Mémoire présenté le : 24 Mai 2022

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

Par : Karl BLANQUART

Titre : Modélisation de la dynamique des carrières des salariés du particulier employeur dans le cadre de la mise en place d'une indemnité de fin de carrière

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

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

Membre présents du jury de l'Institut des Actuaires signature

D. Vallée

Membres présents du jury de l'ISFA

D. Dorobantu

S. Loisel

Entreprise :

Nom : *IRCEM*

Signature : IRCEM

Directeur de mémoire en entreprise :

Nom : Thibaut CHARPENTIER

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

Résumé

Ce mémoire s'inscrit dans un contexte de déploiement d'une garantie « Indemnité de départ volontaire à la retraite » (IDR) au sein d'une institution de prévoyance, l'IRCEM Prévoyance, spécialisée dans l'assurance prévoyance-santé des salariés du particulier employeur (SPE). Cela concerne des métiers tels que les employés de ménage à domicile, les assistants maternels etc.

Dans cet objectif, une connaissance détaillée de la garantie et des spécificités du portefeuille est nécessaire. En effet, les SPE présentent des profils très particuliers par rapport à la population générale : une présence volatile dans la branche, des salaires assez faibles, la possibilité du multi-employeur... Ces caractéristiques, pour la plupart très rares, impliquent un taux d'acquisition de droit pour l'IDR de droit commun très faible (car la possibilité de carrière chez un même employeur est limitée). Ainsi, le taux de Turn-Over important et la non-continuité de la présence sont deux variables qui nécessitent de développer des méthodes pour capter le plus précisément la possibilité qu'un individu reste en portefeuille un certain nombre d'années.

Dans un premier temps, nous ne prendrons pas en compte ces caractéristiques. La première approche qui sera développée lors de notre étude sera l'analyse de survie. En prenant comme hypothèse que toute sortie du portefeuille d'assuré est définitive, on peut se rapprocher d'une étude de loi de maintien. Nous nous servirons de l'estimateur de Kaplan-Meier pour déterminer les taux bruts et nous lisserons les résultats obtenus selon la méthode de Whittaker-Henderson.

Nous nous intéresserons ensuite à deux autres modèles souvent utilisés dans l'analyse des données de survie : le modèle de Cox, à risque proportionnel, et le modèle additif d'Aalen.

Dans un second temps, nous considérerons la non-continuité de la présence des assurés en portefeuille. Nous modifierons la forme de nos données en séquences. Pour cela, nous transformerons notre base pour qu'elle représente le parcours d'un assuré au sein de la branche du salarié du particulier employeur. A l'aide d'une méthode dite d'Optimal Matching (OM), et de la classification selon le critère de Ward, nous tenterons de grouper ces données pour avoir des parcours types. Nous analyserons ensuite ces parcours pour les décrire et nous verrons l'impact de différentes variables sur chaque groupe.

Finalement, nous introduirons les modèles de Markov. Tout d'abord nous rappellerons les caractéristiques de ces modèles dans un cadre homogène, puis inhomogène. Afin de modéliser le cadre inhomogène, nous nous baserons sur une méthode d'intensité constante par morceaux. Nous analyserons alors nos résultats et les comparerons pour déterminer le modèle le plus adapté.

Nous conclurons sur la tarification du produit et les impacts possibles de chocs qui viendraient modifier et dégrader le résultat.

Mots-clés : IDR, SPE, Turn-Over, table de maintien, Kaplan-Meier, Cox, Aalen, Whittaker-Henderson, Optimal-Matching, critère de Ward, modèle de Markov, multi-états

Abstract

This thesis follows the deployment of a « Voluntary Retirement Indemnity » guarantee within IRCEM Prévoyance, a French health insurance specialized in the protection of the employees and employers in cases in which the employer is an individual. This concerns professions such as home maids, childminders etc.

For this purpose, a good knowledge of the guarantee and the portfolio specificities are necessary. Indeed, our policyholders have very specific profiles compared to the general population: a volatile presence in the sector, the possibility of multi-employer… These characteristics implies that a tiny portion of the workers will get the retirement indemnity as it is defined in the common law, due to the low chance to accumulate 10 years of career with the same individual employer. Thus, the high turnover rate and the presence discontinuity are two variables that require the development of methods to capture more precisely the possibility that an individual remains in the portfolio for years.

At first, we will not take these characteristics into account. The first approach that will be developed during our study will be survival analysis. Assuming that any exit from the portfolio is definitive, we can approach a study of the duration distributions. We will use the Kaplan-Meier estimator to determine the raw rates and smooth the results obtained using the Whittaker-Henderson method.

We will then focus on two other models often used in the analysis of survival data: the Cox model, with proportional risk, and the additive Aalen model.

Secondly, we will consider the non-continuity of the presence of policyholders in the portfolio. We will modify the shape of our data into sequences. For this, we will transform our database so that it represents the career of a worker within the sector. Using a method called Optimal Matching (OM), and classification according to Ward's method, we will try to group these data to have typical paths. Then, we will analyze these paths to describe them, and see the impact of different variables on each group.

Finally, we will introduce Markov models. First, we will recall the characteristics of these models in a homogeneous situation, then in an inhomogeneous one. To model the inhomogeneous Markov model, we will rely on a piecewise constant intensity method. We will then analyze our results and compare them to determine the most suitable model.

We will conclude on the distribution of « Voluntary Retirement Indemnity » claims modeled thanks to the different methods and the possible impacts of shocks that would modify and degrade the result.

Keys words : Voluntary Retirement Indemnity, Turn-Over, duration distributions, Kaplan-Meier, Cox, Aalen, Whittaker-Henderson, Optimal-Matching, Ward's method, Markov models, multi-states

Remerciements

Je remercie tout d'abord Jean-Charles GROLLEMUND, Directeur Général de l'IRCEM, de m'avoir permis d'utiliser les données et les moyens de l'IRCEM pour mener à bien ce mémoire

Je remercie également Alain KOUTOUAN, Directeur Actuariat et Finance, pour son accompagnement et son aide technique lors de cette étude.

Je remercie mes collègues, pour les encouragements et la convivialité de ces dernières années. Les concernés sauront se reconnaitre, que nous soyons encore collègues ou maintenant amis.

Je tiens à remercier tout particulièrement Julien DEJONGHE et Thibaut CHARPENTIER, respectivement Directeur Adjoint Actuariat et Finance et Manager de l'équipe inventaire de l'IRCEM Prévoyance, pour leur soutien, leurs encouragements et leur aide au quotidien. C'est grâce au temps qu'ils m'ont consacré, ainsi qu'à la relecture et les échanges tant sur le fond que sur la forme, que ce mémoire a été rendu possible.

Je remercie Frédéric PLANCHET d'avoir accepté la charge de Tuteur pédagogique.

Enfin, je remercie grandement mes parents, ma sœur et mes amis pour leur aide, leur soutien sans faille et les encouragements dans les grands projets de ma vie.

Table des matières

Introduction

Depuis sa création en 1973, le groupe de protection sociale IRCEM est dédié aux emplois de la famille.

Il est constitué :

- d'IRCEM Retraite, institution de retraite complémentaire AGIRC-ARRCO ;
- d'IRCEM Mutuelle (garanties facultatives Santé, Obsèques, …) ;
- d'IRCEM Prévoyance.

IRCEM Prévoyance assure la protection sociale collective et obligatoire des salariés du particulier (SPE) et des assistants maternels (AM) contre différents risques :

- Incapacité ;
- Invalidité ;
- Décès ;
- Rente éducation ;
- Maladies redoutées.

Le secteur du particulier employeur est très spécifique dans son cadre juridique et dans l'exercice de son activité. De ce fait, il n'est pas rare que des aspects du droit commun ne soient pas directement transposables au salarié du particulier employeur. On pourra prendre l'exemple de la rémunération des congés payés CESU, qui dépend du nombre d'heures effectuées mensuellement chez le particulier employeur, qui ne peut pas exister pour un CDI classique. Les spécificités de ces publics sont nombreuses, et la population était historiquement couverte par deux conventions nationales collectives : les assistants maternels (AM) et les salariés du particulier employeur (SPE).

A titre d'illustration, ces particularités entraînent certains risques inédits comme le décès d'un employeur. Pour cela, IRCEM Prévoyance développe des garanties dédiées, comme « Rupture du contrat de travail » qui prévoit, en cas de décès du particulier employeur, le versement d'un capital permettant de faire office d'indemnité de licenciement ou d'indemnité de préavis.

On résume ces deux branches :

	Effectifs	Employeurs	Masse Salariale 2021	Garanties	
AM	260 653	968 284	4 984 M€	- Arrêt de travail (AT) - Décès/Rente Éducation - Maladie Redoutée	
SPE	1 046 466	2 313 858	6 398 M€	- AT - Décès/Rente Éducation - Maladie Redoutée	(A partir de 2022)

Tableau 0.1 - Présentation des branches historiques du salarié du particulier employeur

Depuis plusieurs années, une convergence des deux branches assurées historiquement par l'IRCEM est négociée :

- La branche des salariés du particulier employeur (régie anciennement par la convention collective nationale des salariés du particulier employeur du 24 novembre 1999)
- La branche des assistants maternels du particulier employeur (régie anciennement la convention collective nationale des assistants maternels du particulier employeur du 1er juillet 2004)

Cette nouvelle branche est nommée « branche du secteur des particuliers employeurs et de l'emploi à domicile », et est mise en place en janvier 2022.

Cette convergence a pour but de mettre à jour les textes encadrant les salariés de ces secteurs et de construire un socle de droits collectifs, afin de protéger au mieux les différents acteurs.

Des nouvelles garanties ont aussi été négociées pour les salariés du particulier employeur qui viennent élargir la protection des assurés et qui s'inscrivent dans une vision prospective du secteur.

On y retrouve un maintien de garanties décès/rente éducation, ainsi que maladies redoutées pour les salariés du particulier employeur, anciennement seulement présent dans l'accord obligatoire des assistants maternels.

Enfin, il a été décidé de la mise en place d'une garantie « Indemnité de départ volontaire à la retraite » (IDR), qui nous intéresse dans le cadre de ce mémoire.

I. Statistiques descriptives et spécificités de l'emploi à domicile

Depuis le 1er janvier 2022, une nouvelle convention collective nationale vient refaçonner la protection sociale des salariés du particulier employeur. Au terme d'une procédure d'appel d'offres lancée en 2021, les partenaires sociaux ont renouvelé leur confiance envers l'IRCEM Prévoyance pour gérer et assurer les garanties Prévoyance et IDR pendant 5 ans (2022-2026). Pour le groupe IRCEM, il s'agit d'un nouveau défi avec la fusion de ces 2 branches historiques, celle des assistants maternels et celle des salariés du particulier employeur.

a. Contexte juridique de l'IDR

i. Cadre général

L'indemnité de fin de carrière est un avantage postérieur à l'emploi qui est défini dans le code du travail : articles D1237-1 à D1237-2-3, comme le versement d'un capital au moment du départ de l'employé de l'entreprise.

Elle est conditionnée au mode de départ de l'employé de l'entreprise :

- Si le salarié est mis à la retraite par son employeur, alors l'obtention de droits pour l'indemnité est non conditionnée par l'ancienneté et se calcule (hors convention plus avantageuse) comme ceci :

Ancienneté du salarié	**Montant de l'indemnité**
Jusqu'à 10 ans	*Le quart d'un mois du salaire de référence par année d'ancienneté*
Après 10 ans	*Le quart d'un mois du salaire de référence par année d'ancienneté pour les 10 premières années et* *Un tiers d'un mois du salaire de référence par année d'ancienneté à partir de la 11ème année*

Tableau 1.1.1.1 - Montant d'indemnisation en cas de mise à la retraite par son employeur

- Si le salarié part volontairement de l'entreprise, alors l'indemnité est conditionnée à l'ancienneté de ce dernier. Il sera alors nécessaire d'avoir 10 ans d'ancienneté pour être éligible à l'indemnité de fin de carrière.

Le montant est détaillé dans le tableau ci-dessous :

Ancienneté du salarié	Montant de l'indemnité
Au moins 10 ans	Un demi-mois du salaire de référence*
Au moins 15 ans	Un mois du salaire de référence*
Au moins 20 ans	Un mois et demi du salaire de référence*
Au moins 30 ans	Deux mois du salaire de référence*

Tableau 1.1.1.2 - Montant d'indemnisation en cas de départ volontaire à la retraite

**Dans les deux cas, le salaire de référence quant à lui est calculé comme le maximum entre :*

a. *Un douzième de la rémunération brute (salaire, primes, et autres) des douze derniers mois qui précèdent la notification de la mise à la retraite,*
b. *Un tiers des trois derniers mois de rémunération brute précédant la notification ou la fin du contrat de travail.*

ii. Spécificité pour le salarié du particulier employeur

Dans le cadre singulier des SPE, et de la nouvelle garantie IDR souhaitée par les partenaires sociaux de la branche à effet 2022, le salarié doit, pour prétendre au versement de l'indemnité, attester à date de son départ volontaire à la retraite de deux conditions basées sur l'ancienneté, présentées ci-dessous.

Cependant, la prise en compte de la spécificité de la population entraîne que le calcul de l'ancienneté dépend de la présence en branche, et non chez un employeur en particulier. Ainsi, il ne revient pas au dernier employeur de payer : tous les employeurs cotisent, et c'est l'IRCEM Prévoyance qui paie la prestation.

Les conditions sont les suivantes :

Ancienneté dans les 7 dernières années précédant la date de retraite	Ancienneté totale du salarié	Montant de l'indemnité
Moins de 5 ans	**Sans distinction**	-
5 ans ou plus	**Moins de 10 ans**	-
	Au moins 10 ans	Un mois du salaire de référence
	Au moins 15 ans	Un mois et demi du salaire de référence
	Au moins 20 ans	Deux mois du salaire de référence
	Au moins 30 ans	Deux mois et demi du salaire de référence

Tableau 1.1.2.1 - Montant d'indemnisation prévue par la nouvelle CCN de la branche du secteur des particuliers employeurs et de l'emploi à domicile

Le salaire de référence est calculé comme étant la moyenne mensuelle des salaires bruts perçus par le bénéficiaire au cours des 5 dernières années d'emploi précédant son départ en retraite.

Dans le cas de suspension du salaire pour arrêt de travail, chômage partiel ou congé de formation, une reconstitution des salaires qui auraient été reçus est effectuée pour calculer le salaire mensuel de référence. Concernant ce dernier point, le motif de non-présence n'est pas connu dans nos bases. On se contentera alors de ne prendre en compte que les salaires connus, sans réintégration de salaire recalculé pour les raisons ci-dessus.

NB : Il n'existe pas de cadre particulier de mise à la retraite par l'employeur. On s'appliquera dans ce mémoire à l'étude de l'indemnité de départ volontaire à la retraite (IDR).

En 2018, une commission des études comptables de la CNCC[1] et du CSOEC[2] a établi que l'évaluation des indemnités de fin de carrière doit être effectuée en tenant compte des seules prévisions de démission, en excluant tout autre hypothèse de départ à la retraite, comme la rupture conventionnelle.

Cependant, cette hypothèse entraîne une importante surévaluation des engagements car le turn-over serait clairement sous-évalué. Il est donc nécessaire de l'évaluer autrement pour obtenir une étude aussi précise que nos données nous le permettent.

De plus, les SPE présentent une présence dans leur branche particulière. Un assuré peut ainsi avoir plusieurs contrats chez des employeurs différents, il peut aussi ne pas travailler certaines semaines/mois (saisonnalité de l'emploi) sans que cela implique une sortie évidente de la branche, de notre périmètre d'observation. De plus, l'information sur le motif de sortie de l'assuré est très limitée dans nos bases de données, ainsi nous ne sommes pas en mesure de déterminer avec précision la raison de la dernière sortie de l'assuré.

[1] Compagnie nationale des commissaires aux comptes
[2] Conseil supérieur de l'ordre des experts comptables

iii. Notations actuarielles

Cette partie vise à introduire les différentes notations actuarielles liées à l'indemnité de départ volontaire à la retraite.

On va alors définir le montant de l'indemnité comme étant :

$$IFC_{Théorie} = \frac{S_t}{12}(1 + a_t)^{z-x} * M * {}_{z-x}P_x * {}_{z-x}P'_x * \frac{1}{(1+i)^{z-x}}$$

Avec :

- t, date à laquelle on se place pour l'estimation du montant
- S_t, le salaire annuel de l'individu à t
- a_t, le taux d'augmentation du salaire à t
- M, le nombre de mois de paiement (fonction de l'ancienneté)
- x, l'âge de l'individu
- z, l'âge à la retraite
- i, le taux d'actualisation
- ${}_{z-x}P_x$, la probabilité que l'individu soit vivant à l'âge x, et le soit toujours à l'âge z
- ${}_{z-x}P'_x$, la probabilité que l'individu soit vivant à l'âge x, et soit bénéficiaire à l'âge z

Dépendant de l'ancienneté en branche, **on cherchera dans notre étude à estimer M pour un individu**. Nous nous baserons sur la répartition selon le nombre de mois de paiement pour justifier de la bonne conformité ou non de nos modèles. Il sera impératif pour l'estimation de la provision associée à l'indemnité de regarder les éléments descriptifs, et de prendre certaines hypothèses comme pour le taux d'actualisation ou l'évolution salariale et de les estimer.

Afin d'étudier au mieux la mise en place de cette garantie au sein de l'IRCEM, nous présenterons dans un premier temps les données utilisées.

b. Données et retraitements

Avant de commencer toute analyse, nous allons présenter les données à disposition et utilisées.

i. Description des données disponibles

La qualité et la gestion de la donnée ont toujours été des enjeux majeurs pour les assureurs, et cela a encore été renforcé depuis 2016 par la mise en place de Solvabilité 2. Des données avec un défaut de qualité peuvent entraîner des erreurs majeures de tarification ou d'appréhension du risque. Dans le cadre de notre étude, on s'efforcera de contrôler la qualité de la donnée utilisée à tout moment, ainsi que la pertinence de cette dernière. De plus, les données à caractère personnel doivent respecter un certain nombre de règles au titre du RGPD. Afin de protéger les clients et leurs données personnelles, le règlement européen sur la protection des données met en place des obligations à la charge de toute entité amenée à traiter des données.

Dans ce contexte, la récupération d'un historique exhaustif des données de masses salariales, permettant la description et l'analyse des carrières de nos assurés a été nécessaire.

Il a été mis à disposition une table se présentant comme ci-dessous :

NOM	TYPE	
CIVNUR	CHAR	Identifiant Employeur
CIVNEP	CHAR	N° Employeur
CIVFAP	CHAR	Période d'appel
CIVNOP	CHAR	N° Ordre période
CIVOP0	CHAR	Identifiant salarié
CIVKPI	CHAR	Catégorie période
CIVSAD	DECIMAL	Salaire déclaré
CIVSAS	DECIMAL	Salaire soumis
CIVMCI	DECIMAL	Cotisation IRCEM
CIVEDC	CHAR	Présence données spécifiques (option de calcul des cotisations)
CIVCCM	CHAR	Date calcul montant
CIVENQ	CHAR	Numéro d'enquête CNAV

Tableau 1.2.1.1 - Variable de notre base de salaire

La base contient les masses salariales des publics de l'IRCEM, par trimestre, depuis 1973.

Ces dernières sont décomposées par assuré, par employeur, par période d'appel, par date d'intégration en base et par catégorie d'emploi. Les 2 grandes catégories historiques sont les salariés du particulier employeur (SPE) et les assistants maternels (AM). Malgré la fusion de leurs branches respectives à effet 2022, la décomposition est conservée à des fins de suivi.

Des variables provenant d'autres bases ont été ajoutées, comme une variable géographique, la sinistralité antérieure sur la garantie arrêt de travail ou encore la situation familiale. Cependant, au vu des résultats apportés ou de la fiabilité de ces dernières, il n'a pas semblé nécessaire de les conserver.

ii. Limitation de l'étude

L'ensemble de ces données représentent un trop grand volume (plus de 130 000 SPE pour la génération née en 1958), il n'est pas nécessaire ni pertinent de s'intéresser à l'ensemble du portefeuille de l'existence de l'IRCEM à nos jours.

Ainsi, on se restreint à l'observation d'une partie des données, en essayant de faire une sélection logique vis-à-vis de nos besoins. Le maintien en activité de notre population suppose d'avoir un historique des salaires sur plusieurs années d'un même assuré. Si l'on vient à sélectionner un individu, il nous faudra nous assurer de récupérer l'ensemble de sa carrière observable à date. Nous verrons par la suite qu'une carrière incomplète (dont l'âge actuel est inférieur à l'âge de départ en retraite), qui peut être considérée comme une censure à droite de nos données, n'est pas un obstacle à l'estimation de la présence dans le portefeuille au fil des années. Cependant, pour profiter de l'antériorité forte de nos données, on se concentrera seulement sur des générations récentes.

Dans le contexte de l'étude, à savoir la mise en place d'une garantie indemnisation de fin de carrière, il apparait alors nécessaire de cibler des données pouvant prendre en compte le risque associé. Il est nécessaire de récupérer l'ancienneté totale dans la branche, mais aussi des 7 dernières années qui sont fondamentales pour l'acquisition de droits (voir critères de versement présentés sur la partie ii. du contexte juridique de l'IDR). De plus, on se concentrera ici sur l'analyse de la population des salariés du particulier employeur. L'analyse de la présence des assistants maternels correspond à un tout autre sujet: en effet, leurs caractéristiques diffèrent sur de nombreux points. La séparation de ces deux corps a pour objet de mieux analyser les données et les résultats obtenus.

Enfin, du fait du grand nombre de données, nous n'avons pas besoin de prendre un historique générationnel trop important. Un nombre conséquent de données est très apprécié en théorie, et dans la majorité de la pratique, mais par la nature des approches qui suivront, nous nous retrouverions avec un volume beaucoup trop important et des échecs sur les tentatives de modélisation.

La base sera dans un premier temps composée des générations nées en 1958, 1968, 1978. On prend aussi pour chaque individu ayant eu un salaire dans la profession entre sa naissance et aujourd'hui une ligne par âge (à partir de 18 ans, nous fixons 63 comme étant l'âge maximum pris en compte pour la garantie, 64 sera alors l'état associé à la retraite). Une approximation est alors faite qu'un salaire sur l'année équivaut à une année d'ancienneté (70% des cas), toujours dans le but de réduire les données et ne pas surmultiplier les lignes avec une maille trimestrielle.

iii. Retraitement

Afin de s'assurer de ne prendre en compte que des données fiables, la modification ou la suppression de certaines données sont des étapes essentielles à la constitution d'une table.

On recense dans nos bases différents motifs de non-fiabilité, que l'on peut résumer dans le tableau ci-dessous :

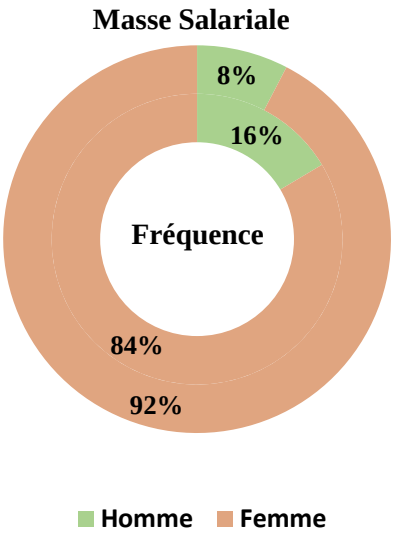
Type de non fiabilité	Part des données concernées	Résolution
Mois sécurité sociale supérieur à 12	0.6%	Comme l'étude se fait par Age entier, on a fixé le mois par défaut à Janvier
Salaire pré-2002 en Franc	15%	On passe tous les salaires en équivalent euros
Salaire nul ou inférieur à 0	0.1%	On supprime les lignes concernées
Age inférieur à 18ans ou supérieur à 63	0.1%	On supprime les périodes concernées

Tableau 1.2.3.1 - Retraitements de notre base de salaire

c. Statistiques descriptives

Nous présenterons ici notre population sur une génération précise, la génération née en 1958. Ce choix est motivé par le fait que ce soit une génération très récemment passée en retraite. Nous avons donc un historique complet et une première vision de notre portefeuille. Pour comparer les salaires entre eux, l'inflation sera prise en compte dans l'ensemble de notre étude.

i. Sexe

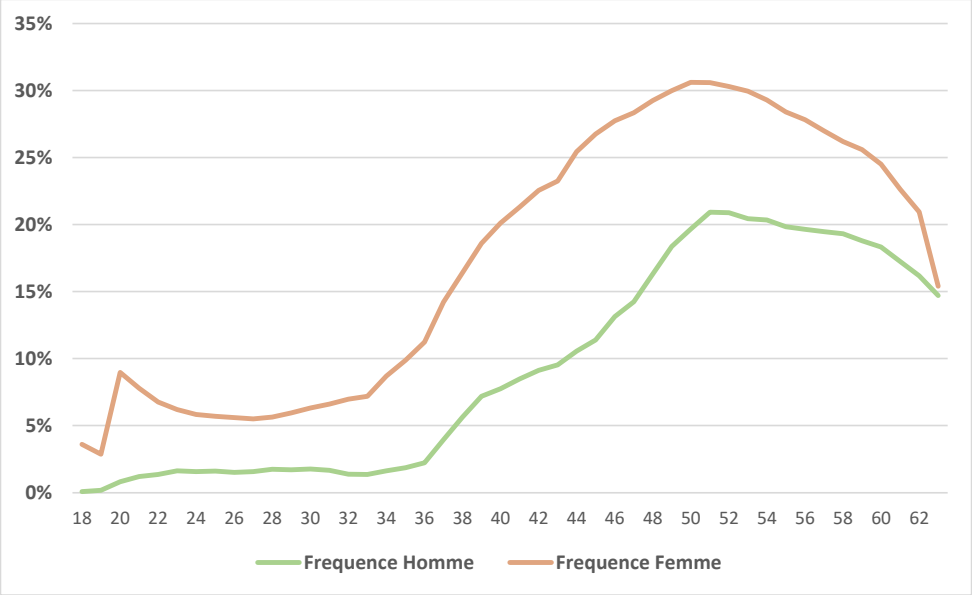


Graphique 1.3.1.1 - Répartition de la masse salariale et du nombré d'assuré selon le sexe

La première donnée qu'on observe est la répartition selon le sexe de l'assuré. Le portefeuille de l'IRCEM possède une particularité dans le sens où il est constitué à très grande majorité d'individu de sexe féminin, dans le sens de leur enregistrement à la sécurité sociale. Au vu de la masse salariale écrasante, considérer que notre portefeuille est essentiellement féminin ne serait pas une hypothèse aberrante. Néanmoins, on souhaitera étudier l'apport de cette variable à nos modèles, et dans le cas où celui-ci est significatif mesurer l'effet associé quand cela est possible.

