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

Par : Monsieur Mathieu Berguig

Titre du mémoire : Sensibilité de l'engagement actuariel au turnover : lissages non paramétriques et approche prédictive grâce aux méthodes de classification

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

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

Membres présents du jury de la filière : Signature :

Membres présents du jury de l'Institut des Actuaires : Signature :

Entreprise :

Nom : Mercer France
Signature :

Directeur de mémoire en entreprise

Nom : Benjamin Sanson
Signature :

Invité :
Nom :
Signature :

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

Signature du responsable entreprise :

Signature du candidat :

Remerciements

Je tiens tout d'abord à remercier l'ensemble du pôle Wealth de Mercer pour son accueil, pour la confiance accordée dans les missions réalisées et pour toute la formation prodiguée tant dans les évaluations actuarielles d'engagements sociaux, qu'en retraite et en investissement.

Je tiens à remercier mon tuteur en entreprise Benjamin Sanson, principal chez Mercer, pour son encadrement et sa bienveillance.

Je remercie aussi mon tuteur académique Olivier Lopez, directeur de l'ISUP, pour son suivi, son recul sur le sujet et ses conseils pertinents.

Enfin, je remercie toute l'équipe enseignante de l'ISUP pour la qualité de l'enseignement dispensé.

Table des matières

Résumé

Mots clés : turnover, norme IAS 19, IFRIC, indemnité de fin de carrière, régime de retraite à prestations définies, lissages de tables, classification, apprentissage supervisé.

Ce mémoire traite de la problématique de l'hypothèse du taux de turnover ou taux de renouvellement du personnel intervenant dans le calcul des engagements sociaux d'une entreprise. Il s'agit d'une hypothèse délicate à établir car elle n'est pas clairement définie dans la norme IAS 19 qui encadre la valorisation et la comptabilisation des engagements. La norme demande à chaque entreprise de refléter le plus fidèlement possible la réalité des taux de sorties observés à travers la construction de tables de taux de démissions. La duration des régimes considérés (indemnité de fin de carrière, médailles du travail et régime de retraite à prestations définies) étant de plusieurs années, les tables construites doivent donner une tendance long terme de la probabilité de présence du salarié au moment de la prestation.

Deux sources d'incertitudes interviennent dans la probabilité de présence : le turnover et la mortalité. Pour la population active considérée, l'engagement se révèle plus sensible au turnover qu'à la mortalité, celle-ci relevant d'évènements rares. Le choix s'est donc porté sur la modélisation du turnover d'une grande entreprise. Les taux bruts de démissions par âge, par ancienneté et par catégorie socio-professionnelle ont été construits. Les irrégularités et aspérités de ces taux dues aux fluctuations d'échantillonnage ont été traitées grâce à plusieurs techniques de lissages.

La méthode utilisée par Mercer est l'ajustement d'une table de référence par rapport aux taux de démissions des 3 dernières années. Bien qu'ayant l'avantage d'être simple et de tenir compte des démissions sur plusieurs années, elle suppose une allure de la courbe de turnover invariante au fil des ans et commune à toutes les entreprises. Les lissages non paramétriques qui sont mis en œuvre ont l'avantage de ne pas donner d'a priori à l'allure de la courbe et d'utiliser au mieux l'information contenue dans les données. L'impact du choix de la table lissée sur l'engagement est mesuré. Des sensibilités de l'engagement à la table lissée retenue sont appliquées grâce aux intervalles de confiance pour les taux estimés.

Enfin, une approche prédictive du turnover est présentée dans l'idée d'utiliser les prédictions individuelles de démission pour le calcul de l'engagement et de déterminer les variables les plus influentes pour sa prédiction parmi les caractéristiques des individus.

Abstract

Keywords : turnover, IAS 19 standard, IFRIC, retirement indemnities, defined benefit pension plan, smoothing tables, classification, supervised machine learning.

This thesis deals with the issue of the turnover assumption in the calculation of the employee benefit obligations. This assumption is quite difficult to establish since it is not clearly defined in IAS 19 standard. The standard asks each company to reflect faithfully the reality of the turnover rates observed within the company through resignation rates tables. The duration of the retirement indemnity plan, jubilees plan and defined benefit pension plan is over several years, therefore tables have to describe a long-term trend of the likelihood of not leaving the company at the moment of the benefit.

Two sources of uncertainty intervene in the likelihood of not leaving the company at the moment of the benefit : turnover and mortality. However, for the population of actives that we consider, the obligation is more sensitive to turnover than mortality. That is why we chose to model the turnover of a large company. The gross rates of resignations by age, seniority and socio-professional category have been set up. To face the irregularities and roughness of the curve due to sampling fluctuations, we applied different smoothing techniques.

The method used by Mercer consists in the adjustment of an initial table with the average resignation rates of the last three years. Although it is a simple and takes into account the resignations over several years, it assumes that the shape of the curve will remain the same over the years since the initial shape is only vertically shifted. Then, non-parametric smoothing are implemented, they have the advantage to make the best use of the information contained in data. The impact of every smoothing table on the obligation of the three plans is evaluated. Thanks to confidence intervals for the turnover rates, sensitivities of the obligation to these tables are presented.

Finally, we present a predictive approach of the turnover in order to use individual predictions in the calculus of the obligation and also to determine the most significant variables for its prediction among the characteristics of individuals.

Introduction

Aujourd'hui, on compte en France environ 30 millions d'actifs et 17 millions de retraités. L'évaluation des engagements sociaux représente un réel enjeu pour les sociétés.

Le passif social désigne l'ensemble des engagements différés pris par une entreprise à l'égard de ses salariés. Il en existe de différentes natures parmi lesquels les indemnités de fin de carrière, les médailles du travail, les régimes supplémentaires de retraite.

L'entreprise peut choisir de gérer son passif social en interne ou en externe. L'externalisation est souvent considérée comme l'option à privilégier pour :
- simplifier l'estimation et l'étalement du passif social, des activités chronophages pour les services financiers et des ressources humaines,
- valoriser l'entreprise : baisse de la charge à reporter au bilan grâce à la constitution d'un actif de régime,
- piloter la trésorerie avec l'anticipation des prestations à payer : dotation libre ou lissage dans le temps afin d'éviter les à-coups dans la trésorerie de l'entreprise,
- bénéficier d'avantages, notamment fiscaux : les dotations sont exonérées de charges sociales et fiscalement déductibles du résultat imposable et les produits financiers réalisés par la capitalisation des fonds sont exonérés d'imposition et de charges sociales,
- la mise en conformité pour l'article 39 et la directive sur la sécurisation des rentes.

Dans le cas d'une gestion en interne, il est important d'ajuster au mieux les montants à provisionner qui peuvent représenter des sommes considérables pour l'entreprise. De plus, depuis la révision de la norme IAS 19 en 2013 et la disparition de la méthode du Corridor pour étaler les écarts actuariels dans le temps, les entreprises cherchent à limiter leurs écarts actuariels, la volatilité de la provision et des fonds propres.

La valorisation et la comptabilisation des avantages au personnel sont définies, depuis le 1er janvier 2005, au sein de la norme IAS 19. L'engagement pour chaque participant du régime se calcule comme la prestation estimée probable et actualisée. Il fait intervenir différentes hypothèses actuarielles.

Parmi ces hypothèses, le turnover ou renouvellement du personnel en constitue l'une des plus délicates à établir. Contrairement à d'autres hypothèses comme le taux d'actualisation,

la norme IAS 19 ne donne pas d'indications précises quant à sa détermination. Cependant, elle n'en reste pas moins importante, nous nous attacherons dans ce mémoire à l'estimer et à l'anticiper en fonction des caractéristiques de la population.

Dans un premier temps, nous présenterons le cadre réglementaire fixé par la norme IAS 19, les méthodes de valorisation des engagements, les hypothèses sous-jacentes ainsi que les méthodes de comptabilisation prévues.

Dans un second temps, nous nous concentrerons sur le turnover et challengerons la méthode d'ajustement actuellement utilisée par Mercer pour construire la table de turnover à travers l'application de plusieurs méthodes de lissages non paramétriques aux taux bruts de démissions à notre disposition.

Dans un troisième temps, nous présenterons quatre méthodes d'apprentissage supervisé pour la prédiction du turnover et les métriques d'évaluation de la performance d'un modèle.

Dans un quatrième temps, nous comparerons les montants d'engagements obtenus selon les différentes tables lissées pour les régimes suivants : indemnité de fin de carrière, médailles du travail et régime de retraite supplémentaire à prestations définies. Nous appliquerons aussi une sensibilité à la table retenue pour chaque régime.

Enfin, nous mettrons en avant les modèles de classification binaire les plus performants ainsi que les variables les plus intéressantes pour la prédiction du turnover.

1 Cadre réglementaire : IAS 19

La première version de la norme comptable IAS 19 ≪ Avantages du personnel ≫ a été publiée par l'IASB (International Accounting Standards Board) en février 1998. L'IASB a pour principaux objectifs d'élaborer et de publier des normes comptables internationales pour la présentation des états financiers, de promouvoir leur utilisation au niveau mondial et de publier des interprétations développées par l'IFRS IC (International Financial Reporting Standards Interpretations Committee). Depuis le 1er avril 2001, les normes édictées par cet organisme se nomment IFRS (International Financial Reporting Standards).

La norme IAS 19 a été adoptée par la Commission Européenne en septembre 2003 et est applicable depuis le 1er janvier 2005 à titre obligatoire par les sociétés faisant appel public à l'épargne, préparant et publiant des comptes consolidés. Elle a été amendée en juin 2011 et adoptée le 5 juin 2012 par l'Union Européenne, pour une application obligatoire au 1er janvier 2013. Des amendements subséquents ont été apportés à la norme IAS 19 révisée.

La norme IAS 19 révisée doit être appliquée pour la comptabilisation, par l'employeur, de tous les avantages du personnel, toute rémunération versée en contrepartie de services rendus à l'exception des rémunérations versées sous forme de capitaux propres auxquelles s'applique IFRS 2 ≪ Paiement fondé sur des actions ≫. Un avantage social naît d'une promesse d'avantage faite par l'employeur à ses salariés, susceptible de lui faire courir un risque.

La norme IAS 19 distingue 4 catégories d'avantages au personnel qui peuvent être mis en place par décision unilatérale de l'employeur, accord collectif, référendum ou simple usage :

1) Avantages court terme : avantages dus intégralement dans les douze mois suivant la fin de l'exercice au cours duquel les membres du personnel ont rendu les services correspondants. Leur montant est comptabilisé en charge.

2) Avantages long terme : avantages dont le règlement intégral est attendu au-delà de douze mois suivant la clôture de l'exercice au cours duquel les membres du personnel ont rendu les services correspondants. Ils génèrent des passifs sociaux.

3) Indemnités de cessation d'emploi : avantages fournis en contrepartie de la

cessation d'emploi d'un membre du personnel résultant soit de la décision de l'entité de mettre fin à l'emploi du membre du personnel avant l'âge normal de départ en retraite soit de la décision du membre du personnel d'accepter une offre d'indemnités en échange de la cessation de son emploi. Elles génèrent des passifs sociaux.

4) Avantages postérieurs à l'emploi : avantages payables après la cessation de l'emploi du membre du personnel (autres que les indemnités de cessation d'emploi).

On distingue 2 types de régimes d'avantages postérieurs à l'emploi :

a) les régimes à cotisations définies : l'engagement porte sur un montant de cotisations convenues versées par l'employeur et/ou le salarié à une entité juridiquement distincte (fonds, contrat d'assurance). Le montant de la rente viagère est ensuite déterminé en fonction des cotisations effectivement versées et des produits financiers générés, le salarié n'a pas de garantie sur le montant des prestations versées. Les cotisations sont comptabilisées en charge. Le PERO ou anciennement Article 83 et le PERECO ou anciennement PERCO sont des régimes à cotisations définies.

b) les régimes à prestations définies : l'engagement porte non plus sur un montant de cotisations mais sur un niveau de prestations ou un rendement minimum sur les cotisations prédéterminé. En principe, l'employeur finance seul le régime. Le coût de financement réel de l'avantage est aléatoire, l'entreprise assume le risque financier. Ces régimes génèrent des passifs sociaux.

Les contrats de retraite à prestations définies et en particulier le contrat IFC (Indemnités de fin de carrière) sont des contrats collectifs d'assurance vie relevant des branches 20 (Vie-Décès) et 22 (Assurances liées à des fonds d'investissement) du Code des assurances, souscrits entre l'entreprise et l'assureur dans le cas d'une externalisation du passif social pour couvrir leurs engagements.

Voici une liste exhaustive des avantages qui entrent dans le cadre d'IAS 19 :

Avantages à court terme	Avantages à long terme	Indemnités de cessation d'emploi	Avantages postérieurs à l'emploi
Salaires Cotisations de Sécurité sociale Congés annuels payés Congés maladie Intéressement et primes Avantages en nature (assistance médicale, logement, voiture et autres biens et services gratuits ou subventionnés)	**Médailles du travail et autres avantages liés à l'ancienneté** Congés liés à l'ancienneté et congés sabbatiques Indemnités d'incapacité de longue durée Comptes épargne-temps Intéressement, primes et salaires différés payables 12 mois ou plus après la fin de l'exercice	Indemnités de licenciement Indemnités versées dans le cadre de plans de départ en préretraite ou de plans de départ volontaire	**Indemnités de fin de carrière** **Pensions de retraite et autres prestations postérieures à l'emploi** Assurance-vie postérieure à l'emploi Assistance médicale postérieure à l'emploi Avantages en nature maintenus pour les retraités

FIGURE 1 – Table des avantages sociaux

Parmi les 3 régimes traités dans ce mémoire, un régime est un avantage long terme et les deux autres sont des avantages postérieurs à l'emploi.

2 Evaluation actuarielle

Une évaluation actuarielle consiste en l'évaluation annuelle, par un actuaire, des engagements sociaux d'une société pour tout le personnel en CDI et en la comptabilisation des avantages générant des passifs sociaux dans les normes comptables locales ou internationales prévues.

2.1 Terminologie

2.1.1 Engagement

La **Defined Benefit Obligation (DBO)** correspond à l'engagement actuariel de l'entreprise à la date d'évaluation.

Les **Actual Benefit Payments (ABP)** correspondent aux prestations effectives de l'année et les **Expected Benefit Payments (EBP)** correspondent aux prestations attendues pour l'année.

Le **Service Cost (SC)** correspond au coût des services rendus au cours de l'exercice, c'est-à-dire à l'accroissement annuel de l'engagement du fait de l'année de service supplémentaire des participants. C'est la charge opérationnelle.
Soit i le taux d'actualisation, on écrira $SC = NC(1+i)$ avec **NC le Normal Cost**.

L'**Interest Cost (IC)** correspond au coût de la désactualisation sur l'année, c'est-à-dire à l'accroissement annuel de l'engagement du fait du rapprochement de la date de règlement des prestations d'un exercice (le montant d'engagement comptabilisé étant un montant actualisé). C'est la charge financière d'intérêt.
Soit i le taux d'actualisation, on écrira $IC = i(DBO - \frac{EBP}{2})$. Le facteur 1/2 traduit l'hypothèse qu'en moyenne un individu recevra sa prestation au milieu de l'année.

Les **Gains et pertes actuariels (G&L)** correspondent aux écarts entre l'engagement réel et l'engagement estimé. Ils sont répartis entre les écarts d'expérience dus à la mise à jour des données, les écarts d'hypothèses financières validées par la direction financière et les écarts d'hypothèses démographiques validées par la direction des ressources humaines.

Des écarts actuariels positifs constituent une perte car ils font augmenter l'engagement, des

écarts actuariels négatifs constituent un gain car ils font diminuer l'engagement. Une analyse de ces écarts est réalisée pour vérifier que les hypothèses de l'évaluation sont cohérentes.

Hors évènements spéciaux, l'engagement se réconcilie de la façon suivante :

$$DBO_N = DBO_{N-1} + SC_N + IC_N - ABP_N + G\&L_N$$

2.1.2 Actif de couverture

L'**Actif de couverture** correspond aux contrats d'assurance éligibles et aux actifs détenus par un fonds conférant des avantages à long terme. Sa juste valeur est :
- soit déduite de la DBO au passif si l'actif de couverture remplit un certain nombre de conditions : émetteur du contrat d'assurance indépendant de l'entreprise, sommes utilisées uniquement pour payer les avantages, aucun droit des créanciers de l'entreprise
- soit admise en créance (droits à remboursements) à l'actif si les conditions ne sont pas remplies et donc non déduite de la DBO.

Dans le cas où la valeur de l'actif de couverture est supérieure à celle de l'engagement à couvrir, un actif doit être comptabilisé. La norme IFRIC 14 intitulée "Le plafonnement de l'actif au titre des régimes à prestations définies, les exigences de financement minimal et leur interaction" précise le traitement comptable de l'excédent de couverture.

Les **Cotisations employeur** correspondent aux cotisations versées sur le fonds, dédiées à la couverture de l'engagement et augmentant la valeur de l'actif de couverture.

L'**Interest Income (II)** correspond au rendement attendu des actifs du régime c'est-à-dire aux intérêts, dividendes et autres produits tirés des actifs ainsi que les profits ou pertes réalisés ou latents relatifs à ces actifs, après déduction des coûts d'administration du régime et de l'impôt à payer.
Soit r le taux de rendement attendu, on écrira $II = r(Actif - \frac{EBP}{2})$.
En IAS 19, le taux de rendement attendu est égal au taux d'actualisation. En normes américaines, il est défini en fonction de la composition du fonds. En normes françaises, les 2 options sont possibles.

Les **Prestations par le fonds** correspondent aux prestations payées au titre du régime,

prélevées sur le fonds constitué en couverture.

Les **Gains et pertes actuariels sur l'actif de couverture ($G\&L_{Actif}$)** correspondent à la différence entre le rendement attendu et le rendement réel des fonds.

Hors évènements spéciaux, la juste valeur de l'actif de couverture se réconcilie ainsi :

$$Actif_N = Actif_{N-1} + Cotisations_N + II_N - Prestations_N + G\&L_{ActifN}$$

2.2 Hypothèses

La norme IAS 19 précise que les hypothèses doivent être objectives et mutuellement compatibles, c'est-à-dire justes et cohérentes. Elle ne régit pas les hypothèses endogènes liées aux comportements sociaux et démographiques de la population étudiée. Les hypothèses exogènes sont déterminées par référence aux attentes du marché. Seul le taux d'actualisation est spécifiquement réglementé par la norme.

L'employeur est responsable des hypothèses, la norme recommande de s'appuyer sur l'avis d'un actuaire et les auditeurs les valident.

2.2.1 Hypothèses économiques

Le **taux d'actualisation** doit être égal au taux de rendement attendu des obligations de haute qualité du secteur privé, de même duration et de même devise que les engagements à la date d'évaluation.

Plusieurs références sont possibles :
- iBoxx Corporate AA
- Bloomberg
- OAT (référence de taux sans risque)
- Fourchettes du SACEI (Syndicat des Actuaires Conseils et Experts Indépendants).

Pour un engagement social, la maturité est la durée moyenne résiduelle de vie active des salariés et la duration est la durée moyenne résiduelle de vie active des salariés pondérée par les flux probables de prestations à payer. La duration est en général plus courte que la maturité car plus la durée moyenne résiduelle de vie active est faible plus les flux probables

de prestations à payer sont élevés. La duration est une fonction décroissante du taux d'actualisation.

Fournir des sensibilités de la DBO au taux d'actualisation (+/-50bp) permet d'obtenir une approximation de la duration utilisée en pratique.

$$\text{Duration DBO} = \frac{\sum_k DBO_k@DR*k}{\sum_k DBO_k@DR} \approx \frac{1}{2}[\frac{ln(DBO@DR-50bp/DBO@DR)}{ln[(1+DR)/(1+DR-50bp)]} + \frac{ln(DBO@DR+50bp/DBO@DR)}{ln[(1+DR)/(1+DR+50bp)]}]$$

Cela se démontre en utilisant les équivalents $ln(1+x) \sim_0 x$ et $(1+x)^k \sim_0 1+kx$.

La duration permet ensuite d'estimer approximativement l'impact d'un changement de taux d'actualisation ou d'un changement de taux d'augmentation des salaires (rôles symétriques dans le calcul de l'engagement) sur l'engagement au global.

Le **taux de revalorisation des salaires** ou profil de carrière doit refléter la politique de l'entreprise à long terme. Il est à observer sur les 3 à 5 dernières années pour projeter le salaire sur plusieurs années. Il est en général supérieur au **taux d'inflation long terme** donné selon les recommandations de la BCE ou les indices du FMI. L'objectif inflation long terme de la BCE étant fixé à 2%, ce taux est parfois considéré comme le seuil minimal d'augmentation des salaires à prendre en compte. Il peut être discriminé selon l'âge ou la catégorie socio-professionnelle (CSP) du salarié.

Le **taux de revalorisation des médailles** est en général indexé sur le taux d'inflation long terme. Le **taux de revalorisation des rentes** peut être indexé sur le taux d'inflation long terme ou sur le taux d'augmentation du point AGIRC-ARRCO par exemple.

