N°10
Janvier 2016

Les cahiers de la CNAV

Document de travail

Durée du cumul RG / RSI : une application des modèles de durée

I Agathe DARDIER I

Sommaire

Durée du cumul RG / RSI : une application des modèles de durée

Grâce à un rapprochement des données du RSI et de la CNAV sur la population âgée de 55 ans et plus, il est possible d'identifier les indépendants ayant demandé leur retraite du régime général. Au 31 décembre 2012, parmi les 573 471 cotisants de 55 ans et plus du RSI, 146 252, soit 26 %, sont également retraités du régime général. Afin de compléter les travaux déjà réalisés sur le cumul emploi-retraite RSI-RG[1], une analyse de la durée passée en cumul a été mise en œuvre. L'objectif de cette note est de montrer l'intérêt et l'application des modèles de durée à partir de l'exemple du cumul emploi-retraite RSI-RG (présentation des modèles, des programmes et des sorties SAS). D'après les modèles non paramétriques et le modèle de Cox, plus de la moitié des indépendants reste en cumul pendant au moins 4 ans. La durée du cumul emploi-retraite s'explique principalement par la carrière des individus, plus que par l'activité exercée en parallèle de la retraite.

[1] *Bac C., Gaudemer C., 2010 « Actif au RSI et retraité au régime général», Zoom sur, n°41, RSI.*
Bac C., Gaudemer C., 2012 « Actif au RSI et retraité au Régime général - évolution de cette situation de cumul entre 2008 et 2010 », Zoom sur, n°64, RSI.
Dardier A., Gaudemer C., 2014 « Actif au RSI et retraité au général à la fin 2012 », Zoom sur, n°82, RSI.

INTRODUCTION

Depuis 2009, la CNAV et le Régime Social des Indépendants (RSI) rapprochent leurs données sur la population âgée de 55 ans et plus afin d'identifier les cotisants du RSI ayant demandé leur retraite au Régime général. Etant donné que la plupart des cotisants du RSI ont également eu une activité en tant que salarié du secteur privé au cours de leur carrière, ils ont acquis des droits propres à la retraite au Régime général. Au 31 décembre 2012, près de 150 000 cotisants du RSI sont retraités du régime général, ce qui correspond à la moitié des cotisants du RSI de 60 ans et plus. Dans le document, les cotisants du RSI retraités du régime général sont appelés les cumulants RSI-RG, et plus simplement les cumulants.

L'appariement des données du RSI et de la CNAV a permis d'identifier les caractéristiques des cumulants, notamment en fonction de l'âge, de l'activité exercée en tant qu'indépendant, de la situation avant la liquidation d'une retraite du régime général, et des niveaux de revenu ou de pension. L'évolution du cumul d'un emploi d'indépendant et d'une retraite du régime général entre 2008 et 2012 a également pu être étudiée. Afin d'approfondir les connaissances sur ce sujet, il est intéressant d'évaluer la durée de ce cumul. Or, même si l'appariement des données depuis 2008 fournit beaucoup d'information, il ne permet pas de suivre l'ensemble des cumuls pendant toute leur durée. Les modèles de durée ont donc été mobilisés. Ils permettent d'estimer la répartition au cours du temps des sorties de cumul et de rechercher l'influence exercée par différents facteurs sur cette répartition.

Les modèles de durée ont été initialement exploités dans le domaine de la bio-statistique et de la démographie. Néanmoins, ils sont de plus en plus utilisés pour analyser les politiques économiques et sociales, ainsi que le marché du travail. Ainsi, Ariane Pailhé et Anne Solaz[2] mobilisent les modèles de durée pour étudier le retour à l'emploi des mères selon le rang de naissance des enfants. Elles montrent que les interruptions d'activité pour s'occuper des enfants augmentent et s'allongent au fil des naissances, en fonction de l'attachement au travail et au degré d'employabilité. Le travail de Thierry Magnac, Benoît Rapoport et Muriel Roger est un autre exemple du recours aux modèles de durée [3]: ils étudient la transition entre la fin de carrière et le départ à la retraite. Les auteurs concluent que la survenue d'un « accident de carrière » (chômage ou préretraite) durant les dernières années d'activité avance l'âge de départ en retraite de la génération 1930.