On décomposera alors quand cela semble nécessaire par sexe les données suivantes, en l'occurrence quand cela a un lien direct avec la problématique.

ii. Age



Graphique 1.3.2.1 - Présence par âge selon le sexe

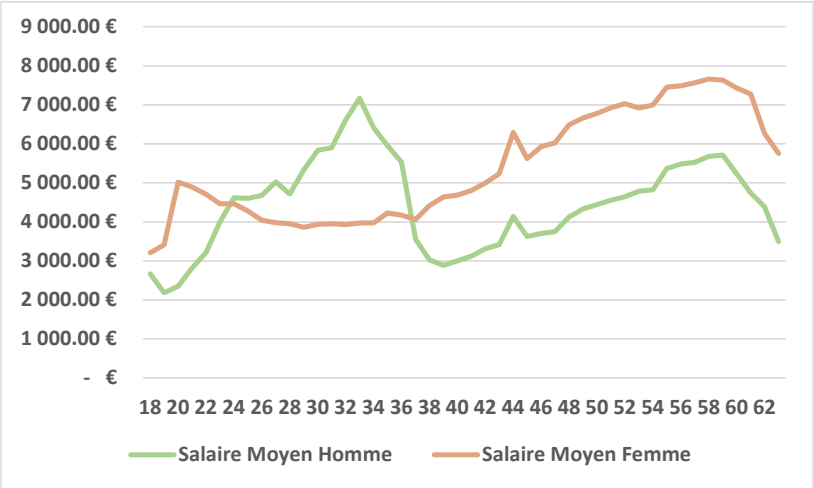
On peut lire ce graphique de manière suivante : parmi l'ensemble des femmes de la génération 1958 ayant un jour travaillé dans la branche, 30% étaient actives à 50 ans.

La part de présence à tout âge dans notre étude est essentielle. En effet, dans notre cadre, pour bénéficier d'une indemnité on prend en compte le nombre d'années passées en portefeuille de deux manières, sur l'ensemble de la carrière puis sur les sept dernières années. Une particularité de notre portefeuille est la présence majoritairement sur la deuxième partie de la carrière des salariés (si on prend en compte une carrière moyenne en France entre 20 et 62 ans). Si on regarde l'ensemble des âges auxquels les salariés du particulier employeur ont eu un salaire, on obtient une moyenne de 47 ans pour les femmes et de 51 ans pour les hommes.

Étant à tout instant au-dessus de cette dernière, la courbe de présence des femmes affiche plus de présence au cours du temps et sur une plus longue durée. Or, si l'on prend en compte un âge de retraite à 62 ans, et des paliers à 10, 15, 20 et 30 années d'ancienneté, on comprend, par la distribution des âges, que les femmes vont acquérir significativement plus de droits, et d'un montant plus conséquent en termes de mois de salaire de référence, que les hommes. On peut expliquer cela par un historique « métier de carrière » pour les femmes, là où dans la branche les hommes ont un profil plus passager dans la profession, étant un métier soit de remplacement soit de complément à leur activité principale.

De plus, on se rend compte de la présence assez faible, même vers l'âge moyen de la catégorie principale, où on dépasse difficilement les 30% de présence. La population présente un turn-over assez élevé comme nous le verrons dans la prochaine partie.

iii. Salaire moyen en fonction de l'âge



Graphique 1.3.3.1 - Évolution du salaire inflaté annuel moyen selon l'âge

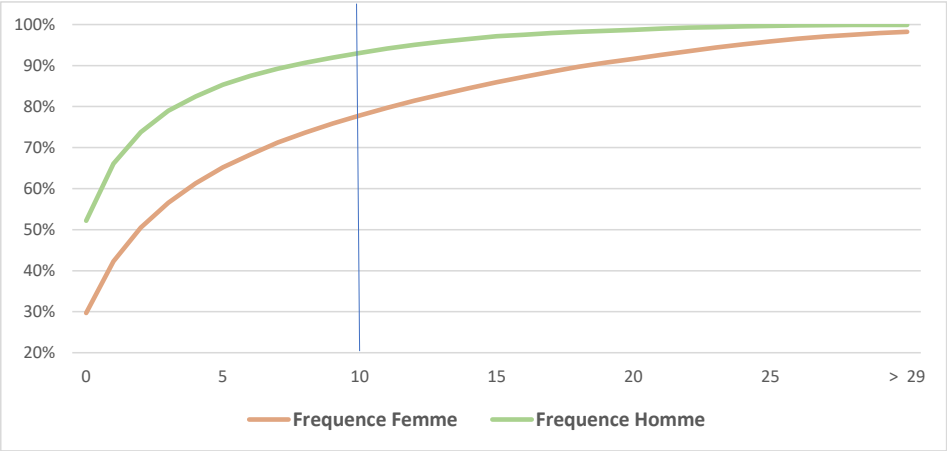
Le graphique 2.2.3 nous permet de préciser plusieurs points. Tout d'abord on peut nettement observer l'écart entre le salaire moyen des femmes et celui des hommes. Ceci va une nouvelle fois rejoindre l'observation faite au point précédent sur le profil type de carrière privilégié par ces deux populations.

De plus, on note que le salaire moyen, bien que croissant avec l'âge, reste très faible même aux âges élevés. Le pic des hommes à 33ans s'explique par le faible nombre d'assurés présents en base, rendant la moyenne très sensible aux très gros salaires, comme un assuré en activité entre 26 et 35 ans avec jusqu'à plus de 10 fois le salaire moyen. Un salaire annuel moyen plus faible s'explique par une non-régularité dans la profession, ou un nombre d'heures plus faible, éléments caractéristiques d'un complément ou d'un remplacement de salaire. Comme discuté précédemment, c'est cette présence en portefeuille qui est le sujet de cette étude.

Il est intéressant de noter le pic lors de l'ancienneté 44. En effet, pour la génération née en 1958, l'ancienneté 44 nous amène à l'année 2002. On a ici l'impact du passage du Franc à l'Euro, non exactement renseigné dans nos bases. La rectification se fait manuellement sans date précise. Il a donc fallu estimer par étude globale et individuelle une date à partir de laquelle on suppose les salaires en €.

On se concentre alors sur l'ancienneté. En effet, si la répartition de l'âge semble correspondre à nos critères avec une présence plus prononcée vers le milieu/fin de carrière, c'est l'ancienneté cumulée à la retraite qui va nous permettre de déterminer l'acquisition de droit. On peut alors confirmer cela avec l'analyse de la distribution de l'ancienneté lors du passage à la retraite de nos populations.

iv. Ancienneté



Graphique 1.3.4.1 - Distribution de l'ancienneté en fin de carrière selon le sexe

On se focalisera sur la distribution cumulée de l'ancienneté. Pour acquérir des droits d'indemnisation de départ volontaire à la retraite, une ancienneté de 10 ans est nécessaire. On remarque une nouvelle fois la différence nette entre les deux populations. Cela se lit facilement dans le graphique 2.2.4 où on a 91.9% des hommes qui n'auront aucune chance d'acquérir des droits, là où le taux chute à 75.9% chez les femmes (au global, ce taux est de 79%). Pour information, la moyenne d'ancienneté pour les femmes est de 6 années, là où celle des hommes est de 2.4 années.

On résume dans le tableau suivant la fréquence et masse salariale en fonction des segments d'ancienneté délimités par la garantie :

Ancienneté totale du salarié	Montant de l'indemnité	Répartition de la génération 58		Masse Salariale Inflatée	
		Homme	Femme	Homme	Femme
Moins de 10ans	-	91.9%	75.9%	38.5%	21.6%
Au moins 10ans	1 mois	4.6%	8.6%	21.2%	14.8%
Au moins 15ans	1.5 mois	2.0%	6.2%	16.2%	17.1%
Au moins 20ans	2 mois	1.4%	7.2%	21.2%	31.4%
Au moins 30ans	2.5 mois	0.1%	2.1%	2.8%	15.1%

Tableau 1.3.4.1 - Acquisition des droits par l'ancienneté de carrière

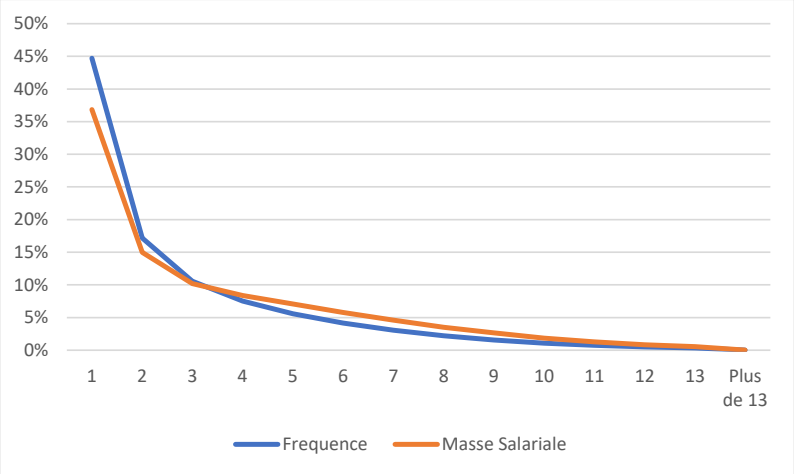
On remarque le très faible nombre d'individus qui ont une chance d'acquérir des droits par leur ancienneté totale : à peine 1/5ème des femmes et moins de 1/10ème des hommes. On comprend que, même si chez les femmes le taux est un peu plus important, il reste que beaucoup des assurés qui travaillent un jour dans la branche ne sont présents que pour de courtes périodes.

Cela s'explique par le caractère de métier de compléments, que ce soit chez des profils jeunes pour un job étudiant, chez des profils plus âgées pour le temps de trouver un autre emploi dans le périmètre voulu, ou encore chez des profils d'individus en fin de carrière pour avoir un complément de revenu pour la retraite.

Comme on peut l'imaginer, la masse salariale est plus équitablement répartie, avec un historique plus important et un salaire moyen plus élevé proportionnellement à l'ancienneté en branche (cf. Graphique *1.3.3.1*).

Dans la partie sur l'approche par répartition on restreint l'étude d'ancienneté aux sept dernières années avant la retraite afin d'observer le nombre réel de cas qui bénéficient d'une indemnisation de départ volontaire à la retraite.

v. Nombre d'employeurs

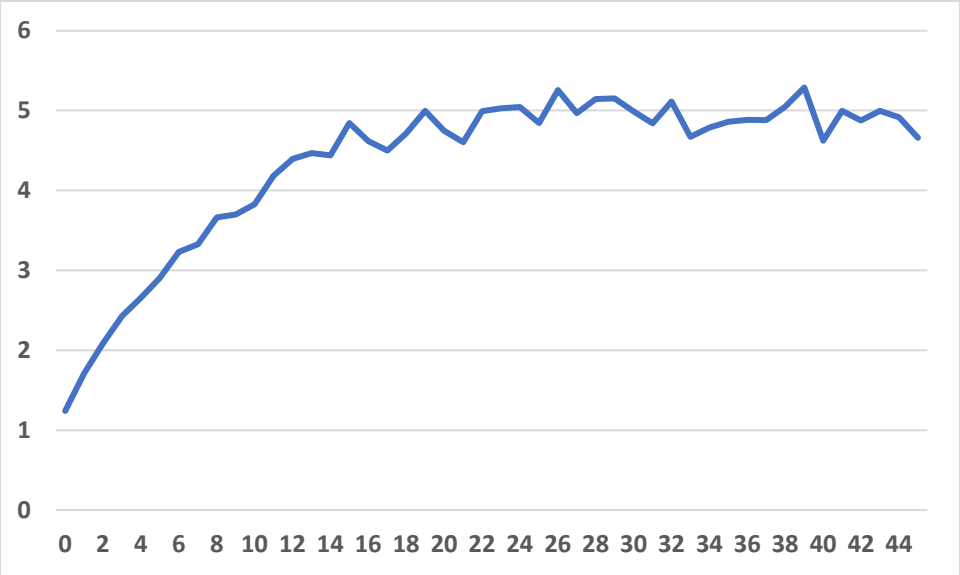


Graphique 1.3.4.1 - Distribution en fonction du nombre d'employeurs par période

Finalement, la dernière donnée étudiée est le nombre d'employeurs distincts d'un assuré.

C'est une spécificité du salarié du particulier employeur, ce dernier peut être employé en simultané par plusieurs particuliers employeurs. On pourra utiliser cette information pour alimenter nos modèles. Même si la modalité la plus fréquente est un seul et unique employeur par période, cela ne représente que 35% de nos individus. La moyenne s'établit elle à 3.4 employeurs par période pour un assuré.

Une autre vue pour analyser cette variable est la moyenne par période en fonction de l'ancienneté de l'individu :



Graphique 1.3.4.2 - Moyenne du nombre d'employeurs par période par ancienneté à la retraite

On remarque que la variable nombre d'employeurs est croissante en fonction du nombre d'années d'ancienneté constaté à la retraite jusqu'à 20 environ. Ensuite cela stagne jusqu'au maximum de 45 années d'ancienneté (de 18 à 63 ans).

Après s'être intéressé aux statistiques descriptives de nos données, et afin de s'intéresser de plus près à notre problématique, on va suivre dans la prochaine partie les différentes évolutions des éléments composant l'IDR.

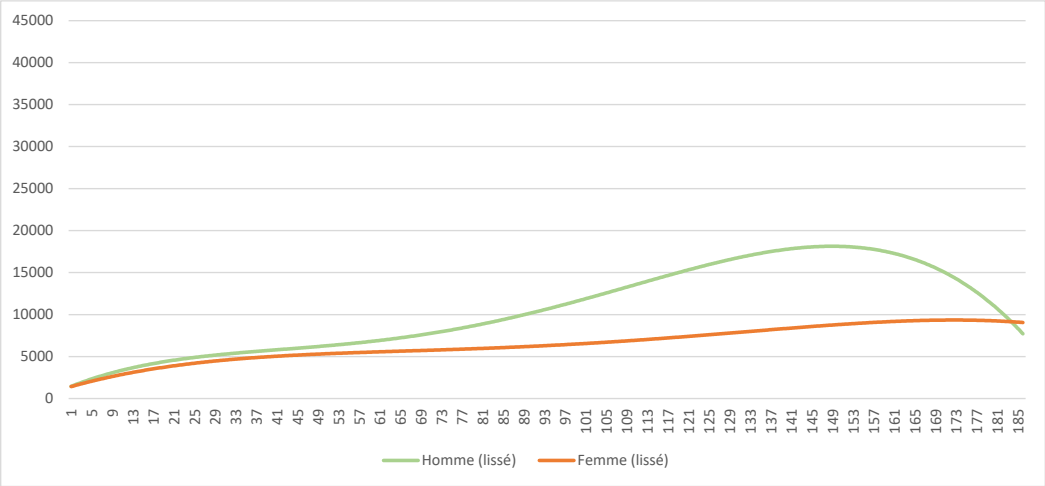
d. Dynamique des carrières

On ne garde ici que la génération née en 1958 et on s'intéresse à sa dynamique au fil des années.

Principalement, on portera un premier regard sur les données qui sont importantes dans l'étude de la garantie indemnité de départ volontaire à la retraite (IDR), à savoir le turn-over et les évolutions salariales.

i. Évolution salariale

On peut s'appliquer à regarder le salaire moyen non plus par âge, mais selon la deuxième variable caractéristique dans l'analyse de survie, par exemple l'ancienneté, toujours décomposée selon le sexe de l'assuré.



Graphique 1.4.1.1 - Évolution du salaire annuel moyen (lissé) selon l'ancienneté

Comme attendu, le salaire moyen est croissant avec l'ancienneté à la retraite. Plus un assuré reste dans la branche, plus on s'attend à un salaire régulier et plus conséquent pour en faire son activité principale. On observe aussi un salaire moyen légèrement supérieur chez les hommes comparativement aux femmes. Cela peut s'expliquer par un type de carrière plus tardive des hommes, et l'ancienneté se calcule alors sur une base plus âgée de la population.

Ancienneté à la retraite	Evolution lors de la 5ème année	Evolution entre la 5ème et la 10ème	Evolution entre la 10ème et la 20ème	Evolution entre la 20ème et la 30ème
5	13%			
10	47%	-19%		
20	80%	13%	-14%	
30	37%	26%	62%	-22%

Tableau 1.4.1.1 - Evolution salariale en fonction de l'ancienneté à la retraite

Ce tableau permet de suivre l'évolution des salaires selon l'ancienneté lors de la retraite. Ainsi, pour les assurés de la génération 1958 ayant 20 années d'ancienneté lors de leur départ en retraite, nous observons des augmentations de salaire de 80% entre leur première leur cinquième année d'activité. Ce taux s'explique par un salaire plus faible, un faible nombre d'heures effectuées et potentiellement moins d'employeurs au début de la carrière d'un SPE.

On représente l'évolution du salaire selon l'ancienneté lors du passage en retraite ainsi qu'en fonction du nombre d'année d'ancienneté observée.

NB : On a exclu volontairement de l'analyse le premier et dernier trimestre. Ceux-ci peuvent être fortement incomplets selon la date exacte de début et de fin, les évolutions observées auraient été aberrantes.

On remarque alors que :

- Les taux d'évolutions sont assez forts dans les cinq premières années,
- On conserve un taux d'évolution positif jusqu'à la dernière période observable,
- Les forts taux d'évolutions à la hausse sont explicables par l'entrée dans le métier. Souvent fait par un premier contrat auprès d'un employeur, l'assuré se développe par la suite et solidifie son activité au fur et à mesure que l'expérience s'accumule.
- Sur les dernières périodes pour les individus ayant 10, 20 et 30 années passées dans la branche à terme, on constate une évolution à la baisse. Ce taux est soit dû à une baisse du temps de travail, soit à une présence plus irrégulière à l'approche de la retraite pour les individus. Cependant, cette tendance n'était pas visible dans le graphique *2.3.6* et *2.4.1* traitant de l'ancienneté. On peut supposer alors que cela est dû à une sortie définitive de la branche, anticipée par une baisse de présence dans la branche. Pour voir si cela se confirme, décomposons en 2 groupes nos individus : ceux qui ont 5, 10, 20 et 30 années d'ancienneté totale dans la branche et dont le dernier salaire obtenu a été obtenu à 62 ou 63 ans, et les autres.

On obtient alors : *(avec dans une case le taux dans le cas du dernier salaire à 62/63 ans et le taux dans le cas contraire)*

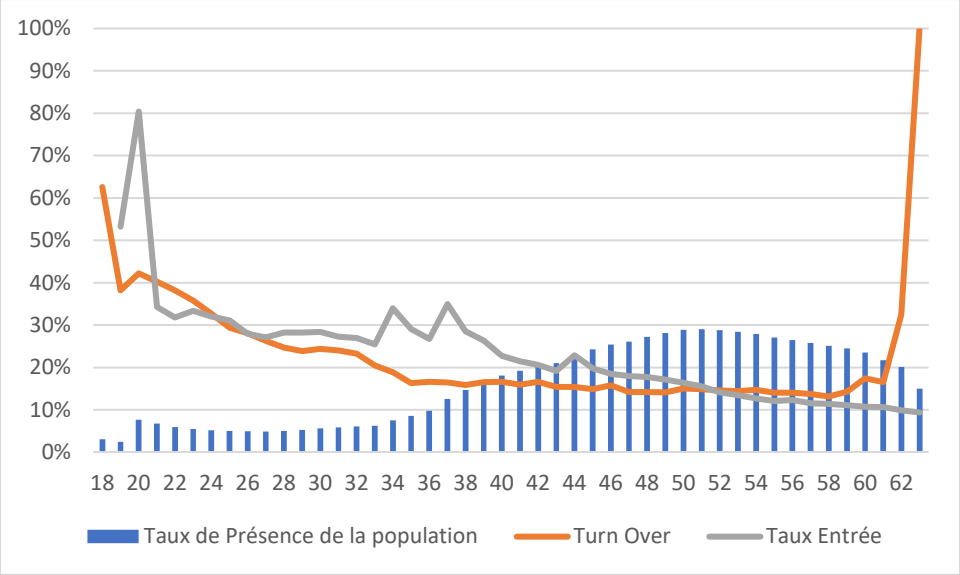
Ancienneté à la retraite	Evolution lors de la 5ème année	Evolution entre la 5ème et la 10ème	Evolution entre la 10ème et la 20ème	Evolution entre la 20ème et la 30ème
Dernier Salaire lors de la retraite	Oui/Non			
5	24% / 11%			
10	58% / 42%	-10% / -23%		
20	81% / 79%	22% / 1%	-8% / -23%	
30	56% / -6%	24% / 32%	69% / 37%	-24% / -12%

Tableau 1.4.1.2 - Evolution salariale en fonction de l'ancienneté à la retraite et de l'âge au dernier salaire

Tout d'abord, pour les taux d'évolutions entre la 20^{ième} et la 30^{ième} année d'ancienneté, il est intéressant de noter que la quasi-totalité de nos données (>95%) d'observation ont un âge à la 30^{ième} année supérieur ou égal à 60 ans, les ¾ étant dans la 1ère catégorie. Tous étant proche de la retraite, il n'est pas pertinent d'analyser la différence. Pour les 2 autres cellules en jaune dans le tableau, notre hypothèse semble se confirmer. On s'aperçoit aussi qu'en règle générale, quelqu'un qui va avoir son dernier salaire dans les âges supposés de retraite va avoir un taux d'évolution salarial plus élevé.

ii. Évolution de la présence en portefeuille

On a pu caractériser l'évolution salariale du notre population, on se penchera dans ce paragraphe à une première vision du turn-over au sein de la branche.



Graphique 1.4.2.1 - Taux de sorties et d'entrées par âge

On représente ici les deux transitions possibles (Entrée[3]/Sortie) entre les deux états (Présent/Non Présent), observés âge par âge. On peut facilement visualiser les moments où les courbes d'entrée et de sortie se croisent et où la tendance passe d'une présence haussière à une présence baissière.

[3] On parlera ici aussi bien de première entrée dans la profession que de retour à la suite d'une sortie

On distingue alors 3 phases :

- Une première entre 18 et 30 ans, qui est assez volatile, alternant entre plus d'entrée et plus de sortie. Cette période est peu intéressante dans le cadre de la mise en place de l'indemnité de départ volontaire à la retraite. Pour l'acquisition des droits sur l'ancienneté sur les sept dernières années, elle n'a aucune incidence. Pour l'acquisition des droits sur l'ancienneté totale elle peut avoir un impact. Cependant, le cas maximum d'ancienneté qui nous intéresse est de 30 ans, ce qui avec une hypothèse forte de non-retour après une sortie, rend non significative toute étude de ce segment.
- Une deuxième entre 30 et 52 ans, où l'entrée en portefeuille est nettement au-dessous de la courbe de sortie. Une personne entrant pour la première fois dans la branche au-delà de ces âges n'aurait d'office pas de droit acquis lors du passage en retraite. Cette période est déterminante pour la condition de droits sur le nombre de mois de salaire de référence versé.
- Une dernière entre 53 et 63 ans, où la tendance s'inverse et le taux de sortie augmente jusqu'à atteindre 100% à 63 ans. A proprement parlé, au début de cette période les taux de sorties ne bougent presque pas, ils restent à peu près les mêmes jusqu'à 59 ans. On observe plutôt une baisse des taux d'entrée. Cette période est primordiale dans l'étude de l'acquisition de droits, car elle porte sur la présence sur les sept années précédant la date de retraite.

La flexibilité offerte par l'emploi est une caractéristique qui est unique et difficilement modélisable tant la diversité de profil et d'organisation du travail est importante. Par ailleurs, pour notre étude nous n'avons pas accès avec précision au type d'activité réalisé par l'assuré, comme le ménage, du jardinage, du soutien scolaire, etc... Cette variable serait à n'en pas douter très intéressante à ajouter et à étudier dans de futurs travaux, apportant avec elle son lot d'informations et d'éclaircissements permettant d'affiner nos résultats, ou de les expliquer en partie.

Après examen des différentes dynamiques relatives à notre cadre, il est alors intéressant de regarder le montant réel d'indemnité de départ volontaire à la retraite, telle que mise en place par les partenaires sociaux de la branche à compter de 2022, si on avait dû la verser pour la génération née en 1958.

e. Garantie IDR : Approche d'un régime par répartition

Dans une approche par répartition, on s'intéresse à ce que l'on doit payer à une génération en finançant par les cotisations de la population active. Pour l'IRCEM, ce système mis en place en 2022 est financé par un % du salaire déclaré. On peut alors mettre en relation la masse salariale de l'ensemble du portefeuille sur une année avec le paiement de prestations d'une génération qui partirait en retraite. La masse salariale est une donnée connue des services actuariat, reste à déterminer les paiements pour la génération étudiée précédemment.

On obtient alors, en reprenant le tableau *1.1.2.1* :

Ancienneté dans les 7 dernières années précèdent la date de retraite	Ancienneté totale du salarié	Montant de l'indemnité	Réparition de la génération 58	Montant d'IDR théorique moyen
Moins de 5ans	-	0	79%	-
5ans ou plus	**Moins de 10ans**	-	3%	-
	Au moins 10ans	1 mois	4%	725 €
	Au moins 15ans	1.5mois	4%	1 180 €
	Au moins 20ans	2mois	7%	1 847 €
	Au moins 30ans	2.5mois	2%	2 667 €

Tableau 1.5.1 - Montant d'indemnisation théorique pour la génération née en 1958

On retrouve l'ordre de grandeur énoncé lors de la partie descriptive relative à l'ancienneté. On a alors 17% des salariés du particulier employeur ayant un jour eu un salaire dans la branche qui auraient eu le droit à une indemnité au titre du départ volontaire à la retraite. Ce chiffre s'explique par les dynamiques que nous avons pu observer, principalement à cause du turn-over élevé et la présence discontinue.

Le nombre de mois d'indemnisation moyen par individu est de 0.3 si on prend en compte l'intégralité du portefeuille, 1.7 si on prend en compte seulement ceux qui ont le droit à au moins 1 mois de salaire de référence.

Le salaire moyen de référence observé est lui de 861€, ce qui est une nouvelle fois cohérent avec le salaire annuel moyen observé au graphique *1.4.1.1* .

f. Synthèse

Nous allons maintenant dresser la synthèse des éléments étudiés dans cette partie :

- Contexte : au 1er janvier 2022, à la suite de la renégociation des conventions collectives assurées par IRCEM Prévoyance, de nouvelles garanties viennent améliorer la protection sociale de ces publics. La garantie IDR en faisant partie, il est nécessaire de pouvoir déterminer avec précision l'ancienneté à la retraite de notre population assurée. Cette ancienneté constitue le paramètre clé pour déterminer la montant de l'indemnité à payer.
- Besoin : connaître notre portefeuille et appréhender la dynamique des carrières et l'ancienneté des assurés lors du départ en retraite.
- Étude descriptive : par les différentes données à notre disposition, et après retraitement, nous avons pu décrire et émettre des premières hypothèses sur le comportement de notre population. La part des femmes est prédominante dans notre portefeuille. Cependant, l'analyse des différences entre hommes et femmes sur les autres variables, dont le salaire et la présence en branche, nous amène à considérer la variable « Sexe » pour nos modèles.
- Nous avons pu mettre en avant certaines caractéristiques singulières de notre population assurée, comme le multi-employeur. Cette caractéristique sera aussi utilisée par la suite comme variable explicative des phénomènes des différents modèles que nous proposerons. Dans un second temps, nous avons pu voir deux dynamiques de notre population : les évolutions salariales et le turn-over. Elles représentent les deux éléments principaux du calcul de l'IDR.
- Première approche de notre problématique : nous avons réalisé une estimation rétrospective du montant de l'IDR pour une génération. Nous avons pu observer la faible part d'assurés qui auront le droit à une indemnité de départ volontaire à la retraite. Le tableau *1.5.1* nous servira alors de base pour les prochains modèles, afin de nous indiquer si les résultats sont éloignés ou non de la réalité.