Le **taux de charges sociales patronales** est en général compris entre 40 % et 60 %. Il peut dépendre de la catégorie socio-professionnelle du salarié.

Les **droits** des indemnités de fin de carrière sont inscrits dans la convention collective nationale (CCN) dont dépend le groupe. Une convention collective est un accord écrit négocié entre les syndicats de salariés et d'employeurs, il y en a plusieurs centaines en France. Les dispositions de la convention collective peuvent être plus favorables pour le salarié que le code du travail et aussi contenir des dispositions que le code du travail ne prévoit pas.

2.2.2 Hypothèses démographiques

Les **tables de mortalité** utilisées sont les tables TH/TF 00-02 vie ou INSEE TH/TF pendant la phase d'activité et les tables générationnelles TGH/TGF 05 pendant la phase de retraite.

Les **tables de turnover** sont en général fonction de l'âge et de la catégorie socio-professionnelle et revues tous les 3 à 5 ans. Cette hypothèse sera développée dans la suite du mémoire.

L'**âge de départ à la retraite** est celui auquel les salariés sortent des effectifs en activité pour liquider leur retraite. L'âge minimum légal de départ à la retraite et l'âge légal de départ à taux plein dépendent de la date de naissance du salarié.

Date de naissance	Age légal de départ en retraite	Age légal de départ à taux plein
Avant le 1er juillet 1951	60 ans	65 ans
Du 1er juillet 1951 au 31 décembre 1951	60 ans et 4 mois	65 ans et 4 mois
1952	60 ans et 9 mois	65 ans et 9 mois
1953	61 ans et 2 mois	66 ans et 2 mois
1954	61 ans et 7 mois	66 ans et 7 mois
1955 et plus	62 ans	67 ans

FIGURE 2 – Table des âges de départ à la retraite

Il peut aussi être défini à partir d'un âge de début de carrière et du nombre de trimestres requis pour le taux plein (40 à 43 annuités en fonction de l'année de naissance). Cet âge de départ à la retraite est souvent supposé plus élevé pour les cadres.

La mise à la retraite est possible lorsque le salarié a atteint l'âge légal de départ à la retraite à taux plein, sous réserve que celui-ci ait formellement accepté de quitter l'entreprise. L'âge de la mise à la retraite d'office (sans l'accord du salarié) est 70 ans.

La modalité de départ à la retraite retenue dans ce mémoire est celle du départ volontaire.

2.3 Méthodes de valorisation

Il y a 3 méthodes de valorisation des engagements qui diffèrent dans le prorata appliqué à la valeur actuelle probable (VAP) des prestations futures. Ce prorata peut être un prorata de droits ou un prorata d'ancienneté, il permet de ramener la dette actuarielle à hauteur des droits acquis ou des services déjà rendus par les salariés.

1) Projected Unit Credit (PUC) Method : DBO = VAP $\frac{droits\ acquis}{droits\ au\ terme}$

C'est la méthode de référence préconisée par l'IAS 19 mais mal adaptée au contexte français.

2) PUC Method with Acquisition Prorate : DBO = VAP $\frac{ancienneté\ actuelle}{ancienneté\ plafond}$

C'est une méthode spécifique à certains régimes (IFC en escalier notamment).

3) PUC Method with Service Prorate : DBO = VAP $\frac{ancienneté\ actuelle}{ancienneté\ au\ terme}$

$NC = \frac{VAP}{ancienneté\ au\ terme} = \frac{DBO}{ancienneté\ actuelle}$ si ancienneté actuelle $<$ ancienneté au terme

Pour un groupe d'individus partant tous à la retraite dans k années, on a : $VAP = DBO + k * NC$.

Cette méthode prospective est mieux adaptée au contexte français. En effet, en France les droits sont garantis mais non acquis : il y a versement d'une prestation sous condition de présence du salarié dans l'entreprise à ce moment. La constitution de la VAP est donc étalée sur la durée de service du salarié et le NC est la portion de la VAP assignée à l'exercice suivant la date de l'évaluation, il est nul dès lors que l'ancienneté au terme est atteinte.

La récente réforme IFRIC vient revoir le prorata d'ancienneté sous certaines conditions.

2.3.1 Réforme IFRIC 2021

L'IFRS IC (International Financial Reporting Standards Interpretations Committee) est un organisme de l'IASB chargé de répondre aux problèmes d'interprétation des normes.

La décision de l'IFRS IC publiée le 21 mai 2021 a apporté des éclaircissements sur l'application de la norme IAS 19 visant à revoir la date de déclenchement de la provision. Au lieu d'étaler l'engagement sur toute la durée de présence dans la société, il est étalé sur les dernières années d'acquisition de droits. Si la différence entre l'ancienneté à la retraite et l'ancienneté au moment où les droits sont gelés vaut n, on ne constitue pas de provision les n premières années et on lisse la provision ensuite.

Les comptes sociaux peuvent s'aligner sur la méthode IFRIC. En revanche, les évaluations en normes américaines ne sont pas impactées. Les plans concernés par cette réforme sont les régimes de retraite (indemnités de départ en retraite inclus) non impactés par la loi PACTE de 2019, avec des droits non gelés, liés à l'ancienneté, plafonnés, avec ou sans paliers et conditionnés à la présence dans la société au moment du départ en retraite.

Cas d'une IFC avec des droits plafonnés à N années d'ancienneté, sans paliers :

Soit A l'ancienneté à la retraite arrondie au supérieur et $a = 1, \ldots, A$ l'ancienneté à l'évaluation arrondie au supérieur.

1) Un salarié ayant $A \leq N$ années d'ancienneté à la retraite ne voit pas son ratio d'ancienneté changer. La DBO et le NC restent inchangés.

2) Un salarié ayant $A \geq N+1$ années d'ancienneté à la retraite verra son engagement étalé entre les années d'ancienneté $A–N+1$ et A :

a) Si $A - a > N$: $DBO = NC = 0$.

b) Si $A - a \leq N$: $Prorata\, d'ancienneté = \frac{a+N-A}{N}$ et $NC = \frac{VAP}{N}$.

Avec cette méthode, la DBO diminue et le NC augmente : on montre aisément que $\frac{a+N-A}{N} \leq \frac{a}{A}$ et $\frac{VAP}{A} < \frac{VAP}{N}$.

Par conséquent, la dette actuarielle au terme ne change pas avec cette méthode, elle est égale à la prestation à payer, mais avant terme, la dette actuarielle est plus faible car la période de linéarisation est plus courte. Le coût de la désactualisation diminue aussi et le coût des services rendus devrait augmenter à long terme. En effet, la période d'acquisition des droits étant raccourcie, les dotations à la provision vont être plus importantes pour pouvoir constituer le bon niveau de provision au moment du départ du salarié.

Cette modification de méthode rend aussi plus sensible l'hypothèse d'âge de départ à la retraite qu'il conviendra peut-être de modifier pour la rendre plus proche de la réalité et des perspectives d'évolution.

Prenons l'exemple d'un individu ayant 44 ans d'ancienneté à la retraite et les droits de sa CCN plafonnés à 30 ans. Voici l'évolution de son prorata d'ancienneté selon les deux méthodes. Il est linéaire de 0 à 44 ans pour la méthode A (sans méthode IFRIC) et nul avant 14 ans pour la méthode B (avec méthode IFRIC).

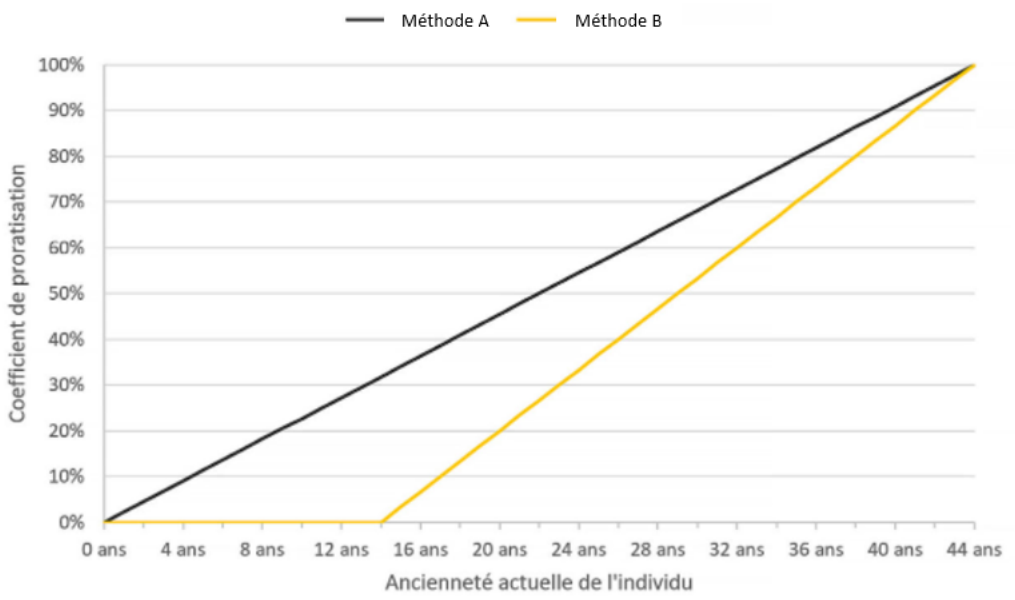

FIGURE 3 – Evolution du prorata d'ancienneté avec et sans la méthode IFRIC

Et ci-dessous l'évolution de sa DBO et de son NC selon les deux méthodes. La DBO constituée au terme est bien la même pour les deux méthodes.

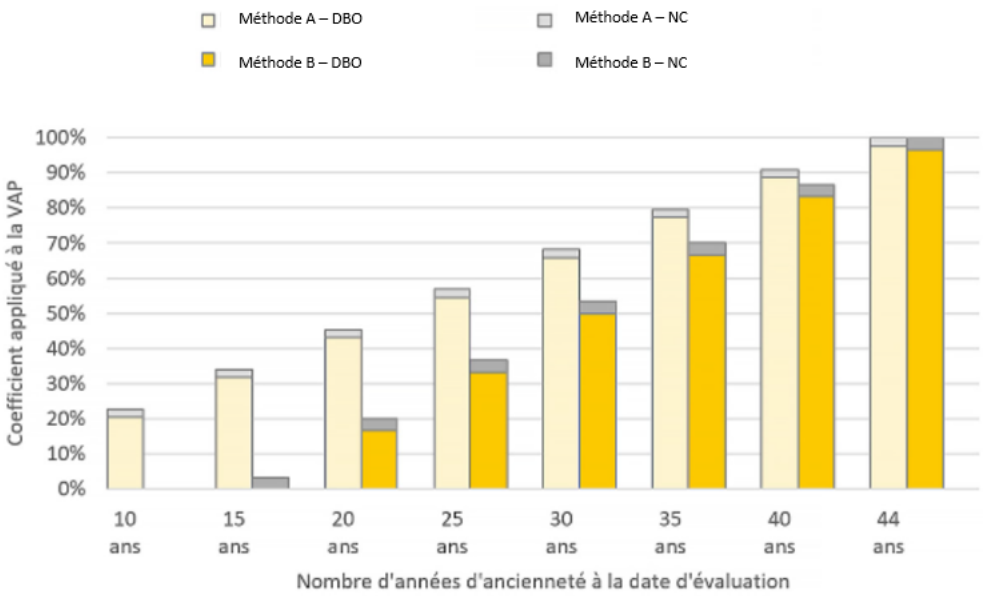

FIGURE 4 – Evolution de la DBO et du NC avec et sans la méthode IFRIC

La première application de ces changements d'interprétations intervient lors des travaux de clôture du 31/12/2021. L'impact IFRIC est reconnu en capitaux propres (OCI).

Nous présentons dans la suite le calcul de l'engagement actuariel individuel pour 3 types d'avantages : les indemnités de fin de carrière, les médailles du travail et les régimes de retraite à prestations définies.

2.3.2 Indemnité de fin de carrière

Toute entreprise a l'obligation de verser des indemnités de fin de carrière à chaque salarié partant à la retraite sous réserve que ce dernier remplisse les conditions d'âge (avoir liquidé ses droits de retraite aux régimes de retraite obligatoires) et d'ancienneté requises. Il s'agit du régime pesant le plus lourd dans le passif social de l'entreprise.

Il faut distinguer l'indemnité versée en cas de départ volontaire et celle versée en cas de mise à la retraite. Dans les 2 cas, l'indemnité versée est égale au maximum entre l'indemnité légale, conventionnelle et contractuelle.

L'indemnité légale de mise à la retraite est équivalente à l'indemnité légale de licenciement, elle est plus élevée que l'indemnité légale de départ à la retraite à l'initiative du salarié. Elle vaut 1/4 de mois du salaire de référence par année d'ancienneté jusque 10 ans d'ancienneté et 1/3 de mois du salaire de référence par année d'ancienneté après 10 ans d'ancienneté. Le salaire de référence est le maximum entre la moyenne des 3 derniers salaires mensuels bruts et la moyenne des 12 derniers salaires mensuels bruts. Une contribution patronale de 50% de l'indemnité de mise à la retraite versée est obligatoire.

Le barème de l'indemnité légale de départ à la retraite à l'initiative du salarié est présenté ci-dessous.

Ancienneté à la rupture du contrat de travail	Montant
Entre 10 et 15 ans	0,5 mois du salaire de référence
Entre 15 et 20 ans	1 mois
Entre 20 et 30 ans	2 mois
Après 30 ans	3 mois

FIGURE 5 – Barème de l'indemnité légale de départ à la retraite à l'initiative du salarié

Calcul de l'engagement individuel pour un régime IFC à l'initiative de l'employé selon la méthode PUC with Service Prorate (hors IFRIC).

Notations :

$\nu = \frac{1}{1+i}$: facteur d'actualisation

ρ : taux d'augmentation des salaires

μ : taux de charges sociales patronales

R : année qui précède l'année de départ à la retraite

x : âge arrondi de l'individu au 31/12/2021

${}_kp_x = \prod_{i=0}^{k-1}(1-q_{x+i})(1-t_{x+i})$: probabilité à l'âge x de présence dans l'entreprise dans k années, q_x étant la probabilité à l'âge x de décéder dans l'année et t_x la probabilité à l'âge x de démissionner dans l'année.

L'engagement individuel calculé à la clôture du 31/12/2021 et projeté au 31/12/2022 s'écrit :

$$\mathbf{DBO}_{2021} = droits_R * salaire_{2021}\ (1+\rho)^{R-2021} * (1+\mu) * {}_{R-2021}p_x * \nu^{R-2021} * \frac{ancienneté_{31/12/2021}}{ancienneté_{31/12/R}}$$

$$\mathbf{DBO}_{2022projetée} = droits_R * salaire_{2021}\ (1+\rho)^{R-2021} * (1+\mu) * {}_{R-2021}p_x * \nu^{R-2022} * \frac{ancienneté_{31/12/2021}+1}{ancienneté_{31/12/R}}$$

L'engagement individuel calculé à la clôture du 31/12/2022 s'écrit :

$$\mathbf{DBO}_{2022} = droits_R * salaire_{2022}\ (1+\rho)^{R-2022} * (1+\mu) * {}_{R-2022}p_{x+1} * \nu^{R-2022} * \frac{ancienneté_{31/12/2022}}{ancienneté_{31/12/R}}$$

L'écart entre la $\mathrm{DBO}_{2022projetée}$ et la DBO_{2022} constitue les écarts actuariels de l'individu.

La juxtaposition de ces formules permet de comprendre que les écarts d'expérience, hors mouvements de population (entrées, sorties, transferts) et corrections de données, s'expliquent par l'écart entre le salaire réel et le salaire projeté à un an et par l'écart d'âge d'un an pour la projection de la probabilité de présence à la retraite.

Enfin, la prestation attendue en $2022+k$, calculée au 31/12/2021, s'écrit :

$$\mathbf{EBP}_{2022+k} = droits_R * salaire_{2021}\ (1+\rho)^k * (1+\mu) * {}_kp_x * \mathbb{1}_{2022+k=R+1}$$

2.3.3 Médailles du travail

Les médailles d'honneur du travail permettent de récompenser un salarié pour l'ancienneté de ses services effectués, elles sont basées sur une ancienneté carrière. Il y a 4 médailles d'honneur du travail : la médaille d'argent pour 20 ans d'ancienneté, la médaille vermeil pour 30 ans d'ancienneté, la médaille d'or pour 35 ans d'ancienneté et la médaille grand or pour 40 ans d'ancienneté.

Certaines entreprises prévoient, par accord ou par usage, la remise d'une gratification à l'occasion de la remise d'une médaille d'honneur, ces gratifications sont alors exonérées de charges sociales dans la limite d'un mois de salaire. Les autres gratifications remises en dehors d'une médaille d'honneur, en général basées sur une ancienneté entreprise, sont entièrement soumises aux charges sociales.

Calcul de l'engagement individuel pour un régime médailles d'honneur du travail selon la méthode PUC with Service Prorate.

L'entreprise s'engage à verser une gratification (que l'on suppose inférieure ou égale à un mois de salaire projeté par simplification) à l'occasion de la remise de chaque médaille d'honneur, on note $A_1, ..., A_4$ les 4 années de remise.

$$\mathbf{DBO}_{2022} = \sum_{k=1}^{4} droits_k * \text{salaire}_{A_k projeté} * {}_{A_k-2022}p_x * \nu^{Ak-2022} * \frac{ancienneté_{31/12/2022}}{ancienneté_{31/12/A_k}}$$

L'engagement d'un régime médailles du travail est toujours croissant avec l'hypothèse d'âge de départ à la retraite. Pour un régime IFC, ce n'est pas forcément le cas : l'augmentation des droits peut être compensée par l'actualisation, la probabilité de présence et le prorata d'ancienneté.

2.3.4 Régime de retraite supplémentaire à prestations définies

Les régimes à prestations définies L137-11-1 du code de la Sécurité sociale ou Article 39 à droits aléatoires sont mis en place par décision unilatérale de l'employeur, par référendum ou par un accord collectif.

Ce type de régime présente deux risques majeurs : le risque de longévité et le risque lié aux investissements. Ainsi, les régimes de retraite à prestations définies créés depuis le 1er janvier 2010 doivent obligatoirement être gérés en externe.

En cas d'externalisation du régime, les organismes gestionnaires peuvent être des organismes d'assurance ou des FRPS (Fonds de Retraite Professionnelle Supplémentaire) pour lesquels les normes prudentielles s'appliquent. En cas de gestion en interne, l'engagement est inscrit au passif des comptes de l'entreprise et la norme IAS 19 s'applique.

Ces régimes constituent, particulièrement pour les cadres et dirigeants qui ont un faible taux de remplacement lors du départ en retraite (rapport entre la pension obligatoire et les derniers revenus d'activité), une solution pour améliorer significativement le niveau de leur pension de retraite.

On distingue deux types de régimes à prestations définies : les régimes additionnels et les régimes différentiels.

Les régimes additionnels garantissent au bénéficiaire un pourcentage fixe du dernier salaire, indépendamment des autres régimes de retraite, l'engagement porte donc sur un niveau de pension. Ces régimes ont un poids sur le passif social indépendant des rendements des régimes obligatoires.

Les régimes différentiels (ou chapeau) garantissent au bénéficiaire un niveau global de revenus qui inclut le régime de base et les régimes complémentaires de retraite, l'engagement porte donc sur la différence. Ces régimes ont une forte dépendance à l'évolution du rendement des régimes obligatoires.

Jusqu'à la loi Pacte de 2019, les régimes à prestations définies conditionnaient le plus souvent l'obtention de la retraite supplémentaire à l'achèvement de la carrière dans l'entre-

prise, d'où le caractère aléatoire. La loi Pacte transpose une directive européenne de 2014 qui supprime cette condition et exige notamment l'acquisition de droits après une période de présence dans l'entreprise d'au plus 3 ans (l'acquisition de droits peut être soumise à une condition d'âge du bénéficiaire, qui ne peut dépasser 21 ans), ce sont les nouveaux régimes L137-11-2 ou Article 39 à droits certains. Les droits acquis sont limités à 3% du salaire de référence par année et le total des droits acquis sur la carrière est limité à 30%. Ces droits sont soumis à des conditions de performance pour les mandataires sociaux et les salariés percevant plus de 8 PASS. Ces droits peuvent être revalorisés annuellement, dans la limite de l'évolution du PASS (revalorisation obligatoirement identique, que le bénéficiaire soit ou non encore dans l'entreprise).

Les régimes L137-11-1 ne peuvent plus accepter de nouveaux bénéficiaires depuis la publication de l'ordonnance n°2019-697 du 3 juillet 2019 relative aux régimes professionnels de retraite supplémentaire et les droits sont gelés au 31 décembre 2019. Comptablement, à la date du gel, on enregistre un coût des services passés égal à VAP – DBO ; après le gel, l'engagement est égal à la VAP et le NC est nul. Seuls les régimes fermés aux nouveaux entrants au plus tard le 20 mai 2014 peuvent continuer à fonctionner comme avant.