L'objectif de cette note est de montrer l'intérêt des modèles de durée et leur mise en œuvre sous SAS à travers l'étude de la durée passée en cumul emploi-retraite RSI-RG. Après avoir présenté brièvement les principales caractéristiques des cumulants, l'étude analyse la durée du cumul emploi-retraite à partir de 2 types de modèle de durée (les modèles non paramétriques, le modèle semi-paramétrique de Cox), pour lesquels chacune des étapes de leur réalisation est explicitée.

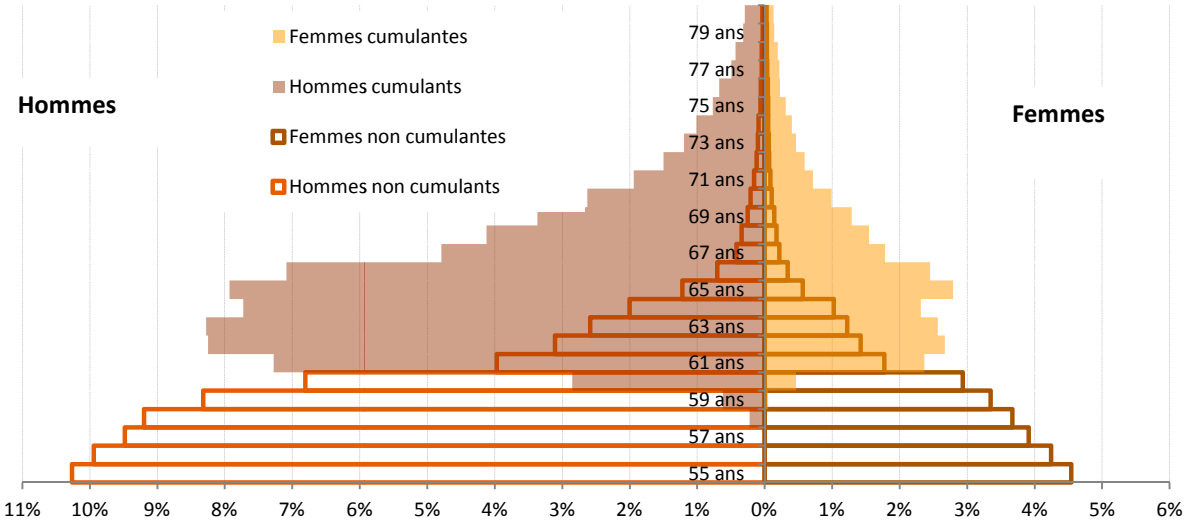
2 Pailhé A., Solaz A., 2012, « Durée et conditions de retour à l'emploi des mères après une naissance », Retraite et société, n°63, p53-75
3 Magnac T., Rapoport B., Roger M., 2006, « Fins de carrière et départs à la retraite : l'apport des modèles de durée », Solidarité et santé, n°3, p101-117.

1. PRESENTATION DU CUMUL EMPLOI-RETRAITE RSI-RG

Depuis le 1er janvier 2004, la poursuite ou la reprise d'une activité relevant de certains régimes (RSI, CNRACL, MSA principalement), ne s'oppose plus au service de la retraite du régime général. Ainsi, une personne qui exerce une activité artisanale, commerçante ou de profession libérale, et qui, avant d'être affiliée au régime social des indépendants (RSI), a cotisé au régime général, peut demander sa retraite de salarié tout en poursuivant une activité non salariée, sans être soumise à aucune contrainte[4].

Au 31 décembre 2012, parmi les 573 471 cotisants de 55 ans et plus du RSI (au titre de la maladie ou de la vieillesse, voir encadré 1), 146 252, soit 26 %, sont également retraités du régime général. Les personnes cumulant activité indépendante et retraite au Régime général sont plus âgées que les cotisants du RSI de 55 ans et plus (graphique 1). En raison de la législation, les actifs du RSI bénéficient d'une retraite du régime général le plus souvent à partir 60 ans. Les retraités ayant également une activité d'indépendant sont de jeunes retraités du régime général : 25% ont entre 60 ans et 62 ans, et plus de 8 sur 10 ont moins de 70 ans.