On a pu voir ce qui va constituer le socle pour la suite de l'étude des carrières de la population assurée. On va maintenant introduire des méthodes permettant de modéliser notre probabilité d'acquérir de l'ancienneté et de ce fait des droits à l'indemnisation. Dans ce but, on détaillera les méthodes utilisées, les raisons de leur sélection, les hypothèses et caractéristiques ainsi que les limites de leur utilisation. On pourra, dans le cas où c'est possible et pertinent, les appliquer à nos données pour observer le comportement des particuliers employeurs dans le temps. L'objectif se rapporte donc à une problématique de durée dans un état.

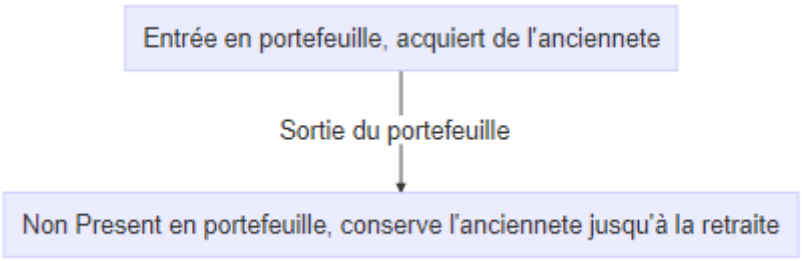
Dans un premier temps, on supposera que toute sortie de la profession est définitive (i.e. 0 salaire sur une année complète). Notre problématique sera alors celle de la durée de survie dans un état (ici la présence en portefeuille), et ce avant de passer dans un autre état définitivement (ici la non-présence en portefeuille, une retraite). On pourra dans cet objectif se rapprocher de **l'estimateur non-paramétrique de Kaplan-Meier**, et lisser les résultats à l'aide de la méthode de Whittaker Henderson. Afin de prendre en compte l'hétérogénéité de nos assurés, on utilisera **le modèle semi-paramétrique de Cox** (dit à hasard proportionnel). On cherchera à vérifier l'hypothèse sous-jacente et à contourner la possible non-vérification de la constance des rapports de risque. On se penchera alors sur **le modèle non-paramétrique d'Aalen** (dit à hasard additif). L'avantage du modèle d'Aalen est sa capacité à prendre en compte l'effet du temps sur les variables explicatives.

Dans un second temps, on ne supposera plus la non-réversibilité des états. Un individu étant sorti de l'activité pendant un nombre indéterminé d'années peut faire son retour dans la branche. Dans ce cas, le calcul de son ancienneté reprend là où celle-ci s'était arrêtée avant sa précédente sortie. Un type d'analyse se prête particulièrement bien à ce genre de modélisation : **l'analyse de séquence**. On verra alors comment représenter la carrière complète d'un individu en prenant en compte ses années de présence et de non-présence. On va ainsi venir définir des groupes-types de trajectoires professionnelles et les caractériser.

Enfin, on cherchera à modéliser la carrière à l'aide **des modèles multi-états de Markov**. On présentera les modèles homogènes, traditionnels de l'approche markovienne, puis les modèles dits inhomogènes, qui prennent en compte une évolution des intensités de transition dans le temps.

II. Modélisation par analyse de survie

L'objectif de cette partie est de modéliser le maintien en branche des individus. Nous prendrons comme hypothèse pour la modélisation qu'il n'y a pas de retour après une sortie de la profession.



Nous étudierons alors la loi de maintien dans la profession comme une loi de survie. L'analyse de survie est présente dans de nombreux domaines, du médical à l'industriel, en passant par le domaine assurantiel. L'objectif est de mesurer une durée, le temps avant un sinistre (ici une sortie). Pour cela, on transformera nos données pour se rapprocher de l'étude de survie. On obtient une ligne qui correspond à une observation de sortie du portefeuille par individu, comme montré dans le tableau 2.0.1 ci-dessous :

Id Individu	Entree en profession	Sortie	Duree	Generation	Sexe	Nombre Employeur Moyen
1	38	TRUE	2	78	F	2
2	47	TRUE	6	58	H	5
3	34	FALSE	19	68	F	1
...	...	...	...	...	...	...

Tableau 2.0.1 - Données organisées pour l'analyse de survie

Avec par individu :

- L'âge d'entrée dans la profession
- La sortie, qui indique par TRUE si la sortie est constatée, FALSE sinon
- Durée représente en année la présence continue en portefeuille
- La génération à laquelle il appartient (78 = né en 1978)
- Le sexe
- Le nombre d'employeur moyen sur la période

a. Notions Actuarielles

On cherche à étudier une durée, notée T, au terme de laquelle, dans le cadre idéal, on observe l'évènement que l'on cherche à déterminer.

Cependant, il est possible que l'événement ne se produise pas. Deux situations sont possibles, soit on perd de vue l'individu, soit nous n'avons plus d'informations car on est à la date limite de l'étude.

On parle alors de censure de nos données.

i. Censure

Soit un individu i, T_i la durée de survie observée, Y_i la durée de survie réelle et C_i, une variable aléatoire de censure.

Il existe 3 types de censure de données :

- **Censure à droite** : $T_i = \min(Y_i; C_i)$

  La durée de survie est supérieure à la durée de censure. Cette censure est la plus répandue dans les données assurantielles. Dans notre étude, ce cas est observé pour tout individu d'une génération après 1959 encore présent aujourd'hui en portefeuille. On ne constate pas la sortie avant la fin de l'étude, ou du passage à la retraite.

- **Censure à gauche** : $T_i = \max(Y_i; C_i)$

  La durée de survie est inférieure à la durée de censure. Il faudrait que l'individu soit sorti du portefeuille avant la date de début d'observation. Or ici, on considère avoir les données complètes pour tout individu jusqu'à la date d'étude. De ce fait cette censure n'est pas constatée dans nos données.

- **Censure par intervalle**

  La durée de survie est comprise dans un intervalle, composé de 2 valeurs de censure. Ce type n'est pas non plus observé dans nos données.

Toutes ces censures peuvent être classées en 3 grands groupes de censures, en prenant en exemple la censure à droite :

- **Censure de type I** : Aussi appelée censure fixe, cette censure que la variable de censure Ci est la même pour tous les individus. Soit $T_i = \min(Y_i; C)$
- **Censure de type II** : la censure au « k^{ième} événement ». Cette censure se produit lorsque l'on détermine en amont de l'étude un nombre k maximum d'événements à partir duquel on décide de stopper l'étude. $T_i = \min(Y_i; C)$, avec C=Y(k)
- **Censure de type III** : la censure aléatoire. Cette censure suppose la durée de censure aléatoire, indépendante et identiquement distribuée. On a alors l'expression générale premièrement développé, avec les C_i i.i.d.

Étant en possession de l'ensemble des données de la carrière d'un individu jusqu'à la date d'étude, on n'observe pas de troncature sur nos données.

On se place dans notre cas dans le cadre d'une **censure à droite de type III**.

ii. Fonctions et relations usuelles de l'analyse de survie

- **Fonction de répartition** : Soit F la fonction de répartition de T, alors :

$$\forall\, t \in \mathbb{R}_+, F(t) = \mathrm{P}(T \leq t)$$

- **Fonction de survie :** Soit S la fonction de survie de T, alors on a :

$$\forall\, t \in \mathbb{R}_+, S(t) = \mathrm{P}(T > t)$$
$$\text{Soit,}\quad S(t) = 1 - F(t)$$

Et la fonction de survie sachant que la personne est présente à l'âge x est :

$$\forall\, t > x, S_x(t) = \mathrm{P}(T > t \,|T > x)$$

- **Densité de probabilité** : Soit f la densité de probabilité de T, f fonction positive ou nulle intégrale sur l'ensemble des réels, on a :

$$\int_{-\infty}^{+\infty} f(u)du = 1$$

et

$$\forall\, t \in \mathbb{R}_+, F(t) = \int_0^t f(u)du$$

On cherche maintenant à déterminer la probabilité que l'individu reste en portefeuille entre t et $t + \Delta t$, sachant sa présence en portefeuille à t. Le théorème de Bayes donne :

Soit $\forall\, t, \Delta t \in \mathbb{R}_+, P(T > t + \Delta t \,|T > t) = \frac{S(t+\Delta t)}{S(t)}$

$$\Leftrightarrow S(t + \Delta t) = P(T > t + \Delta t \,|T > t) \times S(t) = S_t(t + \Delta t) \times S(t)$$

On possède aussi une autre manière de noter ces probabilités dans le cas discret :

- Soit ${}_tq_x$,la probabilité discrète de sortie de l'individu exactement à $x + t$ sachant qu'il est présent en x.

  Soit :

$${}_tq_x = \frac{P(T = x + t)}{S(x)} = 1 - \frac{P(T > x + t)}{S(x)} = 1 - \frac{S(x + t)}{S(x)} = 1 - {}_tp_x$$

Le théorème de Bayes donne : $S_x(t) = \frac{S(x+t)}{S(x)} = \frac{S(x+t)}{S(x+t-1)} \times \frac{S(x+t-1)}{S(x)}$

$$\Leftrightarrow S_x(t) = {}_1p_{x+t-1} \times S_x(t-1) = {}_1p_{x+t-1} \times {}_1p_{x+t-2} \times S_x(t-2)$$

$$\Leftrightarrow S_x(t) = \cdots = {}_1p_{x+t-1} \times {}_1p_{x+t-2} * \ldots \times {}_1p_x$$

Soit

$$S_x(t) = \prod_{1 \le i \le t} (1 - {}_1q_{x+t-i})$$

b. Estimation Kaplan Meier

L'objectif de cette partie est donc d'estimer la fonction de survie $S(t) = \mathrm{P}(T > t)$.

En absence de censure, pour tout t l'estimateur empirique serait :

$$\widehat{S_{emp}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \geq t}$$

Avec $\widehat{S_{emp}}(t)$ convergeant vers S(t) par la loi des grands nombres.

En présence de censure, cet estimateur serait cependant biaisé. Nous prendrons ici comme hypothèse forte que la censure n'apporte aucune information sur la sortie de nos assurés en portefeuille, donc que les T_i sont indépendants des C_i.

Afin de prendre en compte les effets de censures et de troncatures, on va se tourner vers l'estimateur de Kaplan-Meier.

i. Présentation de l'estimateur

L'estimateur de Kaplan Meier, (product-limit estimator), est une méthode non-paramétrique d'estimation d'une fonction de survie. Il est le plus souvent utilisé dans le cadre de constructions de table d'expérience en arrêt de travail

L'estimateur de la fonction de survie, en l'absence d'ex aequo, pour n individus présents en portefeuille à l'âge x est donné par :

$\forall\, i \in [\![1, n]\!]$,

Soit $T_{x,(i)} = \min(Y_{x,i}; C_{x,i})$, le temps avant sortie de l'individu i

Soit $p_{x,i}$ la probabilité de maintien sur l'intervalle $]\, T_{x,(i-1)}, T_{x,(i)}]$, avec $T_{x,(i-1)} < T_{x,(i)}$, sachant la présence à $T_{x,(i-1)}$, et $p_{x,i} = 1 - q_{x,i}$,

$$\widehat{S_x}(t) = \prod_{i \mid T_{x,(i)} < t} (1 - \widehat{q_{x,i}})$$

Avec

$$\widehat{q_{x,i}} = \frac{s_{x,(i)}}{n - i + 1}$$

$- s_{x,(i)} = S_x(T_{x,(i)})$, nombre d'individus sortis de notre portefeuille à $T_{x,(i)}$

ii. Estimateur de la variance

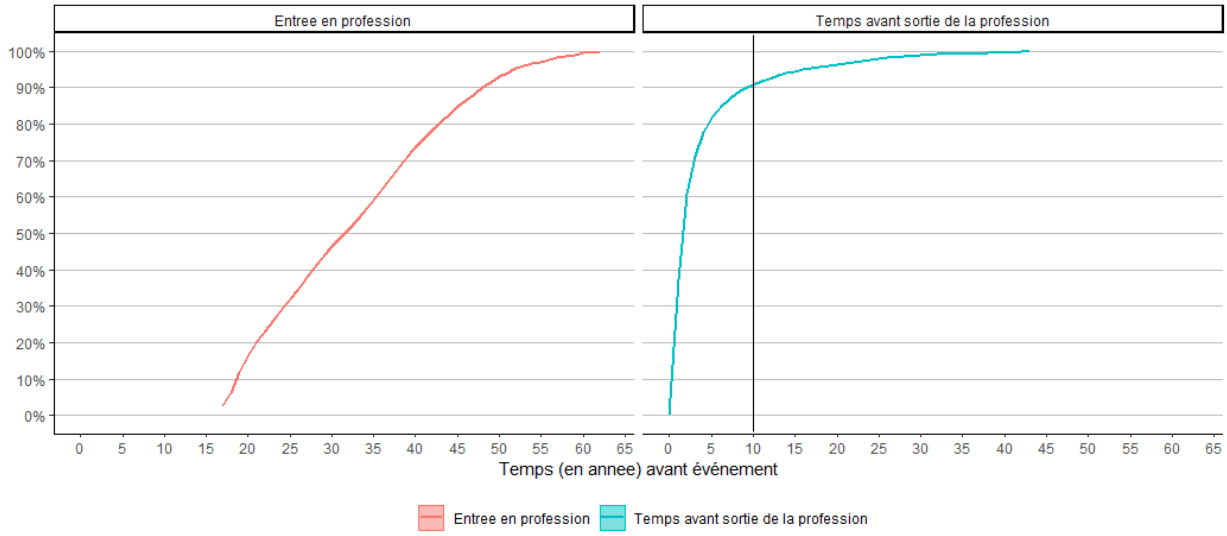
L'estimateur de Greenwood est convergeant pour la variance de l'estimateur de Kaplan-Meier.

Il nous donne :

$$\widehat{\sigma^2_{KM}}(t) = \widehat{S^2_x}(t) * \sum_{i \mid T_{x,(i)} < t} \frac{s_{x,(i)}}{(n-i+1) * (n-i+1-s_{x,(i)})}$$

iii. Présentation des Résultats

On va dans un premier temps s'intéresser à l'étude des taux bruts d'entrée et de sortie du portefeuille.

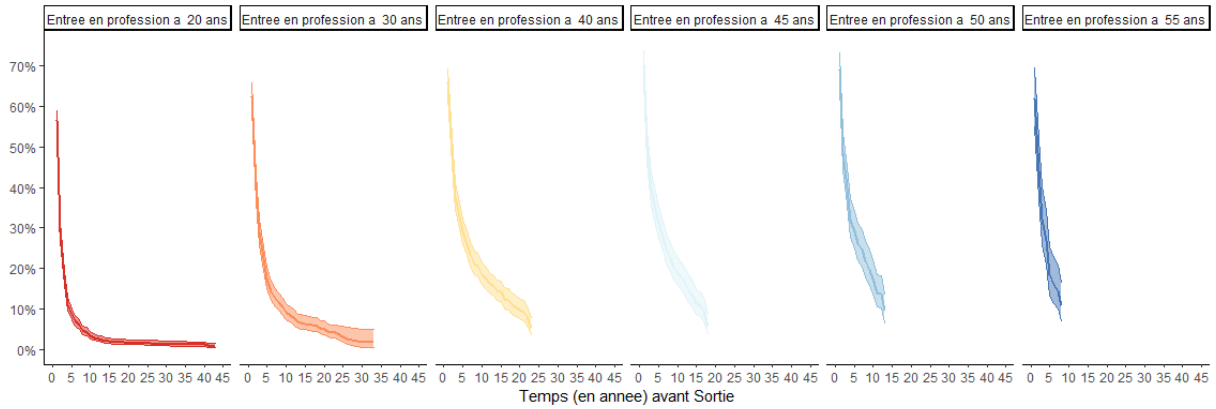


Graphique 2.2.3.1 - Modèle d'entrée en profession (à gauche) et de la sortie (à droite)

On retrouve avec ces graphiques des valeurs vues dans la partie de statistiques descriptives, comme l'ancienneté en portefeuille. La courbe d'entrée en profession nous donne la répartition de l'âge auquel un assuré a reçu son premier salaire dans la branche. On a une première entrée en portefeuille assez régulière, la médiane étant à 33 ans. On a aussi beaucoup d'entrée à des âges plus avancés, un quart ayant leur premier salaire après 41 ans, et 10% ne seront pas en portefeuille avant leurs 49 ans. Cela met bien en avant la disparité des profils au sein de l'IRCEM, et explique la courbe très tassée des sorties de la profession.

En effet, beaucoup d'assurés ayant une entrée tardive dans la profession, combiné avec le turn-over élevé (voir Graphique *1.4.2.1*) et la non-prise en compte des retours, on observe moins de 10% des salariés du particulier employeur dépassant les 10 ans d'ancienneté. Ce taux se dégrade exponentiellement pour passer à 5.5% qui atteignent les 15 années, 3.5% les 20 et seulement 1% les 30. On est loin des taux estimés dans la partie « Approche par répartition ». Cependant, cela s'explique facilement par la non-prise en compte des retours. Néanmoins, la structure reste pertinente à étudier.

Dans une analyse classique de maintien, on cherche à exprimer les taux bruts en fonction de l'âge et de l'ancienneté. On va recalculer les sorties du portefeuille en modélisant par âge à l'entrée en profession :



Graphique 2.2.3.2 - Kaplan Meier selon l'âge d'entrée dans la profession

On s'aperçoit sur le graphique 3.1.2 de la présence croissante en portefeuille en fonction de l'âge d'entrée dans la profession. Les intervalles de confiance sont des plus en plus large, du fait du nombre de données décroissantes par âge d'entrée (voir Graphique *2.2.3.1*).

On a maintenant :

		Âge d'entrée dans la profession					
		20	30	40	45	50	55
Ancienneté	5	9%	17%	29%	31%	29%	18%
	10	3%	9%	19%	19%	17%	
	13	2%	7%	15%	15%	9%	
	15	2%	6%	14%	11%		
	18	2%	6%	11%	6%		
	20	2%	5%	10%			
	23	2%	4%	5%			
	30	1%	2%				

Tableau 2.2.3.1 - Probabilité d'être présent en portefeuille en fonction de l'âge d'entrée et de l'ancienneté

On a récupéré les taux importants dans notre cadre d'étude selon certains âges. Ils nous permettent une analyse plus précise en remarquant la croissance des taux bruts en fonction de l'âge d'entrée jusque 40-45 ans. Ensuite ils vont avoir tendance à légèrement décroitre (voir annexe pour la table complète).

iv. Lissage Whittaker Henderson

Le lissage de Whittaker Henderson est la méthode de lissage usuelle pour l'estimateur de Kaplan Meier. L'objectif d'un lissage est de donner une forme plus présentable, plus droite, aux taux et de réduire l'erreur entre valeur estimée et valeur réelle. On cherche à s'éloigner des effets dus au portefeuille et de se rapproche d'une approche plus générale.

Le lissage de Whittaker Henderson cherche à minimiser 2 critères :

- F, critère de fidélité, qui a pour but de mesurer la qualité de l'ajustement, noté

$$F = \sum_{x_{min}}^{x_{max}} w_x(q_x - \widehat{q_x}))^2$$

- Rv, le critère de régularité vertical, et Rh le critère de régularité horizontal :

$R_z = \sum_{x_{min}}^{x_{max}-z}(\Delta^z q_x)^2$

Avec :

- $\Delta^z f(x) = \sum_{j=0}^{z} \binom{z}{j} (-1)^{z-j} f(x+j)$, opérateur de différenciation obtenu par récurrence de $\forall x, \Delta f(x) = f(x+1) - f(x)$
- z paramètre contrôlant le lissage des taux
- $q_x \; et \; \hat{q}_x$ respectivement taux lissés et taux bruts
- w_x, les poids, associés au nombre d'assurés exposés au risque

On cherche alors à minimer $M = F + h * R_z$, par rapport à q_x, de la manière suivant (Planchet [8]) :

- Notons W =diag (w_x)
- $q = (q_i)_{x_{min} \leq i \leq x_{max}}$ et $\hat{q} = (\hat{q}_i)_{x_{min} \leq i \leq x_{max}}$, respectivement vecteur des taux lissés et vecteur des taux bruts

On peut réécrire :

$$F = (q - \hat{q})' * W * (q - \hat{q})$$

$$R = (\Delta^z q)' \Delta^z q$$

On introduit K_z la matrice de taille $(p-z, p)$ telle que $\Delta^z q = K_z q$

M se reformule : $M = (q - \hat{q})' * W * (q - \hat{q}) + h \times q' * K_z' * K_z * q$

$$= (\hat{q}' * W * q) - (2 \times q * W * \hat{q}) + (\hat{q}' * W * \hat{q}) + (h \times q' K_z' K_z q)$$

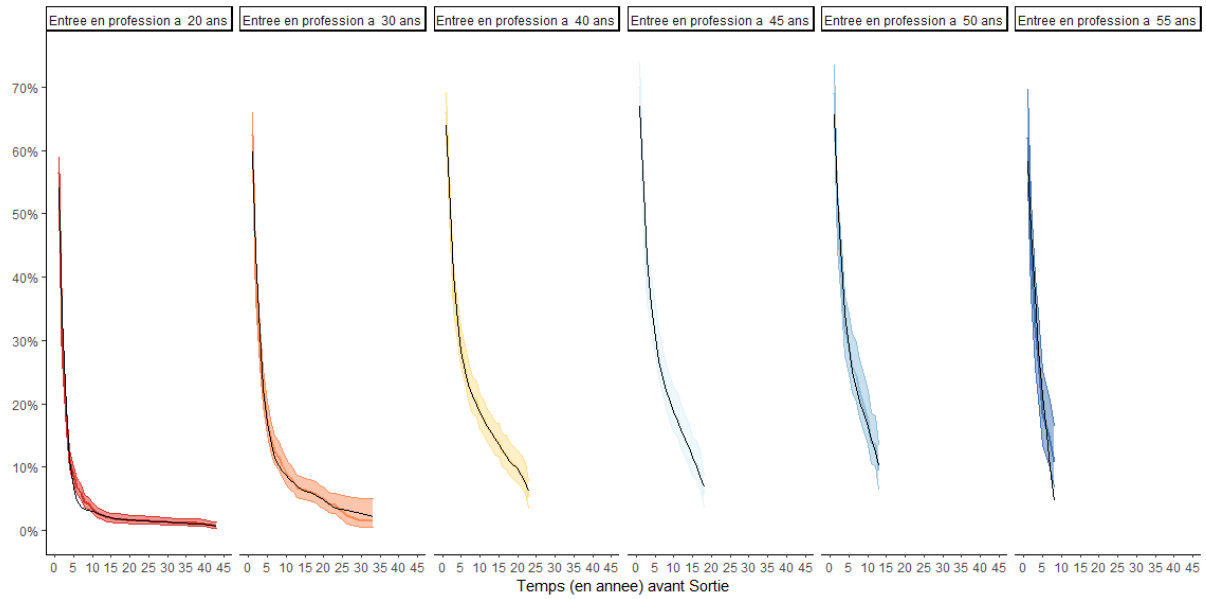
Trouver le minimum revient à résoudre : $\frac{\partial M}{\partial q} = 0 = 2 * W * q - 2 * W * \hat{q} + h \times K_z' K_z q$

On a donc, dans le cas où $W + h * K_z' K_z * q$ est inversible,

$$q = (W + h * K_z' K_z)^{-1} * W * \hat{q}$$

v. Application à nos données

On décidera, après lecture graphique, de choisir les paramètres z=2 et h=1. Les poids seront déterminés comme l'exposition à l'âge x rapportés à l'exposition moyenne par âge.



Graphique 2.2.5.1 - Résultat du lissage par la méthode de Whittaker-Henderson par âge

Dans l'ensemble, le lissage suit plutôt bien les taux bruts dans la limite des intervalles de confiance, sauf légèrement entre 5 et 10 ans pour l'entrée en profession à 20 ans.

vi. Segmentation

Afin de mieux mesurer notre risque de maintien en activité pour nos assurés, il est important de constituer des profils de risques. Ces derniers vont permettre, à défaut de construire un modèle par assuré qui n'aurait pas de sens, de grouper selon une ou plusieurs caractéristiques les salariés du particuliers employeurs. On cherche ainsi à créer une réponse différente selon les groupes étudiés. On peut alors faire une différenciation selon le sexe, la catégorie socio-professionnelle, la situation familiale etc… Dans notre cas, on utilisera la variable sexe, couramment utilisée lors de la segmentation en assurance, ainsi qu'une spécificité de notre population, à savoir le nombre d'employeur.

La comparaison du maintien dans plusieurs groupes peut s'effectuer à l'aide du test du log-rank ou du test de Wilcoxon. On utilisera dans notre étude le test du log-rank.

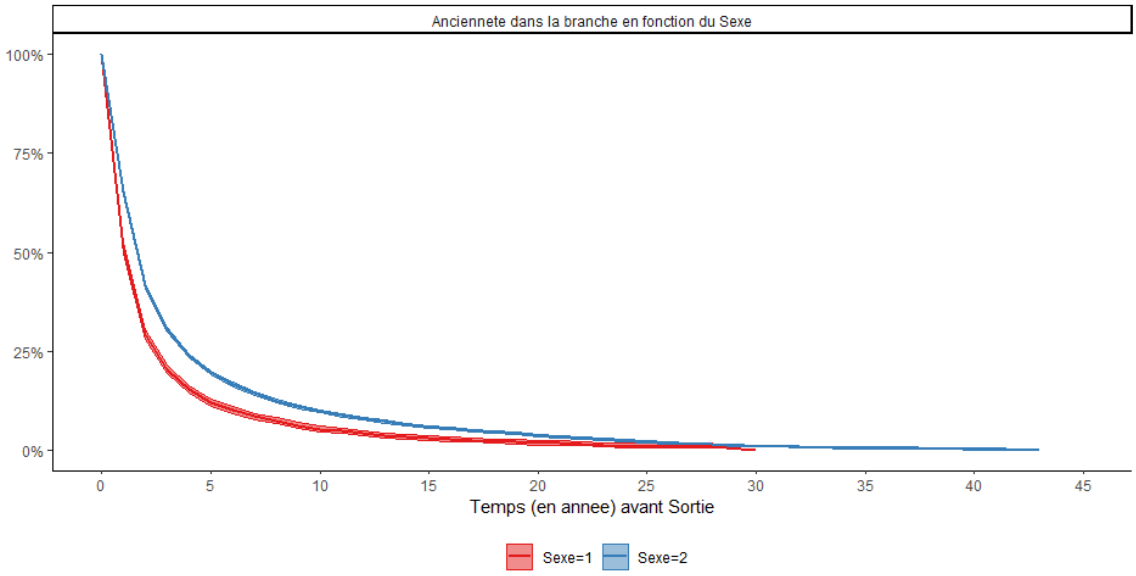
Le test du log-rank consiste à tester l'égalité entre les fonctions de maintien (hypothèse nulle) :

$$LR = \left(\frac{Obs_i - Pred_i}{Pred_i}\right)^2 + \left(\frac{Obs_j - Pred_j}{Pred_j}\right)^2 \sim \chi^2(1)$$

L'hypothèse nulle est rejetée si la p-value observée est inférieure au seuil α ($ici\ = 0.05$).