L'ordonnance sur la sécurisation des rentes prise par le gouvernement le 9 juillet 2015 prévoit l'obligation de sécuriser une quote-part de l'engagement relatif aux rentes en service au titre des régimes à prestations définies mentionnés à l'article L137-11 d'ici 2030 afin de protéger les salariés bénéficiaires contre le risque d'insolvabilité de l'employeur. Cette quote-part est de 50% des rentes en service, dans la limite de 1,5 fois le PASS, par bénéficiaire et par année. Les taux intermédiaires de 10%, 20% et 40% doivent être atteints à compter, respectivement, de la clôture des comptes immédiatement postérieure au 1er janvier 2017, 1er janvier 2020 et 1er janvier 2025.

La sécurisation peut être faite par la souscription d'un contrat auprès d'une compagnie d'assurance, d'une institution de prévoyance ou d'une mutuelle ; par des fiducies ; par des suretés réelles ou personnelles. En cas de non-respect de ces nouvelles obligations, l'ordonnance prévoit une pénalité annuelle de 30% de la différence entre le montant qui devrait être sécurisé et le montant effectivement sécurisé.

Enfin, les régimes L137-11-1 sont soumis, depuis le 1er janvier 2004, à une taxation

sur les rentes versées ou sur le financement, sur option irrévocable de l'employeur (cf. Annexe 1). Cette contribution patronale, aussi appelée taxe Fillon, est payée directement par l'organisme gestionnaire de la rente à l'URSSAF.

Pour les régimes L137-11-2, le financement de l'employeur (cotisations versées à l'organisme en charge de l'assurance du régime) est soumis à une contribution au taux de 29,7% contre 24% pour les régimes L137-11-1. Ces nouveaux régimes offrent désormais la possibilité de passer d'une taxation sur les rentes à une taxation sur le financement.

Calcul de l'engagement individuel pour un régime additionnel de retraite à prestations définies selon la méthode PUC with Service Prorate (hors IFRIC).

L'entreprise s'engage à verser au salarié ou au réversataire une rente annuelle égale à $k\%$ du salaire au moment de départ à la retraite du salarié, si ce dernier fait toujours partie de l'entreprise à cet instant. Les rentes sont supposées annuelles, à terme échu, revalorisées annuellement au taux r et réversibles d'un coefficient α.

Notations :
R_x : âge du salarié au $31/12/R$
R_y : âge du conjoint au $31/12/R$
$\ddot{a}_x = \sum_{k=1}^{\infty} {}_kp_x(\frac{1+r}{1+i})^k$
$\ddot{a}_{x|y} = \sum_{k=1}^{\infty} {}_kp_y(1 - {}_kp_x)(\frac{1+r}{1+i})^k$

$\mathbf{DBO}_{2022} = k\% * salaire_{2022}\ (1+\rho)^{R-2022} * (1+\mu) * {}_{R-2022}p_x * (\ddot{a}_{R_x} + {}_{R-2022}p_y * \alpha * \ddot{a}_{R_x|R_y})$
$* \nu^{R-2022} * \frac{ancienneté_{31/12/2022}}{ancienneté_{31/12/R}}$

Pour rappel, le taux technique $(\frac{1+i}{1+r} - 1)$ maximum en vie (hors branche 26) est égal à min $(3,5\%$; 60% TME moyen des 6 derniers mois). Le TME est le taux de rendement sur le marché secondaire des emprunts d'Etat à taux fixe supérieurs à 7 ans.

Le coefficient de rente $\ddot{a}_x$ correspond au capital nécessaire pour avoir une rente viagère immédiate, à terme échu, revalorisée annuellement, de 1 €. Dans le cas où le taux technique est nul, le coefficient $\ddot{a}_x$ représente le nombre d'années moyen pendant lequel la rente va être versée au bénéficiaire. L'inverse de $\ddot{a}_x$ correspond au taux de conversion : en le multipliant par le capital constitué à la liquidation, il permet de connaître la rente annuelle servie.

2.4 Evènements spéciaux

Les différents évènements spéciaux faisant l'objet d'un traitement particulier dans l'évaluation des engagements sociaux sont présentés ci-dessous.

Evènement spécial	Exemples	Impact sur l'engagement
Modification de régime (Plan amendment)	Changement des droits accordés par la CCN d'un régime IFC. Changement de barème ou des dates anniversaires de paiement dans un régime de médailles du travail. Changement de la rente garantie d'un régime à prestations définies.	+/-
Réduction de régime (Curtailement)	Plan social (PSE) ou nombre de départs significatif (plus de 10% de l'engagement). Changement des termes d'un régime à prestations définies de sorte qu'une partie des services futurs des bénéficiaires ne leur donnera plus de droits à prestations.	-
Liquidation de régime (Settlement)	Transfert du risque à un assureur (externalisation, sécurisation des rentes).	+/-
Acquisition / cession	Acquisition / cession d'une entité, d'une ligne d'activité.	+/-
Fusions / transferts	Fusion d'entités ou transferts massifs d'employés inter-entités.	+/-

FIGURE 6 – Table des différents évènements spéciaux

Un Past Service Cost (PSC) est comptabilisé dès lors qu'il y a une modification de régime.

Par ailleurs, en IAS 19 et en French GAAP, les évènements spéciaux sont comptabilisés à la date de signature alors qu'en US GAAP, ils sont comptabilisés à la date de survenance.

2.5 Comptabilisation en normes IAS 19, French GAAP et US GAAP

Selon les entreprises, la comptabilisation des engagements sociaux s'effectue en normes françaises French GAAP, américaines US GAAP et/ou internationales IFRS. Ces normes se diffèrent principalement dans la terminologie, la reconnaissance des gains et pertes actuariels et la comptabilisation des évènements spéciaux.

En France, les normes comptables sont édictées par l'Autorité des normes comptables (ANC). L'ANC résulte de la fusion du Conseil national de la comptabilité (CNC) et du Comité de la réglementation comptable (CRC) en 2009. La mise en œuvre des normes IFRS pour l'établissement des comptes consolidés des sociétés cotées sur le marché boursier français est gérée et contrôlée par l'ANC.

Les passifs sociaux sont comptabilisés dans les comptes consolidés et ils doivent être mentionnés en annexe des comptes sociaux. En comptes sociaux, il n'est pas nécessaire de provisionner les avantages similaires à des engagements de retraite. En revanche, le provisionnement des engagements afférents aux médailles du travail et gratifications d'ancienneté est obligatoire.

La recommandation R.2013.02 de l'ANC abroge la précédente recommandation du CNC n°2003-R.01 à l'exception des sections 7 et 8 de son annexe relative aux autres avantages long terme et aux indemnités de cessation d'emploi. Elle permet aux entités qui provisionnent en totalité leurs engagements de retraite, au choix, de continuer à appliquer les dispositions de l'ancienne recommandation ou de se rapprocher au maximum des nouvelles dispositions de la norme IAS 19 révisée notamment pour les entreprises faisant partie d'un groupe établissant ses comptes consolidés en normes IFRS. Enfin, pour les entreprises comptant moins de 250 salariés, il est toujours possible de recourir à une méthode simplifiée d'évaluation de leurs engagements de retraite.

Le traitement comptable des différents éléments intervenant dans la réconciliation de la DBO et de la juste valeur de l'actif de couverture dans les 3 normes citées est présenté ci-dessous.

Réconciliation de la DBO

	IAS 19		**French GAAP**		**US GAAP**		
	P&L	*OCI*	*P&L*	*Hors bilan*	*P&L*	*OCI*	*AOCI*
DBO BOY							
Service Cost	X		X		X		
Interest Cost	X		X		X		
Prestations de l'employeur							
Prestations par le fonds							
Modification de régime	X (PSC)		X (1)	X (Stock PSC)	X (1)	X (PSC de l'année)	X (Stock PSC)
			X (3)				
Réduction de régime	X (PSC)		X (4)		X (4)		
Liquidation de régime	X		X		X		
G&L Avantages postérieurs à l'emploi		X	X (2)	X (Stock G&L)	X (2)	X (G&L de l'année)	X (Stock G&L)
			X (3)		X (3)		
G&L Autres avantages long terme	X		X		X		
DBO EOY							

Réconciliation de l'Actif de couverture

	IAS 19		French GAAP		US GAAP		
	P&L	*OCI*	*P&L*	*Hors bilan*	*P&L*	*OCI*	*AOCI*
Actif BOY							
Cotisations employeur							
Interest Income	X		X		X		
Prestations par le fonds							
G&L Avantages postérieurs à l'emploi		X	X (2)	X (Stock G&L)	X (2)	X (G&L de l'année)	X (Stock G&L)
			X (3)		X (3)		
G&L Autres avantages long terme	X		X		X		
Actif EOY							

(1) Amortissement linéaire sur la durée résiduelle d'activité des salariés (égale au maximum au barycentre des durées résiduelles d'activité des salariés pondérées par les probabilités de présence à la retraite) ou sur l'espérance de vie résiduelle pour les retraités.

(2) Amortissement par la méthode du Corridor.

(3) Autre méthode de comptabilisation possible, une fois choisie elle ne doit plus changer.

(4) La réduction impacte le stock de gains et pertes actuariels (seulement dans le cas où il s'agit d'un stock de pertes en US GAAP) ainsi que le stock de PSC. La somme de ces impacts et du montant de la réduction passe dans la charge de l'année.

Méthode du Corridor : Les écarts actuariels ne sont pas tous reconnus immédiatement. Le stock des écarts excédant le Corridor est amorti sur la durée de vie résiduelle des actifs via le compte de résultat, ce qui n'excède pas le Corridor reste en hors-bilan.

$$\text{Amortissement}_{N+1} = \frac{(Stock\ G\&L_N - Corridor_N)_+}{Durée\ résiduelle}$$

$$\text{Corridor}_N = 10\% * max(\text{Actif}_N\ ;\ \text{DBO}_N)$$

Ainsi, le stock est augmenté des gains et pertes actuariels de l'année et diminué des amortissements. Cette méthode permet de différer la comptabilisation des écarts actuariels et de réduire la volatilité du passif en particulier si les variations année par année des engagements se compensent.

L'OCI (Other Comprehensive Income) est un compte de capitaux propres servant à lisser la volatilité du résultat. En French GAAP et en US GAAP, certains éléments sont initialement reconnus en hors-bilan ou OCI puis recyclés en P&L (compte de résultat). En US GAAP, il y a une distinction entre OCI où sont reconnus les gains et pertes actuariels et le coût des services passés de l'année et AOCI (Accumulated Other Comprehensive Income) qui correspond au stock d'OCI.

Finalement, il y a 2 façons d'obtenir la provision de l'année, l'une étant la réconciliation par rapport à celle de l'année dernière, l'autre étant obtenue à partir de la DBO de l'année.

Réconciliation de la provision

Provision N-1	**DBO N**
+ Charge N	- Actif de couverture N
+ OCI N (1)	*+ Stock G&L N (2)*
- Cotisations et prestations employeur N	*+ Stock PSC N (2)*
Provision N	**Provision N**

(1) OCI en IAS 19, AOCI en US GAAP et Hors-bilan en French GAAP.
(2) N'existe pas en IAS 19.

3 Contexte de l'étude

Le turnover ou rotation du personnel est un indicateur décrivant le rythme de renouvellement des effectifs dans une organisation, il peut fortement fluctuer en fonction du secteur. Dans le contexte des engagements sociaux, il peut être considéré comme une mortalité administrative puisque la prestation est en général conditionnée à la présence du salarié dans l'entreprise.

C'est un indicateur essentiel en gestion des ressources humaines.
Plusieurs éléments encouragent la diminution du turnover : coût du départ et du remplacement, rupture des tâches, formation des entrants et productivité faible au début, risque personne clé (personne jouant un rôle déterminant dans le fonctionnement de l'entreprise), risque de perte de savoirs, de compétences, d'expérience et impact sur l'image et le climat de l'entreprise.
D'autres encouragent l'augmentation du turnover : erreur de recrutement, diminution de la masse salariale en remplaçant des salariés à forte rémunération par des salariés à rémunération moindre, c'est l'effet de noria.

Dans les engagements sociaux en particulier, on s'intéresse aux démissions. Le taux de démission est un indicateur cyclique. Il est bas durant les crises et il augmente en période de reprise, d'autant plus fortement que l'embellie conjoncturelle est rapide. Ainsi, pendant la crise sanitaire de Covid 19, notamment en 2020 et 2021, le nombre de démissions a chuté avant de repartir à la hausse. Au 1er trimestre 2022, le nombre de démissions a atteint un niveau historiquement haut avec près de 470 000 démissions de CDI (taux de démission de 2,7%). Le record précédent datait du 1er trimestre 2008, avec 400 000 démissions de CDI (taux de démission de 2,9%).

Le risque d'une "Grande démission" fut évoqué en France après la crise, faisant référence à une expression décrivant la situation du marché du travail américain courant 2021 : "Big Quit" ou "Great Resignation", un phénomène qui s'illustre par la propension des salariés à quitter leur poste.

Voici un graphique de l'évolution des démissions en France sur les 15 dernières années :

Champ : ensemble des établissements français employeurs de France métropolitaine sur le champ privé hors agriculture, intérim et particuliers employeurs.
Source : Dares, Mouvements de main-d'œuvre.

FIGURE 7 – Evolution des démissions en France

Par ailleurs, avec un taux de chômage au plus bas depuis des décennies, la sécurité de l'emploi n'attire plus autant et pourrait engendrer une hausse du turnover. Dans un tel contexte, il est important de bien appréhender le facteur turnover dans le calcul de l'engagement actuariel des entreprises.

Le turnover est classiquement défini comme suit :

$$Turnover_N = \frac{nombre\ de\ sorties\ l'année\ N + nombre\ d'entrées\ l'année\ N}{2 * \ effectif\ en\ début\ d'année\ N}$$

Un turnover à 0% signifie qu'aucun salarié n'est arrivé ni parti ; à 100%, l'intégralité des postes a été renouvelée.

Néanmoins dans le cadre des engagements sociaux, les entrées ne sont pas prises en compte dans le turnover afin de pouvoir définir une probabilité de sortie pour chaque individu. Ainsi, l'engagement lié à chaque entrée est comptabilisé en pertes d'expérience.

L'hypothèse de turnover n'étant pas clairement définie pour les sorties, la Compagnie Nationale des Commissaires aux Comptes (CNCC) a rédigé en 2018 une note qui précise que seuls les motifs de démissions doivent être pris en compte.

La CNCC tient le raisonnement suivant :
Qu'il s'agisse du départ à la retraite, du licenciement ou de la rupture conventionnelle, l'entité est tenue de payer une indemnité au salarié qui la quitte. Le seul cas de départ du salarié n'engendrant le paiement d'aucune indemnité est celui de la démission. Dans la mesure où tout autre cas de départ avant l'âge de la retraite engendre pour l'entité un paiement au moins aussi important que l'indemnité de fin de carrière, la Commission est d'avis que l'évaluation des indemnités de fin de carrière doit être effectuée en tenant compte des seules prévisions de démission.

Le raisonnement développé selon les normes IFRS est le suivant :
En procédant à des licenciements, l'entreprise s'exonère de son obligation de payer des indemnités de départ à la retraite, mais elle ne peut le faire qu'en lui substituant une autre obligation, en général plus onéreuse, celle de payer des indemnités de licenciements. Or, les licenciements étant sous le contrôle de l'entreprise, ils ne peuvent être provisionnés qu'à la date d'annonce des licenciements conformément aux conditions imposées par IAS 19.165-167. En conséquence, tenir compte des futurs licenciements dans le calcul du taux de rotation retenu pour calculer l'engagement de retraite aboutirait à sous-évaluer les provisions reconnues au bilan au titre des indemnités de départ à la retraite.

Par conséquent, les motifs de départ exclus dans le turnover sont les décès, les licenciements économiques, les licenciements individuels, les mutations, les invalidités et inaptitudes, les fins de période d'essai et les ruptures conventionnelles. Il n'anticipe pas non plus les évènements spéciaux. Les salariés en congé maladie ou congé parental ne sont pas comptés comme des sortants, leur rémunération est annualisée. Ainsi le montant projeté de l'engagement de chaque sortant pour une autre raison qu'un départ en retraite est compté en gains d'expérience.

Le turnover est à discriminer selon l'âge et parfois la CSP. Les pratiques marché lui imposent d'être décroissant en fonction de l'âge et nul après 55 ans. Les variables ancienneté, sexe et salaire seront aussi prises en compte dans les modèles de prédiction des démissions.

D'autres facteurs tels que le statut marital, le nombre d'enfants à charge, la localisation géographique, la distance domicile - lieu de travail pourraient avoir une influence sur cette variable mais nous ne disposons pas de ces informations.

La problématique de ce mémoire est d'approfondir la sensibilité de l'engagement actuariel au turnover. Pour ce faire, il s'agit dans un premier temps de comparer l'impact sur l'engagement de différentes techniques de lissage de tables et d'appliquer dans un second temps certains modèles de classification pour la prédiction du turnover.

4 Présentation des données

4.1 Extraction et traitement des données

Mercer, spécialiste des grands comptes, accompagne de nombreuses entreprises du CAC 40 dans l'évaluation de leurs engagements sociaux. Je dispose ainsi des données 2019, 2020 et 2021 de l'effectif d'un grand groupe : matricule, CSP, date de naissance, date d'ancienneté, salaire et sexe. Je dispose aussi des données de sorties des années 2019, 2020, 2021 et ne regarde que les démissions qui figurent dans les effectifs en début d'année.

L'étude de turnover se concentre sur les données 2019. En effet, les années 2020 et 2021 ont été marquées par la crise sanitaire et notre but est de décrire une tendance des taux de démissions applicable aux évaluations actuarielles des 5 prochaines années. Au vu du graphique de l'évolution des démissions présenté en partie 3 et des données dont nous disposons, baser notre étude sur l'année 2020 notamment aboutirait à surestimer l'engagement du fait de la chute situationnelle des démissions.

La période d'observation est d'un an, du 1er janvier 2019 au 31 décembre 2019. Par simplification, les données sont supposées complètes, au sens où les individus ont atteint l'âge x après le début de la période d'observation (pas de troncature à gauche) et ont démissionné ou bien atteint l'âge $x + 1$ avant la fin de la période d'observation (pas de censure à droite).

Pour l'évaluation actuarielle qui suivra, on se placera au 31/12/2021. Les données utilisées seront les données 2021 obtenues après l'étape du traitement de données. Cette étape essentielle d'une évaluation actuarielle consiste à lever toutes les incohérences des données avant de lancer les calculs et d'évaluer les écarts d'expérience.

Il s'agit tout d'abord de vérifier qu'il n'y a pas de données manquantes. Ensuite, des contrôles sur l'âge sont réalisés : un salarié ayant moins de 18 ans pourrait ne pas être en CDI et un salarié ayant plus de 67 ans pourrait déjà être parti à la retraite. Pour les salaires annuels transmis, une explication ou une annualisation est demandée s'ils sont inférieurs au SMIC brut annuel. Aussi, une réconciliation par rapport aux données de l'exercice précédent est faite pour interroger le client si une variation à la hausse ou à la baisse du salaire est très importante ou si une rétrogradation de CSP, un changement de sexe, de date de

naissance ou d'ancienneté a eu lieu. Enfin, le cas d'un salarié absent l'année dernière mais ayant plus de 2 ans d'ancienneté ou d'un salarié désormais absent des données mais n'apparaissant pas dans les sorties de l'année réclame aussi une question. Cette liste de questions posées au client n'est pas exhaustive mais donne un aperçu des différents contrôles effectués.

4.2 Etude descriptive

L'effectif 2019 du groupe s'élève à 36 520 personnes réparties entre 27 616 cadres et 8 904 non cadres, 27 441 hommes et 9 079 femmes. Le nombre de démissions est de 740 personnes réparties entre 698 cadres et 42 non cadres, 582 hommes et 158 femmes. L'effectif contient une majorité d'hommes cadres, ce sont aussi ceux qui démissionnent le plus. Le nombre de démissions sur l'année représente une proportion assez faible de l'effectif global. Cette problématique sera traitée avant application des modèles de prédiction.

Nous disposerons de 5 variables explicatives pour la prédiction de la variable binaire démission : 2 variables catégorielles (sexe et CSP), 2 variables discrètes (âge et ancienneté) et 1 variable continue (salaire).