Graphique 1 : Répartition par âge des cotisants au RSI de 55 ans et plus selon qu'ils aient ou non liquidé leur pension du Régime général au 31 décembre 2012



Source : Panel des cumulants RSI-RG 2008-2012
Champ : cumulants RSI-RG au 31/12/2012 ayant liquidé leur retraite du Régime général depuis 2004

Le cumul emploi-retraite RSI-Cnav est principalement masculin (71%). Les caractéristiques des hommes et des femmes cumulants sont proches, et c'est pourquoi, le reste de l'étude du cumul n'est pas détaillée par sexe.
41 % des cumulants sont des commerçants, 36 % sont des professions libérales[5] et 23 % sont des artisans (tableau 1). Les cotisants du RSI ayant liquidé leur retraite au Régime général sont plus fréquemment auto-entrepreneurs (4 fois sur 10), que ceux qui n'ont pas liquidé de pension (1 fois sur 5). Toutefois, plus d'un quart des cumulants ayant opté pour le statut d'auto-entrepreneur a déclaré un chiffre d'affaires nul au titre de l'année 2012.

4 Il faut toutefois noter que pour les personnes parties en retraite après le 1er janvier 2015, la liquidation d'une retraite au régime général n'entraînera pas l'ouverture de nouveau droit à retraite au RSI, en cas de poursuite ou de nouvelle activité d'indépendant. Cette nouvelle législation ne s'applique pas aux cumulants RSI-RG de cette étude, car ils ont tous liquidé leur retraite avant le 31 décembre 2011.
5 Il s'agit de toutes les professions libérales sauf des médecins du secteur 1, de certains médecins du secteur II, et de quelques artistes-auteurs.

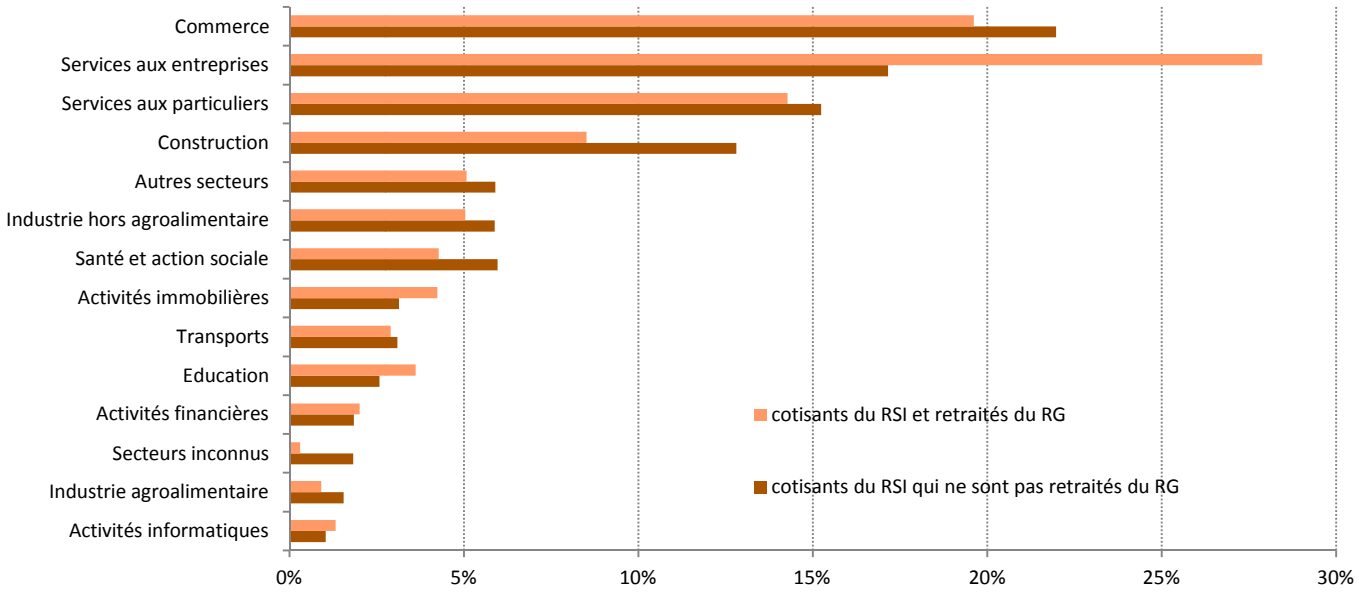
Tableau 1 : Cotisants au RSI de 55 ans et plus au 31 décembre 2012 selon leur groupe professionnel et le statut d'auto-entrepreneur