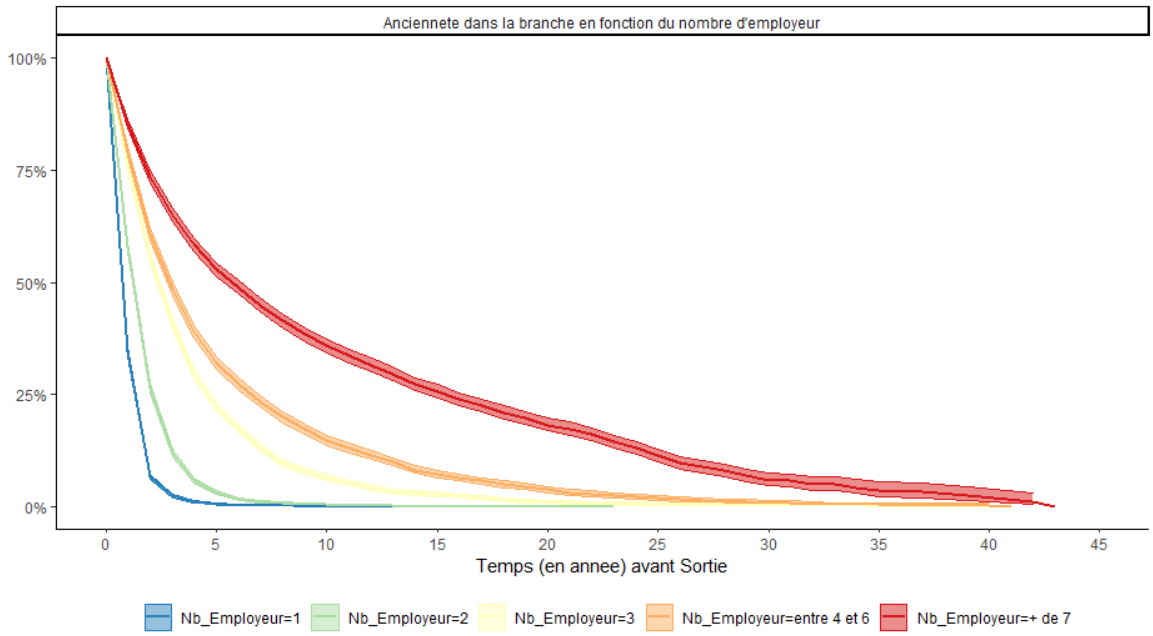
Pour les deux variables, le sexe et le nombre d'employeur, la p-value est <2 e-16.

On rejette donc l'hypothèse nulle et les taux bruts nous donnent :



Graphique 2.2.6.1 - Estimation des courbes de sorties par sexe

Les taux de maintien des femmes (sexe=2) sont en tout temps supérieurs à ceux des hommes. On retrouve l'observation faite dans la partie descriptive que les femmes ont plus tendances à rester dans la profession et à avoir un profil de carrière long.



Graphique 2.2.6.2 - Estimation des courbes de sorties par le nombre d'employeur

De même, on observe que les taux sont croissants en fonction du nombre d'employeur.

On peut, dans un second temps, s'intéresser au croisement de ces deux variables. Ceci permettra de détecter des profils différents que l'analyse séparée ne permet pas de mettre en avant.

Tout d'abord, on regarde la corrélation entre les deux variables. Le tau de Kendall est un test statistique de corrélation basé sur un test de rang. Il peut être utilisé lorsque les données ne proviennemnt pas d'une distribution normale.

Il est calculé comme suit :

Soit l'ensemble des observations uniques possibles des variables X et Y : $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. On définit les paires comme :

- Concordantes si $\forall i, j \in [1, n]$ $\frac{x_i - x_j}{y_i - y_j} > 0$
- Discordantes si $\forall i, j \in [1, n]$ $\frac{x_i - x_j}{y_i - y_j} < 0$
- Si $x_i = x_j \; ou \; y_i = y_j$, on ne prend pas en compte le couple.

On définit alors le tau de Kendall comme :

$$\tau = 2 * \frac{(nombre\ paires\ concordantes) - (nombre\ de\ paires\ discordantes}{n * (n-1)}$$

Et sa variance est donnée par :

$$Var(\tau) = \frac{2 * (2 * n + 5)}{9 * n * (n - 1)}$$

La statistique du test de nullité du coefficient de corrélation donne :

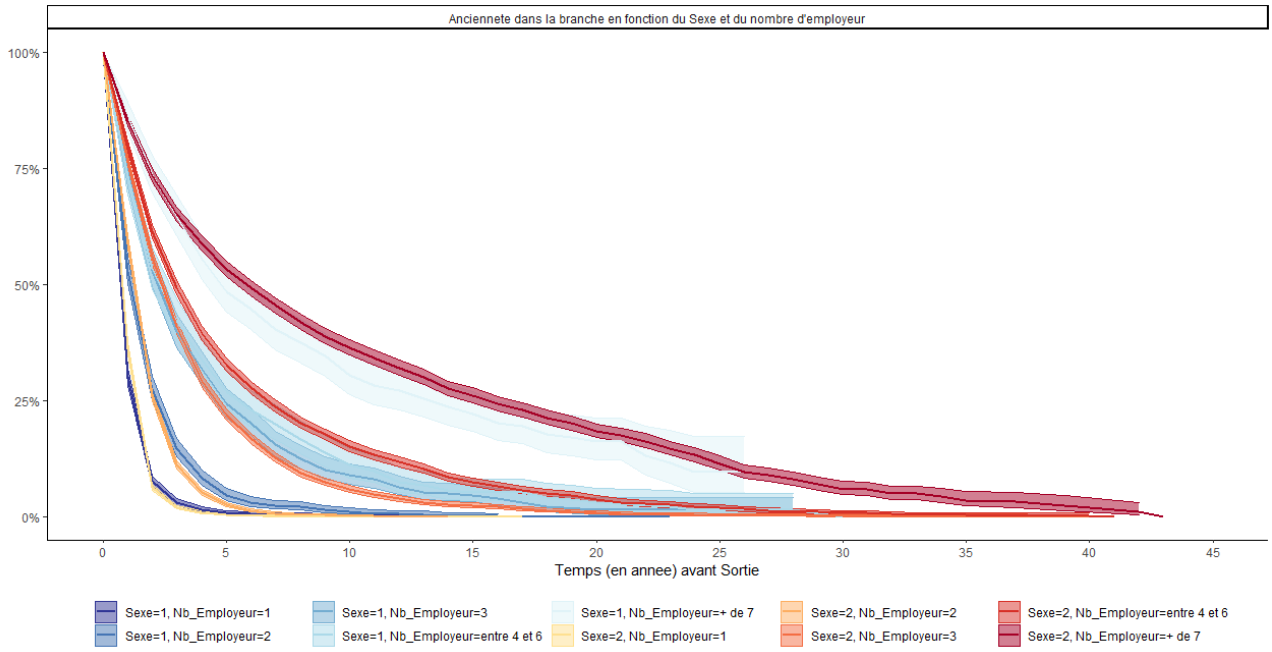
$$z = \frac{\tau}{\sqrt{Var(\tau)}} \sim N(0; 1)$$

Le résultat du test donne :

```
        Kendall's rank correlation tau

data:  as.numeric(Surv_Sortie$nb_employ_Moy_gr) and as.numeric(Surv_Sortie$Sexe)
z = 30.647, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.162069
```

L'hypothèse nulle est rejetée, on est donc dans l'hypothèse alternative où tau n'est pas significativement égal à 0. La corrélation observée est néanmoins faible.

On décide d'afficher la distribution en fonction de ces deux variables :



Graphique 2.2.6.3 - Estimation des courbes de sorties par sexe et nombre d'employeur

Les courbes respectent l'ordre pour le nombre d'employeur, où toutes les courbes avec un nombre d'employeur plus élevée sont au-dessus de celles avec un nombre plus faible.
Cependant, ce n'est pas le cas pour la variable sexe. En effet, individuellement, la modalité femme (Sexe=2) était en tout point supérieur à la modalité homme (Sexe=1). Or ici, on observe la tendance inverse pour un nombre inférieur ou égal à 3 employeurs, avec une courbe des hommes bien en dessous pour une ancienneté courte (<5) mais qui passe au-dessous après 4 années d'ancienneté.

Cette différence vient principalement de la répartition du nombre d'employeur moyen chez les femmes, avec 35% qui ont plus de 3 employeurs, là où les hommes sont à 19%.

On cherche alors à modéliser l'impact des covariables sur le maintien de la présence de notre population dans la branche. Pour cela, on utilisera dans une première partie un modèle semi-paramétrique, le modèle de Cox dit à « risques proportionnels ». Puis on déterminera si l'hypothèse sous-jacente à l'utilisation de ce modèle, à savoir l'hypothèse de proportionnalité des risques, est vérifiée. Dans un second temps, on s'emploiera à utiliser un modèle à « risques additifs », le modèle d'Aalen, qui permet de prendre en compte un effet des covariables dépendant du temps.

c. Cox

Le but de ces parties est d'expliciter l'impact des différentes variables sur le temps de maintien dans la branche. Le modèle de Cox à risques proportionnels est un des modèles les plus utilisés dans l'analyse de survie classique pour prendre en compte l'hétérogénéité. L'hétérogénéité prend en compte les particularités d'un individu et de groupes d'individus dans notre étude. La non-prise en compte de l'hétérogénéité induit des biais dans la modélisation. Son observation est donc indispensable dans un cadre assurantiel compétitif.

On peut distinguer deux types d'hétérogénéités :

- L'hétérogénéité observée, par des covariables, qui permet d'expliquer les différences de survie (ici de maintien) d'une population selon des facteurs dits observables (géographique, sexe…).
- L'hétérogénéité non-observée, ou cachée. On part du principe que l'hétérogénéité ne peut pas être seulement expliquée par des facteurs observables.

On supposera ici seulement l'hétérogénéité observée. La non-observée est cependant très intéressante et pourra être proposée en axe d'amélioration de nos modèles.

iii. Modèle à risques proportionnels

On introduit ici :

- Soit le taux de hasard h, la probabilité de sortie un court laps de temps après une présence en t :

  Soit

$$\forall\, t \in \mathbb{R}_+, \qquad h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t |\, T \geq t)}{\Delta t} = -\frac{\partial S(t)}{\partial t}$$

Dans le cas discret, on a : $h(t) = P(T = t \mid T \geq t)$

- Soit le taux de hasard cumulé Υ,

$$\forall\, t \in \mathbb{R}_+, \qquad \Upsilon(t) = \int_{-\infty}^{t} h(u)\partial u = -\ln\left(S(t)\right)$$

Dans le cas discret, on obtient :

$$\forall\, t \in \mathbb{R}_+, \qquad \Upsilon(t) = \sum_{i|u_i \leq t} h(u)$$

- Dans le cadre des modèles à hasard proportionnel, soit Z les covariables, on a :

$\forall\, t \in \mathbb{R}_+,\ h(t|Z = z) = \phi(z, \beta) * h_0(t)$, avec :

- β paramètre d'intérêt pour chaque covariable sur la fonction de hasard,
- ϕ représente l'effet du vecteur des covariables par rapport au profil de référence. On a alors : $\phi(0, \beta) = 1$

iv. Présentation du Modèle

Le modèle de Cox est une application des modèles à risques proportionnels à base exponentielle.

On note alors :

$$\forall\, t \in \mathbb{R}_+, \qquad h(t|Z) = h_0(t) * \ exp(\beta' * z),$$

Avec :

- β vecteur des paramètres d'intérêt, ne dépendant pas du temps
- z vecteur de variables explicatives

L'hypothèse de hasard proportionnel, indispensable à la bonne application du modèle, est visible en faisant le rapport de deux sous-groupes différents selon les covariables considérés.

On a dans le cadre du modèle de Cox :

$$\forall\, i, j \in \mathbb{R}_+, \qquad \frac{h(t|Z_i)}{h(t|Z_j)} = exp\big(\beta' * (z_i - z_j\big)$$

La méthode proposée par Cox [1973] permet d'éliminer la fonction de hasard h_0 du terme de la vraisemblance partielle :

$$L_{Cox}(\beta) = \prod_{i=1}^{r} \left(\frac{e^{\beta' z_{(i)}}}{\sum_{k \in R_{(i)}} e^{\beta' z_{(k)}}}\right)$$

,avec r le nombre de décès observés

Une estimation de β se fait par dérivation de la log-vraisemblance,

$$\frac{\partial L_{Cox}(\beta)}{\partial \beta} = 0$$

Enfin, le taux de hasard cumulé de base se détermine grâce à l'estimateur de Breslow :

$$\widehat{\Upsilon}_0(t) \ = \sum_{i|u_i \le t} \frac{d_i}{\sum_{k \in R_{(i)}} e^{\widehat{\beta}' z_{(k)}}}$$

Et la relation entre taux de hasard et fonction de survie donne :

$$\hat{S}(t) = \exp\left(-\hat{\Upsilon}_0(t) * \exp\left(\hat{\beta}' z\right)\right) = \hat{S}_0(t)^{\exp(\hat{\beta}' z)}$$

L'effet d'une covariable est exp(β_i), mesurant l'impact à la hausse ou à la baisse (ou même impact si exp(β_i)=1) selon la modalité choisie et la modalité de référence.

v. Application à nos données

On cherche alors à étudier l'impact des différentes covariables sur nos données. On crée alors cinq modèles combinaisons des covariables suivantes :

- Le Sexe
- Le nombre d'employeur moyen
- L'âge d'entrée en profession

De plus, le regroupement des modalités, comme pour le nombre d'employeurs, permet de ne pas prendre en compte un trop petit nombre d'observation par sous-groupe de notre population et d'améliorer la qualité de nos modèles.

La sortie (*Tableau 3.2.1*) de la fonction coxph du package survival sous R donne la p-value pour les modalités, où l'hypothèse nulle est que les coefficients β_j du modèle sont significativement différents de zéro.

```
                             coef  exp(coef)   se(coef)       z Pr(>|z|)
Sexe2                   -0.0517806  0.9495371  0.0165825  -3.123  0.00179 **
Nb_Employeur2           -0.5513734  0.5761580  0.0175601 -31.399  < 2e-16 ***
Nb_Employeur3           -1.2860864  0.2763502  0.0198981 -64.634  < 2e-16 ***
Nb_Employeurentre 4 et 6 -1.5632846  0.2094470  0.0214272 -72.958  < 2e-16 ***
Nb_Employeur+ de 7      -2.1757158  0.1135269  0.0244028 -89.158  < 2e-16 ***
Entree_Profession       -0.0092299  0.9908125  0.0006156 -14.993  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                        exp(coef) exp(-coef) lower .95 upper .95
Sexe2                      0.9495      1.053    0.9192    0.9809
Nb_Employeur2              0.5762      1.736    0.5567    0.5963
Nb_Employeur3              0.2764      3.619    0.2658    0.2873
Nb_Employeurentre 4 et 6   0.2094      4.774    0.2008    0.2184
Nb_Employeur+ de 7         0.1135      8.808    0.1082    0.1191
Entree_Profession          0.9908      1.009    0.9896    0.9920

Concordance= 0.737  (se = 0.002 )
Likelihood ratio test= 11405  on 6 df,   p=<2e-16
Wald test            = 10740  on 6 df,   p=<2e-16
Score (logrank) test = 12226  on 6 df,   p=<2e-16
```

Tableau 2.3.5.1 - Sortie du modèle de Cox prenant en compte nos trois variables

Tout d'abord, l'ensemble des modalités ont une p-value suffisamment faible pour rejeter l'hypothèse nulle. On remarque que celle du sexe est plus élevée que les autres. On s'intéresse donc à l'intérêt de la conserver par rapport à son gain dans notre modèle.

On crée alors un nouveau modèle de Cox sans cette variable, et on récupère les critères AIC et BIC pour les deux modèles. Le critère d'information d'Akaike (AIC) et le critère d'information Bayésien (BIC) permettent de tester l'impact des variables sur le modèle.

$$AIC = 2k - 2ln(L)$$

Avec k le nombre de paramètre du modèle et L le maximum de la fonction de vraisemblance du modèle

$$BIC = -2\ln(L) + \ln(n)\, k$$

Avec n le nombre d'observations de l'échantillon Ces critères bien que proches ne mesurent pas l'impact des variables de la même manière. Le AIC va moins pénaliser le nombre de paramètre que le BIC, ce qui a pour effet de limiter le surapprentissage

	~ Age d'entrée dans la profession + Nombre d'employeur	~ Age d'entrée dans la profession + Nombre d'employeur + Sexe
AIC	473881	473873
BIC	473921.9	473922.4

Tableau 2.3.5.2 - Critère AIC et BIC selon le modèle utilisé

Le critère d'Akaike est plus faible dans le cas d'utilisation du sexe, mais le BIC est plus faible dans le cas de sa non-sélection. Pour départager, on va étudier la contribution de ces variables sur la vraisemblance de notre modèle selon le test du Chi-Deux :

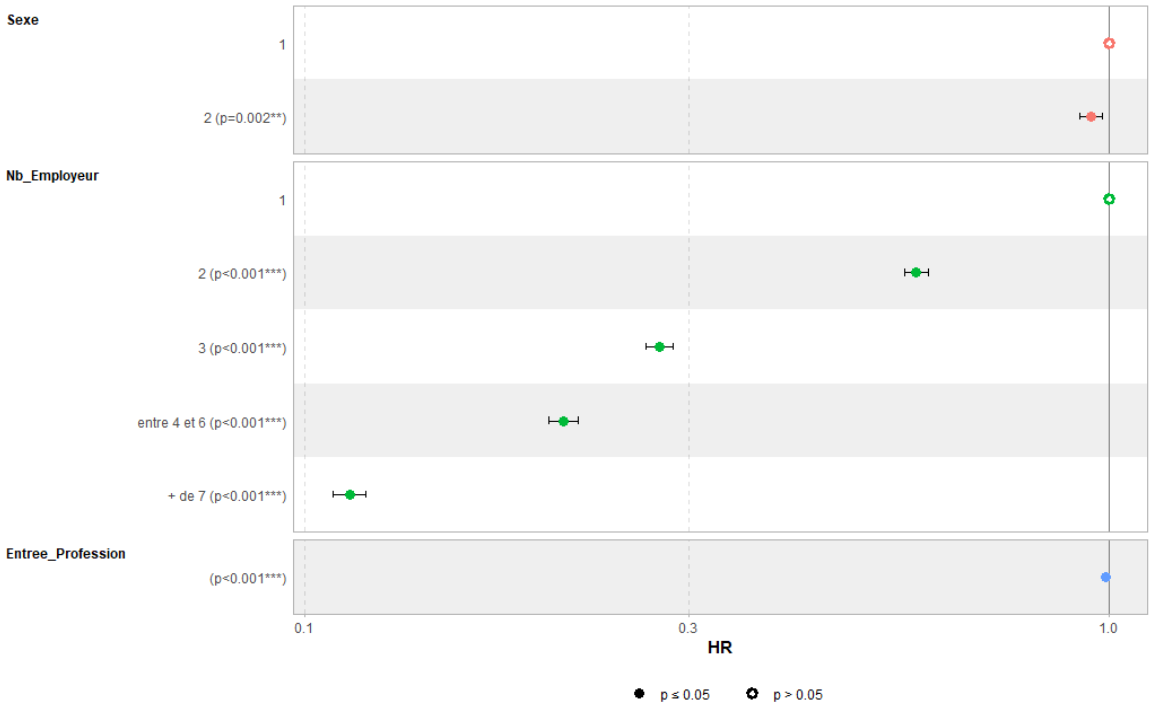
	Chisq	P-Value
Sexe	323.04	< 2.2e-16
Nb_Employeur	10854	< 2.2e-16
Age Entrée Profession	227.72	< 2.2e-16

Tableau 2.3.5.3 - Contribution des covariables au modèle

La variable nombre d'employeurs est celle qui contribue le plus à la vraisemblance du modèle. On remarque que la variable Sexe est la deuxième contributrice à notre modèle, devant l'âge d'entrée en profession. On décidera finalement de conserver les trois covariables pour notre modélisation.

D'autres modèles ont été essayés, sans amélioration de nos critères.

Une analyse du modèle de Cox possible se situe au niveau des ratio de risques instantanées.



Graphique 2.3.5.1 - Ratio de risques instantanées (Hazard Ratio) du modèle de Cox

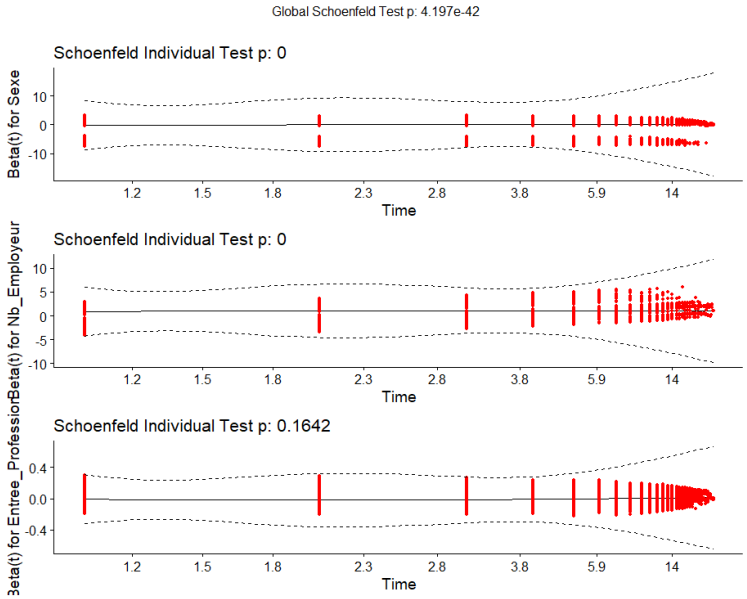
L'individu de référence dans ce cas est un homme avec une entrée en profession de 18 ans et qui possède un employeur. On garde bien que l'ensemble des modalités ont un coefficient β différent de 0 (donc un exponentiel β différent de 1). On retrouve bien les tendances observées par les courbes de maintien de la partie segmentation. Dans le cas de la covariable Sexe, l'exponentiel du coefficient associé aux femmes est de 0.949. Cela signifie que les hommes (modalités de références) ont un risque 5.1% plus élevé que les femmes, ce ratio étant supposé constant dans le temps.

On doit maintenant vérifier notre hypothèse de risques proportionnels dans le temps. C'est cette vérification qui permettra de déterminer si le modèle de Cox est utilisable ou non. On s'intéresse alors au résidu de Schoenfeld.
Le test des résidus de Schoenfeld pose comme hypothèse nulle :

$$H_0 : z_j(t) = z_j \text{ contre } H_1 : z_j(t) \neq z_j$$

Le but est, pour chaque sortie s_i, de mesurer la différence entre *la valeur de la covariable j en* T_{s_i} et *une moyenne pondérée des valeurs de cette covariable sur l'ensemble des sujets à risque en* T_{s_i}.

Les résidus de Schoenfeld de notre modèle de Cox donnent :

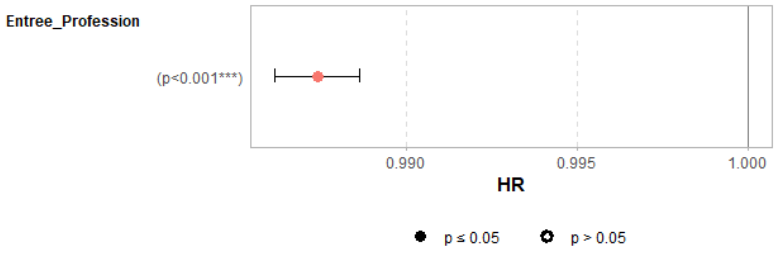


Graphique 2.3.5.2 - Analyse des résidus du modèle de Cox

On a bien pour chaque covariable étudiée une p-value associée aux résidus. Cette p-value permet de rejeter ou non l'hypothèse nulle. Ainsi, on peut rejeter l'hypothèse nulle pour les variables sexe et nombre d'employeurs. Le risque associé à ces covariables n'est donc pas proportionnel dans le temps, comme le montre la tendance à la baisse en fin de répartition des résidus pour le sexe.

La variable d'âge d'entrée en profession elle a une p-value>0.05, elle est donc significativement à risque proportionnel dans le temps. Le modèle de Cox peut donc s'appliquer dans le cas de la segmentation par âge d'entrée en profession, mais pas pour la différenciation Homme/Femme ou par nombre d'employeurs.

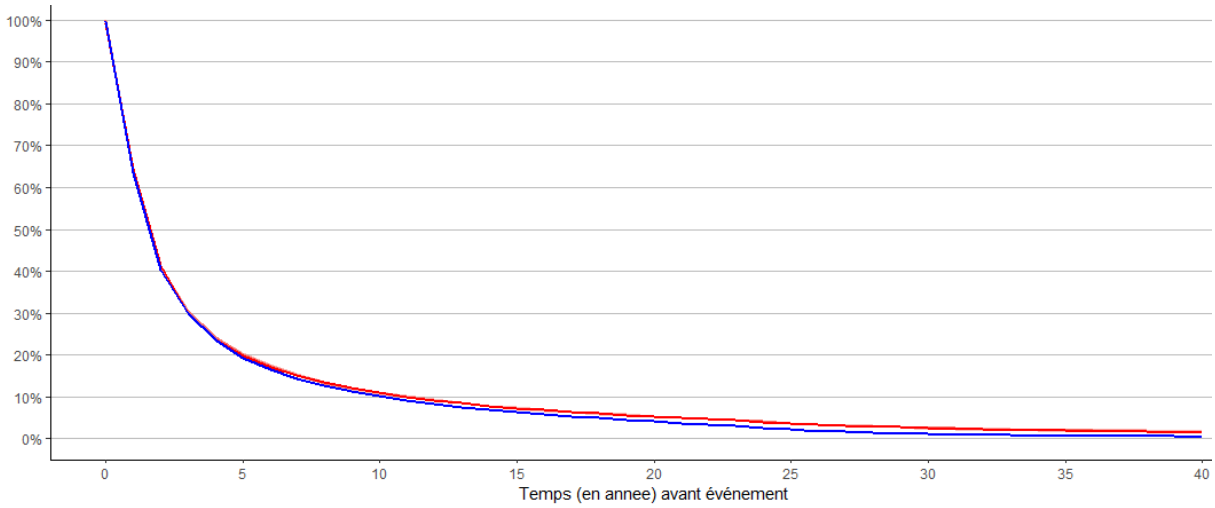
On se concentre sur la seule covariable d'âge d'entrée en profession. Le ratio de risques instantanés est alors :



Graphique 2.3.5.3 - Ratio de risques instantanées (Hazard Ratio) du modèle de Cox avec âge d'entrée en profession

Cela signifie que pour chaque année d'entrée plus tardive dans la profession, le risque est multiplié par 0.987, proportionnellement selon l'ancienneté, par rapport à une entrée à 18 ans.

La courbe de maintien appliquée à nos données donne :



Graphique 2.3.5.4 - Courbe de survie modèle de Cox (en bleu) et de Kaplan-Meier (rouge)

On a un maintien toujours en dessous de ce à quoi on s'attend et semblable à la courbe de maintien par l'estimateur de Kaplan Meier.

L'hypothèse de proportionnalité du risque au cours du temps n'est pas toujours vérifiée et trop restrictive. Il nous faut alors partir sur un modèle qui permet de prendre en compte un effet dépendant du temps.

Dans cet objectif, il existe des modèles de Cox avec effets dépendant du temps (implémenté dans le package timereg avec la fonction timecox()).

On peut aussi utiliser une méthode de partition du temps et estimer les risques proportionnels par morceaux (Gamerman (1991)).

