à la retraite "extrêmement" tôt ou tard, sur la plan "axe 2 - axe 3"	85
4.15	V de Cramer entre chaque variables prédictives qualitatives	86
4.16	Répartition des genres selon le domaine du métier	86
4.17	Représentation des niveaux de vie à 15 ans en fonction de l'âge d'entrée dans l'entreprise	87
4.18	Rapports de corrélation des couples de variables de natures différentes	88
4.19	Répartition des niveaux de vie à 15 ans selon la catégorie	88
4.20	Répartition des niveaux de vie à 15 ans en fonction de l'écart "âge retraite - âge minimal"	89
4.21	Evolution de la RMSE en fonction du niveau de complexité	94
4.22	Dispersion des coefficients des variables	95
4.23	Dispersion des RMSE	95
4.24	Valeurs des coefficients de la régression linéaire retenue	96
4.25	Comparaison des écarts prédits et des écarts réels, sur l'échantillon test	98
4.26	Évolution de la RMSE en fonction du nombre de variables prédic- tives effectives	102
4.27	Écart-type de l'importance des variables, selon sa moyenne	103
4.28	Ecart-type	103
4.29	Représentation simplifiée de l'arbre de régression retenu pour mo- déliser l'âge de retraite	104
4.30	Comparaison entre les écarts prédits et les écarts réels.	105
4.31	Comparaison des écarts 'âge retraite - âge minimal' du modèle de référence	107

Liste des tableaux

2.1	L'âge d'ouverture des droits à la retraite	16
2.2	Durée de cotisation pour obtenir la retraite à taux plein	18
2.3	Âge du taux plein automatique	18
3.1	Politique salariale de l'entreprise étudiée	54