Visuellement, la répartition de l'effectif selon la CSP et le sexe est la suivante :

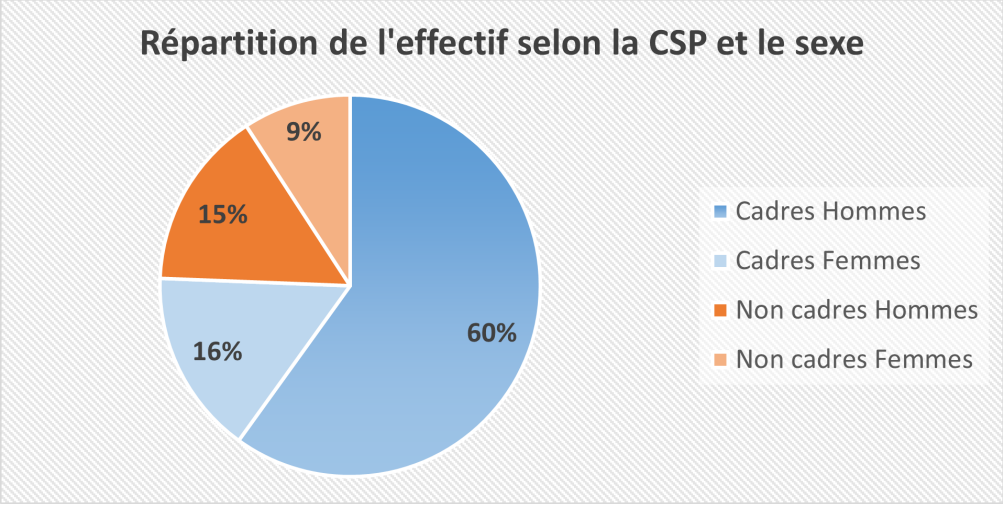

FIGURE 8 – Répartition de l'effectif

Visuellement, la répartition des démissions est la suivante :

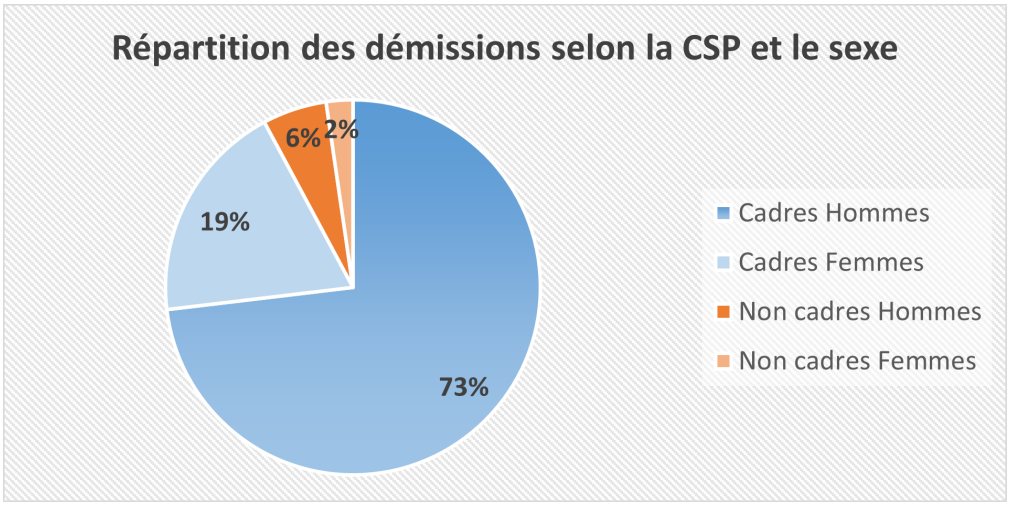

FIGURE 9 – Répartition des démissions

La matrice des V de Cramer pour les variables catégorielles est présentée ci-dessous :

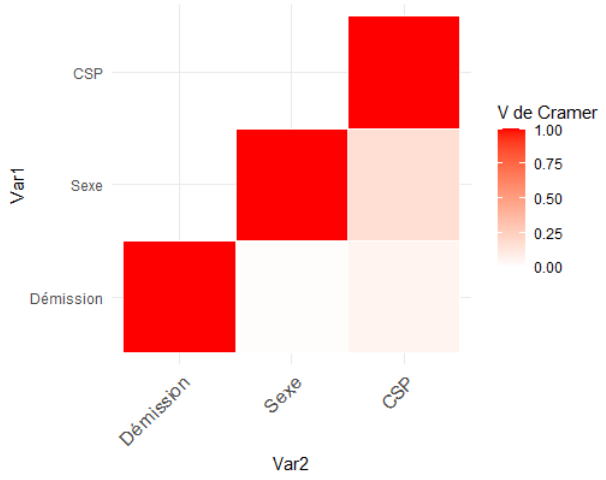

FIGURE 10 – Matrice des V de Cramer

Le V de Cramer est une mesure d'association entre deux variables qualitatives, construite à partir de la statistique de test du Khi-2 d'indépendance. Plus sa valeur est proche de 0, moins les variables sont dépendantes. La variable démission semble avoir plus de liens de dépendance avec la CSP qu'avec le sexe des individus.

Les principales statistiques des variables quantitatives sont résumées ci-dessous :

Variable	Moyenne	Minimum	1er Quartile	Médiane	3e Quartile	Maximum
Age	46	20	37	46	55	76
Ancienneté	17	0	6	16	27	50
Salaire annuel	63 271 €	21 645 €	43 683 €	56 682 €	74 017 €	714 566 €

FIGURE 11 – Statistiques des variables quantitatives

Les corrélations linéaires de Pearson entre variables quantitatives et la variable d'intérêt présentées ci-dessous témoignent de l'importante corrélation linéaire positive entre âge et ancienneté. La variable salaire est aussi logiquement corrélée positivement aux variables âge et ancienneté. La variable démission apparaît décorrélée des salaires et corrélée négativement à l'âge et l'ancienneté.

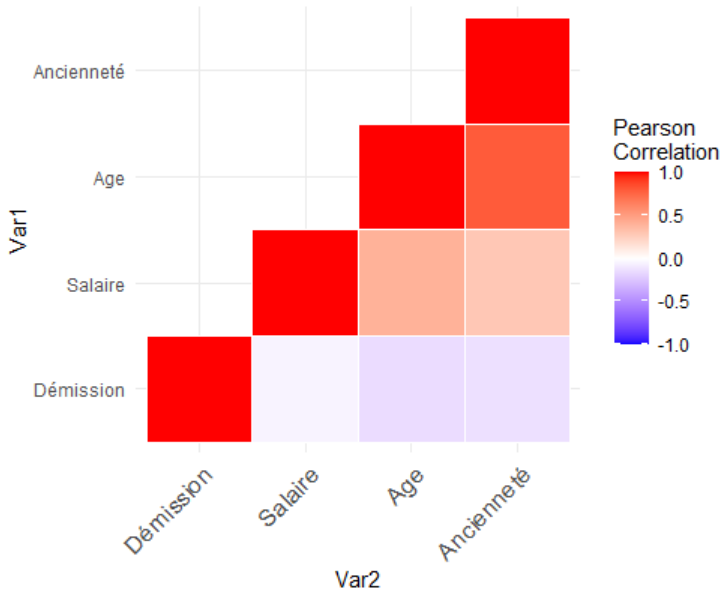

FIGURE 12 – Matrice des corrélations linéaires de Pearson

4.3 Taux bruts de démissions

L'estimateur utilisé pour les taux bruts de turnover est l'estimateur binomial de Hoem.

Le taux de démission réel à l'âge x, t_x est estimé par $\hat{t_x} = \frac{nombre\ de\ démissions\ à\ l'âge\ x}{nombre\ de\ personnes\ d'âge\ x} = \frac{d_x}{n_x}$.

On suppose qu'il y a indépendance entre individus et que le nombre de démissions à l'âge x suit une loi binomiale : $d_x = \sum_{i=1}^{n_x} \mathbb{1}_{i\ démissionne} \sim Bin(n_x, t_x)$.

Par le théorème central limite : $\sqrt{n_x}(\hat{t_x} - t_x) \xrightarrow[n_x \to \infty]{\mathcal{L}} \mathcal{N}(0, t_x(1-t_x))$.

Ce théorème combiné au lemme de Slutsky permet de construire l'intervalle de confiance asymptotique de niveau α suivant : $t_x \in [\hat{t_x} \pm q_{1-\alpha/2}(\mathcal{N}(0,1))\sqrt{\frac{\hat{t_x}(1-\hat{t_x})}{n_x}}]$.

Enfin, on peut montrer que $\hat{t_x}$ est l'estimateur du maximum de vraisemblance de t_x.

Les taux bruts de turnover par âge discriminés selon la CSP sont les suivants :

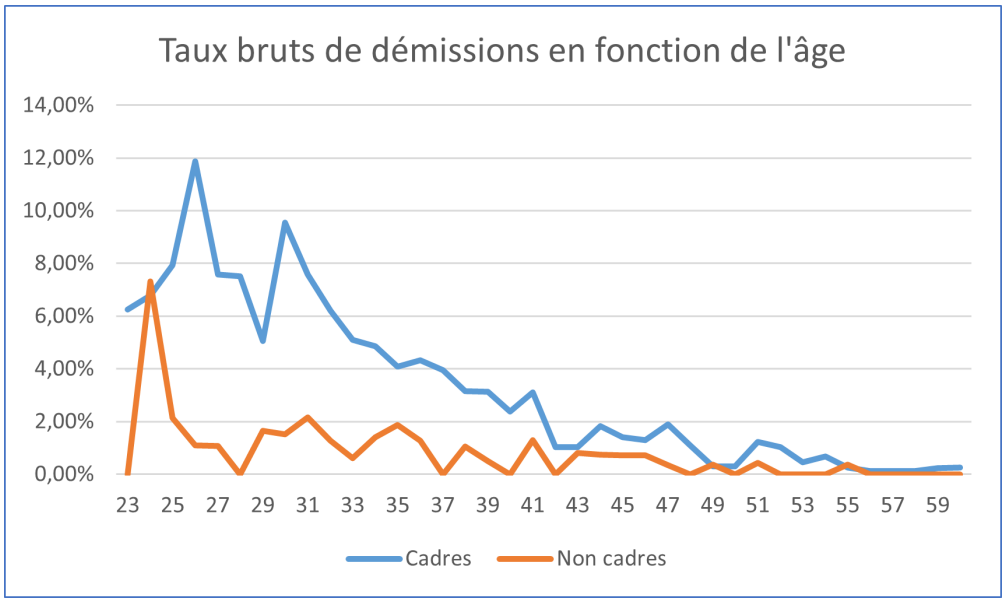

FIGURE 13 – Taux bruts de démissions

La plage des âges retenue s'étend de 23 à 60 ans, le volume de données étant jugé trop faible pour les âges inférieurs à 23 ans et les taux de démissions étant quoi qu'il en soit fixés à 0% au-delà de 55 ans selon l'usage du marché.

Cet usage marché est discutable et ce d'autant plus avec la réforme des retraites attendue pour 2023. De plus, les taux de turnover de fin de table, associés aux âges les plus élevés, impactent toute la population, la qualité de leur ajustement est d'autant plus importante. Toutefois, cet usage n'a pas été challengé dans ce mémoire étant donné les très faibles taux de démission constatés empiriquement à partir de 56 ans (inférieurs à 2‰).

Par ailleurs, on note que séparer cadres et non cadres a du sens pour cette entreprise. La courbe des taux bruts est assez irrégulière bien qu'elle ait tendance à décroître avec l'âge. Il est légitime de penser que ces irrégularités ne reflètent pas le phénomène sous-jacent que l'on cherche à mesurer. Ces aspérités sont dues aux fluctuations d'échantillonnage et il s'agit désormais d'appliquer différentes techniques de lissage aux courbes obtenues.

4.4 Méthode d'ajustement à partir d'une table de référence

La méthode d'ajustement de la table de turnover utilisée chez Mercer consiste à partir d'une table de référence initialement mise en place à partir des données d'une centaine de clients et à l'ajuster tous les 3 à 5 ans. Une valeur cible est utilisée pour obtenir le coefficient multiplicateur à appliquer à la table de référence pour retrouver le taux de démission moyen observé sur les 3 dernières années.

Les taux lissés obtenus à partir de la table de référence sont les suivants :

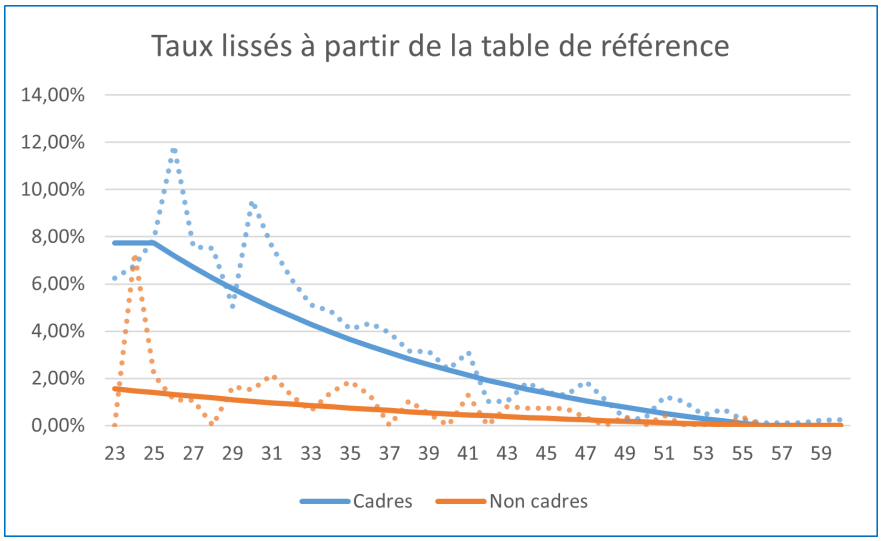

FIGURE 14 – Taux lissés à partir de la table de référence

Cette méthode a l'avantage d'être simple et de tenir compte des démissions sur 3 années mais elle suppose une allure de la courbe turnover commune à tous les clients et invariante au fil des années étant donné qu'elle sera simplement translatée verticalement par rapport à la courbe initiale.

Les lissages non paramétriques qui sont étudiés dans la suite ont l'avantage de ne pas donner d'a priori à l'allure de la courbe de turnover et d'utiliser au mieux l'information contenue dans les données.

5 Lissages non paramétriques

Le choix d'une procédure de révision des données brutes fait intervenir deux types de contraintes qui devront être prises en considération de manière conjointe :
- la précision ou fidélité : il est naturel d'attendre des taux révisés qu'ils soient proches des taux initiaux
- la régularité : la suite des taux ajustés sera recherchée aussi régulière que possible.

La conciliation et l'importance relative donnée à ces deux objectifs contradictoires passe par le choix d'un paramètre de lissage.

L'objet de cette partie est le lissage non paramétrique, ainsi contrairement à l'ajustement d'un modèle paramétrique, on ne fait pas d'hypothèse sur la distribution de la courbe de turnover. A ce titre, les courbes de lissage ne sont pas prévues pour obtenir une estimation des taux de turnover en dehors de la plage de lissage. D'ailleurs, une extrapolation est par nature impossible pour les méthodes de lissages non paramétriques étant donné que les taux de turnover ne sont pas représentés à l'aide d'une fonction mathématique, exception faite aux splines de lissage qui coïncident avec les splines naturelles dont on connaît une paramétrisation.

Par conséquent, bien que moins pratique pour la manipulation et l'interprétation, le non paramétrique faisant moins d'hypothèses, est censé diminuer le biais de modèle.

Dans tous les lissages qui seront réalisés, un retraitement sera effectué pour fixer à 0% le turnover au-delà de 55 ans et des choix forts pourront être faits sur les paramètres pour avoir une table décroissante avec l'âge.

Enfin, le test usuel du Khi-2 d'adéquation largement utilisé dans le cadre des ajustements paramétriques pour vérifier la qualité globale des taux révisés est ici plus délicat à mettre en œuvre car la détermination du nombre de degrés de liberté pour la loi du Khi-2 pose problème.

5.1 Méthode des moyennes mobiles pondérées

La méthode des moyennes mobiles pondérées est l'une des premières méthodes de lissage non paramétrique à avoir été développée. La valeur lissée du taux à l'âge x est obtenue en prenant la moyenne mobile pondérée de $2h+1$ taux bruts centrés en x.

$$MA_h(\hat{q}_x) = \frac{1}{2h+1} \sum_{i=-h}^{h} \alpha_i \hat{q}_{x+i}$$

Le choix de h détermine la taille de la plage de lissage et permet d'arbitrer entre lissage et proximité aux données brutes.

L'avantage de la méthode des moyennes mobiles pondérées est qu'elle est très simple à mettre en œuvre. Toutefois, elle présente deux inconvénients majeurs, d'une part la moyenne est sensible aux valeurs extrêmes, d'autre part l'utilisation de la méthode pose problème aux extrémités de la plage d'âge et ce d'autant plus que h est grand.

Voici les lissages obtenus en utilisant une moyenne mobile simple avec $h = 7$:

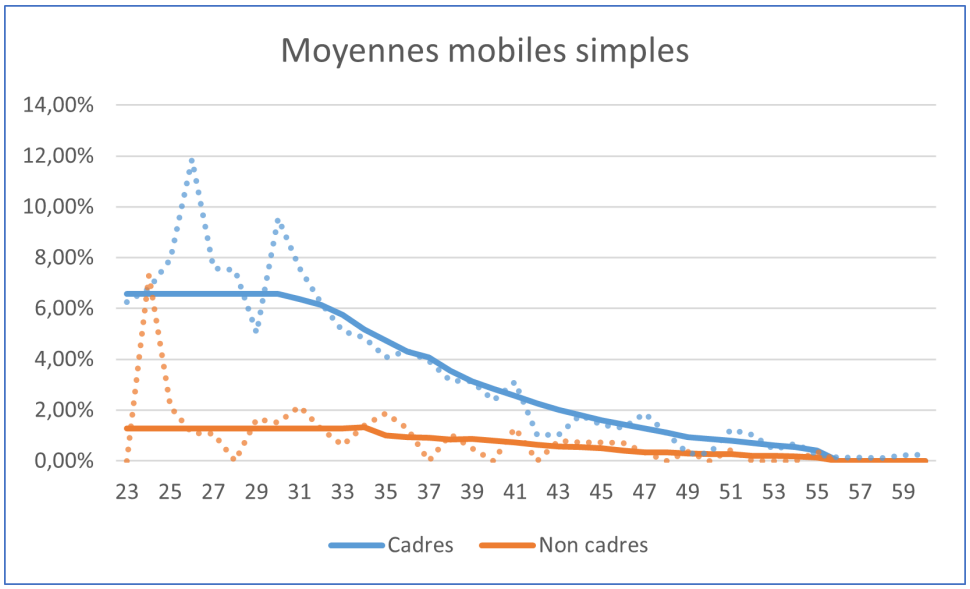

FIGURE 15 – Taux lissés selon la méthode des moyennes mobiles

5.2 Méthode de Whittaker-Henderson

La méthode de Whittaker-Henderson introduite en 1923 est la combinaison linéaire d'un critère de fidélité et d'un critère de régularité, une bonne courbe devant à la fois être proche de la courbe des taux bruts et suffisamment régulière.

5.2.1 Méthode en une dimension

On dispose des estimations $\hat{q} = (\hat{q_i})_{1 \leq i \leq p}$ des p taux bruts pour chaque âge.

Critère de fidélité : $F(q) = \sum_{i=1}^{p} w_i (q_i - \hat{q_i})^2 = (q - \hat{q})' W (q - \hat{q})$

En général, on choisit $w_x = \frac{n_x}{\overline{n}}$ ou $w_x = 1$. Le premier choix permet de limiter le poids donné aux points aberrants.

Critère de régularité : $S(q) = \sum_{i=1}^{p-m} (\Delta^m q_i)^2 = q' K_m' K_m q$

On note $\Delta^m q_i = \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} q_{i+j}$ avec $m < p$ généralement égal à 2 ou 3.

On introduit la matrice K_m de taille $(p - m)$ x p, dont les termes sont les coefficients binomiaux d'ordre m dont le signe alterne et commence positivement pour m pair.

Cela permet finalement d'écrire le critère de Whittaker-Henderson à minimiser :

$$WH_\lambda(q) = F(q) + \lambda S(q) = (q - \hat{q})' W (q - \hat{q}) + \lambda q' K_m' K_m q$$

On résout $\frac{dWH_\lambda(q)}{dq} = 2W(q - \hat{q}) + 2\lambda K_m' K_m q = 0 \iff (W + \lambda K_m' K_m) q = W\hat{q}$.

La solution de problème est : $q^* = (W + \lambda K_m' K_m)^{-1} W \hat{q}$ avec λ le paramètre de lissage.

Il faut veiller au fait que $\lambda K_m' K_m$ n'est pas inversible, l'addition de W diagonale de coefficients diagonaux non nuls rend la matrice à inverser inversible mais l'inversion peut être délicate. En pratique, on peut utiliser la décomposition de Cholesky de $W + \lambda K_m' K_m$ qui est symétrique définie positive pour l'inverser.