On construit alors n intervalles tel que $\forall\, i \in [1, n], t_i > t_{i-1}$ et $t_0 = 0$.

On note I_i l'intervalle $]\, t_{i-1}; t_i]$.

La fonction de hasard se note alors :

$$\forall\, t \in I_i, h(t) = exp({\beta_i}' * z)$$

Et

$$\beta_i = \beta_{i-1} + \mathcal{E}_i$$

avec

$$\mathcal{E}_i \sim \mathcal{N}(0, \sigma_i^2)$$

Néanmoins dans notre cas, on se tournera vers le modèle à risques dits additifs.

d. Modèle Aalen Additif

Un des atouts du modèle non paramétrique additif d'Aalen est de définir l'effet de chaque covariables de manière additive en fonction du temps.

i. Présentation du modèle

On se base dans cette partie sur les travaux de Teodorescu, Van Keilegom et Cao [2005]

Le taux de hasard est défini comme :

$$\forall\, t \in \mathbb{R}_+, \qquad h\big(t|Z = z_j(t)\big) = \beta_0(t) + \sum_{k=1}^{p} \big(\beta_k(t) * z_{j,k}(t)\big),$$

Avec : $z_j(t) = (\, z_{j,1}(t), \dots, z_{j,\mathrm{p}}(t))$ un vecteur de variables pouvant dépendre du temps.

- β vecteur des paramètres d'intérêt, dépendant du temps
- z_j vecteur de variables explicatives, dépendant aussi du temps

On nomme ce modèle additif car le taux de hasard au temps t est somme de $\beta_0(t)$ et de la combinaison linéaire des covariables z_j avec les coefficients β.

La fonction de survie est estimée par :

$$S(t) = \exp\,(-\int_0^t (\beta_0(u) + \sum_{k=1}^{p} \beta_{\mathrm{k}}(u) * z_{j,k}(u))\partial u$$

Il reste à déterminer $\forall\, t \in \mathbb{R}_+ \; les\; \beta_{\mathrm{k}}(t)$.

Posons : $B_{\mathrm{k}}(t) = \int_0^t \beta_{\mathrm{k}}(u)\, \partial u$ et $B(t) = (B_0(t), \dots, B_{\mathrm{p}}(t))$.

On admet qu'il est montré qu'un estimateur de B est :

$$\hat{B}(t) = \sum_{T_i \le t} [X'(T_i)X(T_i)]^{-1}\, X'(T_i)I(T_i)$$

Avec :

- $X(t)$ matrice de dimension n*(p+1) (n étant le nombre d'individu à risques, p le nombre de covariables). La ligne i de X(t) vaut $(1, z_{i,1}(t), \dots, z_{i,p}(t))$ si l'individu i est sous risque à la date t, 0 sinon.
- $\mathrm{I}(T_i)$ un vecteur de dimension n*1 avec l'élément i valant 1 si le sujet est sous risque à T_i, 0 sinon.

ii. Application à nos données

On reprend le modèle calculé pour la méthode de Cox, avec les covariables suivantes :

- Le sexe
- Le nombre d'employeur moyen
- L'âge d'entrée en profession

On suppose dans un premier temps ces trois variables comme ayant un effet dépendant du temps.

Le package TimeReg de R permet l'étude des coefficients significatifs et de l'effet variant avec le temps.

On a :

```
Test for non-significant effects
                      Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                                     Inf                   0
Nb_Employeur2                                   Inf                   0
Nb_Employeur3                                   Inf                   0
Nb_Employeurentre 4 et 6                        Inf                   0
Nb_Employeur+ de 7                              Inf                   0
Sexe2                                           Inf                   0
Entree_Profession                               Inf                   0

Test for time invariant effects
                           Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                                6.0600                        0.000
Nb_Employeur2                              1.7200                        0.077
Nb_Employeur3                              3.3800                        0.000
Nb_Employeurentre 4 et 6                   4.0500                        0.000
Nb_Employeur+ de 7                         4.7000                        0.000
Sexe2                                      0.1290                        0.441
Entree_Profession                          0.0266                        0.000
                            Cramer von Mises test p-value H_0:constant effect
(Intercept)                              6.27e+02                        0.000
Nb_Employeur2                            2.48e+01                        0.002
Nb_Employeur3                            2.05e+02                        0.000
Nb_Employeurentre 4 et 6                 2.88e+02                        0.000
Nb_Employeur+ de 7                       4.02e+02                        0.000
Sexe2                                    6.75e-02                        0.357
Entree_Profession                        5.69e-03                        0.000
```

Tableau 3.3.1 : Sortie du modèle d'Aalen prenant en compte nos trois variables

Toutes les variables présentes ont un effet non-nul sur le modèle. Le test de Kolmogorov-Smirnov (de validité d'une fonction de distribution) rejette l'impact relatif au temps de la variable Sexe (comme vu pour le modèle de Cox) ainsi que pour la modalité 2 du nombre d'employeurs. On utilisera alors dorénavant la covariable sexe comme invariante par rapport au temps.

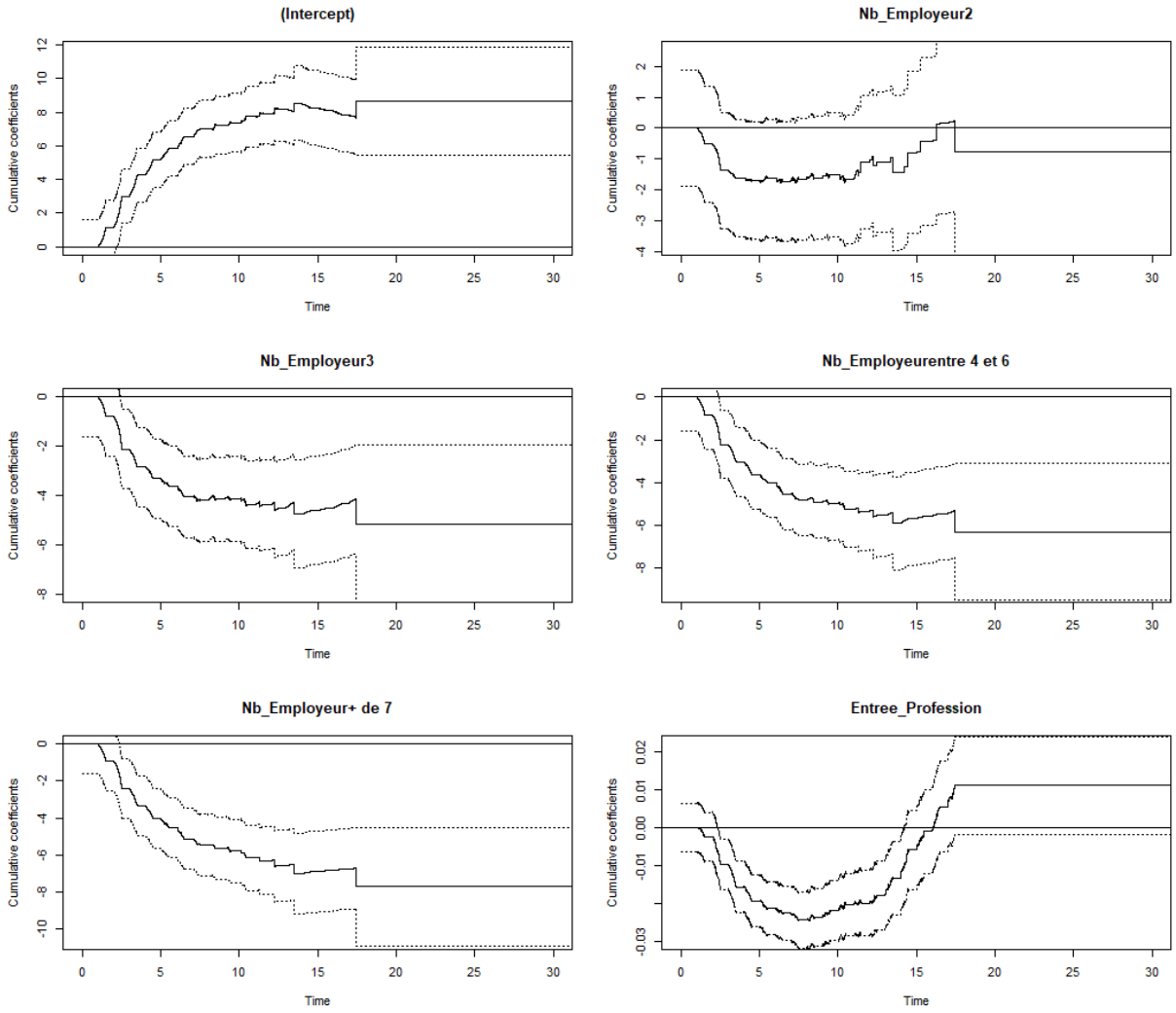
On obtient :

```
Test for non-significant effects
                        Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                                      57.6                    0
Nb_Employeur2                                    29.1                    0
Nb_Employeur3                                    47.5                    0
Nb_Employeurentre 4 et 6                         50.6                    0
Nb_Employeur+ de 7                               57.1                    0
Entree_Profession                                14.8                    0

Test for time invariant effects
                                 Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                                       5.8300                         0.00
Nb_Employeur2                                     1.6700                         0.01
Nb_Employeur3                                     3.3600                         0.00
Nb_Employeurentre 4 et 6                          3.9400                         0.00
Nb_Employeur+ de 7                                4.6000                         0.00
Entree_Profession                                 0.0268                         0.00
                                   Cramer von Mises test p-value H_0:constant effect
(Intercept)                                    5.96e+02                        0.000
Nb_Employeur2                                  2.54e+01                        0.003
Nb_Employeur3                                  2.01e+02                        0.000
Nb_Employeurentre 4 et 6                       2.85e+02                        0.000
Nb_Employeur+ de 7                             3.99e+02                        0.000
Entree_Profession                              6.13e-03                        0.000

Parametric terms :
              Coef.       SE Robust SE    z    P-val lower2.5% upper97.5%
const(Sexe)2 0.0158 0.00407   0.00452 3.49 0.000476   0.00782     0.0238

  Call:
aalen(formula = Surv(Duree, Sortie) ~ Nb_Employeur + const(Sexe) +
    Entree_Profession, data = Surv_Sortie)
```

Tableau 3.3.1 - Sortie du modèle d'Aalen prenant en compte nos trois variables, dont le sexe constant par rapport à t

On a bien nos tests significatifs pour toutes nos covariables. La variable sexe a maintenant un effet constant par rapport au temps avec $\hat{\beta}$ = 0.0158 et ce coefficient est significatif (p-value <0.0005). On remarque que cette transformation rend la modalité nombre d'employeurs=2 significativement variante avec le temps.

On veut maintenant observer la tendance de nos autres modalités par rapport au temps.



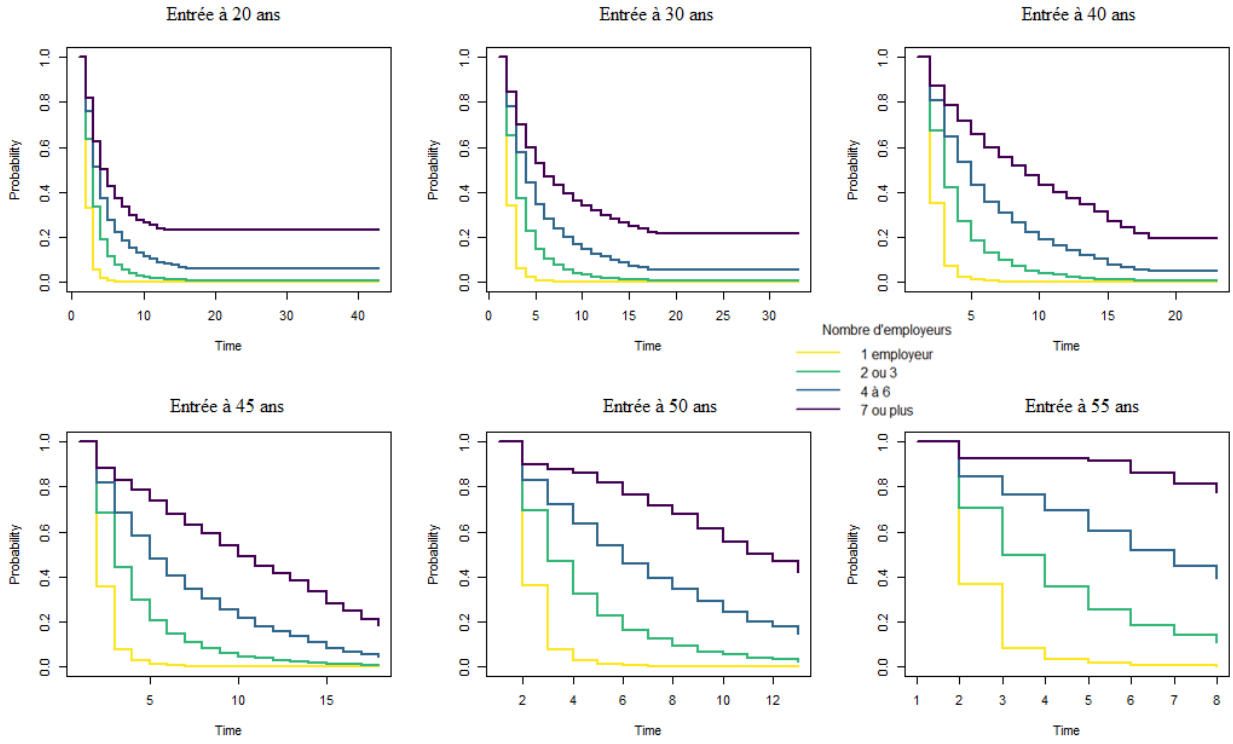
Graphique 3.3.1 - Effet par rapport au temps des modalités du modèle d'Aalen

Les intervalles de confiance sont ceux de Hall-Wellner, à 95%.

On remarque que pour le nombre d'employeur, hormis la modalité égale à deux, on a bien un effet au début de baisse de l'incidence. Cependant, plus on avance dans le temps, plus cet effet va stagner, généralement vers 15 années.

Pour le nombre d'employeurs égal à deux, on a des intervalles de confiance plutôt large. On remarque que l'incidence par rapport à la modalité de référence baisse dans les premières années. A partir de 5années les coefficients stagner, puis s'annuler pour faire remonter la courbe cumulée entre 12 et 17 années.

On peut observer les courbes de survie suivantes pour le nombre d'employeurs :



Graphique 3.3.2 - Courbe de survie des modalités du nombre d'employeur du modèle d'Aalen

Les courbes de survie donnent les résultats intéressants observés dans la première partie de l'analyse de survie. On constate une hausse du maintien plus l'entrée en portefeuille se fait tardivement, et plus le nombre d'employeurs est élevé.

e. Conclusion

L'analyse de survie a permis de mettre en avant le maintien dans la branche pour notre population. Grâce aux modèles développés, on a mis en perspective l'impact de covariables sur nos données, ainsi que les variations de ces effets dans le temps.

La méthode par l'estimateur de Kaplan Meier permet une première approche du maintien en activité. Il permet d'observer indépendamment les courbes de survie en prenant en compte une décomposition par sexe ou par nombre d'employeur.

Le modèle de Cox permet quant à lui l'ajout de covariables directement dans le modèle. Cependant, l'hypothèse de proportionnalité du risque au cours du temps n'est pas vérifiée pour toutes nos variables explicatives. On a néanmoins pu voir un modèle avec l'âge d'entrée en profession en variable explicative.

Le modèle d'Aalen a permis l'intégration de l'ensemble des variables explicatives. En définissant l'effet de l'âge d'entrée en profession comme constant par rapport au temps, on a pu ajouter le sexe et le nombre d'employeur.

Nous verrons dans la dernière partie de ce mémoire l'application des différents modèles sur une génération, et ainsi la cohérence avec les données observées.

Une des limites de l'analyse de survie dans notre cas d'étude est la non prise en compte d'éventuels retours dans la profession, ce qui conduit à une sous-estimation manifeste du nombre de mois acquis à la retraite pour le calcul de l'indemnité de départ volontaire en retraite. Afin de contourner cette limite, nous mettrons en place dans la suite l'étude des trajectoires de notre population et non plus une étude de survie.

III. Modélisation par analyse de séquence

Dans la première partie de la modélisation, on s'est entièrement concentré sur les modèles de survie, sans prendre en compte le retour en profession. Cependant, afin de proposer un modèle reflétant au mieux nos données et le parcours de vie des salariés du particulier employeur, il est indispensable de considérer comme réversibles nos deux états : « Présent en branche » et « Non-présent ».

On introduira alors un autre type d'analyse de données, l'analyse de séquence. Celle-ci se base sur l'étude des données reflétant soit la carrière complète, soit une partie de cette dernière sur un intervalle précis. On peut alors décrire et classifier les trajectoires des individus, en considérant chaque intervalle comme une succession d'états observés.

On s'intéressera à une méthode permettant de quantifier la différence entre deux carrières, l'Optimal Matching. L'intérêt est de mettre en avant l'homogénéité des profils, en prenant en compte différentes caractéristiques, comme la longueur de chacune, ou le nombre d'observations dans un état semblable ou au contraire un état différent. Ceci a pour but de créer des profils-types de carrière.

On va créer un score, appelé coût, qui servira dans le regroupement de toutes les carrières en groupe distincts. On utilise alors deux méthodes de classification permettant de créer des groupes de profils semblables d'un point de vue de ce coût. La première sera la classification ascendante hiérarchique (CAH). Cette classification est souvent utilisée dans le domaine assurantiel. On challengera la CAH avec une autre méthode de clustering, le Partitioning Around Medoids (PAM), un algorithme d'apprentissage automatique non supervisé.

Ces deux méthodes donnant des résultats différents, on déterminera dans un premier temps graphiquement, puis à l'aide de tests statistiques, la plus significative pour notre étude. On obtiendra alors des clusters, et on voudra décrire les individus qui les composent. On proposera pour cela une étude descriptive, selon nos variables « Sexe » et « Nombre d'employeurs moyen », puis on utilisera une régression logistique multinomiale pour expliciter l'effet de ces variables sur les différents clusters.

a. Introduction à l'analyse de séquence

Depuis plusieurs années, la collecte des données est un enjeu majeur dans nombre de secteurs. Le médical, les sciences sociales, la finance, les nouvelles technologies ou encore l'assurance sont tant de domaines où l'accumulation de données les plus précises et les plus fréquentes possibles sont indispensables à leur évolution. Parallèlement, différentes manières d'aborder les données sont mises en place, toujours dans une perspective d'avancée technique. Parmi elles, l'analyse de données dites longitudinales, qui sera l'objet de cette partie.

On mesure alors de manière répétée dans le temps plusieurs fois la même variable pour un même individu. Ainsi, on regroupe à chaque instant mesuré les informations sur le salaire, le nombre d'employeur, mais aussi sa présence ou non en portefeuille (i.e si la personne a reçu un salaire ou non dans notre cas).

L'analyse de ces données peut alors être vue de deux manières :

- On peut considérer les données pour un individu comme un seul bloc. Cette vision peut se justifier par l'aspect projet de vie des individus. Pour un bon nombre de personnes, les grands projets de vie sont planifiés, et les actions réalisées ou les états dans lesquels on se trouve vise à servir ce grand projet de vie. On parle ici de perspective holiste.
- Dans un second temps, on peut aussi considérer des évènements imprévus, ou un parcours de vie moins anticipé, ce qui semble être de plus en plus le cas pour les générations plus récentes. On aborde alors les données comme une succession de séquences, qui vont alors former une trajectoire qui sera dépendante des événements, la période à laquelle ils se produisent, l'ordre dans lequel ils se produisent ainsi que la durée de ces différents événements.

La difficulté réside dans le nombre infini de possibilité qu'on peut obtenir par de telles études. Selon le nombre d'événements considérés, selon s'ils soient réversibles ou non, ou selon le nombre de points d'observations, on peut rapidement se retrouver avec un nombre exponentiel d'informations disponibles.

L'analyse devient intéressante lorsqu'on dispose d'assez de données pour reconstituer une trajectoire de vie d'un individu. La longueur de la trajectoire dépend du risque que l'on veut étudier. Si on se base sur un médicament, on peut vouloir suivre l'évolution de l'état d'un individu mois par mois pendant quelques années. Si on se base dans notre situation où l'on cherche à étudier l'ancienneté dans la branche d'un individu de son premier salaire à la retraite, alors nous avons besoin de données bien plus large, se basant sur des dizaines d'années. Il est cependant à noter que la longueur des séquences peut être variable au sein d'une même étude.

Afin d'intégrer l'analyse de séquence dans notre étude, les données sont présentées de la manière suivante, une ligne par pas de temps (ici une année) et par individu :

Id Individu	Age	Sexe	Presence	Salaire	Nombre d'employeur
1	18	F	0	0	0
1	19	F	0	0	0
...		F			
1	43	F	1	819	2
1	44	F	1	1002	2
1	45	F	0	0	0
1	46	F	1	988	1
...		F			
1	64	F	0	0	0
2	18	H	1	730	4
...		H			

Tableau 3.3.1.1 - Type de données pour l'analyse de séquence

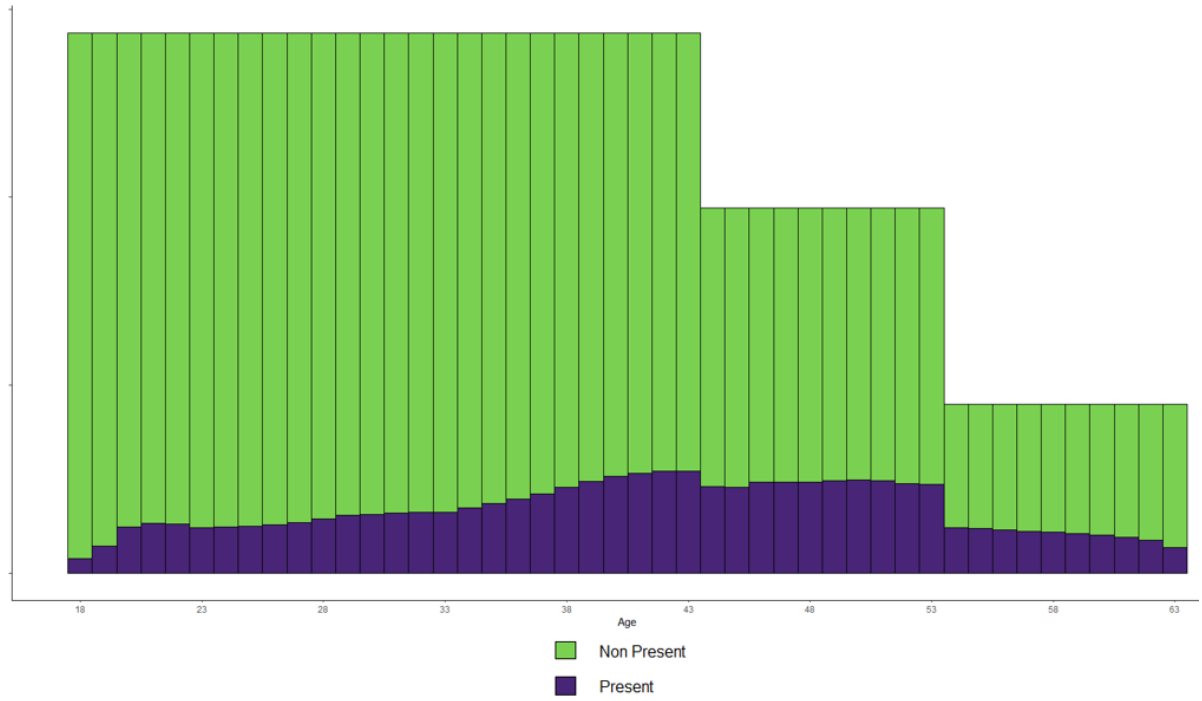
On considère dans notre analyse que nous disposons de l'ensemble des salaires pour une personne, ce qui est confirmé par l'analyse des masses salariales de la table et comparée aux données dans nos comptes techniques. On va alors établir la table avec une ligne pour chaque individu et par âge (entre ses 18 et 63 ans). On se concentrera une nouvelle fois sur les générations nées en 1958, 1968 et 1978. On indiquera alors si l'individu a été présent ou non ainsi que son salaire annuel et son nombre d'employeur unique.

Id Individu	18ans	19ans	...	62ans	63ans	Sexe	Nombre d'employeur moyen
1	0	1		1	1	1	2
2	1	1		0	0	2	4
...							

Tableau 3.3.1.1 - Type de données pour l'analyse de séquence

Grâce à cette forme de données, on va pouvoir tracer une séquence par individu. On pourra se référer à la séquence comme étant la carrière de l'individu.

La représentation graphique donne :



Graphique 3.2.1 - Distribution de la présence dans la branche pour les trois générations

Le graphique nous montre la présence à chaque âge (en abscisse) de toute notre population (en ordonnée). On remarque les sauts aux âges 43 et 53. Cela correspond à la censure de nos données pour les générations nées en 1968 et 1978. A partir de 53 ans il ne reste plus que la génération née en 1958 en portefeuille. Si on ramène en proportion de population observable, on a bien la répartition vue dans la partie descriptive avec une plus grande présence entre 43 et 53 ans.

Il nous faut alors une méthode qui permet de représenter les carrières et d'associer entre-elles celles qui se ressemblent, de les classifier.

b. Optimal Matching

L'objectif est d'identifier les profils-types, de mettre en avant les similitudes entre les carrières. Il est nécessaire de définir une méthode qui va quantifier la différence entre individus, de regarder la distance qui sépare une séquence d'une autre séquence.

L'Optimal Matching (OM) [Abbott, 1995] se rapporte au minimum de la somme des coûts pour passer d'une séquence (une carrière ici) à une autre. Afin d'établir cette somme, il nous faut prendre en compte les différentes opérations qu'il peut exister pour passer d'une trajectoire à l'autre.

On prend en exemple nos données. Les deux états possibles sont la présence (on notera P) et l'absence en portefeuille (N).

On définit les trois opérations suivantes :

- L'insertion d'un élément dans la séquence soit NNPN ← N = NNPNN
- La suppression soit NNP~~N~~ = NNP
- La substitution, on remplace un élément par une autre état soit NNP → NPP

La distance entre deux séquences est alors définie comme le nombre d'opérations minimales nécessaires pour passer d'une séquence à l'autre. Un cadre simple serait de considérer que chacune des opérations au même coût que les autres. Dans ce cas, il suffit de comptabiliser le nombre d'opérations minimum à opérer pour obtenir la mesure de la distance.

Soit $d_{i,j}$ la distance entre le profil i et le profil j, I le nombre d'insertion, D le nombre de suppression et S le nombre de substitution, on a simplement :

$$d_{i,j} = I + D + S$$

En pratique, on préféra différencier les coûts liés à la substitution de ceux liés à l'insertion et à la suppression. La mesure de la distance devient alors :

$$d_{i,j} = I \times c_i + D \times c_D + S \times c_S$$

Avec c le coût associé à l'opération.

Il faut maintenant établir la matrice de coût pour chacune de ces transformations.

Soit deux chaînes de caractères AAP et AAPA. Il vient logiquement que AAP ← A = AAPA et AAP~~A~~ = AAP. On a donc les opérations d'insertion et de suppression qui correspondent à la même opération dans l'absolu. On déterminera donc un coût pour l'insertion, égal à celui de la suppression (nommé indel dans la littérature de l'analyse de séquence), et un coût pour la substitution. Plus qu'un coût, c'est en réalité un rapport que l'on définit entre les deux types d'opérations. Ainsi, fixons un coût pour l'indel et nous en déduirons un coût pour la substitution.