	Cotisants au RSI au 31/12/2012	Cotisants RSI et retraités au Régime général au 31/12/2012	Répartition des cotisants au RSI au 31/12/2012	Répartition des cotisants au RSI et retraités au Régime général au 31/12/2012	Taux de cumulants
Artisans	163 253	33 889	28,5%	23,2%	20,8%
Commerçants	240 704	60 391	42,0%	41,3%	25,1%
Professions libérales[5]	169 514	51 972	29,6%	35,5%	30,7%
Auto-entrepreneurs	135 676	54 852	23,7%	37,5%	40,4%
Non auto-entrepreneurs	437 795	91 400	76,3%	62,5%	20,9%
Ensemble	573 471	146 252	100,0%	100,0%	25,5%

Source : Panel des cumulants RSI-RG 2008-2012
Champ : cumulants RSI-RG au 31/12/2012 ayant liquidé leur retraite du Régime général depuis 2004

En termes de secteur d'activité, les cotisants du RSI de 55 ans et plus travaillent principalement dans les domaines du commerce (21 %), des services aux entreprises (20%) ou aux particuliers (15 %), et dans la construction (12 %). Les cumulants s'engagent moins dans la construction, mais plus souvent dans les services aux entreprises (28 %) et aux particuliers (14 %).

Graphique 2 : Répartition des cotisants au RSI de 55 ans et plus selon le secteur d'activité au 31 décembre 2012



Source : Panel des cumulants RSI-RG 2008-2012
Champ : cumulants RSI-RG au 31/12/2012 ayant liquidé leur retraite du Régime général depuis 2004

Encadré 1 : Sources

Le RSI assure la couverture maladie et vieillesse des artisans et des commerçants. Il assure également la couverture maladie-maternité de toutes les professions libérales sauf des médecins du secteur 1, de certains médecins du secteur II, et de quelques artistes-auteurs.
Afin de mieux connaître la population cumulant un emploi et une retraite, une base statistique a été constituée par le RSI et la CNAV pour quantifier le nombre de cotisants du RSI qui ont fait valoir leurs droits à la retraite au Régime général.

Avec l'accord de la CNIL, le RSI et la CNAV procèdent à un appariement, sur données individuelles, des informations carrière et retraite des deux régimes. La base constituée est anonymisée et contient, au 31 décembre 2012, 788 952 observations d'assurés cotisants au RSI âgés de 55 ans et plus.

La première opération entre la CNAV et le RSI a eu lieu au second semestre de l'année 2009 pour une situation arrêtée au 31 décembre 2008. L'opération a été renouvelée début 2011 afin de prendre en compte les situations des cotisants du RSI des années 2009 et 2010 ; puis de nouveau en 2013 sur les données au 31 décembre 2012. A compter de ce dernier appariement, la base est constituée sous forme de panel afin de suivre individuellement la situation de cumul emploi-retraite.

Une personne est reconnue comme cumulant RSI-Cnav lorsqu'elle cotise au RSI après avoir liquidé sa retraite du régime général. Pour compléter cette définition de cumulant, il aurait été intéressant de prendre en compte une condition de revenu d'activité. Cependant, les indépendants ont des revenus très variables, et parfois des revenus nuls (surtout les auto-entrepreneurs). Il ne semble donc pas pertinent d'utiliser les revenus pour définir les cumulants.

Les dates précises concernant le début et la fin d'activité sont disponibles dans la base. Par rapport à la partie 2 et pour ne pas fausser l'estimation de la durée du cumul emploi-retraite, dans la suite de l'étude, les personnes dont la durée de cumul est inférieure à 1 mois ont été retirées (243 personnes). Il a été considéré que la situation de cumul résultait uniquement du délai pour mettre fin à leur activité relevant du RSI. La population prise en compte dans l'étude de la durée du cumul emploi-retraite n'est donc pas identique à celle retenue pour les statistiques descriptives de la partie 2.

En 2012, 51 % des cotisants du RSI ayant liquidé une retraite au Régime général après 2003 exerçaient déjà une activité d'indépendant. 39 % étaient salariés et ont débuté une activité d'indépendant après leur départ en retraite, de même que 9 % des cumulants qui n'étaient ni salariés, ni indépendants (tableau 2).
La moitié des cumulants qui exercent une profession libérale[6] étaient salariés avant la liquidation de leur retraite. Inversement, plus de la moitié des commerçants ayant une pension au Régime général exerçait déjà une activité d'indépendant, lors de la liquidation.