Les taux lissés selon la méthode de Whittaker-Henderson implémentée sous R avec pour paramètres $m = 2$ et $h = 20$ sont les suivants :

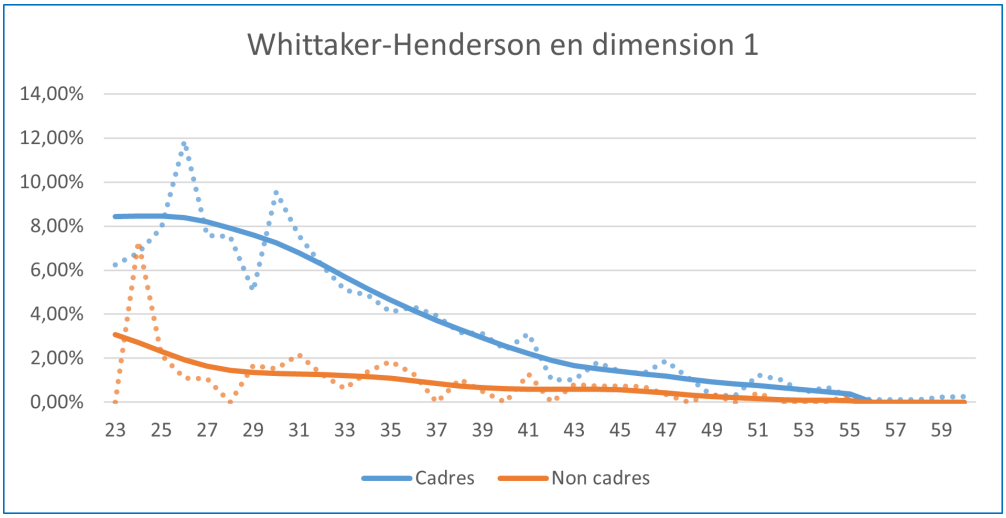

FIGURE 16 – Taux lissés selon la méthode de Whittaker-Henderson en dimension 1

5.2.2 Méthode en deux dimensions

On dispose des estimations $\hat{q} = (\hat{q_{ij}})_{1\leq i\leq p, 1\leq j\leq q}$ des pq taux bruts pour chaque âge et chaque ancienneté.

Critère de fidélité : $F(q) = \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij}(q_{ij} - \hat{q_{ij}})^2$

Critère de régularité verticale : $S_v(q) = \sum_{j=1}^{q}\sum_{i=1}^{p-m}(\Delta_v^m q_{ij})^2$ $\quad m < p$ l'ordre vertical

Critère de régularité horizontale : $S_h(q) = \sum_{i=1}^{p}\sum_{j=1}^{q-n}(\Delta_h^n q_{ij})^2$ $\quad n < q$ l'ordre horizontal

Critère de Whittaker-Henderson : $WH_{\alpha,\beta}(q) = F(q) + \alpha S_v(q) + \beta S_h(q)$

La résolution du problème d'optimisation s'effectue en réarrangeant les éléments pour se ramener au cas unidimensionnel.

On définit le vecteur colonne u de taille pq tel que $u_{q(i-1)+j} = \hat{q_{ij}}$ et la matrice des poids W^*, carrée de taille pq, telle que $w^*_{q(i-1)+j,q(i-1)+j} = w_{ij}$.

K_m^v est une matrice de taille $(p-m)q$ x pq et K_n^h est une matrice de taille $(q-n)p$ x pq.

Le critère à minimiser s'écrit : $(q-u)'W^*(q-u) + \alpha q' K_m^{v\prime} K_m^v q + \beta q' K_n^{h\prime} K_n^h q$.

La solution de ce problème est : $q^* = (W^* + \alpha K_m^{v\prime} K_m^v + \beta K_n^{h\prime} K_n^h)^{-1} W^* u$.

Les taux bruts fonction de l'âge et de l'ancienneté sont les suivants :

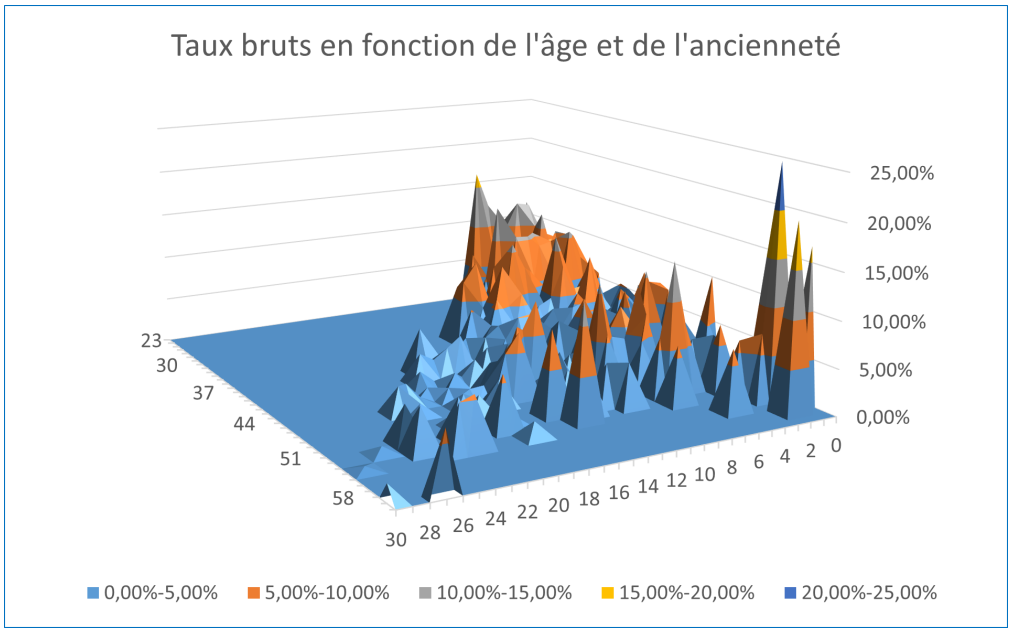

FIGURE 17 – Taux bruts de démissions en 2 dimensions

La partie triangulaire de la table correspondant à des âges bas pour des anciennetés élevées est logiquement vide.

Les taux lissés selon la méthode de Whittaker-Henderson avec pour paramètres $m = n = 3$ et $\alpha = \beta = 20$ sont les suivants :

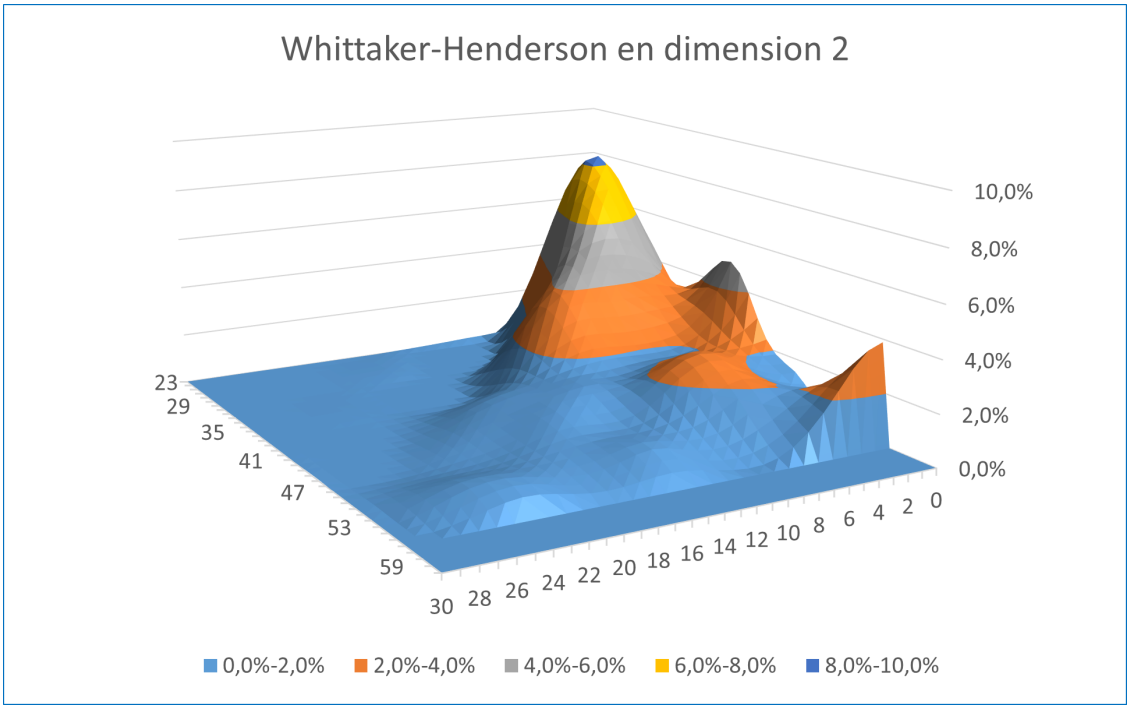

FIGURE 18 – Taux lissés selon la méthode de Whittaker-Henderson en dimension 2

Cette méthode permet un lissage conjoint dans les deux directions, elle a l'avantage de tenir compte de l'ancienneté et donc de la trajectoire de la carrière du salarié au sein de l'entreprise. Elle est plus efficace que le lissage séparé selon chaque variable mais réclame plus de puissance de calcul.

Le taux de démission le plus élevé observé après lissage est atteint pour les salariés ayant 27 ans et 2 ans d'ancienneté avec 8,3% de démissions dans l'année. Il s'agit donc d'un profil jeune avec un peu d'expérience, certainement assez recherché dans le domaine dans lequel il exerce et disposant de nombreuses opportunités de poursuite de carrière.

5.3 Méthode des splines de lissage

La troisième méthode de lissage non paramétrique utilisée est la méthode des splines de lissage. La théorie des splines polynomiales est d'abord présentée.

5.3.1 Splines polynomiales

Définition : Pour un intervalle $[a, b]$ et une suite de K points $z_1 < ... < z_K$ dans $[a, b]$, on appelle spline polynomiale d'ordre $r \geq 1$ ayant pour nœuds simples les points $z_1, ..., z_K$, toute fonction $f : [a, b] \to \mathbb{R}$ telle que :
(i) f est continûment dérivable jusqu'à l'ordre $r - 2$,
(ii) sur chaque sous-intervalle $[a, z_1], ..., [z_i, z_i + 1], ..., [z_K, b]$, f coïncide avec un polynôme de degré $r - 1$.

L'ensemble de ces fonctions sera noté $S_r(z_1, ..., z_K)$.

Définition : On appelle spline polynomiale naturelle d'ordre $2r$ une spline polynomiale d'ordre $2r$ qui coïncide avec un polynôme de degré $r - 1$ en dehors de $[z_1, z_K]$.

L'ensemble de ces fonctions sera noté $NS_r(z_1, ..., z_K)$.

En particulier, une spline cubique naturelle est une fonction continûment différentiable d'ordre 2 telle que la restriction à chaque sous-intervalle de $[z_1, z_K]$ est un polynôme de degré 3 et coïncide avec une fonction linéaire sur $[a, z_1]$ et sur $[z_K, b]$.

Remarque : Les splines polynomiales permettent de faire face aux grandes oscillations présentes dans l'interpolation polynomiale de Lagrange. Le polynôme d'interpolation de Lagrange est le polynôme de degré (minimal) inférieur ou égal à n passant par les $n + 1$ points distincts des données $(x_0; y_0), ..., (x_n; y_n)$. Il s'écrit : $L(x) = \sum_{j=0}^{n} y_j (\prod_{i=0; i \neq j}^{n} \frac{x - x_i}{x_j - x_i})$.

Il serait donc possible d'interpoler nos 38 points via un polynôme de degré inférieur ou égal à 37, cependant l'interpolation des points par un polynôme de degré élevé est délicate et peut mener à des oscillations de grande amplitude. L'idée d'augmenter le nombre de points d'interpolation pour éviter les grandes oscillations n'améliore guère l'approximation globale : c'est le phénomène de Runge.

Paramétrisation d'une spline cubique :

Soit f_i la restriction de f à $[z_i, z_{i+1}]$. Chaque fonction f_i est définie par 4 coefficients et peut s'écrire $f_i(x) = a_i(x - z_i)^3 + b_i(x - z_i)^2 + c_i(x - z_i) + d_i$. Il y a $4(K-1)$ inconnues.

Par définition des splines cubiques naturelles, les contraintes sont les suivantes :
(i) $f_i(z_i) = y_i$ $\quad K-1$ équations
(ii) $f_i(z_{i+1}) = y_{i+1}$ $\quad K-1$ équations
(iii) $f_i'(z_{i+1}) = f_{i+1}'(z_{i+1})$ $\quad K-2$ équations
(iv) $f_i''(z_{i+1}) = f_{i+1}''(z_{i+1})$ $\quad K-2$ équations
(v) $f_1''(z_1) = f_{K-1}''(z_K) = 0$ $\quad 2$ équations

Il y a donc $4(K-1)$ contraintes, le système admet une unique solution.

Une autre paramétrisation de la spline cubique est utilisée en pratique : celle définie par ses valeurs et les valeurs de sa dérivée seconde aux nœuds. En notant $h_i = z_{i+1} - z_i$:

$$f_i(x) = \frac{z_{i+1}-x}{h_i} y_i + \frac{x-z_i}{h_i} y_{i+1} + \frac{(z_{i+1}-x)^3/h_i - h_i(z_{i+1}-x)}{6} f_i''(z_i) + \frac{(x-z_i)^3/h_i - h_i(x-z_i)}{6} f_i''(z_{i+1})$$

Théorème (Caractérisation des splines cubiques naturelles) :

Soient $g = (f(z_1), ..., f(z_K)) \in \mathbb{R}^K$; $\gamma = (f''(z_1), ..., f''(z_K)) \in \mathbb{R}^{K-2}$.
Un couple (g, γ) définit une spline cubique naturelle si et seulement si $Q'g = R\gamma$.

Q définie par $q_{j-1,j} = \frac{1}{h_{j-1}}; q_{j+1,j} = \frac{1}{h_j}; q_{j,j} = \frac{1}{h_{j-1}} - \frac{1}{h_j}$ *et* $q_{i,j} = 0$ *si* $|i-j| > 1$.
R définie par $r_{i,i} = \frac{h_{i-1}+h_i}{3}; r_{i,i+1} = r_{i+1,i} = \frac{h_i}{6}$ *et* $r_{i,j} = 0$ *si* $|i-j| > 1$.
De plus, $|r_{i,i}| > \sum_{i \neq j} r_{i,j}$.
R est une matrice symétrique tridiagonale et à diagonale dominante donc inversible.

Enfin, sous cette condition : $\int_a^b f''(t)^2 dt = \gamma' R \gamma = g' Q R^{-1} Q' g$.

Théorème : L'ensemble $S_r(z_1, ..., z_K)$ est un sous-espace vectoriel de l'espace des fonctions dérivables jusqu'à l'ordre $r-2$ sur $[a, b]$ de dimension $r + K$.

De plus, l'ensemble $NS_r(z_1, ..., z_K)$ est un espace vectoriel de dimension K, indépendant de l'ordre de la spline.

Par conséquent, toute spline de $S_r(z_1, ..., z_K)$ peut s'écrire $f(t) = \sum_{i=1}^{r+K} \beta_i N_{i,r}(t)$.
Deux bases sont couramment utilisées, la base des puissances tronquées et la base B-splines.

Base des puissances tronquées :

$N_{i,r}(t) = t^{i-1} \quad i = 1, ..., r$
$N_{r+j,r}(t) = (t - z_j)_+^{r-1} \quad j = 1, ..., K$

Cette base est simple et explicite, néanmoins c'est la base B-splines qui est le plus souvent retenue pour des raisons numériques.

Base B-splines :

$N_{i,1}(t) = \mathbb{1}_{t_i \leq t < t_{i+1}}$
$N_{i,j}(t) = \frac{t - t_i}{t_{i+j-1} - t_i} N_{i,j-1}(t) + \frac{t_{i+j} - t}{t_{i+j} - t_{i+1}} N_{i+1,j-1}(t) \quad i = 1, ..., m - j = K + j = n \quad j \geq 1$

Avec la suite augmentée de nœuds uniformément espacés :

$$\begin{cases} t_1 = ... = t_r = a \\ t_{r+j} = z_j \quad j = 1, ..., K \\ b = t_{r+K+1} = ... = t_{K+2r} = t_m \end{cases}$$

Une fonction de base d'ordre j est ainsi définie par récurrence à partir de 2 fonctions de base d'ordre $j - 1$.

Propriétés :
(i) $N_{i,j}$ est un polynôme de degré $j - 1$ sur $[t_i, t_{i+j}[$ et nul en dehors.
(ii) $0 < N_{i,j}(t) \leq 1 \ \forall t \in]t_i, t_{i+j}[$
(iii) $\sum_{i=1}^{n} N_{i,j} = 1$

5.3.2 Splines de lissage

Nous considérons le modèle de régression non paramétrique suivant : $y_i = f(x_i) + \epsilon_i$.

Aucune hypothèse n'est faite sur la forme de f, les ϵ_i sont supposés décorrélés, centrés et de variance constante. Le critère des moindres carrés s'écrit : $min_f \sum_{i=1}^{n}(y_i - f(x_i))^2$.

Toute fonction interpolant les données comme le polynôme d'interpolation de Lagrange ou les splines polynomiales est solution de ce problème. Or, nous cherchons une tendance générale plutôt qu'une représentation des variations locales, l'interpolation n'est donc pas adaptée.

Une première façon d'y remédier est de restreindre l'espace des fonctions à un espace de dimension finie, on peut par exemple imposer à f d'appartenir à l'espace des splines d'ordre r de nœuds $(z_1, ..., z_K)$, l'estimateur est dans ce cas appelé spline de régression des moindres carrés. Cela devient un modèle de régression paramétrique classique.

Une seconde approche, appelée régularisation, consiste à ajouter un terme de pénalité de sorte que le critère devient : $min_{f \in W^m[a;b]} \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)})^2 dx$.

On note $\mathrm{W}^m[a;b] = \{f : f^{(j)}$ absolument continues $j = 0, ..., m-1$ et $f^{(m)} \in L_2[a,b]\}$.

L'estimateur est dans ce cas appelé une fonction de lissage et la pénalité ajoutée est une mesure de la courbure de la fonction de lissage. Pour $m = 2$ on parle de pénalité de la rugosité. $\lambda > 0$ contrôle le compromis entre la fidélité aux données mesurée par le premier terme et la régularité de la fonction mesurée par le second terme. Lorsque λ tend vers 0 on se ramène au problème initial, lorsque λ tend vers $+\infty$ on retombe sur une régression polynomiale de degré $m-1$ annulant la pénalisation.

Théorème : Etant donnés n points (x_i, y_i) d'abscisses distinctes dans $[a;b]$, un entier $m \leq n$, et un réel $\lambda > 0$, il existe une unique fonction solution du problème :

$$min_{f \in W^m[a;b]} \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx$$

De plus, cette fonction est une spline polynomiale naturelle d'ordre $2m$ ayant pour nœuds

les points $x_1, ..., x_n$.

Démonstration : On se restreindra au cas $m = 2$ dans la suite.

Notons f une spline cubique naturelle et g une autre fonction d'interpolation pour les points (x_i, y_i). Définissons $h(x) = g(x) - f(x)$. Montrons que $\int_a^b (g''(x))^2 dx \geq \int_a^b (f''(x))^2 dx$.

$\int_a^b (g''(x))^2 dx = \int_a^b (h''(x) + f''(x))^2 dx = \int_a^b (f''(x))^2 dx + 2\int_a^b f''(x)h''(x)dx + \int_a^b (h''(x))^2 dx$

En intégrant par parties le deuxième terme on a :

$\int_a^b f''(x)h''(x)dx$
$= f''(b)h'(b) - f''(a)h'(a) - \int_a^b f'''(x)h'(x)dx$
$=$ - $\int_a^b f'''(x)h'(x)dx$ comme $f''(a) = f''(b) = 0$ par définition de la spline cubique naturelle
$=$ - $\sum_{i=1}^{n-1} c_i \int_{x_i}^{x_{i+1}} h'(x)dx$ comme f''' est constante par morceaux
$=$ - $\sum_{i=1}^{n-1} c_i (h(x_{i+1}) - h(x_i))$
$= 0$ car $h(x_i) = 0 \ \forall i$ par définition

D'où $\int_a^b (g''(x))^2 dx = \int_a^b (f''(x))^2 dx + \int_a^b (h''(x))^2 dx \geq \int_a^b (f''(x))^2$.

On vient de montrer que la spline cubique naturelle est la fonction d'interpolation des points (x_i, y_i) la plus lisse qui soit. Supposons qu'une fonction f^* (possiblement différente d'une fonction d'interpolation des points (x_i, y_i)) minimise effectivement le critère. On peut décider d'interpoler tous les points $(x_i, f^*(x_i))$ à l'aide d'une spline cubique naturelle f. Ainsi, f et f^* donneront la même valeur pour le terme de gauche du critère et f aura nécessairement une valeur moindre pour le terme de droite d'après ce qu'on a démontré précédemment. On en déduit que $f^* = f$: la spline de lissage est une spline cubique naturelle interpolant les points $(x_i, f^*(x_i))$.

Par conséquent, contrairement aux splines de régression où l'on suppose que f est une spline, ici la forme spline de f résulte d'un problème d'optimisation. De plus, avec les splines de lissage, le problème du choix de nombre de nœuds et de leur position ne se pose pas puisque les nœuds correspondent aux points d'observation.