Si on s'intéresse maintenant à la transformation en elle-même, il vient que la substitution ne modifie que l'enchainement des événements, là où l'indel modifie la structure en elle-même de la séquence, en ajoutant/supprimant un élément dans celle-ci. On serait tenté alors de donner un fort coût indel par rapport au coût de substitution. Cependant, dans le cas de séquences de longueurs égales, comme on l'étudiera dans la prochaine partie, si le coût indel est trop élevé, on lui préféra systématiquement des substitutions. Et inversement dans le cas où son coût est trop faible.

Il n'existe pas de bonne méthode permettant de déterminer avec précision ses coûts. Ici nous sommes en présence de longueurs inégales de séquence dues aux différentes générations. On se placera dans un cadre classique, celui de la LCS (Longuest Common Subsequence). On attribue alors un coût indel égal à 1, et un coût de substitution égal à 2. Ce choix est effectué car il suppose en réalité le même coût pour une opération de substitution que d'insertion et suppression. Si on reprend notre exemple, pour passer de AAP à AAA, on effectue une opération de substitution, ce qui coûte 2. Mais on peut aussi le voir comme AA~~P~~=AA puis

AA←A=AAA. On a donc effectué une opération d'insertion et une de suppression. On a donc un coût de 2 aussi.

NB : En pratique, on peut définir un coût différent pour chaque substitution. Cela peut être utile dans le cadre d'une hiérarchisation des états par exemple. Dans notre cas, on considère le cas simple où les deux événements réversibles « entrée en portefeuille » et « sortie du portefeuille » ont le même coût.

c. Classification

On a déterminé dans la partie précédente les différents coûts que nous allons associer à nos opérations. On détermine alors la matrice des distances suivantes :

Individus	1	2	3	4
1	0	4	22	8
2	4	0	13	15
3	22	13	0	6
4	8	15	6	0

Tableau 3.2.3.1 - Matrice des distances entre individus

Dans l'exemple ci-dessus, les individus 1 et 2 ont le profil le plus proche, là où l'individu 1 et 3 sont les plus éloignés.

Maintenant que nous avons toutes les distances dont nous avons besoin, il nous faut choisir une méthode pour partitionner nos résultats. L'intérêt de la classification, ou clustering, est de créer des sous-ensembles homogènes de notre population de base. On cherche alors à établir des profils pour observer ou relever des comportements types de chacun des sous-ensemble.

Une première méthode est la classification ascendante hiérarchique (CAH). Cette technique repose sur une matrice de dissimilarité entre tous les individus présents, que nous avons obtenue grâce à l'optimal matching, et le principe itératif suivant :

- Sélectionner les deux individus qui possèdent la distance la plus faible parmi tous les individus (possibilité dans le cas d'égalités d'en associer plusieurs) et grouper ensemble les individus qui ont la plus faible distance entre eux.
- Sélectionner le/les groupes obtenus, et calculer la distance minimale selon un critère de dissimilarité entre ces groupes et les autres individus/groupes, pour former un/des nouveaux groupes.
- Recommencer l'opération jusqu'à avoir groupé tous les individus.

La dernière classe obtenue est ainsi le groupe avec l'ensemble des individus.

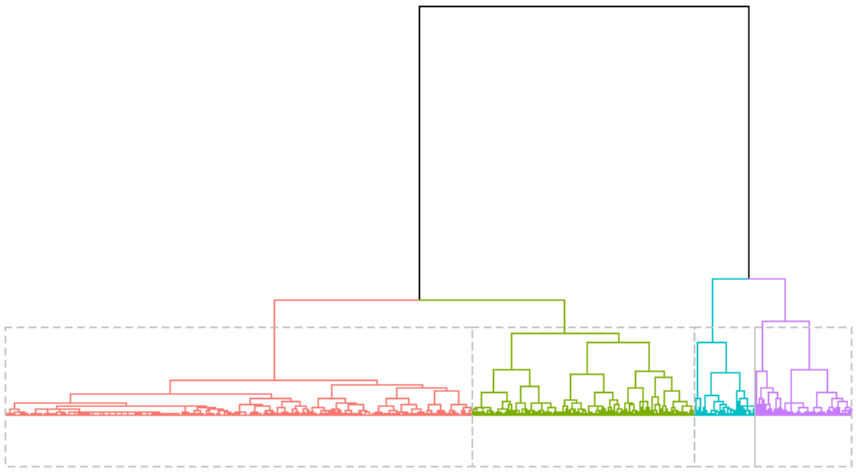
Une alternative à cette méthode serait la classification descendante hiérarchique, qui prend le même principe mais l'on part de l'ensemble des individus, et cette fois on exclut l'individu, ou groupe d'individus, selon un critère préalablement défini.

Dans notre cas, on sélectionne comme critère de dissimilarité à minimiser le critère de Ward, qui s'exprime comme étant :

$$Ward_{i,j} = \frac{n_i * n_j}{n_i + n_j} * d^2(g_i, g_j),$$

Avec n le nombre d'effectifs de la classe et g son centre de gravité.

On peut alors représenter les différentes classes dans un dendrogramme de la forme suivante :



Graphique 3.2.3.2 - Dendrogramme avec sélection de quatre groupes

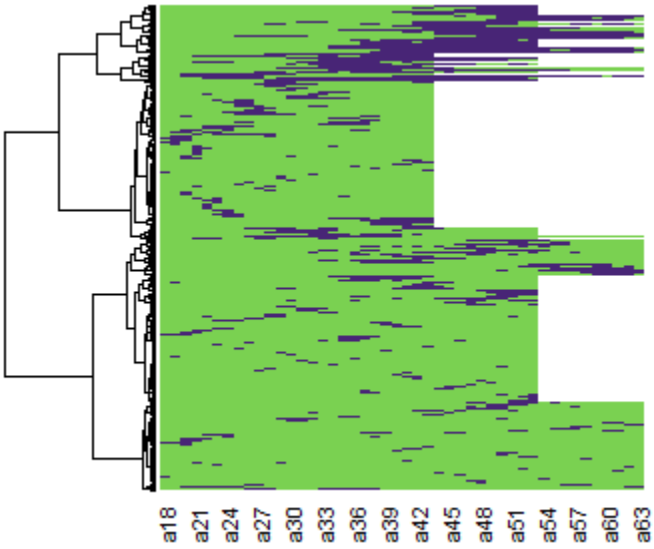
Chacun des individus est représenté par une branche en bas du dendrogramme, et le haut du dendrogramme correspond au dernier regroupement avant celui composé de l'ensemble des individus.

Plusieurs groupes de couleurs sont dessinés sur notre exemple. Ceux-ci correspondent aux groupes choisis si l'on sélectionne un nombre de sous-groupes homogènes égal à quatre. Pour les déterminer, on s'intéresse à la hauteur de chaque regroupement. Plus cette hauteur est basse, moins la distance était importante entre nos individus/groupes. On peut définir la hauteur comme :

$$Hauteur = \sqrt{2 * Ward_{i,j}}$$

Pour savoir comment on regroupe les sous-ensembles, il nous suffit alors de partir du sommet du dendrogramme, et de parcourir les branches jusqu'à avoir le nombre de groupes voulus.

Une autre représentation est celle du tapis de données. Avec les sous-ensembles définis ci-dessus, on représente maintenant de gauche à droite, avec au bout de la branche l'individu représenté par sa carrière.



Graphique 3.2.3.4 - Tapis de données toutes générations

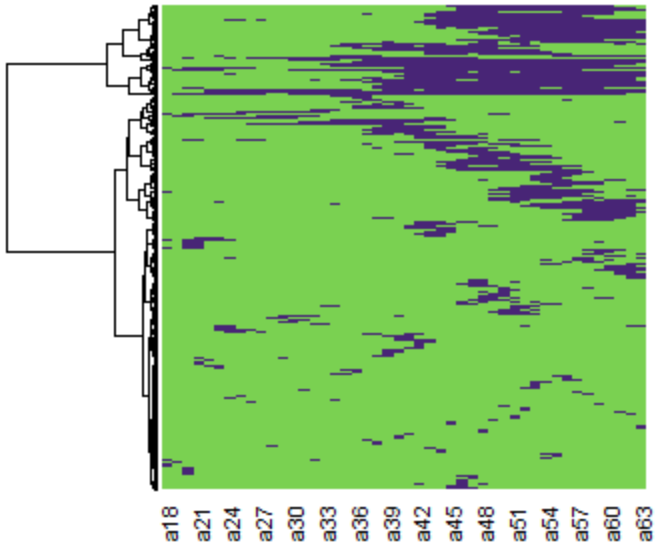
On observe très clairement un pattern dans la constitution des groupes. De gauche à droite nous avons la présence par âge, « a18 » étant la présence à 18 ans, « a63 » la présence à 63 ans. Les sous-ensembles de la partie supérieure semblent avoir des carrières (en violet) plus complètes et plus longues. La différence de longueur des séquences s'explique par la génération de l'individu. Pour les individus nés en 1978, la longueur s'arrête alors à 43 ans, là où celle d'un individu né en 1958 s'arrête à 63 ans.

Malheureusement, les critères choisis n'empêchent pas le regroupement principalement selon la longueur de la séquence, et non selon les enchainements de séquences similaires. Des solutions peuvent être apportées, soit par une sélection moins restrictives sur les générations sélectionnées, soit avec un rapport coût substitution/indel moins en faveur de l'indel.

Le choix a été pris pour la suite de sélectionner les données seulement de la génération la plus intéressante dans notre étude, la génération née en 1958, par ses carrières entières et supposées finies à ce jour. Ainsi la longueur des séquences sera égale et nous pourrons observer plus aisément l'enchainement des événements au cœur des séquences, ce qui est le cœur de notre étude.

d. Génération 1958

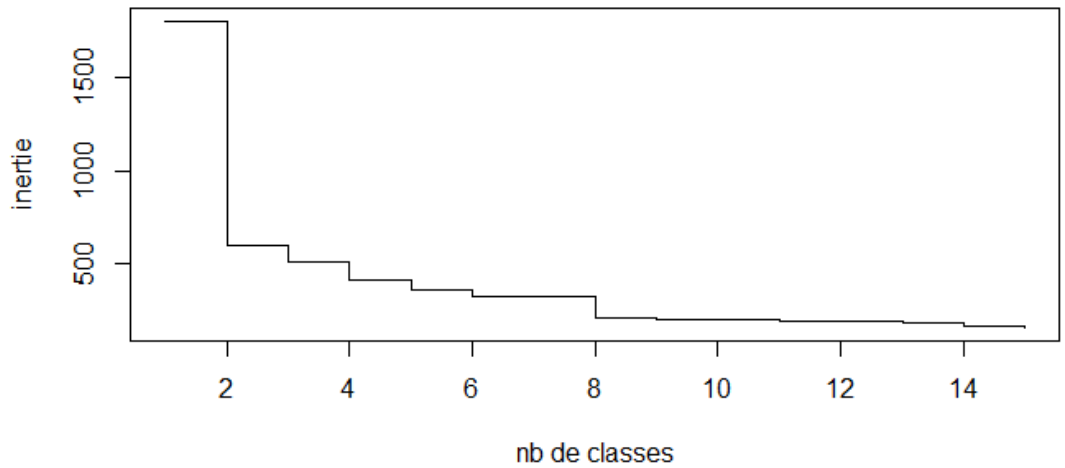
Le tapis de données par sous-ensembles est alors :



Graphique 3.2.3.4 - Tapis de données pour la génération née en 1958

Il apparaît des profils plutôt distincts. En partant du haut vers le bas, on observe des carrières longues mais débutant tardivement (~45 ans), puis des carrières longues démarrant plus tôt dans la vie des individus (~37 ans). Suit un enchainement d'individus avec des carrières plus courtes démarrant à des instants distincts, et enfin la dernière moitié du tapis concerne des individus présents en portefeuille de manière passagère.

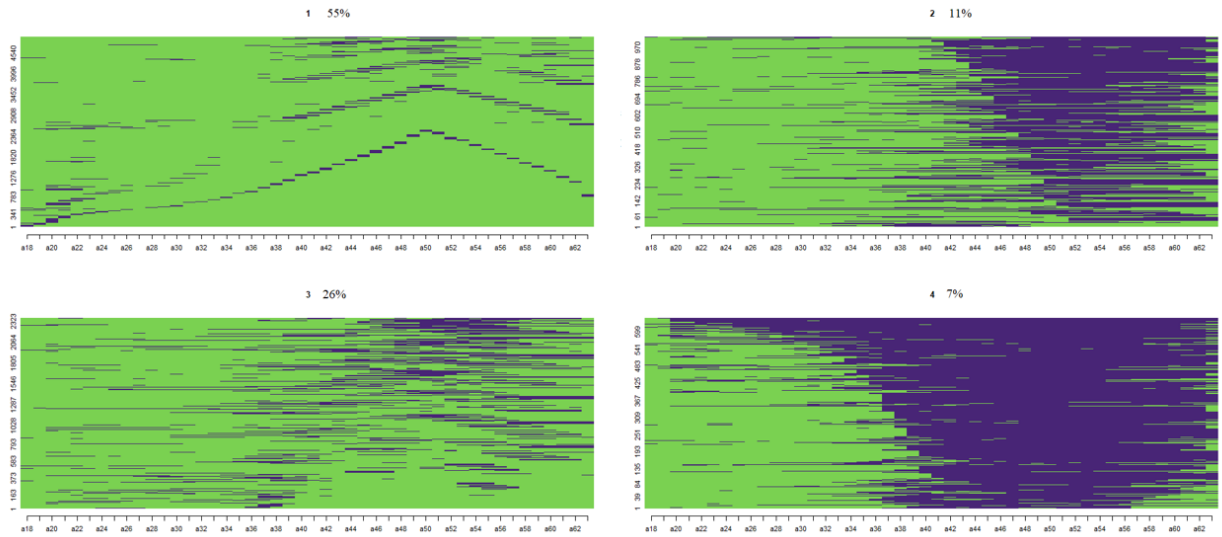
On doit maintenant choisir un nombre de classes pour segmenter notre portefeuille. Le choix du nombre de sous-ensemble est libre. Soit on a un nombre prédéfini, pour des raisons spécifiques à nos données par exemple, soit on peut représenter les sauts d'inerties liés au nombre de sous-ensembles.



Graphique 3.2.3.3 - Représentation des sauts d'inerties selon le nombre de classes

Là encore, le choix est libre. On peut utiliser la technique dite du coude, cependant ici le nombre de classes données est de 2. On peut aussi remarquer une réduction de la hauteur du saut d'inertie en 4, puis en 8. En pratique, 2 classes distinctes peuvent suffire. Nous allons nous intéresser ici aux profils dégagés en fonction du nombre de classes pour nos données.

On remarque des profils intéressants à 4 groupes :



Le premier groupe se compose essentiellement de passages très courts dans la profession, la grande majorité ne pourront jamais acquérir les droits nécessaires à l'indemnisation de départ volontaire à la retraite. Ce groupe peut s'apparenter à des gens présents de manière « passagère » en branche.

Le second groupe se compose lui de personnes arrivant tardivement en portefeuille, mais présentant pour la plupart une ancienneté conséquente. Cependant on remarque une présence moindre sur les dernières années, potentiellement empêchant les individus d'acquérir les droits à l'IDR. Il correspond alors à des carrières dites « tardives ».

La troisième catégorie est plus floue. On observe des carrières plus hétérogènes, peu présentes à l'approche de la retraite. Les carrières observées sont plus longues que celles dans le groupe 1. On parlera alors de « carrières prématurées ».

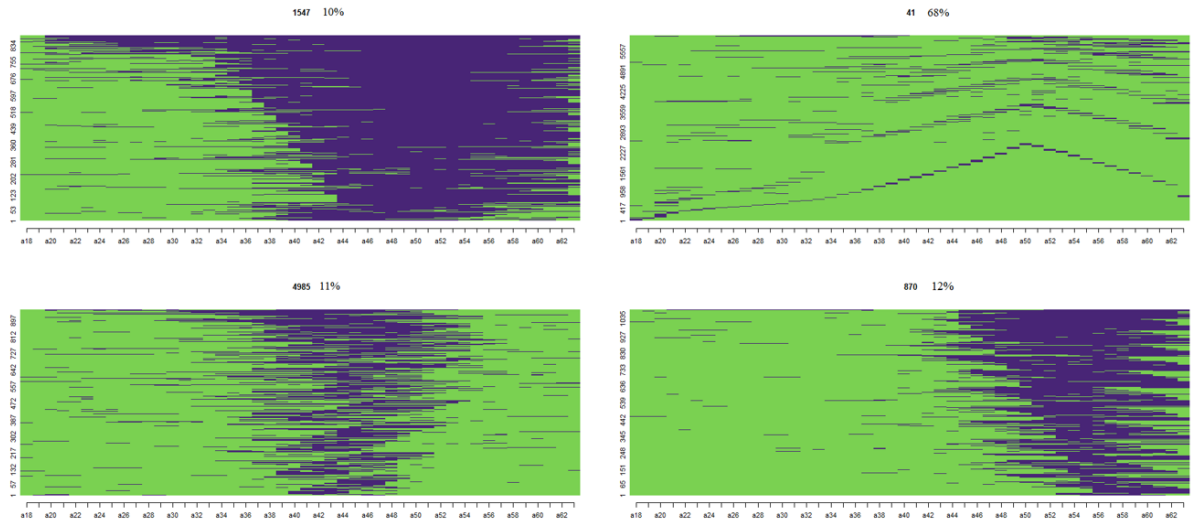
La dernière catégorie présente les carrières les plus longues et les plus complètes. Ces derniers sont présents en moyenne dès les 35 ans voir avant, et jusqu'à la retraite. Ceux seront les « carrières longues ».

Ainsi, la classification nous présente des sous-ensembles qui ont des caractéristiques intéressantes pour notre cadre d'indemnité de fin de carrière.

Le groupe 2 et le groupe 3 néanmoins sont un peu moins évidents à catégoriser du point de vue de l'IDR.

Nous essaierons une autre méthode de classification faisant appel à un algorithme des k-médoïdes : la méthode PAM (Partition Around Medoids). Le médoïde est défini comme le point qui minimise la somme des distances aux autres données. On conserve ici la distance de base définie dans la partie de l'optimal matching. PAM vise alors à minimiser la somme des distances au médoïde de chaque groupe.

On obtient les sous-ensembles suivant :



Visuellement, les classes créées semblent plus significatives qu'avec la classification ascendante hiérarchique.

Si on veut effectuer des tests statistiques, on obtient :

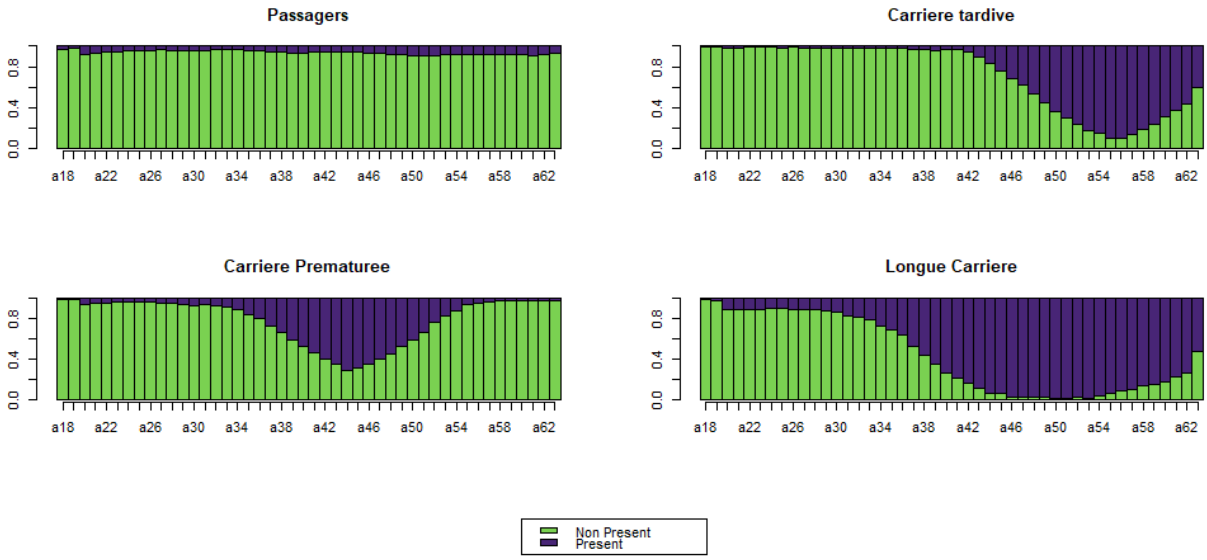
Test	CAH	PAM
ASW	0.36	0.49
HG	0.84	0.93
CHsq	8206	9607

Avec :

- **ASW** : Average Silhouette Width, mesure la moyenne des distances moyennes d'un point avec d'autres points d'un autre groupe que le sien. Le coefficient se situe entre [-1 ;1], 1 étant associé à un meilleur modèle.
- **HG** : Hubert's Gamma, mesure utilisée pour tester la similarité entre deux clusters différents. Le coefficient se situe entre [-1 ;1], 1 étant associé à un meilleur modèle.
- **CHsq** : Indice de Calinski-Harabasz au carré, mesure le rapport entre la variance inter-groupes et la variance intra-groupe. Le coefficient se situe entre [0 ;+∞[, plus il est grand, meilleur le modèle en est.

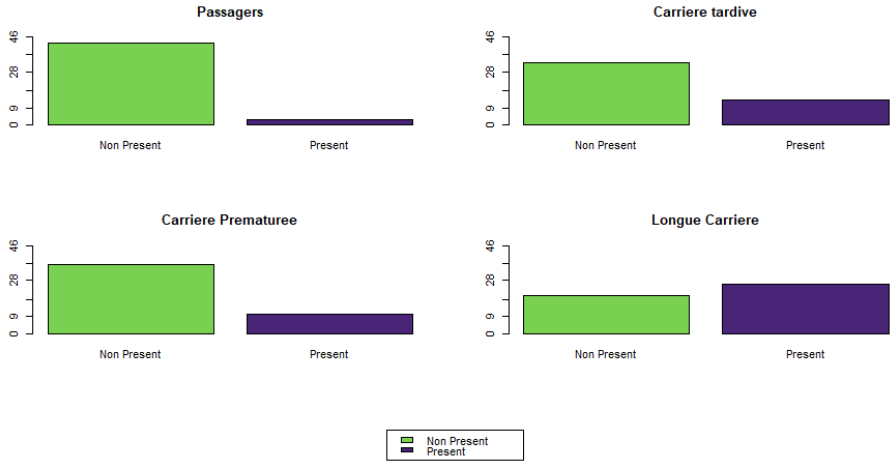
Dans notre cas, on a alors les trois tests qui indiquent une meilleure classification pour la méthode de PAM et on la sélectionnera pour la suite de l'étude en renommant les groupes.

La présence à l'intérieur de ces groupes est ressemblante à celle obtenue par les sous-ensemble de la méthode CAH. Le groupe en haut à gauche sera le groupe des « carrières longues », en haut à droite les « passagers », le groupe en bas à gauche les « carrières prématurées » et enfin les « carrière tardive ».



Graphique 3.2.4.2 - Présence en portefeuille par sous-ensemble de la méthode PAM

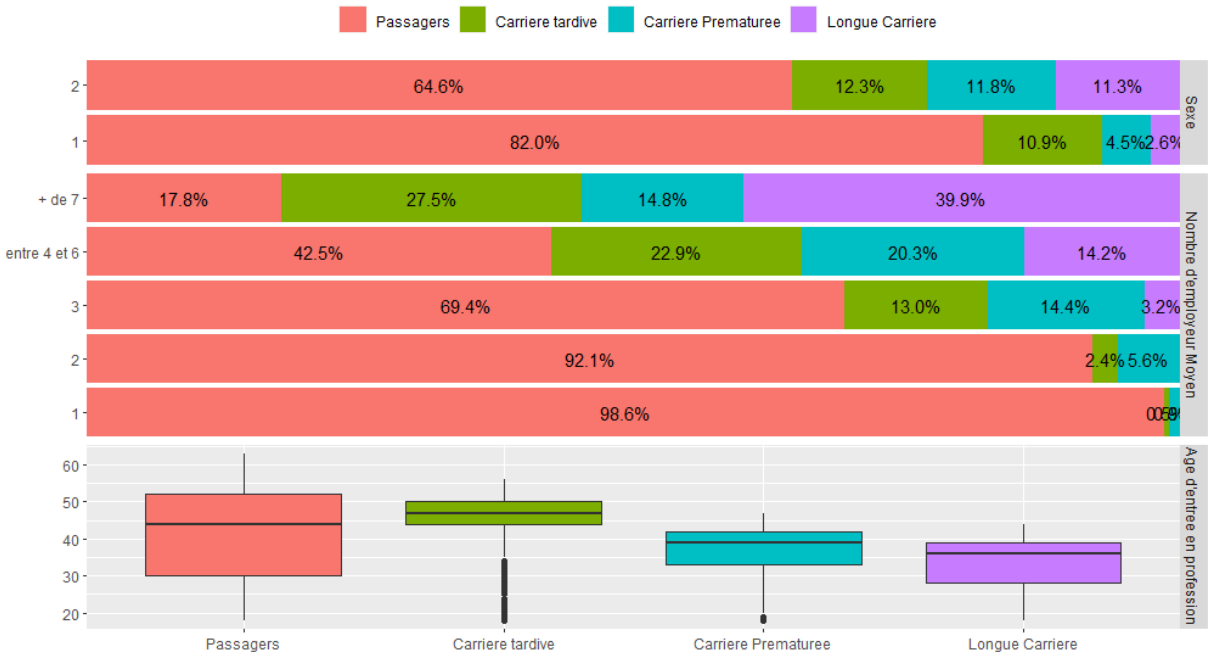
La moyenne d'ancienneté (en année) selon le profil donne :



Graphique 3.2.5.2 - Présence moyenne selon le profil de carrière déterminé

e. Descriptif des groupes obtenus

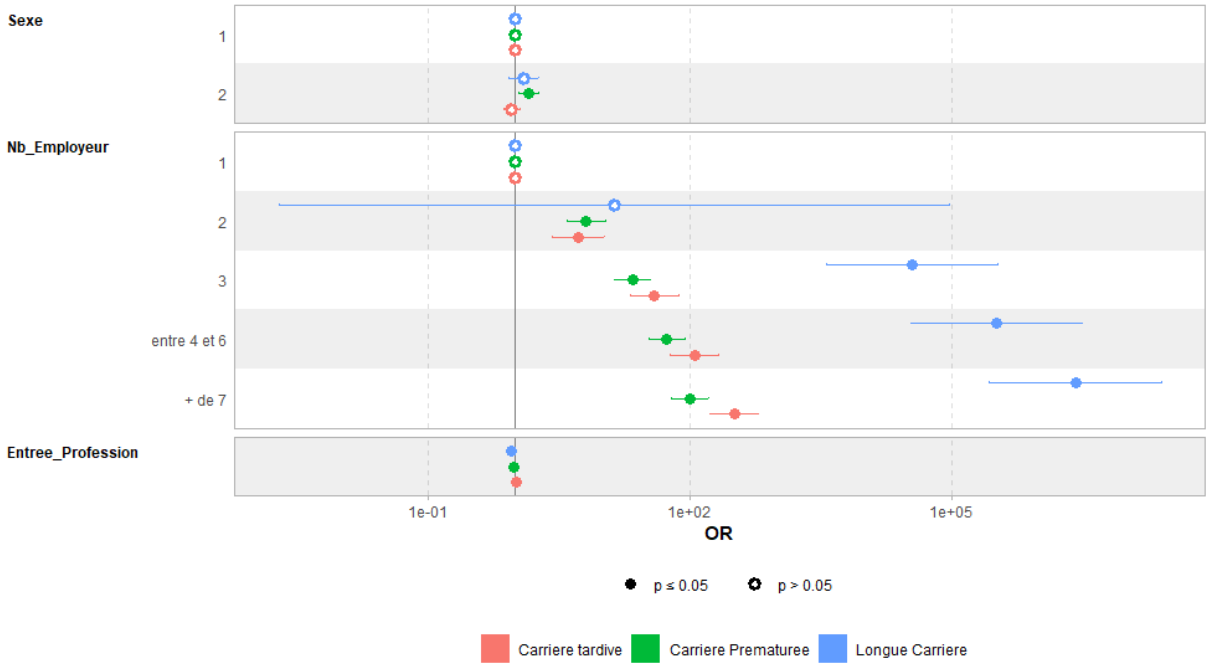
Afin d'identifier les individus qui composent les groupes, on s'intéresse à l'influence de nos co-variables utilisées dans l'analyse de survie sur chacun des groupes.