Tableau 2 : Répartition des cumulants selon la situation au moment de la liquidation RG et la profession actuelle des indépendants

	Artisans	Commerçants	Professions libérales	Ensemble
Salarié au moment de la liquidation RG	32,0%	32,0%	52,3%	39,2%
Indépendant au moment de la liquidation RG	60,2%	58,3%	37,6%	51,4%
Non salarié du privé et non indépendant au moment de la liquidation RG	7,7%	9,8%	10,2%	9,4%
Total	100,0%	100,0%	100,0%	100,0%

Source : Panel des cumulants RSI-RG 2008-2012
Champ : cumulants RSI-RG au 31/12/2012 ayant liquidé leur retraite du Régime général depuis 2004

[6] *Il s'agit de toutes les professions liberales sauf des médecins du secteur 1, de certains médecins du secteur II, et de quelques artistes-auteurs.*

2. PRESENTATION GENERALE DES MODELES DE DUREE ET CREATION DES VARIABLES NECESSAIRES A LEUR MISE EN OEUVRE

Les statistiques descriptives permettent de connaître les caractéristiques des personnes en cumul emploi-retraite. Comme l'un des objectifs de l'étude est de comprendre quel est l'usage de ce dispositif, il est intéressant d'étudier sa durée d'utilisation. Si le cumul emploi-retraite était observé en intégralité pour chaque individu, il serait possible de calculer parfaitement des indicateurs comme la durée moyenne. Or, il n'est pas observé en entier pour tous les individus ; les modèles de durée permettent alors de calculer des indicateurs de durée non biaisés. Par ailleurs, nous cherchons aussi à expliquer cette durée. Il est donc nécessaire d'utiliser des modèles de régression qui prennent en compte une dimension temporelle : les modèles de durée offrent cette possibilité.

Il existe 3 types de modèles de durée : non paramétrique, paramétrique, et semi-paramétrique. L'étude du cumul emploi-retraite repose sur la mise en œuvre de modèles non paramétriques et semi-paramétriques[7]. Avant de présenter ces deux types de modèle, nous créons les variables nécessaires à la réalisation d'une analyse de durée.