Le calcul de la spline de lissage peut se faire de 2 façons :

a) Avec la représentation $g - \gamma$:

$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx$
$= (y-g)'(y-g) + \lambda g' K g$
$= g'(I + \lambda K)g - 2y'g + y'y$

$K = QR^{-1}Q'$ est de taille n x n, symétrique semi-définie positive donc $(I+\lambda K)$ symétrique définie positive et $\hat{g} = (I + \lambda K)^{-1}y$. La matrice chapeau est $S_\lambda = (I + \lambda K)^{-1}$.

L'algorithme numérique de Reinsch présenté en Annexe 2 permet de rapidement calculer g.

b) Dans la base puissances tronquées ou dans la base B-splines :

$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx$
$= (y - N\beta)'(y - N\beta) + \lambda \beta' \Omega \beta$
$= y'y - 2y'N\beta + \beta' N' N \beta + \lambda \beta' \Omega \beta$

N est la matrice n x n de rang n contenant les $N_j(x_i), i = 1, ..., n,\ j = 1, ..., n$ et Ω la matrice symétrique n x n contenant les éléments $\int_a^b N_i''(x) N_j''(x) dx$.
$N'N$ est symétrique définie positive car N de rang plein.

On résout $-2N'y + 2N'N\beta + 2\lambda\Omega\beta = 0 \iff (N'N + \lambda\Omega)\beta = N'y$.

D'où $\hat{\beta} = (N'N + \lambda\Omega)^{-1}N'y$ (forme de la solution d'une régression Ridge). La matrice de lissage est $S_\lambda = N(N'N + \lambda\Omega)^{-1}N'$.

La trace de cette matrice correspond au nombre de degrés de liberté (équivalent non paramétrique du nombre de paramètres à estimer). Elle tend vers 2 quand λ tend vers $+\infty$ synonyme d'un modèle linéaire simple et tend vers n quand λ tend vers 0 synonyme d'interpolation de tous les points observés.

Le lissage obtenu par la méthode des splines de lissage cubiques avec $\lambda = 0,05$ est le suivant :

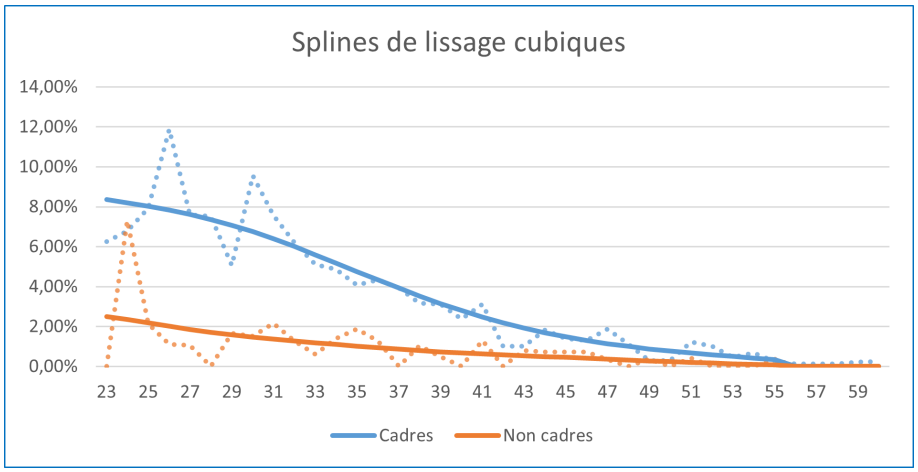

FIGURE 19 – Taux lissés selon la méthode des splines de lissage

5.3.3 P-splines

La méthode des splines de régression consiste en la résolution du problème suivant :

$$min_{f \in S_r(z_1,\ldots,z_K)} \sum_{i=1}^{K} (y_i - f(z_i))^2$$

Elle nécessite dans un premier temps une sélection du nombre de nœuds et de leur position, puis une fois que la suite de nœuds est fixée, on estime les coefficients de la spline par minimisation des moindres carrés. Cependant, le fait de n'ajouter aucune contrainte sur les coefficients de la spline entraîne un surlissage des données (overfitting) et donc l'obtention d'une courbe non lisse dès que le nombre de nœuds est important.

Au contraire, avec la méthode des splines de lissage vue précédemment, on place un nœud en chaque point d'observation et on contrôle le phénomène de surlissage par l'ajout d'une pénalité de la rugosité. Toutefois, les splines de lissages peuvent présenter des difficultés numériques dès que le nombre d'observations est trop grand. En effet un nœud est placé en chaque point d'observation ce qui rend la matrice de lissage de taille n x n.

La méthode des splines pénalisées ou P-splines, développée par Eilers et Marx en 1996, combine les deux approches. Elle consiste à choisir un nombre de nœuds K relativement grand mais inférieur au nombre n d'observations, pas forcément équirépartis, et de contraindre leur influence par un terme de pénalité plus général que le terme de pénalité de la rugosité.

On considère des splines d'ordre r de nœuds $(z_1, ..., z_K)$, ainsi $f(t) = \sum_{j=1}^{r+K} \beta_j N_{j,r}(t)$.

Le critère des moindres carrés pondérés avec pénalisation par les différences finies d'ordre m s'écrit :

$$min_{f \in S_r(z_1,...,z_K)} \sum_{i=1}^{K} w_i (y_i - f(z_i))^2 + \lambda \sum_{j=1}^{r+K-m} (\Delta^m \beta_j)^2$$

En notant W la matrice des poids et D la matrice des différences finies, on peut montrer que $\hat{\beta} = (N'WN + \lambda D'D)^{-1} N'W'y$ contre $\hat{\beta} = (N'N)^{-1} N'y$ pour une régression splines.

Le lissage non paramétrique de Whittaker-Henderson se révèle être un cas particulier du lissage paramétrique P-splines dans le cas où on utilise une base B-splines d'ordre 1 avec des nœuds équirépartis, la solution étant : $\hat{\beta} = (W + \lambda D'D)^{-1} W'y$.

La méthode P-splines est flexible et représente un bon compromis entre splines de régression et splines de lissage. Aussi, la taille des matrices à inverser peut être largement réduite grâce au choix réduit du nombre de nœuds. Néanmoins, contrairement aux précédentes méthodes, il s'agit d'un modèle paramétrique qui n'est donc que présenté théoriquement. La régression a quoi qu'il en soit donné des résultats très proches de ceux obtenus par la méthode des splines de lissage.

5.4 Choix du paramètre de lissage

Ces lissages non paramétriques ont l'avantage de permettre d'arbitrer sur leur forme via les paramètres du modèle. Toutefois, cela constitue aussi un inconvénient, car bien que des techniques de choix de paramètres existent, il n'est pas toujours aisé de les choisir au mieux pour décrire la tendance des démissions tout en satisfaisant les contraintes de décroissance de la courbe de turnover.

Deux critères sont souvent utilisés pour estimer la valeur optimale du paramètre de lissage λ dans la méthode de Whittaker-Henderson et la méthode des splines de lissage : le critère de validation croisée et le critère de validation croisée généralisé.

1) Le critère de validation croisée consiste à minimiser :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2$$

$\hat{f}^{-i}$ est l'estimateur de f construit sans la ième observation.

Pour un estimateur linéaire, il peut s'écrire :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}(\lambda)})^2$$

Les $S_{ii}(\lambda)$ sont les éléments diagonaux de la matrice chapeau $S(\lambda)$. Cette écriture permet d'éviter de résoudre, pour chaque valeur du paramètre de lissage λ, n problèmes de lissage.

2) Le critère de validation croisée généralisée consiste à minimiser :

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{f}(x_i)}{1 - n^{-1}tr(S(\lambda))})^2$$

Ce deuxième critère attribue le même poids à chaque observation. C'est ce critère qui a été utilisé pour sélectionner le paramètre de lissage adéquat dans la méthode de Whittaker-Henderson en dimension un et la méthode des splines de lissage.

5.5 Lissage retenu

En conclusion, hormis le lissage par moyennes mobiles qui n'est pas à privilégier, il n'y a pas de différence majeure sur la forme des différents lissages réalisés. D'ailleurs cela sera constaté au moment de la valorisation des engagements sociaux selon les différentes tables de turnover. Le lissage de Whittaker-Henderson en dimension deux a l'avantage d'être plus poussé et plus fin, mais il n'intègre pas la dimension catégorie socio-professionnelle et paraît difficile à mettre en œuvre dans la durée.

La méthode de lissage non paramétrique retenue est celle des splines de lissage (cf. Annexe 4). Elle donne des résultats proches de la méthode de Whittaker-Henderson en dimension 1. Elle est simple à mettre en œuvre, le paramètre de lissage optimal est estimé par une méthode de validation croisée et bien que découlant d'un problème non paramétrique, on connaît une paramétrisation de la forme du lissage.

L'intervalle de confiance asymptotique présenté dans la partie 4.3 sera utilisé au moment des calculs pour mieux appréhender la sensibilité de l'engagement au turnover. Avec un niveau de confiance à 95%, il donne les résultats suivants sur la méthode des splines de lissage :

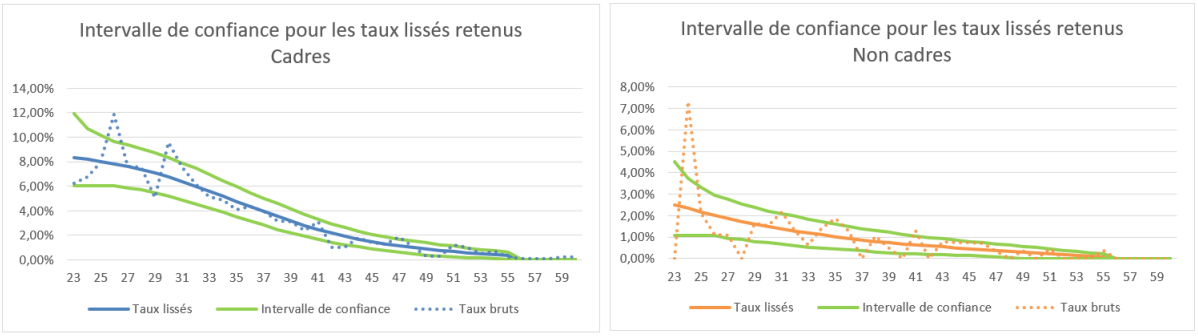

FIGURE 20 – Intervalles de confiance pour les taux lissés retenus

6 Modèles de prédiction

Dans cette partie, nous allons présenter le cadre théorique de 4 méthodes d'apprentissage supervisé qui seront utilisées pour la prédiction de la variable binaire démission. Nous parlerons des modèles linéaires et additifs généralisés, de l'arbre de décision et des forêts aléatoires. Le modèle linéaire généralisé est paramétrique, le modèle additif généralisé est semi-paramétrique, l'arbre de décision et les forêts aléatoires sont non paramétriques.

6.1 Cadre de la classification binaire

La classification binaire consiste à classer les éléments d'un ensemble en deux groupes sur la base d'une règle de classification. On note Y la variable démission à valeurs dans $\{0; 1\}$.

On considère le classifieur $t : \mathcal{X} \longrightarrow \mathcal{Y}$ dont le risque est défini par $R(t) = \mathbb{E}(\mathbb{1}_{t(X)\neq Y})$.

L'estimateur de Bayes pour une perte binaire est défini par $t^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)>1/2}$.

On peut montrer que pour tout classifieur t, $R(t^*) \leq R(t)$.

6.2 Modèles linéaires et additifs généralisés

La théorie des modèles linéaires généralisés (GLM) a été développée par Nelder et Wedderburn en 1972 et celle des modèles additifs généralisés (GAM) par Hastie et Tibshirani en 1986.

Un modèle additif généralisé est un modèle linéaire généralisé dans lequel les prédicteurs linéaires $\sum_{j=1}^{l} \beta_j x_{i,j}$ sont remplacés par des fonctions des variables explicatives $\sum_{j=1}^{l} f_j(x_{i,j})$. Parmi ces fonctions, on trouve des fonctions linéaires et des fonctions lisses parmi lesquelles les splines cubiques naturelles et les P-splines présentées précédemment.

Ils vérifient : i) $Y|X = x$ de loi appartenant à la famille exponentielle
ii) $g(E[Y_i|X_i = x_i]) = g(\mu_i) = \eta_i = \beta_0 + f_1(x_{i,1}) + ... + f_l(x_{i,l})$ $\quad g$ fonction de lien bijective

Pour rappel, en notant θ le paramètre canonique et ϕ le paramètre de dispersion, la famille exponentielle contient les lois admettant une densité par rapport à une mesure dominante

de la forme :

$$f_{\theta,\phi}(y) = c_\phi(y) exp(\frac{y\theta - a(\theta)}{\phi})$$

.
En outre, $E(Y) = a'(\theta)$ et $Var(Y) = \phi a''(\theta)$.

Parmi les lois appartenant à cette famille, on trouve la loi binomiale, la loi binomiale négative, la loi de Poisson, la loi Gamma, la loi normale et la loi inverse-normale.

6.2.1 Régression logistique

Dans le cas d'un GLM, le vecteur β des p paramètres est estimé en maximisant la log-vraisemblance : $l(\beta) = \sum_{i=1}^n ln(f(y_i; \beta, \phi)) = \sum_{i=1}^n ln(c_\phi(y_i)) + (\frac{y_i\theta_i - a(\theta_i)}{\phi}) = \sum_{i=1}^n l_i(\theta_i)$.

La maximisation implique la résolution du système des p équations du score :

$$\frac{\partial l(\beta)}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \frac{x_{ij}(y_i - \mu_i)}{g'(\mu_i)a''(\theta_i)} = 0 \quad j = 1, ..., p$$

Matriciellement, le système s'écrit $X'D(y - g^{-1}(X\beta)) = 0$ avec D la matrice diagonale ayant pour ième terme diagonal $\frac{1}{\phi g'(\mu_i)a''(\theta_i)}$.

Si g est la fonction de lien canonique, les équations s'écrivent $\sum_{i=1}^n x_{ij}(y_i - \mu_i) = 0$, relations qui traduisent l'orthogonalité entre les variables explicatives et les résidus.

Démonstration : En appliquant la règle de dérivation en chaîne, on a :

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i(\beta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \quad \text{avec} \quad \frac{\partial l_i(\beta)}{\partial \theta_i} = \frac{y_i - a'(\theta_i)}{\phi}\,; \quad \frac{\partial \mu_i}{\partial \theta_i} = a''(\theta_i)\,; \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Ce système d'équations étant en général non linéaire, β est estimé numériquement grâce à l'**algorithme de Newton-Raphson** ou sa variante l'**algorithme du score de Fisher** :

1) Initialisation : $\beta^{(0)} = \beta_0$
2) Newton-Raphson : $\beta^{(m+1)} = \beta^{(m)} - (\nabla^2 l(\beta^{(m)}))^{-1} \nabla l(\beta^{(m)})$
Fisher : $\beta^{(m+1)} = \beta^{(m)} - [E(\nabla^2 l(\beta^{(m)}))]^{-1} \nabla l(\beta^{(m)}) = \beta^{(m)} + [E(\nabla l(\beta^{(m)}) \nabla l(\beta^{(m)})')]^{-1} \nabla l(\beta^{(m)})$
3) Stabilisation : dès que $|\beta^{(m_*+1)} - \beta^{(m_*)}| < \epsilon$ pour un epsilon choisi, $\hat{\beta} = \beta^{(m_*)}$.

Proposition :

La hessienne $\nabla^2 l(\beta)$ a pour terme général :

$$\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_k} = \frac{1}{\phi}\sum_{i=1}^{n}[\frac{-1}{g'(\mu_i)^2 a''(\theta_i)} + (y_i-\mu_i)\frac{\partial^2\theta_i}{\partial\eta_i^2}]x_{ij}x_{ik}$$

La matrice d'information de Fisher $\mathcal{I}_n(\beta)$ a pour terme général :

$$[\mathcal{I}_n(\beta)]_{jk} = -E(\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_k}) = \frac{1}{\phi}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{g'(\mu_i)^2 a''(\theta_i)} \quad j,k = 1,...,p$$

i.e. $\mathcal{I}_n(\beta) = X'W_\beta X$ avec W_β la matrice diagonale ayant pour ième terme $\frac{1}{\phi g'(\mu_i)^2 a''(\theta_i)}$.

De plus, si g est la fonction de lien canonique, $\eta_i = \theta_i$ et les méthodes de Newton-Raphson et de Fisher coïncident.

Démonstration : On a que $\frac{\partial^2 l(\beta)}{\partial\beta_j\partial\beta_k} = \sum_{i=1}^{n}\frac{\partial^2 l_i(\beta)}{\partial\eta_i^2}x_{ij}x_{ik}$ avec

$\frac{\partial^2 l_i(\beta)}{\partial\eta_i^2} = \frac{\partial}{\partial\eta_i}(\frac{\partial l_i(\beta)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i}) = \frac{\partial^2 l_i(\beta)}{\partial\eta_i\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i} + \frac{\partial l_i(\beta)}{\partial\theta_i}\frac{\partial^2\theta_i}{\partial\eta_i^2} = \frac{\partial^2 l_i(\beta)}{\partial\theta_i^2}(\frac{\partial\theta_i}{\partial\eta_i})^2 + \frac{\partial l_i(\beta)}{\partial\theta_i}\frac{\partial^2\theta_i}{\partial\eta_i^2}$
$= \frac{-a''(\theta_i)}{\phi}(\frac{\partial\theta_i}{\partial\mu_i})^2(\frac{\partial\mu_i}{\partial\eta_i})^2 + \frac{(y_i-\mu_i)}{\phi}\frac{\partial^2\theta_i}{\partial\eta_i^2} = \frac{-1}{\phi a''(\theta_i)}(\frac{\partial\mu_i}{\partial\eta_i})^2 + \frac{(y_i-\mu_i)}{\phi}\frac{\partial^2\theta_i}{\partial\eta_i^2}$

En retenant la méthode de Fisher : $\beta^{(m+1)} = \beta^{(m)} + (X'W_{\beta^{(m)}}X)^{-1}X'D(y - g^{-1}(X\beta^{(m)}))$.

Théorème : Sous les hypothèses β ouvert convexe et $g \in C^2$, on a :

1) $\hat\beta$ existe et est consistant
2) $\hat\beta$ asymptotiquement normal : $\mathcal{I}_n(\beta)^{1/2}(\hat\beta - \beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_p(0, I_p)$
3) Par ailleurs, $(\hat\beta - \beta)'\mathcal{I}_n(\beta)(\hat\beta - \beta) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(p)$

On peut alors construire des régions de confiance asymptotiques et des tests en remplaçant $\mathcal{I}_n(\beta)$ par $\mathcal{I}_n(\hat\beta)$.

Nous allons appliquer ce modèle à la prédiction du turnover, un problème de classification binaire (démission dans l'année ou non), à partir des caractéristiques dont on dispose. Nous considérons donc la régression logistique binaire pour laquelle la loi de $Y|X = x$ est la loi de Bernoulli et la fonction de lien canonique est le logit.

Le modèle s'écrit : $ln(\frac{\mathbb{P}(Y_i=1|X_i=x_i)}{1-\mathbb{P}(Y_i=1|X_i=x_i)}) = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij}$.

L'équation $\beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} = 0$ définit un hyperplan de direction orthogonale le vecteur $\beta = (\beta_1, ..., \beta_l)'$. La règle de classification s'écrit : $P[Y_i = 1|X_i = x_i] > 1/2$ si et seulement si $\beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} > 0$.

La vraisemblance est donnée par : $\prod_{i=1}^{n} (\frac{exp(<\beta,x_i>)}{1+exp(<\beta,x_i>)})^{y_i} (\frac{1}{1+exp(<\beta,x_i>)})^{1-y_i}$.

La log-vraisemblance est donnée par : $l(\beta) = \sum_{i=1}^{n} [y_i < \beta, x_i > -ln(1 + exp(< \beta, x_i >))]$.

Le gradient est le suivant : $\nabla l(\beta) = \sum_{i=1}^{n} [y_i - \frac{exp(<\beta,x_i>)}{1+exp(<\beta,x_i>)}] x_i$.

La matrice d'information de Fisher s'écrit : $\mathcal{I}_n(\beta) = -H_n(\beta) = \sum_{i=1}^{n} \frac{exp(<\beta,x_i>)}{(1+exp(<\beta,x_i>))^2} x_i x_i'$.

L'ellipse de confiance de niveau $1 - \alpha$ est donnée par : $\mathcal{E}_{n,\alpha} = \hat{\beta} + \mathcal{I}_n(\hat{\beta})^{-1/2} B(0, r_{1-\alpha})$ avec $B(0, r_{1-\alpha})$ la boule de confiance de niveau $1-\alpha$ pour une normale multivariée $\mathcal{N}_p(0, I_p)$.

Pour tester la significativité d'une variable explicative, on peut pénaliser la log vraisemblance avec une pénalité de type Lasso, Ridge ou Elastic Net ou bien réaliser un test de nullité de coefficient à l'aide du résultat de normalité asymptotique.

Contrairement au modèle de régression linéaire, le modèle de régression logistique n'admet pas de solution analytique, la solution est donc obtenue à l'aide d'algorithmes numériques comme ceux évoqués précédemment.