Graphique 3.2.5.1 - Proportion d'individu selon les co-variables dans chaque groupe

On a la probabilité, pour chaque modalité, d'appartenir à un des quatre groupes définis précédemment. On confirme des schémas vus dans les parties précédentes. Ainsi les femmes (sexe=2) sont en proportion plus présentes que les hommes dans les carrières longues (35.4% contre 18% pour les hommes). Concernant le nombre d'employeur, on constate que le nombre d'employeur est plus important pour des carrières plus longues et surtout qui se terminent souvent à la retraite.

Si l'on veut faire une analyse de l'effet de ces covariables sur ces sous-ensembles, on peut se rapprocher d'une régression logistique multinomiale. Ce modèle est utile dans notre cas car il permet d'analyser une variable expliquée avec plusieurs modalités selon plusieurs variables explicatives.



Graphique 3.2.5.2 - Odds Ratio de nos covariables sur l'appartenance aux groupes d'individus

La covariable sexe n'est pas à conserver dans notre modèle. Avec en profil de référence les carrières « passagères », le coefficient n'est pas significativement différent pour les carrières tardives et les longues carrières. On aura alors comme covariables le nombre d'employeur et l'entrée en profession.

Les coefficients sont ici des Odds Ratio. On peut les interpréter comme : on a un coefficient de 100 pour la modalité [entre 4 et 6] du nombre d'employeur pour l'appartenance à la carrière tardive. La modalité de référence étant 1 employeur, on a alors une côte d'appartenir au sous-ensemble « carrière tardive » 100 fois supérieure à celle d'appartenir au sous-ensemble « passagers » si on a entre 4 et 6 employeurs que 1 seul.

Ainsi, on peut avoir des idées, selon nos covariables, de l'appartenance à un type de profil. On cherche alors à déterminer les droits obtenus par les individus à l'intérieur de ces sous-ensembles, en analysant l'ancienneté totale d'abord :

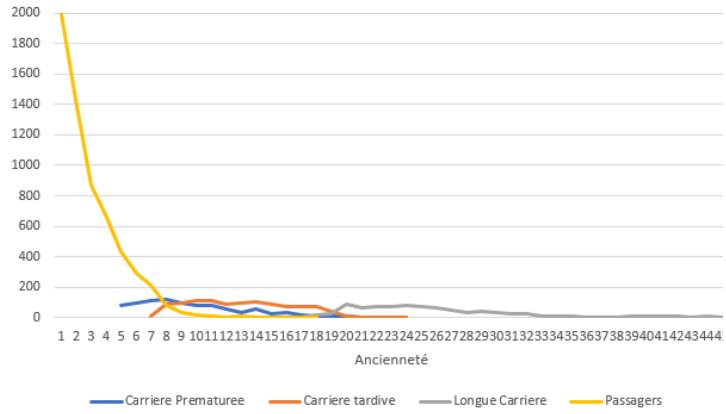


Tableau 3.2.5.1 - Ancienneté à la retraite selon le profil déterminé par l'analyse de séquence

L'ancienneté est bien prise en compte dans notre méthode de classification. On a un décalage des courbes des carrières. Le corps des « passagers » est le premier et est dominant sur l'ensemble. Il correspond bien aux courtes carrières avec 97% qui ont moins de 7années d'ancienneté. Puis commence à 5 années le groupe des « carrières prématurés », jusqu'environ 20 années, puis la même courbe démarre à 7 années pour le groupe des « carrière tardives ». Enfin, on a les « carrières longues » qui ont un minimum de 17 années d'ancienneté.

Pour compléter notre analyse, reprenons le tableau *2.5.1* qui donne les droits dans le cadre de l'indemnisation de départ volontaire à la retraite :

Ancienneté dans les 7 dernières années précèdent la date de retraite	Ancienneté totale du salarié	Profil de présence en branche				
		Montant de l'indemnité	Passagers	Carriere Prematuree	Carriere tardive	Longue Carriere
Moins de 5ans	-	0	96%	99%	16%	12%
5ans ou plus	**Moins de 10 ans**	0	4%	0%	13%	0%
	Au moins 10 ans	1 mois	0%	1%	38%	0%
	Au moins 15 ans	1.5 mois	0%	0%	30%	2%
	Au moins 20 ans	2 mois	0%	0%	2%	66%
	Au moins 30 ans	2.5 mois	0%	0%	0%	19%

Tableau 3.2.5.1 - Droits IDR selon le profil déterminé par l'analyse de séquence

Les groupes sont alors bien représentés et bien nommés, notamment les carrières prématurées qui, malgré la forme semblable aux carrières tardives, bien que plus tôt de deux ans, ont nettement moins d'individus dans les groupes possédant des droits à l'IDR.

f. Conclusion

Dans cette partie, nous avons pu déterminer différents profils homogènes en termes de carrière de nos assurés.

Nous avons utilisé deux méthodes de classification, CAH et PAM, basées sur la même méthode de calcul des distances (Optimal Matching). Grâce à un critère visuel et des critères statistiques, nous avons convenu de la meilleure précision de la méthode PAM.

Ensuite, nous avons pu mettre en avant différents profils relatifs à la présence en portefeuille en fonction de nos variables, selon une analyse de la répartition des sous-ensembles par rapport à nos covariables, puis grâce à une régression logistique multinomiale.

Enfin, nous avons, en fonction des classes d'individus, obtenu le tableau de répartition de la population selon les critères de l'IDR.

Les limites rencontrées ont été :

- Le calcul de distance trop influencé par le coût choisi de substitution et d'insertion/suppression. Il pourra être réévalué pour prendre en compte différentes générations.
- Les calculs parfois lourds, comme lors du calcul de la matrice regroupant les distances, obligeant à réduire le nombre d'individus observés.
- L'analyse de séquence est pour l'instant une méthode descriptive de nos données. Il n'a pas été développé ici une véritable approche prédictive de la présence discontinue en branche des assurés.

IV. Modélisation Multi-États

Afin de pouvoir modéliser nos deux états de présence en branche, et de prévoir les états futurs, nous nous rapprocherons de la théorie markovienne. Cette approche est nécessaire en complément de l'analyse de séquence. En effet, l'analyse de séquence nous permet de quantifier les retours en activité de salariés du particulier employeur. 29% d'entre eux reviennent après une année sans salaire, et parmi ces derniers, 21% sont éligibles à une indemnité de départ volontaire à la retraite.

Il nous faut alors prendre en compte les deux états de manière récurrente et réversible, là où l'analyse de survie considère l'état non présent comme absorbant. Un état i est considéré comme absorbant dans la théorie markovienne si :

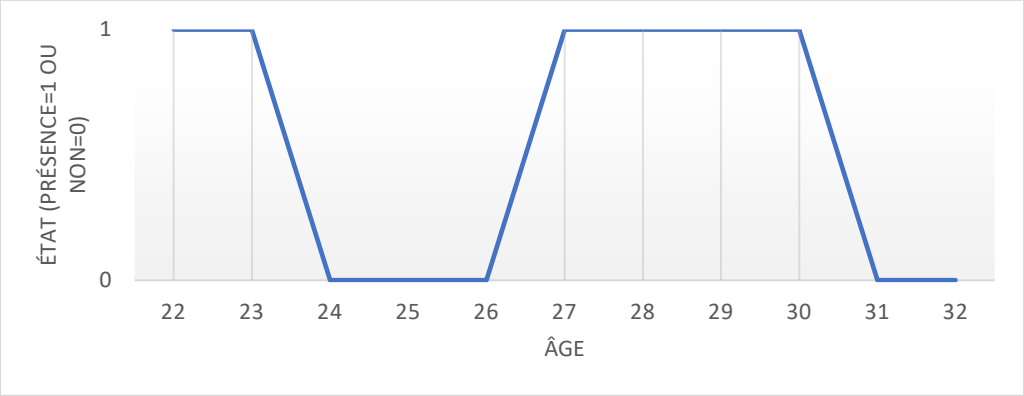
Soit $p_{i,j}$ la probabilité de passer d'un état i à un état j,

$$\forall\, j \neq i, p_{i,j} = 0 \; et \; p_{i,i} = 1$$

Ainsi, une entrée dans un état absorbant est définitive, on ne peut donc plus revenir dans un autre état quel que soit le temps passé dans cet état.

Un état réversible est lui définit comme : si $p_{i,j} > 0$ alors $p_{j,i} > 0$.

Le but de cette analyse est de décrire le procédé suivant :



Graphique 4.0.1 - Schéma d'un modèle multi-état réversible

On a bien nos deux états représentés selon l'âge de notre individu et à chaque observation. L'ancienneté est alors le cumul des états de présence en branche.

a. Modèle de Markov

i. Présentation Modèle de Markov

La propriété générale des modèles de Markov est la suivante : « Le futur n'est dépendant que de l'état présent ».

Ainsi, on a les propriétés ci-dessous :

- Soit E un ensemble d'états, fini ou infini, nécessairement dénombrable
- Soit $(M_i)_{i\geq 0}$, suite de variables aléatoires à valeurs dans E. On dit que M est une chaîne de Markov d'espace d'états E si et seulement si :

  $\forall x_j \in E, état\ au\ moment\ j,$

  $$\mathbb{P}(X_i = x_i, \dots, X_0 = x_0) > 0$$

  Et,

  $$\mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i, \dots, X_0 = x_0) = \mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i)$$

On a alors bien la propriété générale qui est respectée, l'état i+1 ne dépend que de l'état i, quelques soient les anciens états.

Ce cadre markovien de base peut cependant être modifié tout en conservant certaines propriétés.

Soit l'intensité de transition α l'intensité de transition de notre processus. Dans le cadre des modèles multi-états markoviens, l'intensité α peut dépendre de deux facteurs :

- La durée du suivi (t)
- La durée de séjour (d)

Alors on nommera :

- **Modèle de Markov homogène**, un modèle où $\alpha(t,d) = \alpha$
- **Modèle de Markov non-homogène**, un modèle où $\alpha(t,d) = \alpha(t)$
- **Modèle semi-Markovien homogène**, un modèle où $\alpha(t,d) = \alpha(d)$
- **Modèle semi-Markovien inhomogène**, un modèle où $\alpha(t,d) = \alpha(t,d)$

Pour notre étude nous verrons dans l'application les deux premiers types de modèles, à savoir markov homogène et markov non-homogène, ne dépendant pas de la durée de séjour.

ii. Intensité de transition

Pour décrire le phénomène observé, les modèles multi-états vont chercher à déterminer les intensités de transition entre les différents états.

Dans un cadre continu, on a :

- Soit **P**(s,t)={ $p_{i,j}(s,t)$} la matrice de probabilités de transition avec :

$$p_{i,j}(s,t) = \mathbb{P}(X(t) = j \,|X(s) = i)$$

- La propriété de Chapman-Kolmogorov donne pour les probabilités de transition :

$$\forall\, i,j \in E, \forall\, 0 < s < u < t,$$

$$p_{i,j}(s,t) = \sum_{k \in E} p_{i,k}(s,u) \times p_{k,j}(u,t)$$

Une autre forme, la forme matricielle, de cette équation donne :

$$\mathbf{P}(\mathrm{s},\mathrm{t}) = \mathbf{P}(\mathrm{s},\mathrm{u}) \times \mathbf{P}(\mathrm{u},\mathrm{t})$$

- L'intensité de transition entre les états, défini par la matrice **Q**(t)={ $\alpha_{i,j}(t)$}, avec, pour i≠j :

$$\alpha_{i,j}(t) = \lim_{\Delta t \to 0} \frac{p_{i,j}(t, t+\Delta t)}{\Delta t}$$

Et

$$\alpha_{i,i}(t) = -\sum_{i \neq j} \alpha_{i,j}(t)$$

b. Modèle de Markov homogène

Dans un premier temps, on se placera dans le cadre dit Markovien homogène. On se réfère ici à une partie des travaux de P. Saint Pierre [2005]. De plus amples détails et descriptions des méthodes et des modèles pourront y être trouvés.

Une chaîne de Markov est désignée comme homogène si la probabilité de passage d'un état à un autre est constant dans le temps.

Ainsi, pour tout i et j appartenant à E, on a :

$$p_{i,j}(s,t) = \mathbb{P}(X(t) = j \,|X(s) = i) = \mathbb{P}(X(t-s) = j \,|X(0) = i)$$

$$\Leftrightarrow p_{i,j}(s,t) = p_{i,j}(0, t-s) = \mathrm{P}_{i,j}(t-s)$$

On peut définir toute succession d'état comme étant dépendant d'un état initial (x_0) et des probabilités de transitions entre états :

$$\mathbb{P}(X_i = x_i, \dots, X_0 = x_0) = \mathbb{P}(X_0 = x_0) \times p_{x_0,x_1}(1) \times p_{x_1,x_2}(1) \times \dots \times p_{x_{i-1},x_i}(1)$$

La matrice des probabilités s'écrit en temps homogène : $\mathbf{P}(\mathrm{s}, \mathrm{s} + \Delta\mathrm{t}) = \mathbf{P}(0, \Delta\mathrm{t})$

Comme vu précédemment, l'intensité de transition devient : $\alpha_{i,j}(t) = \alpha_{i,j}$ et $\mathbf{Q}(\mathrm{t}) = \mathbf{Q}$

La relation vue précédemment, donnée par l'équation de Chapman-Kolmogorov, permet d'obtenir la relation suivante dans le cas homogène :

$$\frac{\partial \mathbf{P}(0, \Delta t)}{\partial t} = \mathbf{P}(0, \Delta\mathrm{t}) \times \mathbf{Q} \quad \boldsymbol{et} \quad \mathbf{P}(\mathrm{s}, \mathrm{s}) = \boldsymbol{Id}$$

Avec pour solution :

$\mathbf{P}(0, \Delta\mathrm{t}) = \exp(\mathbf{Q} \times \Delta\mathrm{t})$, avec exp la matrice exponentielle

On peut alors noter le temps de séjour dans l'état s comme $-\frac{1}{\alpha_{s,s}}$.

La probabilité de passer d'un état r à un état s est de $-\frac{\alpha_{r,s}}{\alpha_{s,s}}$.

Enfin, on s'intéresse à l' estimateur du maximum de vraisemblance. La contribution individuelle (l_h) est égale à :

$$l_h = \mathbf{P}_0[x_{h,0}] \times \prod_{j=1}^{n_h} P_{x_{h,j-1},x_{h,j}}(T_{h,j} - T_{h,j-1})$$

Et la vraisemblance totale :

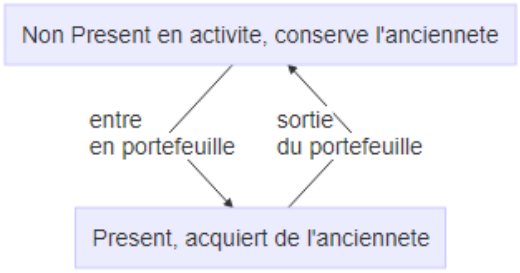
$$L: \prod_{h=1}^{n} l_h$$

Pour déterminer l'estimateur du maximum de vraisemblance, on utilisera la méthode de quasi-newton présente dans la fonction optim() de R. La méthode de quasi-Newton est un algorithme qui recherche un point stationnaire pour une fonction. C'est un moyen efficace lorsque le calcul pour déterminer l'optimum est coûteux et complexe.

i. Application d'un modèle simple de Markov à nos données

Pour modéliser à l'aide d'un modèle de Markov homogène, nous utiliserons la même forme de données que lors de l'analyse de séquence, à savoir une ligne par individu et par pas de temps.

La matrice de transition peut se décrire comme :



Graphique 4.2.1.1 : États et transitions entre états

Ce modèle est le plus simple possible pour nos données. Nous n'avons pas de plus amples informations concernant le motif de sortie du portefeuille, nous ne pouvons donc pas décrire avec plus de précision les différents états du salarié.

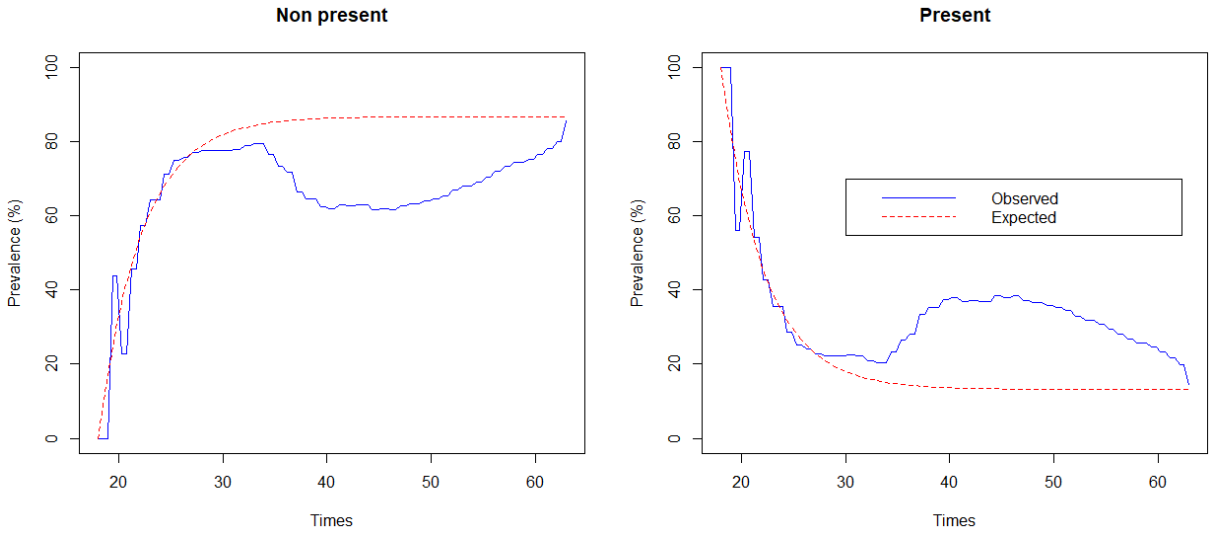
Pour ne pas considérer un nombre trop important de non-présence, nous utiliserons les données à partir de la première présence en branche pour chaque individu, et non à partir de ses 18 ans.

- Les transitions brutes entre les deux états nous donnent :

		État d'arrivée	
		Non Présent	Présent
État initial	Non Présent	96%	4%
	Présent	23%	77%

Tableau 4.2.1.1 - Transition entre états dans le cadre markovien classique

- Le modèle nous donne alors :



Graphique 4.2.1.1 - Résultat du modèle de Markov simple

Le graphique ci-dessus nous donne les présences en branche de notre portefeuille. En bleu on retrouve les données observées et en pointillé rouge nos estimations, par âge et selon l'état. Comme on considère les probabilités de transitions invariantes sur l'ensemble des âges, et sans variables explicatives, la prédiction suit la tendance globale mais ne va pas s'adapter à nos données. On a une estimation plutôt correcte de la présence dans les premières années, de 18 à 27 ans. Puis nous avons une sous-estimation importante de la présence à tout âge au-dessus de 35 ans jusqu'à la retraite, où l'écart va se resserrer. Cette partie de la carrière étant la plus importante dans notre étude, comme nous avons pu le constater notamment lors de l'analyse par répartition, les résultats seront ici très sous-estimés.

ii. Ajout de Covariable

Afin d'améliorer la précision de notre modèle multi-état, on pourra intégrer les variables explicatives vu dans les analyses précédentes. On se place alors dans un cadre simplifié de modèle à intensités proportionnelles. Ainsi :

$$\alpha_{i,j}(t, Z) = \alpha_{i,j,0} \times \exp\left(\beta'_{i,j} \times Z\right)$$

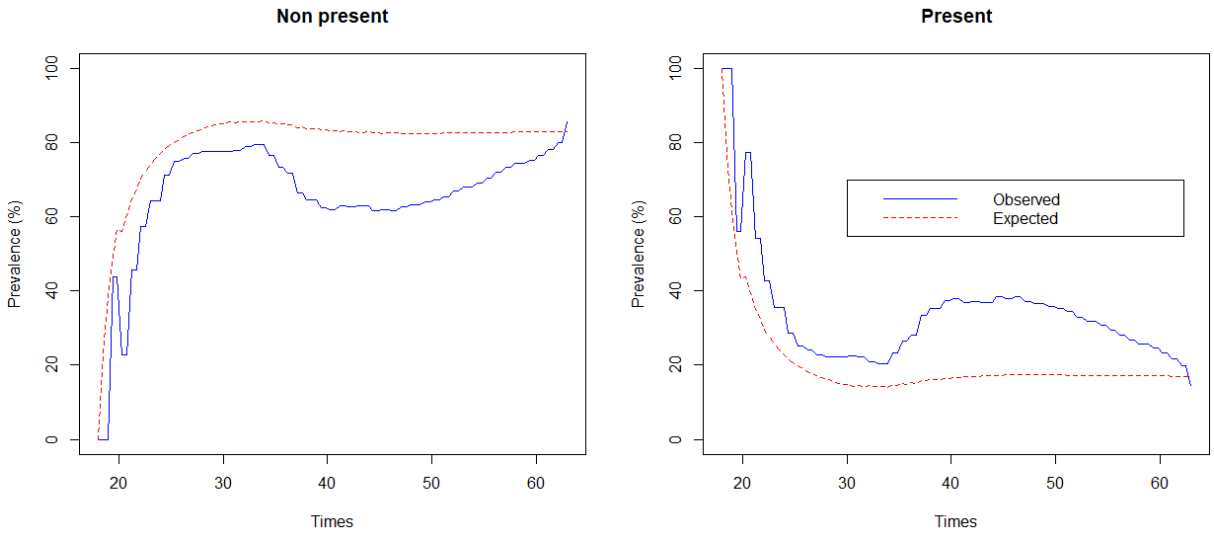
Avec :

- $\alpha_{i,j,0}$ l'intensité de transition de base
- Z=$(Z_1, \dots, Z_k)^T$ vecteur de k covariables
- $\beta_{i,j} = (\beta_{i,j,1}, \dots, \beta_{i,j,k})^T$ vecteur de coefficients de régression associé à la transition de l'état i vers l'état j

Dans les faits, les covariables peuvent être dépendantes du temps. L'intensité de transition entre [t1 ; t2[est alors dépendante de la valeur de la covariable en t1.

Les covariables utilisées dans notre modélisation sont, comme précédemment, le sexe, le nombre d'employeur et l'âge d'entrée en profession.

On observe alors :



Graphique 4.2.2.1 - Résultat du modèle de Markov avec ajout de covariables

Dans les faits, l'ajout de covariables ici n'a que peu de conséquences. On sous-estime la présence dans les premiers âges (18-27 ans). On peut néanmoins observer une légère amélioration entre 35 et 63 ans. L'hypothèse d'intensité constante sur l'ensemble de la durée de suivi est trop restrictive.

Nous avons vu ici l'expression la plus simple des modèles de Markov. Toutefois, on se rend compte que la propriété de base Markovienne n'est pas respectable dans notre modélisation, comme dans de nombreux cas en assurance. Il nous faudra alors prendre en compte l'impact du temps sur les probabilités de transitions et l'ajouter à notre modélisation. Par conséquent, le recours aux processus inhomogènes semble nécessaire.

c. Modèle de Markov à Modélisation à intensité constante par morceaux

Nous décidons de nous orienter vers la modélisation de modèle de Markov non-homogène. Comme vu précédemment, cela revient à prendre en compte une notion de durée de suivi pour les intensités de transition. Dans notre cas, et étant une approche généralement utilisé pour simplifier l'étude non-homogène, nous étudierons le cas d'homogénéité par morceaux (Jones, 1993, 1994).

i. Présentation

On définit alors les intensités de transitions, sur r+1 intervalles $[t_{k-1},t_k[$:

- $\alpha_{i,j}(t) = \alpha_{i,j,0}$, si $t_0 \leq t < t_1$
- $\alpha_{i,j}(t) = \alpha_{i,j,k} = \alpha_{i,j,0}\exp\left(\sum_{h=1}^{k} \eta_{i,j,h}\right)$, si $t_k \leq t < t_{k+1}$
- $\alpha_{i,j}(t) = \alpha_{i,j,0}\exp\left(\sum_{h=1}^{r} \eta_{i,j,h}\right)$, si $t \geq t_r$

Avec :

- $t_{r+1} = +\infty$
- $\alpha_{i,j,k}$ l'intensité de base sur l'intervalle $[t_k,t_{k+1}[$
- $\eta_{i,j} = (\eta_{i,j,1}, \dots, \eta_{i,j,k})$ vecteur de coefficients de régression associé à la transition de l'état i vers l'état j

On a bien alors l'intensité constante sur une période, mais différente entre deux périodes distinctes.

De ce fait, la vraisemblance doit-elle être modifiée. Considérons deux temps consécutifs d'observations, T_1 et T_2, et X_1 et X_2 les états associés. Si T_1 et T_2 appartiennent à une période avec les mêmes intensités de transition, alors on se trouve dans le cadre classique.