[7] *Les modèles paramétriques sont brièvement explicités en annexe 1*

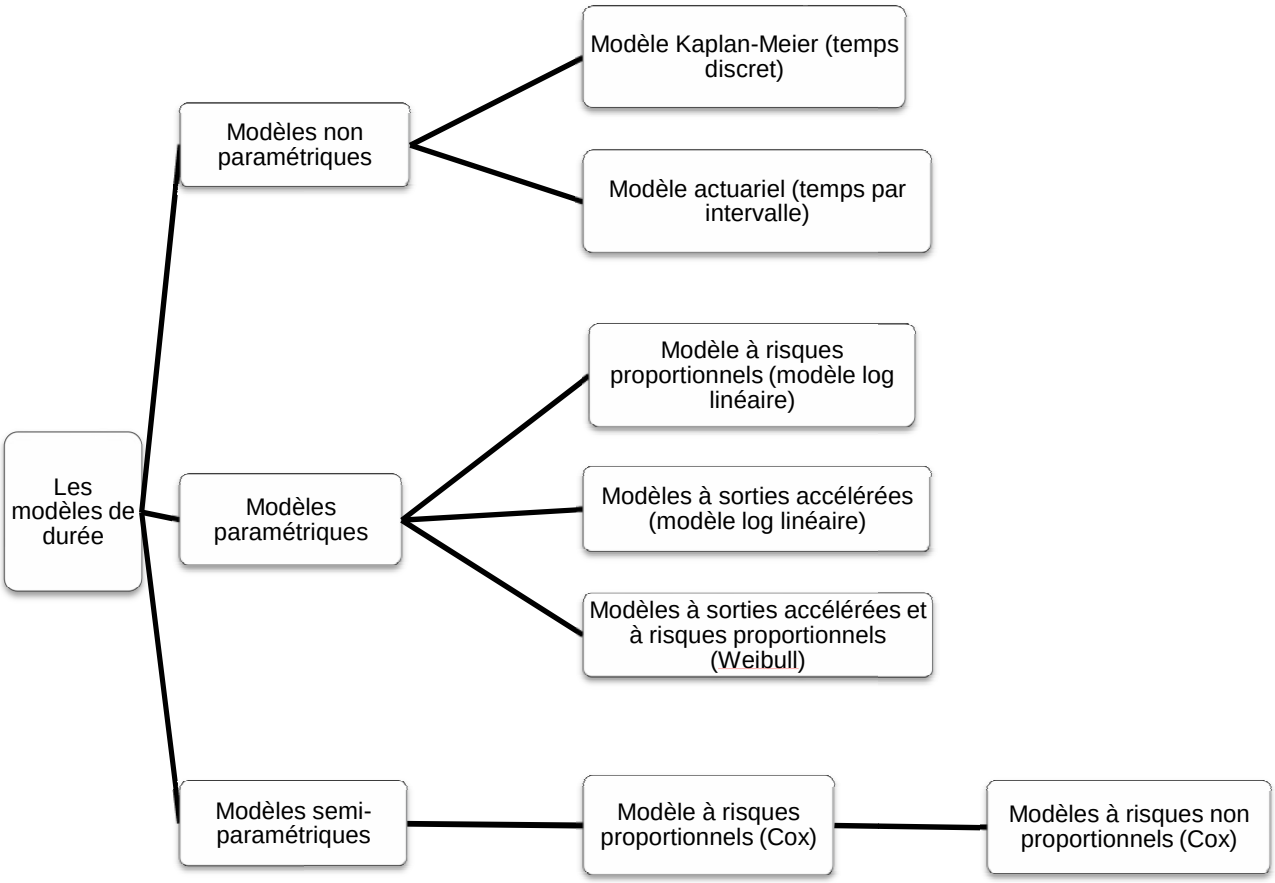


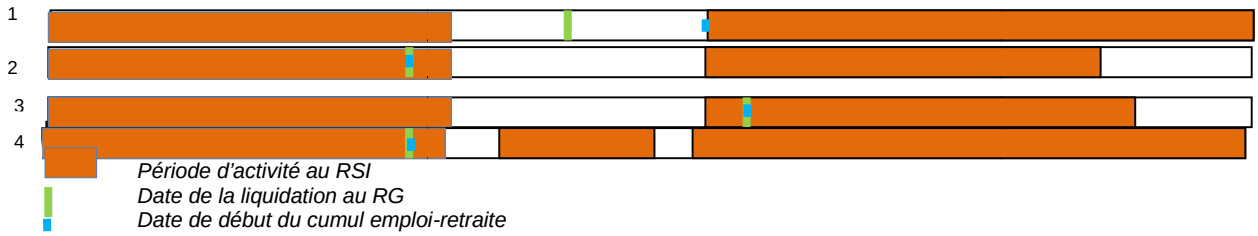
Figure 1 : Les différents modèles de durée

2.1. Définir la durée

Les modèles de durée reposent sur l'étude d'une variable continue et positive qui représente la durée. Pour définir cette durée, il faut au préalable déterminer une date initiale et l'évènement étudié.

L'évènement étudié correspond à la fin du cumul emploi-retraite, qui se traduit par la fin de l'activité d'indépendant.

La date initiale correspond au début de l'exposition au risque que l'évènement se produise. Il s'agit rarement de la même date pour tous. Dans notre cas, la date initiale correspond à la date de début du cumul emploi-retraite, soit la date à laquelle l'individu est pour la première fois simultanément en activité au RSI et retraité au régime général. Voici différents cas possibles :



La durée d'une période de cumul emploi-retraite est la différence entre la date d'origine (la date d'entrée dans le cumul) et la survenue de l'évènement (la sortie du cumul emploi-retraite). Pour les personnes ayant plusieurs périodes de cumul, c'est la somme des durées de ces périodes qui est retenue. Ainsi, par rapport au cas n°4 ci-dessus, la durée de cumul ne prend pas en compte les 2 périodes sans activité (en blanc) qui ont eu lieu au cours du cumul. Ces cumuls interrompus représentent moins de 4% de l'ensemble des cumulants RSI-RG, et sont difficiles à expliquer (exemple : auto-entrepreneur qui arrête temporairement son activité).

Dans notre étude, la variable correspondant à la durée du cumul emploi-retraite est appelée ***duree***. Elle est calculée de date à date, et exprimée en année. Comme nous connaissons précisément (jour, mois, année) la date d'effet de la pension, la date de début et de fin d'activité en tant qu'indépendant, il est possible de calculer précisément la durée de la période de cumul emploi-retraite.

Par exemple, pour un cumul débutant le 1er mai 2008 et s'achevant le 15 juillet 2008, la variable *duree* correspond à 0,208 (76 jours/365 jours).