6.2.2 Régression logistique additive

Dans le cas d'un GAM avec splines de lissage, le vecteur β des paramètres des splines dans leur base est estimé en maximisant la log-vraisemblance pénalisée pour éviter le surajustement :

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_{j=1}^{l} \lambda_j \int (f_j''(x_j))^2 dx_j$$

On impose la contrainte d'identifiabilité $\sum_{i=1}^{n} f_j(x_{ij}) = 0 \quad \forall j$ et la résolution du problème se fait par la méthode P-IRLS (Penalized Iteratively Re-weighted Least Squares) développée par Wood en 2006.

Les modèles additifs généralisés ont l'avantage d'offrir beaucoup de flexibilité dans la représentation du lien entre prédicteurs et variable réponse étant donné qu'on ne fait pas d'hypothèse forte sur la forme des prédicteurs, en particulier ils permettent de modéliser des effets non linéaires et vont donc plus loin que les modèles linéaires généralisés.

Aussi, le fait que la variable réponse soit une fonction additive des variables explicatives permet d'isoler l'effet de chacun des prédicteurs. Cela rend les GAM plus facilement interprétables que d'autres modèles encore plus souples comme les forêts aléatoires. Néanmoins, comme la spline n'a pas d'équation paramétrique simple, seule une visualisation de la fonction permet de comprendre les effets estimés.

Une extension intéressante au modèle GAM consiste en l'utilisation de fonctions de lissages bivariées (via une base de lissage par produit tensoriel) tenant compte des interactions entre les différentes variables.

6.3 Arbre de décision et forêts aléatoires

6.3.1 Arbre de décision

L'arbre de décision est couramment utilisé en machine learning pour sa simplicité et pour son caractère facilement interprétable. Il s'agit d'un classifieur $t : \mathcal{X} \longrightarrow \mathcal{Y}$. On distingue les arbres de classification ($\mathcal{Y}$ fini) des arbres de régression ($\mathcal{Y}$ infini). Un arbre de décision est une représentation visuelle d'un algorithme de classification de données suivant différents critères : décisions ou nœuds. Mathématiquement, il s'agit d'un partitionnement de $\mathcal{X} \subset \mathbb{R}^d$ en hyperrectangles.

Voici un exemple d'arbre de décision qui pourrait être appliqué aux données.

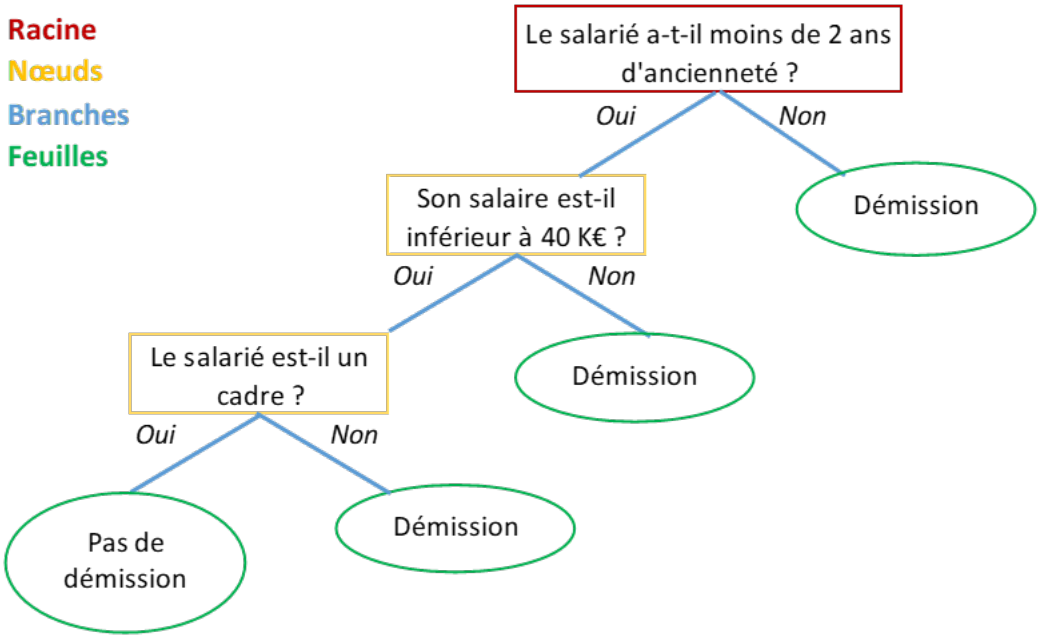

FIGURE 21 – Exemple d'arbre de décision

La racine contient l'ensemble de la population à segmenter, c'est le point de départ.
Les branches contiennent les règles de division qui permettent de segmenter la population.
Les nœuds contiennent les sous-populations homogènes créées et fournissent l'estimation de la quantité d'intérêt.

Les feuilles ou nœuds purs indiquent la classe résultante.

On se place au nœud racine et on considère un ensemble de questions possibles à poser. Chaque question est posée à l'ensemble des individus de la base et ne concerne qu'une variable. Chaque question posée doit être la plus discriminante possible, c'est-à-dire qu'elle doit permettre de séparer les données d'apprentissage en deux classes les plus homogènes possibles.

L'algorithme CART (Classification And Regression Trees) a été développé par Breiman, Friedman, Olshen et Stone en 1984. Nous allons l'utiliser pour générer un arbre de décision binaire (un nœud a exactement zéro ou deux fils) ayant pour critère de segmentation l'indice de Gini ou l'entropie, des mesures d'impureté.

On note $p_i(h)$ la proportion de la classe i parmi la sous-population associée au nœud h.

L'indice de Gini d'un nœud h est la quantité :

$$G(h) = \sum_{i=1}^{n} p_i(h)(1 - p_i(h))$$

En classification binaire : $G(h) = 1 - p_0(h)^2 - p_1(h)^2$

L'entropie d'un nœud h se définit ainsi :

$$H(h) = -\sum_{i=1}^{n} p_i(h) log_2(p_i(h))$$

L'erreur de classification associée à un nœud h se définit ainsi :

$$E(h) = 1 - max_i\, p_i(h)$$

Plus l'indice de Gini ou l'entropie sont faibles, meilleur est le nœud : l'une des deux classes prédomine. Dans le cas d'un nœud pur (ne contenant qu'une seule classe), $G(h) = H(h) = 0$. Néanmoins pour éviter le surapprentissage et ne pas obtenir des arbres trop complexes, on fixe des critères d'arrêt afin de ne pas développer l'arbre jusqu'au bout.

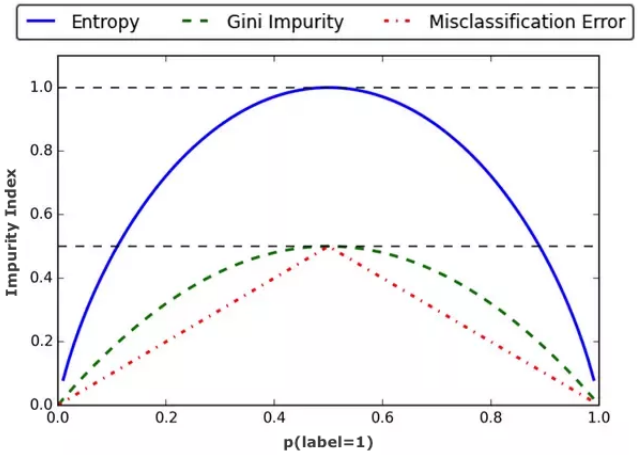

FIGURE 22 – Indices d'impureté

Parmi les conditions d'arrêt les plus courantes, on peut fixer une profondeur d'arbre maximale, un seuil en dessous duquel la population de chaque nœud doit se trouver ou un seuil en dessous duquel l'indice de Gini ou l'entropie doit se trouver.

Algorithme CART : Pour chaque feuille h de l'arbre, les étapes sont les suivantes :

1) Trouver la paire (variable explicative, seuil) qui minimise l'indice de Gini ou l'entropie.
2) Créer deux nouveaux fils.
3) S'arrêter dès lors que le critère d'arrêt est atteint.
4) Sinon continuer.

La performance de l'arbre n'est pas toujours améliorée par la profondeur de l'arbre. Plus l'arbre est petit, plus il est robuste (variance faible), mais le biais sera grand. A l'inverse, plus l'arbre est grand, moins le biais sera important mais plus la variance sera élevée avec une tendance à l'overfitting. En effet un trop grand nombre de branches pourrait conduire à refléter des anomalies dues au bruit ou aux données aberrantes.

Un inconvénient majeur de l'arbre de décision est appelé l'effet papillon : si l'on modifie une variable, tout l'arbre est modifié et si la première variable de segmentation est mal choisie, l'arbre engendré sera un mauvais classifieur. Cela amène à utiliser des techniques plus poussées telles que le Bagging ou le Boosting pour en améliorer les performances.

6.3.2 Forêts aléatoires

Comme son nom l'indique, une forêt aléatoire est constituée de plusieurs arbres de décision. L'algorithme des forêts aléatoires a été formellement proposé en 2001 par Breiman et Cutler. L'algorithme combine deux sources d'aléa : le sous-échantillonnage des variables explicatives et le Bagging (Bootstrap Aggregating).

Le sous-échantillonnage des variables explicatives consiste à retenir aléatoirement un certain nombre de variables explicatives ($m \sim \sqrt{p}$ par défaut) au travers desquelles l'arbre sera construit.

Le Bagging consiste à effectuer plusieurs tirages avec remise de taille n à partir des n observations de l'échantillon initial, à ajuster un classifieur sur chacun des échantillons créés (fractions qui servent à l'entraînement de l'algorithme : "in bag") et à agréger ces différents classifieurs de sorte à en obtenir un seul via un vote à la majorité. Le classifieur peut ainsi prédire une classe pour chacun des individus de la fraction restante ("out of bag").

Cette technique appliquée aux arbres de décisions permet de réduire la corrélation entre les différents arbres et la variance globale du modèle. Le risque de surapprentissage est bien plus faible même en développant jusqu'au bout les arbres, car les réponses individuelles des arbres sont agrégées dans la réponse finale. Il n'est donc pas nécessaire de prévoir un critère d'arrêt.

L'indicateur de performance du modèle est l'erreur OOB (out of bag) qui indique le taux d'erreur des arbres sur les individus laissés "out of bag" par le modèle. Plus le nombre d'arbres de la forêt est grand, plus le modèle sera précis mais plus il mettra de temps à s'entraîner. Une façon de trouver le meilleur compromis entre vitesse et performance est de considérer le nombre d'arbres à partir duquel l'erreur OOB se stabilise.

6.4 Rééchantillonnage en présence de données déséquilibrées

En cas de données déséquilibrées comme c'est le cas dans notre jeu de données avec environ 2% de démissions, il peut être pertinent d'utiliser un rééchantillonnage de l'évènement rare. Deux méthodes sont généralement privilégiées : le sous-échantillonnage (undersampling) et le sur-échantillonnage (oversampling).

Les méthodes d'undersampling fonctionnent en diminuant le nombre d'observations de la classe majoritaire afin d'arriver à un ratio classe minoritaire / classe majoritaire satisfaisant (pas forcément égal à 1). Les méthodes d'oversampling fonctionnent en augmentant le nombre d'observations de la classe minoritaire afin d'arriver à un ratio classe minoritaire / classe majoritaire satisfaisant.

Les méthodes de sous-échantillonnage sont à privilégier sur les très grands jeux de données pour ne pas supprimer d'informations importantes sur les plus petites bases.

Nous nous concentrerons sur une méthode de sur-échantillonnage des observations minoritaires avec l'**algorithme SMOTE (Synthetic Minority Oversampling Technique)**. Ce rééchantillonnage est réalisé de manière réfléchie en utilisant la méthode des k plus proches voisins et le bootstrapping.

Les étapes de l'algorithme sont les suivantes :
1) Sélectionner aléatoirement une observation minoritaire initiale.
2) Identifier ses k plus proches voisins parmi les observations minoritaires (k est un paramètre à définir). La distance utilisée est la distance euclidienne.
3) Choisir aléatoirement l'un des k plus proches voisins.
4) Générer aléatoirement un coefficient $0 < \alpha < 1$.
5) Créer un nouvel individu entre l'observation initiale et le plus proche voisin choisi, selon la valeur du coefficient. Par exemple, si $\alpha = 0.5$, le nouvel individu sera positionné à mi-chemin entre l'observation initiale et le plus proche voisin choisi.

Ces étapes sont répétées jusqu'à ce que le nombre d'individus générés atteigne une valeur définie par l'utilisateur. Un retraitement est enfin nécessaire avant d'appliquer le modèle dès lors que les variables sont discrètes.

6.5 Métriques d'évaluation de modèle

Dans le cadre d'une classification, plusieurs métriques existent pour comparer et évaluer des modèles à partir de la matrice de confusion : l'exactitude, la sensibilité, la spécificité, la précision, le score F1, le coefficient de corrélation de Matthews (MCC) et l'AUC (Area under the ROC Curve).

Une matrice de confusion mesure la qualité d'un système de classification. Chaque ligne correspond à une classe réelle, chaque colonne correspond à une classe estimée. Un classifieur est d'autant meilleur que sa matrice de confusion "s'approche" d'une matrice diagonale.

Le tableau ci-dessous présente 4 métriques usuelles associées à une matrice de confusion.

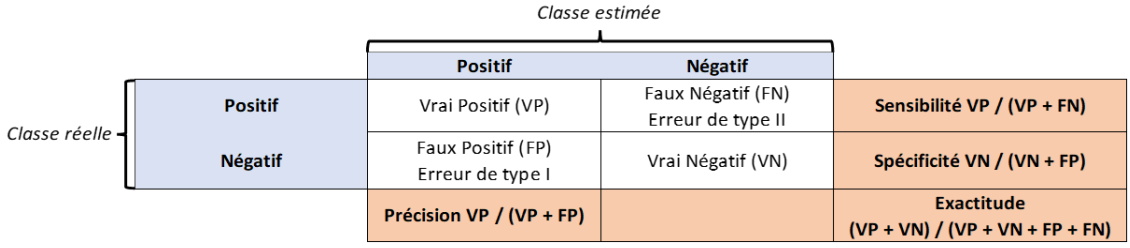

FIGURE 23 – Métriques et matrice de confusion

L'exactitude indique le taux de bonnes prédictions réalisées par le modèle. La sensibilité et la spécificité indiquent la capacité du modèle à prédire correctement tous les positifs et tous les négatifs. La précision indique la capacité du modèle à ne prédire que les positifs.

Le score F1 est une autre métrique régulièrement utilisée. Il s'agit de la moyenne harmonique de la sensibilité et de la précision :

$$F1 = 2 * \frac{Sensibilité * Précision}{Sensibilité + Précision} = \frac{2VP}{2VP + FP + FN}$$

.

Néanmoins en présence de données déséquilibrées, le coefficient de corrélation de Matthews est la métrique la plus pertinente. Elle est d'une part symétrique et ne produit d'autre part un score élevé que si la prédiction a obtenu de bons résultats dans les quatre catégories de la matrice de confusion, proportionnellement à la fois à la taille des éléments positifs et négatifs dans l'ensemble de données. Il est défini par :

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Enfin, la courbe Receiver Operating Characteristic (ROC) représente le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (1 - spécificité) sur un continuum de seuils s. Plutôt que la règle de classification $t^*(x) = \mathbb{1}_{\mathbb{P}(Y=1|X=x)>1/2}$ évoquée en introduction, on considère le score $S : \mathcal{X} \longrightarrow \mathbb{R}$ et un seuil s tels que : $t_s(x) = \mathbb{1}_{S(x)>s}$.

La courbe ROC d'un score S est la courbe paramétrée suivante :

$$\begin{cases} x(s) &= \mathbb{P}(S(X) \geq s|Y = 0) = 1 - spécificité(s) \\ y(s) &= \mathbb{P}(S(X) \geq s|Y = 1) = sensibilité(s) \end{cases}$$

Un score parfait vérifie $x(s^*) = 0$ et $y(s^*) = 1$. Un score aléatoire vérifie $x(s) = y(s)\ \forall s$.

L'aire sous la courbe ROC associée au score parfait serait de 1. A l'inverse pour un score aléatoire, l'AUC serait de 1/2. Plus le score est élevé, plus on s'attend à ce que la proportion de cas bien classés soit grande par rapport à la proportion de cas mal classés.

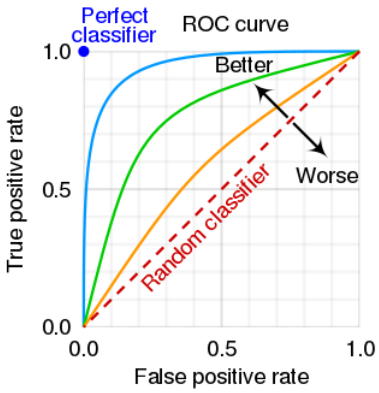

FIGURE 24 – ROC pour différents scores

Théorème : Soit $S^*(x) = \mathbb{P}(Y = 1|X = x)$. Alors $AUC(S^*) \geq AUC(S)$ pour tout S.

L'AUC a l'avantage d'être insensible aux disparités dans les proportions de classes et d'être indépendante d'un changement de seuil.

7 Choix du lissage et impact sur l'engagement

Les lissages non paramétriques présentés en partie 5 ont été appliqués aux évaluations actuarielles au 31/12/2021 des trois régimes suivants : indemnité de fin de carrière, médailles du travail, régime de retraite additionnel.

Les hypothèses utilisées pour les calculs sont les suivantes :

Hypothèses au 31/12/2021	
Taux d'actualisation	1,00%
Taux de revalorisation des salaires	2,50%
Taux de revalorisation des rentes	1,00%
Charges patronales	45%
Taxation sur les rentes	32%
Rente annuelle servie	5% du salaire annuel à la date de liquidation
Table de mortalité	TH/TF00-02 décalée en phase d'activité TGH/TGF05 en phase de retraite
Table de turnover	Variable
Table des âges de départ à la retraite	Fonction de l'âge et de la CSP
Convention collective	Métallurgie

FIGURE 25 – Table des hypothèses utilisées pour les calculs d'engagements

Habituellement, le taux d'actualisation utilisé n'est pas le même pour les régimes médailles du travail et IFC, la duration des engagements étant plus courte pour un régime médailles du travail.

Le taux de revalorisation des rentes a été pris égal au taux d'augmentation 2021 de la valeur du point AGIRC-ARRCO. Le taux technique est ainsi égal à 0% comme c'était le cas au 31/12/2021.

Par ailleurs, la CCN Métallurgie étant concernée par l'impact IFRIC (cf. Annexe 3), celui-ci est calculé pour le régime IFC.

Tous les résultats des parties suivantes 7.1 à 7.3 sont exprimés en M€.

7.1 Résultats IFC

Les résultats IFC hors impact IFRIC tirés d'une macro VBA sont les suivants :

Méthode	DBO 2021	NC 2022	IC 2022	EBP 2022	DBO 2022 projetée
Table de référence	*743*	*38*	*7*	*29*	*760*
Moyenne mobile	733	37	7	29	748
Whittaker-Henderson 1	736	38	7	29	752
Whittaker-Henderson 2	746	38	7	29	763
Splines de lissage	736	37	7	29	751

FIGURE 26 – Tableau des résultats selon les différents lissages

Les résultats IFC y compris impact IFRIC sont les suivants :

Méthode	DBO 2021	NC 2022	IC 2022	EBP 2022	DBO 2022 projetée
Table de référence	*684*	*38*	*7*	*29*	*700*
Moyenne mobile	676	37	7	29	691
Whittaker-Henderson 1	678	37	7	29	693
Whittaker-Henderson 2	687	38	7	29	703
Splines de lissage	**678**	**37**	**7**	**29**	**693**
Borne supérieure	***692***	***39***	***7***	***29***	***708***
Borne inférieure	***666***	***36***	***7***	***29***	***679***

FIGURE 27 – Tableau des résultats selon les différents lissages

Comme indiqué, le poids du régime IFC dans le passif social d'une grande entreprise est considérable.

Les lissages en une dimension réalisés tendent à diminuer le montant de l'engagement

calculé à partir de la table de référence. L'impact est d'environ 1%, ce qui représente néanmoins plusieurs millions d'euros sur des sommes aussi importantes. L'impact sur le coût normal et le coût de la désactualisation est négligeable et il est nul sur les EBP comme le taux de turnover est fixé à 0% après 55 ans pour toutes les tables.

Le lissage en deux dimensions, de maille plus fine bien que ne tenant pas compte de la CSP, engendre une hausse de l'engagement par rapport à la table de référence. Il attribue en effet des taux de démission plus faibles à ceux qui ont une ancienneté plus élevée et donc pour qui la prestation à payer sera plus importante.

Néanmoins, ces lissages n'impactent pas de façon majeure l'engagement comme ça peut être le cas pour un changement de méthode de valorisation tel l'IFRIC 2021. La nouvelle méthode engendre une baisse de 8% de la DBO soit environ 60 M€. Elle n'impacte que très faiblement le coût normal du fait de la compensation entre les individus dont le coût normal passe à 0 et les autres pour lesquels il augmente.