Si T_1 et T_2 sont sur deux périodes différentes, elle se note maintenant :

$$p_{X_1,X_2}(T_1,T_2|\boldsymbol{z}(t),T_1 \leq t < T_2) = \sum_{k_1}\sum_{k_2}\dots\sum_{k_v}\left(\prod_{i=0}^{i=v} p_{k_i,k_{i+1}}^{(I_{T_1}+i)}\left(t_{I_{T_1}+i} - t_{I_{T_1}+i-1}\middle|\boldsymbol{z}\left(t_{I_{T_1}+i-1}\right)\right)\right)$$

Avec :

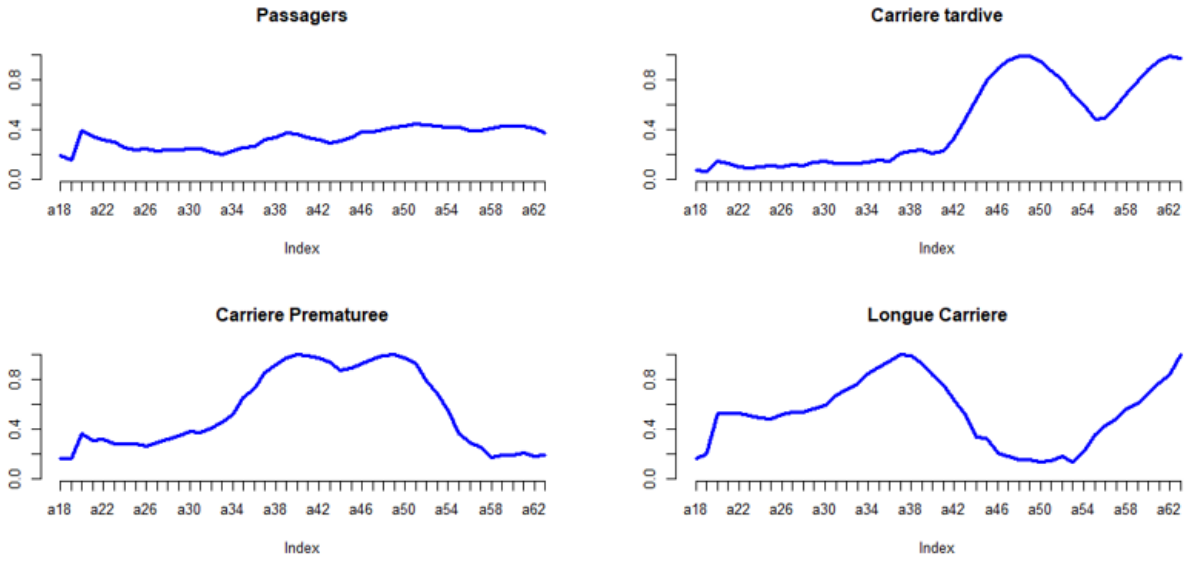
- $v = I_{T_2} - I_{T_1}$
- $k_0 = X_1 \; et \; k_{v+1} = X_2$
- $I_{T_1} + v = I_{T_2}$
- $t_{I_{T_1}-1} = T_1$
- $\boldsymbol{z}(t) = (\boldsymbol{z}_0(t), \dots, \boldsymbol{z}_r(t))'$ un vecteur tel que :
  - $\boldsymbol{z}_0(t) = 0 \, , \forall t$
  - $\boldsymbol{z}_k(t) = \begin{cases} 0 \; si \; t_0 \leq t < t_k \\ 1 \; sinon \end{cases}$

On peut associer le modèle à intensité constante par morceau et le modèle avec covariable pour obtenir un modèle plus précis et qui prend bien en compte l'effet du temps sur la présence et sur les covariables.

ii. Modélisation par intensité constante par morceaux

Pour notre modélisation, plusieurs tests ont été réalisés pour déterminer un modèle dont la décomposition en intensité constante est semblable aux données observées.

Pour obtenir nos intervalles d'intensité constante, on utilisera l'entropie des classes obtenues dans la partie d'analyse de séquence :



Graphique 4.3.2.1 - Entropie selon les classes de profils-types

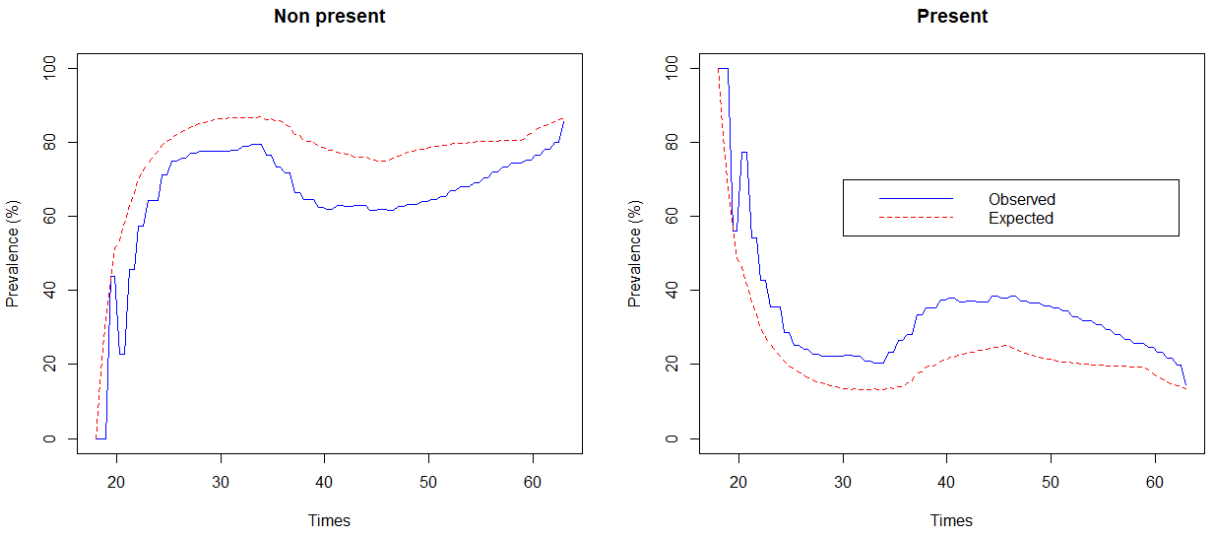
L'entropie sert à mesurer l'écart de la répartition des séquences à chaque âge à l'intérieur de chaque classe. Une entropie proche de 1 nous dit alors que la classe est très hétérogène, alors qu'une entropie proche de 0 nous informe que les individus de cette classe ont un profil très similaire (ici, qu'ils sont tous dans le même état). On peut ainsi définir des âges, ou des tranches d'âges, où l'on observe une forte entrée ou sortie de la profession.

Pour les classes « carrières prématurées » et « longue carrière », le maximum d'entropie se situe vers 36 ans. Le pic pour les « carrières tardives » arrive lui vers 46 ans. Enfin, on observe un nouveau pic pour les « carrières tardives » et les « longues carrières » à l'approche de la retraite. On pourra aussi s'apercevoir que les carrières prématurées ont un pic plus long, les individus qui la composent ont des profils assez variés, néanmoins ils présentent quasi tous une entrée/sortie définitive entre 36 et 52ans.

On découpera ainsi l'intervalle [18 ; 63] en 4 intervalles dans lesquels l'intensité est constante :

- [18 ; 36] , [37 ; 46] , [47 ; 59] , [60 ; 63].

Les résultats obtenus sont :



Graphique 4.3.2.2 - Résultat du modèle de Markov non-homogène

La courbe estimée semble bien plus proche de l'observée qu'auparavant. Notre modèle est meilleur que ceux respectant l'hypothèse de Markov classique. On sous-estime à nouveau la présence en portefeuille sur les âges 18 à 35 ans. Néanmoins, on semble mieux modéliser la présence sur le reste des âges, tout en restant en dessous des chiffres observés.

On risque alors d'avoir une sous-estimation importante selon les critères de l'IDR. En effet, comme développé lors de l'analyse des courbes du modèle simple de Markov et de l'analyse de séquence, c'est la tranche d'âge au-delà de 40 ans qui nous intéresse le plus.

iii. Evaluation du modèle par le test d'ajustement de Pearson

En plus de l'analyse visuelle, nous pouvons déterminer l'ajustement de notre modélisation à intensité constante par morceaux par rapport aux modélisations dans le cadre Markovien homogène.

Pour cela, on peut utiliser le test d'ajustement de Pearson. La statistique de test est alors :

$$T = \sum_{h,t_h,Trans(i,j)} \frac{(observés - estimés)^2}{estimés}$$

Avec h l'individu, t_h la période sur laquelle on observe h, et Trans(i,j) les transitions possibles entre les états i et j.

La sortie de la fonction *pearson.msm()* de R donne :

```
$test
     stat df.lower p.lower df.upper p.upper
 362.8063       -8     NaN       12       0
```

Le test du rapport de vraisemblance nous donne en comparaison :

- Comparaison modèle Markovien Homogène sans covariable (MHSC) et modèle Markovien Homogène avec covariable (MHC) :

```
                    -2 log LR df p
Model_Multi_Cov3  3186.542 12 0
```

- Comparaison MHC et modèle Markovien non-homogène (MNH) :

```
                      -2 log LR df p
Model_Multi_IHMG_bis  139.9447  6 0
```

Le modèle à intensités constantes par morceau est bien celui qui décrit au mieux nos données.

d. Conclusion

Nous avons pu voir dans cette partie la modélisation par des modèles multi-états, dans un cadre Markovien homogène puis non-homogène.

Cette approche permet une modélisation année après année de la présence en portefeuille des individus en prenant en compte les retours, et non plus une approche globale comme vue lors de l'analyse de séquence, ou encore une approche année après année mais sans retour possible comme avec l'analyse de survie.

Néanmoins, les résultats donnés par l'approche homogène sont loin d'être satisfaisants, comme souvent dans un cadre assurantiel.

L'approche non-homogène semble quant à elle être de meilleure qualité, mais une sous-estimation de la présence est visible sur la partie de la carrière importante dans le calcul de l'IDR. D'autres variables pourront être envisagées pour améliorer la qualité de notre modèle. De plus, les méthodes semi-markoviennes, non appliquées ici, sont aussi à envisager pour prendre en compte la durée dans l'état.

Nous verrons dans la dernière partie les conséquences en termes de prestations estimées et le lien avec l'approche par répartition.

V. Application au sein de l'entreprise

Les travaux effectués ci-dessus s'inscrivent dans le cadre de la mise en place de la nouvelle garantie d'indemnité de départ volontaire à la retraite. Il nous faut alors observer les résultats des différents modèles sur l'ancienneté et plus précisément sur les droits acquis lors d'une carrière.

Reprenons le tableau obtenu pour l'approche par répartition vu dans la partie descriptive, mais cette fois pour la génération née en 1957 :

Ancienneté dans les 7 dernières années précèdent la date de retraite	Ancienneté totale du salarié	Montant de l'indemnité	Réparition de la génération 57	Réparition de la génération 58	Montant d'IDR théorique moyen 57	Montant d'IDR théorique moyen 58
Moins de 5 ans	-	0	81%	79%	-	-
5 ans ou plus	**Moins de 10 ans**	0	2.8%	3.1%	-	-
	Au moins 10 ans	1 mois	3.8%	4.0%	699 €	725 €
	Au moins 15 ans	1.5 mois	4.1%	4.3%	1 156 €	1 180 €
	Au moins 20 ans	2 mois	6.6%	7.2%	1 813 €	1 847 €
	Au moins 30 ans	2.5 mois	1.9%	2.1%	2 661 €	2 667 €

Tableau 5.0.1 - Montant d'indemnisation théorique pour la génération née en 1957 et 1958

Les droits relatifs à la garantie IDR sont très semblables entre les deux générations concernées. On peut s'intéresser alors l'estimation des différentes méthodes de modélisation.

a. Comparaison des différentes méthodes et analyse du P/C

On regarde dans un premier temps la répartition selon les conditions d'ancienneté de l'IDR, puis nous verrons l'impact par rapport au montant global estimé par la méthode par répartition.

Ancienneté dans les 7 dernières années précèdent la date de retraite	Ancienneté totale du salarié	Montant de l'indemnité	Réparition de la génération 57	Réparition par Kaplan Meier	Réparition modèle de Cox	Réparition par modèle Aalen	Réparition par modèle multi-état
Moins de 5 ans	-	0	81%	88%	88%	81%	81%
5 ans ou plus	**Moins de 10 ans**	0	2.8%	4.9%	5.9%	5.4%	6.0%
	Au moins 10 ans	1 mois	3.8%	2.6%	2.8%	8.4%	3.0%
	Au moins 15 ans	1.5 mois	4.1%	1.9%	1.7%	3.7%	6.9%
	Au moins 20 ans	2 mois	6.6%	2.2%	1.2%	1.4%	3.2%
	Au moins 30 ans	2.5 mois	1.9%	0.3%	0.1%	0.5%	0.0%

Tableau 5.1.1 - Répartition par méthode selon les conditions sur l'ancienneté

Le tableau ci-dessus présente les répartitions des droits d'IDR pour la population née en 1957. On remarque que l'ensemble des méthodes surestiment le pourcentage ayant 5 années de présence avant la retraite mais moins de 10 années au total. Ceci correspond au maintien en activité pour les personnes rentrant tardivement en branche (entre 5 et 9 années avant la retraite). En revanche, on sous-estime systématiquement le pourcentage pour les catégories « 2mois » et « 2.5 mois ».

Les méthodes par Kaplan Meier et modèle de Cox, deux analyses de survie, ont l'intégralité des catégories où l'on verse un montant d'indemnité non nul fortement en dessous de la répartition observée. Ceci s'explique par le manque de la notion de retour en branche après une sortie, ce qui pourrait se compenser en le combinant avec un modèle du retour en activité.

Le modèle d'Aalen, lui, donne une proportion beaucoup plus élevée de personne ayant droit à un mois d'indemnité (8.4% contre 3.8% pour l'observé). Cependant, on conserve la mauvaise modélisation des anciennetés importantes, avec respectivement 1.4% contre 6.6% pour 2 mois de droits, et 0.5% contre 1.9% pour 2.5 mois.

Enfin, la modélisation par modèle multi-état markovien à intensité constante par morceaux surestime les deux premières catégories de droits d'IDR . De plus, on observe un nombre nul de présence pour la catégorie « 2.5 mois », s'expliquant par la modélisation trop faible de présence sur le graphique *3.3.2.1*.

En termes de prestations simulées, nous prendrons une base 100 pour garder la confidentialité de la donnée. Ainsi on a :

Approche par répartition	Approche par estimateur de Kaplan Meier	Modèle de Aalen	Modèle multi-état	Analyse de séquence
100%	37%	59%	65%	88%

Tableau 5.1.2 - Comparaison de la prestation totale selon les différentes méthodes pour la génération née en 1957

On s'aperçoit que les méthodes prédictives sont très éloignées de la réalité. Ceci peut s'expliquer par plusieurs paramètres. Pour les méthodes de survie, la raison majeure de l'écart est la non prise en compte du retour. Pour le modèle multi-état, une des raisons que l'on peut expliciter est le manque de données, par exemple la connaissance des différents types de sorties de branche ou le type de profession exercée.

Dans le calcul des prestations, on a rajouté l'analyse de séquence bien que la méthode soit plutôt descriptive que prédictive. A l'aide des groupes et de leur décomposition fournie par le graphique *3.2.5.1*, on regarde la probabilité d'être dans chaque groupe pour un individu et pour chaque groupe le nombre de mois de droits moyen. Ceci, associé à un salaire moyen donne une estimation générale de la prestation pour une génération.

Cette méthode est celle qui présente la meilleure estimation pour la génération née en 1957 sur base des données de celle née en 1958. Nous n'avons cependant pas d'idée précise de la forme de la carrière des individus, le mieux que l'on puisse faire étant de lier la probabilité d'être dans un groupe et une forme représentative du groupe, par exemple basée sur la séquence qui revient le plus.

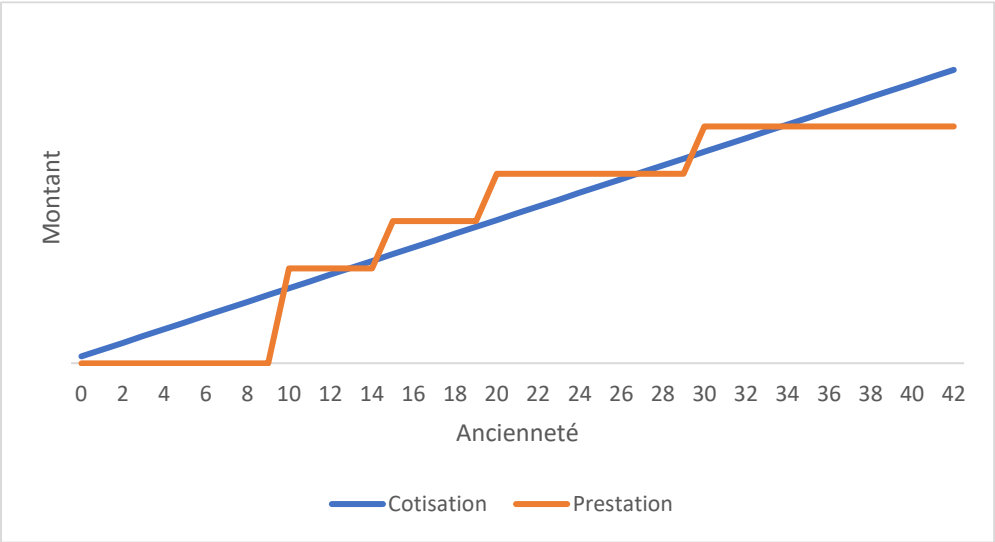
b. Scénarios de stress

On peut maintenant s'intéresser à la sensibilité de ce montant de prestations. L'arrivée de Solvabilité II incite les assureurs à recourir de plus en plus à des stress-tests qui permettent de juger de l'impact d'une aggravation de la situation vis-à-vis d'un ou plusieurs risques.

Ayant un montant de prestations, vu ci-dessus, défini en fonction de l'ancienneté et du salaire, nous nous intéressons alors à la variation de ce montant en cas de variations du salaire des individus.

En effet, la convention mise en application au 1er janvier 2022 informe que la cotisation d'indemnité de départ volontaire à la retraite est définie comme un pourcentage prédéfini du salaire.

On peut représenter sur un graphique l'évolution de la prestation et de la cotisation en fonction de l'ancienneté pour un salaire annuel de 12 000€, sans revalorisation, et en supposant que l'individu valide le critère de présence à la retraite.

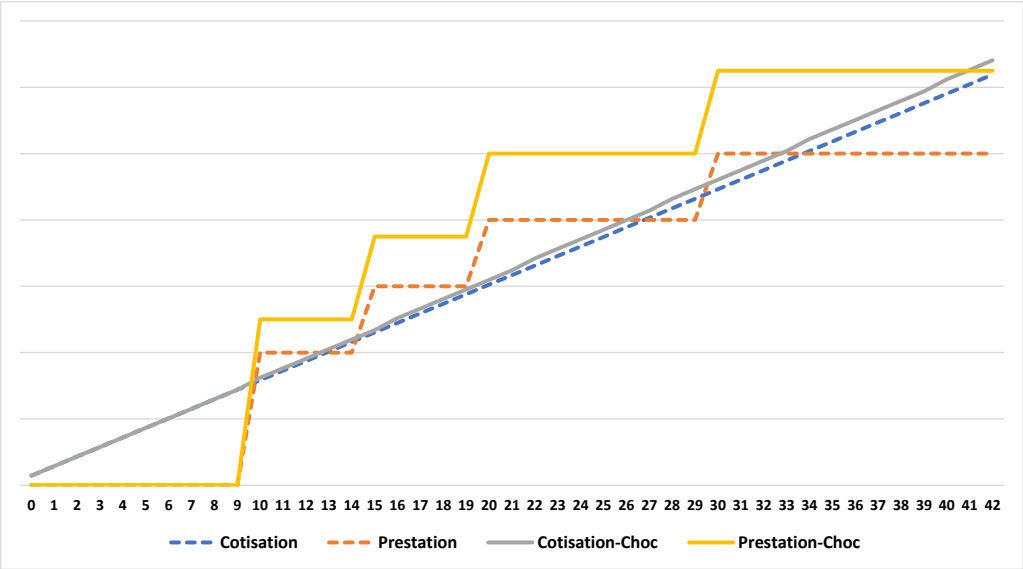


Graphique 5.2.1 - Relation entre la prestation et la cotisation sans revalorisation de salaire dans le cadre de l'IDR

En abscisse, on retrouve l'ancienneté dans la branche. En ordonné, les axes ont été effacés pour des raisons de confidentialité, nous avons le montant en euros.

On va s'intéresser à un scénario de dérive salariale sur les dernières années avant la retraite. Cela aura un impact principalement sur le montant de prestations de par l'augmentation du salaire de référence, mais aussi sur le montant de cotisations.

Si l'on prend maintenant en compte une évolution de 25% du salaire sur les 5 années avant la retraite, la relation se transforme et devient :



Graphique 5.2.2 - Relation entre la prestation et la cotisation avec revalorisation de salaire de référence dans le cadre de l'IDR

On voit que le rapport prestations sur cotisations augmentent avec l'ancienneté si le salaire de l'individu augmente significativement lors de l'approche du passage en retraite. Ainsi, le surplus de cotisations ne permet pas de compenser la perte liée à cette augmentation brusque du salaire de référence.

Si l'on reprend l'exemple de la génération née en 1957, toujours en partant d'une base 100, et que l'on applique une augmentation de 25% du salaire de référence pour le calcul du montant de l'IDR, soit les 5 dernières années, on obtient les résultats suivants :

Prestations supplémentaires dues au choc	10%	25%	50%	100%
Cotisations supplémentaires dues au choc	2%	6%	11%	23%

Tableau 5.2.1 - Surplus de cotisations et de prestations

On observe bien le déséquilibre qui se forme suivant l'augmentation du salaire de référence des individus.

Dans un second temps, nous supposerons que l'ensemble des salariés, présents en branche 5 années avant leur retraite théorique, vont tous rester actif jusqu'à la retraite, et ainsi auront tous pu acquérir des droits concernant la présence avant la retraite.

Ancienneté dans les 7 dernières années précèdent la date de retraite	Ancienneté totale du salarié	Montant de l'indemnité	Réparition de la génération 57	Montant d'IDR théorique moyen	Réparition-stressée de la génération 57	Montant d'IDR stressé moyen
Moins de 5 ans	-	0	81%	- €	55%	- €
5 ans ou plus	**Moins de 10 ans**	0	3%	- €	21%	- €
	Au moins 10 ans	1 mois	4%	699 €	7%	572 €
	Au moins 15 ans	1.5 mois	4%	1 156 €	6%	1 055 €
	Au moins 20 ans	2 mois	7%	1 813 €	9%	1 820 €
	Au moins 30 ans	2.5 mois	2%	2 661 €	2%	2 670 €

Tableau 5.2.2 - Comparaison de la répartition observée et la répartition sous le scénario choc

On obtient alors le montant de prestations et de cotisations supplémentaires suivant :

Prestations supplémentaires dues au choc	33%
Cotisations supplémentaires dues au choc	3%

Tableau 5.2.3 - Surplus de cotisations et de prestations

Comme pour le premier choc, on voit que le surplus de cotisations ne permet pas de combler le surplus de prestations.

Ces chocs permettent de mettre en avant la sensibilité du montant de prestations vis-à-vis de deux hypothèses sur les deux éléments caractéristiques de l'IDR, à savoir le salaire de référence et la présence en branche.

Il nous faudra alors, dans la continuité de notre étude, dans un premier temps améliorer et solidifier les méthodes de prédictions de la présence en branche. Ensuite, un modèle de suivi des salaires semble indispensable au vu de la sensibilité illustrée ci-dessus.

Conclusion

Ce mémoire a permis d'introduire un travail sur les assurés de l'IRCEM. Nous avons pu mettre en lumière les comportements des individus et mettre en avant les caractéristiques spécifiques aux métiers du salarié du particulier employeur. Les méthodes prédictives se révèlent peu utilisables en l'état. L'analyse de survie nous permet d'avoir une bonne approche sans retour, cependant une table d'entrée en présence et une de retour en branche semblent indispensables pour espérer bien approcher notre risque. L'analyse multi-état sous-estime fortement notre présence sur les âges directement concernés par l'IDR, compte tenue de la prise en compte des retours dans ce type de modèle. Le nombre de variables explicatives limitées par nos données ou le manque de contexte de sorties sont des facteurs pouvant expliquer le manque de fiabilité de nos modèles. L'ajout de nouvelles données permettra de compléter et d'améliorer la capacité prédictive de ces modèles.

La visualisation par analyse de séquence a permis de prendre conscience des profils des assurés au sein de la branche. Ce type d'analyse est intéressant pour mieux cerner les comportements de ces derniers et sera de plus utilisé pour de futures analyses de population, en croisant l'analyse avec les types de métiers exercés par les particuliers employeurs par exemple.

Les méthodes citées ci-dessus ont été adoptées dans le cadre de la mise en place de l'indemnité de départ volontaire à la retraite, qui est une nouveauté pour l'IRCEM Prévoyance en 2022. Cependant, au vu des résultats et de la donnée modélisée, la présence en portefeuille, les applications ne se limitent pas à ce contexte. Les travaux d'estimation de masse salariale, essentiels pour un assureur dans le cadre de l'évaluation des cotisations à percevoir, sont un exemple d'utilisation possible. Pouvoir estimer le montant de cotisations, qui dépend essentiellement de la présence en branche et du salaire des individus, permet de déterminer avec plus de précision un ratio de solvabilité.

L'évolution du secteur et le contexte actuel, dans lequel s'inscrit la convergence des deux branches principales du salarié du particulier employeur, est un des enjeux des prochaines années. J'espère sincèrement que les méthodes et applications mises en œuvre dans ce mémoire seront utiles pour le développement de l'IRCEM Prévoyance.

BIBLIOGRAPHIE

Kaplan E. L., Meier P. [1958] « Nonparametric estimation from incomplete observations », Journal of the American Statistical Association, Vol.53, 457-481.

Laurencell [2009] « Le tau et le tau-b de Kendall pour la corrélation de variables ordinales simples ou catégorielle » Tutorials in Quantitative Methods for Psychology 2009, Vol. 5(2), p. 51-58. 51.

Cox D. R. [1972] « Regression Models and Life-Tables », Journal of the Royal Society, Series B (Methodological), Vol.34, 2.

Gamerman D. [1991] « Dynamic bayesian models for survival data. », Journal of the Royal Statistical Society. Series C, 40 : 63–79.

Aalen O. [1978] « Non-parametric inference for a family of counting processes », Annals of Statistics, Vol.6, 701 : 726.

TEODORESCU B.; VAN KEILEGOM I.; CAO R. [2005] «Generalized Time-dependent Conditional Linear Models under Left Truncation and Right Censoring» UCL, Working Paper.

Choukroun M. [2008] « Le modèle additif d'aalen, une alternative au modèle de cox dans le cadre de la construction d'une loi de maintien en incapacité de travail. » Bulletin Français d'Actuariat, 8(16) : 107–138.

Robette N. [2011] « Explorer et décrire les parcours de vie : les typologies de trajectoires » CEPED, pp.86.

Abbott A. [1995]. «A comment on 'Measuring the agreement between sequences.' » Sociological Methods & Research 24(2):232–43.

Kaufman, L. and Rousseeuw, P.J. (1990) «Partitioning around Medoids (Program PAM). In: Kaufman, L. and Rousseeuw, P.J., Eds., Finding Groups in Data: An Introduction to Cluster Analysis» John Wiley & Sons, Inc., Hoboken, 68-125.

Guibert Q. [2015] « Sur l'utilisation des modèles multi-états pour la mesure et la gestion des risques d'un contrat d'assurance » Gestion et management. Université Claude Bernard - Lyon I.

Saint Pierre P., [2005], Modèles multi-états de type Markovien et application à l'asthme, Thèse, Montpellier : Université Montpellier I, 202 p.

Laurans C. [2008] « L'Asset Ceiling en IAS 19, Comptabilisation et sensibilités aux hypothèses actuarielles », Mémoire de fin d'études EURIA.

Planchet F. [2021] « Modèles de durée, Statistiques des modèles non paramètriques » Support de cours, ISFA.

Planchet F. [2020-2021] « Modèle de durée, Arrêt de travail » Support de cours, ISFA.