Enfin, des sensibilités de l'engagement à la table de turnover retenue ont été calculées grâce aux intervalles de confiance asymptotiques présentés en partie 5.4. On note que l'intervalle de DBO construit avec un niveau de confiance à 95% encadre toutes les DBO associées aux lissages réalisés précédemment. L'impact sur l'engagement de l'utilisation de la courbe haute ou de la courbe basse plutôt que la courbe centrale est d'environ 2%.

7.2 Résultats Médailles du travail

Les résultats obtenus pour le régime médailles du travail (non impacté par la réforme IFRIC) sont les suivants :

Méthode	DBO 2021	NC 2022	IC 2022	EBP 2022	DBO 2022 projetée
Table de référence	*228*	*14*	*2*	*16*	*229*
Moyenne mobile	223	14	2	16	223
Whittaker-Henderson 1	224	14	2	16	225
Whittaker-Henderson 2	230	14	2	16	231
Splines de lissage	**224**	**14**	**2**	**16**	**224**
Borne supérieure	***232***	***15***	***2***	***16***	***233***
Borne inférieure	***217***	***13***	***2***	***16***	***216***

FIGURE 28 – Tableau des résultats selon les différents lissages

Pour la même population, le poids du régime médailles du travail dans le passif social de l'entreprise est plus faible. De la même façon que pour le régime IFC, les lissages en une dimension tendent à diminuer légèrement le montant de la DBO et le lissage en deux dimensions tend à l'augmenter.

L'impact sur l'engagement de l'utilisation de la courbe haute ou de la courbe basse plutôt que la courbe centrale est d'environ 3%. Là aussi l'intervalle de DBO construit avec un niveau de confiance à 95% encadre toutes les DBO associées aux lissages réalisés.

7.3 Résultats Régime de retraite supplémentaire

Les résultats obtenus pour le régime de retraite supplémentaire sont les suivants :

Méthode	DBO 2021	NC 2022	IC 2022	EBP 2022	DBO 2022 projetée
Table de référence	*786*	*44*	*8*	*3*	*835*
Moyenne mobile	777	43	8	3	826
Whittaker-Henderson 1	780	43	8	3	828
Whittaker-Henderson 2	785	43	8	3	833
Splines de lissage	**780**	**43**	**8**	**3**	**828**
Borne supérieure	***794***	***45***	***8***	***3***	***845***
Borne inférieure	***766***	***41***	***8***	***3***	***813***

FIGURE 29 – Tableau des résultats selon les différents lissages

Bien qu'en général les participants d'un régime de retraite à prestations définies soient moins nombreux que ceux d'un régime IFC et concerne plus particulièrement les cadres et une population plus âgée, ce type de régime a un poids considérable sur le passif social d'une entreprise.

Nous n'avons pas considéré d'option de réversion car nous ne disposions des informations sur les conjoints. Evidemment, intégrer une pension de réversion au calcul augmenterait le montant de l'engagement.

Les sensibilités montrent que l'impact sur l'engagement de l'utilisation de la courbe haute ou de la courbe basse plutôt que la courbe centrale est d'environ 2%.

Les résultats obtenus sur les 3 régimes montrent que les montants d'engagements calculés varient d'au plus 3% selon le choix de la table lissée. L'engagement apparaît donc moins sensible à un changement de méthode de lissage que ce à quoi on pouvait s'attendre mais il convient toujours d'ajuster au mieux cette hypothèse démographique.

8 Résultats des prédictions

Cette partie reprend les modèles présentés précédemment et implémentés sous R pour les appliquer aux données rééchantillonnées et à la prédiction du turnover. L'idée de départ est d'obtenir, pour chaque individu, une indicatrice de présence au moment de la prestation afin de comparer l'engagement obtenu dans ce cas à celui obtenu classiquement avec l'utilisation d'une table de turnover.

L'algorithme SMOTE a été utilisé pour le rééchantillonnage car le taux de démissions de l'année est trop faible. En effet, avant application du sur-échantillonnage, les modèles ne prédisaient aucune démission.

Dans le cadre de l'application de SMOTE, le choix a été fait de multiplier par 20 la taille de la classe minoritaire et d'identifier à chaque étape les 2 plus proches voisins de l'observation minoritaire.

Les modèles que nous considérons ont été entraînés sur 70% des données de la base rééchantillonnée. Sur cette base d'entraînement, une validation croisée avec 10 blocs est utilisée pour l'apprentissage du modèle.

La méthode de validation croisée est résumée ci-dessous :

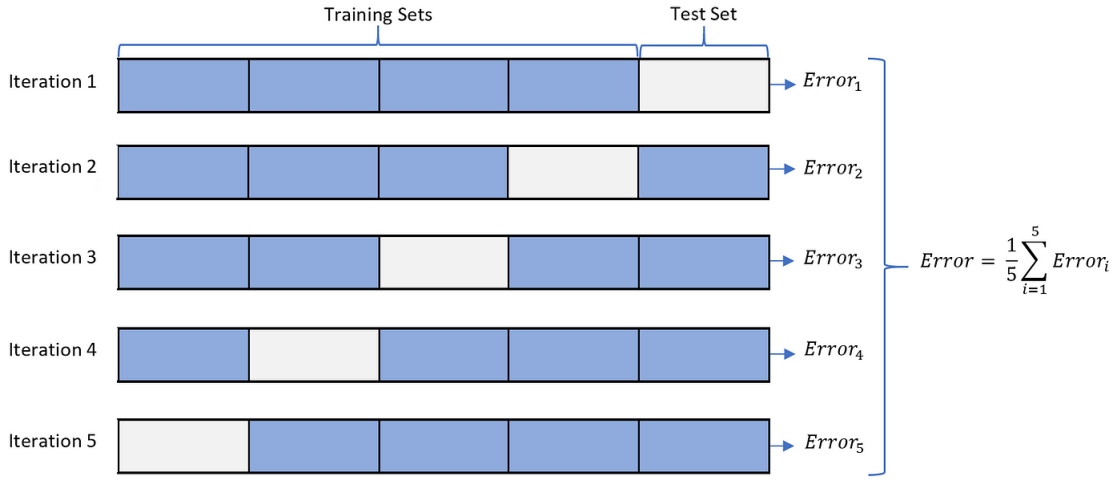

FIGURE 30 – Validation croisée avec 5 blocs

Elle consiste à diviser l'ensemble des données en k ensembles de tailles presque égales. Le premier ensemble est sélectionné comme ensemble de test et le modèle est entraîné sur les $k-1$ autres ensembles. Le taux d'erreur est ensuite calculé après ajustement du modèle aux données de test. Dans la deuxième itération, le deuxième ensemble est sélectionné comme un ensemble de test et les $k-1$ autres ensembles sont utilisés pour l'entraînement et on calcule le taux d'erreur à nouveau. Ce processus se poursuit pour tous les k ensembles. Le score final à attribuer à son modèle peut être retrouvé en moyennant les performances obtenues sur les k itérations.

Les prédictions et performances des modèles sont observées sur le reste de la base, l'échantillon de test, destiné à évaluer les modèles sur des données qu'il ne connaît pas. Les résultats selon les différentes métriques sont présentés ci-dessous :

Modèle	Sensibilité	Spécificité	Précision	Exactitude	Score F1	MCC	AUC
GLM	81 %	61 %	85 %	76 %	83 %	40 %	69 %
GAM	82 %	61 %	84 %	76 %	83 %	43 %	71 %
CART	90 %	95 %	98 %	91 %	94 %	80 %	87 %
Forêts aléatoires	98 %	99 %	100 %	98 %	99 %	96 %	97 %

FIGURE 31 – Comparaison des résultats des différents modèles sur l'échantillon de test

Etant donné le temps de calcul supplémentaire, le GAM n'a pas une grande valeur ajoutée par rapport au GLM sur la performance globale. Cela se voit d'ailleurs clairement sur les courbes ROC obtenues :

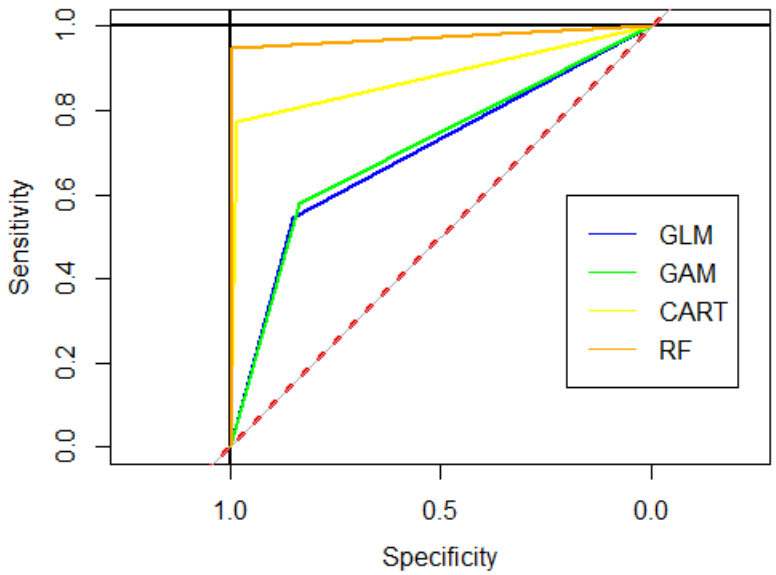

FIGURE 32 – Comparaison des courbes ROC des différents modèles

Sans surprise, l'arbre de décision et les forêts aléatoires, des modèles non paramétriques, sont très performants mais réclament aussi plus de temps de calcul. Pour l'arbre de décision, le critère d'arrêt sélectionné est la profondeur de l'arbre choisie égale à 20. Pour la forêt aléatoire, le nombre de variables explicatives choisi dans le cadre du sous-échantillonnage est pris égal à 2 et le nombre d'arbres a été sélectionné grâce au graphique de l'erreur OOB en fonction du nombre d'arbres : cette erreur se stabilise à 300 arbres.

Par ailleurs, comme on pouvait s'y attendre étant donné la matrice des V de Cramer et la matrice des corrélations linéaires de Pearson, il a été observé en manipulant les modèles que les variables salaire et sexe jouent un rôle moins significatif dans la prédiction des démissions que les variables âge, ancienneté et CSP.

Finalement, l'idée de départ qui était d'obtenir, pour chaque individu, une indicatrice de présence au moment de la prestation n'a pas pu être appliquée que partiellement. En effet, les modèles prédisent sur la base des données d'entraînement qui renseignent sur les démissions dans l'année uniquement. Il serait intéressant d'étendre ce travail à des prédictions de présence au moment de la prestation.

Conclusion

L'étude réalisée nous a permis de mieux appréhender le facteur turnover et la place qu'il occupe dans l'engagement actuariel. Affiner l'hypothèse de turnover revêt une importance particulière car si une table est bien ajustée, les écarts actuariels seront réduits, la volatilité des fonds propres et de la provision d'une année à l'autre également.

Nous avons pu évoquer les faiblesses de la méthode d'ajustement de table actuellement utilisée par Mercer. Cette méthode a l'avantage d'être simple et de tenir compte des démissions sur 3 années mais elle suppose une allure de la courbe de turnover commune à tous les clients et invariante au fil des années.

Les méthodes de lissage non paramétriques appliquées diminuent le biais de modèle, aucune hypothèse n'est faite sur la distribution de la courbe. Elles permettent d'arbitrer sur la forme de la courbe et sur l'importance relative donnée à la fidélité d'une part et à la régularité d'autre part à travers le choix du paramètre de lissage.

La méthode de Whittaker-Henderson en dimension deux se démarque en permettant un lissage conjoint dans les deux directions. Elle a l'avantage de tenir compte de l'ancienneté et donc de la trajectoire de la carrière du salarié au sein de l'entreprise, pourtant elle ne tient pas compte de la CSP et paraît plus difficile à mettre en œuvre sur le long terme. Les méthodes de Whittaker en dimension 1 et des splines de lissage semblent être les méthodes à privilégier.

Les intervalles de confiance asymptotiques calculés permettent de décrire la sensibilité de l'engagement à la table lissée retenue. Il peut être pertinent d'adopter l'habitude de présenter des sensibilités au turnover au même titre que des sensibilités au taux d'actualisation.

Enfin, force est de constater que ces lissages n'impactent pas de façon majeure l'engagement comme ça peut être le cas pour un changement de méthode de valorisation tel l'IFRIC 2021.

Par ailleurs, les modèles de classification apportent un regard innovant sur le turnover. Ils montrent tout d'abord qu'un rééchantillonnage de l'évènement "démission" s'impose.

Le sur-échantillonnage de la classe minoritaire à l'aide de l'algorithme SMOTE a permis d'obtenir des résultats de prédiction pertinents et d'attribuer à tous les individus une probabilité de démission dans l'année sur la base de leurs caractéristiques et de les classer. Au vu des métriques d'évaluation et en particulier des métriques MCC et AUC efficaces en présence de données déséquilibrées, les modèles les plus performants sont l'arbre de décision et les forêts aléatoires.

Aussi, nous avons pu observer que les variables âge, ancienneté et catégorie socio-professionnelle jouent un rôle plus important dans la prédiction du turnover que les variables sexe et salaire. Cela montre que la manière dont ont été discriminés les taux de démissions selon l'âge, l'ancienneté et la CSP dans la partie lissage est appropriée. Comme déjà énoncé, l'ajout d'autres variables non disponibles telles que le statut marital et la zone géographique pourrait encore améliorer la performance des modèles.

Finalement, une limite importante de ces modèles de prédiction vient du fait que la probabilité de présence calculée dans le cadre de l'engagement actuariel concerne une présence au moment de la prestation et non une probabilité de présence dans l'année.

Il serait intéressant d'étendre ce concept de prédiction sur un an à une prédiction sur plusieurs années, peut-être en récoltant des données de suivi de présence sur plusieurs années. Cela permettrait de comparer l'engagement obtenu avec un lissage de table à celui obtenu avec la probabilité de présence au moment de la prestation prédite par le modèle ou encore à celui obtenu avec une indicatrice de présence prédite par la classification binaire. Cela peut faire l'objet de recherches futures.

Bibliographie

[1] Focus IFRS, IAS 19 Avantages du personnel version 2013 (2022).

[2] Jean-Jacques Droesbeke et Gilbert Saporta - Approches non paramétriques en régression (2011). Chapitre 5 par Christine Thomas-Agnan.

[3] C. Andrieu - Thèse : Modélisation fonctionnelle de profils de vitesse en lien avec l'infrastructure et méthodologie de construction d'un profil agrégé (2013).

[4] Lignes directrices mortalité de la Commission d'Agrément (2005).

[5] Frédéric Planchet - Cours Méthodes de lissage et d'ajustement (2022).

[6] Julien Thomas - Algorithme IRWLS avec R.

[7] E. Moulines - Etude de cas régression non paramétrique.

[8] Marion Lainé - Mémoire : Construction de tables prospectives de turnover (2013).

[9] Minh Tu Pham - Mémoire : Construction de tables de turnover par application de l'apporche d'apprentissage automatique dans l'évaluation des IFC en norme IAS 19 (2021).

[10] Adrien Lagouge, Ismaël Ramajo, Victor Barry - Article : "La France vit-elle une "Grande démission" ?" (2022).

[11] Claire Boyer - Cours Machine learning (2021).

[12] Norbert Gautron - Cours Actuariat des retraites (2021).

Annexe 1 : Contributions patronales régimes L137-11-1

Option irrévocable de l'employeur	Assiette	Taux
Taxation sur les rentes	Rentes liquidées à compter du 01/01/2001 jusqu'au 31/12/2012 (au premier euro depuis le 01/01/2011 contre ce qui excédait 1/3 du plafond de la Sécurité sociale auparavant)	16% 8% jusqu'au 31/12/2009
	Rentes liquidées à compter du 01/01/2013	32%
Taxation sur le financement	Primes d'assurance (contributions à l'assureur si le régime est externalisé)	24% à compter du 01/01/2013 12% jusqu'au 31/12/2012 6% jusqu'au 31/12/2009
	Partie de la dotation à la provision ou au montant inscrit en annexe au bilan, correspondant aux services rendus dans l'année (SC)	48% à compter du 01/01/2013 24% jusqu'au 31/12/2012 12% jusqu'au 31/12/2009

FIGURE 33 – Contributions patronales régimes L137-11-1

Annexe 2 : Algorithme de Reinsch

L'algorithme de Reinsch permet de calculer numériquement et rapidement g de la représentation $g - \gamma$ d'une spline de lissage cubique.

En utilisant la caractérisation des splines cubiques naturelles, on a :
$g = y - \lambda K \iff g = y - \lambda Q\gamma \iff R\gamma = Q'y - \lambda Q'Q\gamma \iff (R + \lambda Q'Q)\gamma = Q'y$

Les étapes de l'algorithme sont les suivantes :

1) On calcule $Q'y$.
2) On calcule la décomposition de Cholesky de la matrice $(R + \lambda Q'Q)$.
3) On résout $LDL'\gamma = Q'y$ en γ.
4) On détermine g en utilisant $g = y - \lambda Q\gamma$.

La matrice $(R + \lambda Q'Q)$ est symétrique définie positive de largeur de bande 5 (nombre de diagonales non nulles). Les matrices bandes sont creuses et économiques à stocker.

Annexe 3 : CCN Métallurgie

Ancienneté	Droits (mois)
0	0
1	0
2	0,5
3	0,5
4	0,5
5	1
6	1
7	1
8	1
9	1
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	3
21	3
22	3
23	3
24	3
25	3
26	3
27	3
28	3
29	3
30	4
31	4
32	4
33	4
34	4
35	5
36	5
37	5
38	5
39	5
40 et plus	6

FIGURE 34 – CCN Métallurgie - Départ volontaire à la retraite

Les droits conférés par la CCN Métallurgie fonctionnent par paliers et sont plafonnés à 40 ans d'ancienneté. Le calcul IFC est donc impacté par la réforme IFRIC.

Annexe 4 : Table de turnover retenue

	Taux bruts		*Méthode Mercer*		*Splines de lissage*	
Age	**Cadre**	**Non cadre**	**Cadre**	**Non cadre**	**Cadre**	**Non cadre**
23	6,3%	0,0%	7,7%	1,6%	8,4%	2,5%
24	6,8%	7,3%	7,7%	1,5%	8,2%	2,3%
25	7,9%	2,1%	7,7%	1,4%	8,0%	2,2%
26	11,9%	1,1%	7,2%	1,3%	7,8%	2,0%
27	7,6%	1,1%	6,7%	1,2%	7,6%	1,9%
28	7,5%	0,0%	6,3%	1,2%	7,4%	1,7%
29	5,0%	1,7%	5,8%	1,1%	7,1%	1,6%
30	9,5%	1,5%	5,4%	1,0%	6,7%	1,5%
31	7,6%	2,2%	5,0%	1,0%	6,4%	1,4%
32	6,2%	1,3%	4,6%	0,9%	6,0%	1,3%
33	5,1%	0,6%	4,3%	0,9%	5,6%	1,2%
34	4,9%	1,4%	4,0%	0,8%	5,2%	1,1%
35	4,1%	1,9%	3,7%	0,7%	4,7%	1,0%
36	4,3%	1,3%	3,4%	0,7%	4,3%	0,9%
37	3,9%	0,0%	3,1%	0,6%	3,9%	0,9%
38	3,2%	1,0%	2,8%	0,6%	3,5%	0,8%
39	3,1%	0,5%	2,6%	0,5%	3,2%	0,7%
40	2,4%	0,0%	2,3%	0,5%	2,8%	0,7%
41	3,1%	1,3%	2,1%	0,5%	2,5%	0,6%
42	1,0%	0,0%	1,9%	0,4%	2,2%	0,6%
43	1,0%	0,8%	1,7%	0,4%	1,9%	0,5%
44	1,8%	0,7%	1,5%	0,3%	1,7%	0,5%
45	1,4%	0,7%	1,4%	0,3%	1,5%	0,5%
46	1,3%	0,7%	1,2%	0,3%	1,3%	0,4%
47	1,9%	0,3%	1,1%	0,2%	1,1%	0,4%
48	1,1%	0,0%	0,9%	0,2%	1,0%	0,3%
49	0,3%	0,4%	0,8%	0,2%	0,9%	0,3%
50	0,3%	0,0%	0,6%	0,2%	0,8%	0,2%
51	1,2%	0,4%	0,5%	0,1%	0,7%	0,2%
52	1,0%	0,0%	0,4%	0,1%	0,6%	0,2%
53	0,5%	0,0%	0,3%	0,1%	0,5%	0,1%
54	0,7%	0,0%	0,2%	0,0%	0,4%	0,1%
55	0,2%	0,4%	0,1%	0,0%	0,3%	0,1%
56	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%
57	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%
58	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%
59	0,2%	0,0%	0,0%	0,0%	0,0%	0,0%
60	0,2%	0,0%	0,0%	0,0%	0,0%	0,0%

FIGURE 35 – Table de turnover selon la méthode initiale et la méthode retenue

Table des figures