2.2. Les fonctions sur lesquelles reposent les modèles de durée

La fonction de survie[8] S(t) est la probabilité d'être dans l'état initial, dans notre cas en cumul emploi-retraite, au moins jusqu'en t. Elle donne également la proportion de la population qui n'a pas encore connu l'évènement, soit la fin du cumul emploi-retraite, au bout d'une certaine durée t.

La fonction de risque instantané (appelée souvent fonction hasard) h(t) est la probabilité que l'évènement (la sortie du cumul emploi-retraite) se produise entre t et un intervalle de temps court ðt, sachant que l'évènement ne s'est pas produit avant et en t.

Les fonctions de risque instantané et de survie sont liées et la connaissance de l'une permet d'en connaître l'autre (encadré 2).

Encadré 2 : Fonctions de base des modèles de durée

Soit T, la variable aléatoire positive qui représente la durée

La fonction de répartition = fonction de densité cumulée F(t) :

$F(t) = \int_0^t f(s)ds = P(T < t)$. F(t) est la probabilité que l'épisode se termine au plus tard à l'instant t.

Propriété : F(0)=0 et $\lim_{t\to\infty} F(t) = 1$

La fonction de survie S(t) (complément à 1 de la fonction de répartition), appelée aussi fonction de séjour. $S(t) = 1 - F(t) = 1 - P(T < t) = P(T \geq t)$.

Propriété : S(0)=1 et F(t)+S(t)=1

C'est la probabilité que l'évènement dure au moins jusqu'à l'instant t.

Le taux de hasard ou le risque instantané h(t)

Le risque est construit à partir de la probabilité de survenue de l'évènement durant l'intervalle [t+ ∂t] sachant qu'il ne s'était pas réalisé avant t : $P(t \leq T \leq t + \partial t | T \geq t)$.

[8] *Les modèles de durée ont été en premier lieu mobilisés pour étudier la durée de vie, d'où le nom des fonctions de ces modèles.*

Si on passe alors à une évaluation du risque de connaître l'évènement durant un intervalle de temps considéré : $\frac{P(t \leq T \leq t+\partial t | T \geq t)}{\partial t}$

on peut définir une fonction de risque qui apparaît comme une mesure du risque instantané. (Attention : il ne s'agit pas d'une probabilité, et le risque peut donc prendre des valeurs comprises entre 0 et l'infini).

$$h(t) = \frac{\mathrm{f(t)}}{\mathrm{S(t)}} \sim \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t)}{\partial t P(t \leq T)}$$

La fonction de risque cumulé H(t) (risque instantané h(t) cumulé) est une pseudo probabilité de connaître l'évènement si un individu était constamment soumis au risque.

$$H(t) = \int_0^t h(u)du = \int_0^t \frac{-dLog(S(u))}{du} = -\log(S(t)$$

2.3. La population soumise au risque

Pour réaliser une analyse de durée, il est nécessaire de connaître la population soumise au risque, c'est-à-dire la population qui est potentiellement concernée par l'évènement étudié. Dans notre cas, nous disposons des données sur les cumuls emploi-retraite effectués entre le 1er janvier 2008 et le 31 décembre 2012. Afin de mener une étude avec la plus grande durée d'observation possible, la population choisie correspond aux personnes ayant débuté un cumul emploi-retraite en 2008.

Par ailleurs, il n'est pas possible d'étudier les durées de cumul emploi-retraite des professions libérales. Le RSI assure la couverture maladie des artisans, des commerçants et des professions libérales, et également la retraite des artisans et des commerçants. En raison de la non couverture du risque vieillesse par le RSI des professions libérales, les informations sont moins bien renseignées pour cette population. Il manque notamment la date de fin d'activité au RSI des professions libérales nécessaire au calcul des durées de cumul.

Afin de ne pas fausser l'estimation de la durée du cumul emploi-retraite, les personnes dont la durée de cumul est inférieure à 1 mois ont été retirées (243 personnes). Il a été considéré que la situation de cumul résultait uniquement du délai pour mettre fin à leur activité relevant du RSI.

Ainsi, la population soumise au risque de sortir du cumul emploi-retraite comprend 15 017 personnes.

2.4. Les censures

Une des spécificités de l'analyse de durée est la présence d'individus, au sein de la population soumise au risque, pour lesquels l'évènement ne se produit pas pendant la période d'observation. Ces observations sont dites censurées. Le cas le plus fréquent correspond aux censures à droite : l'évènement ne s'est produit ni avant, ni pendant la période d'observation. Il y a plusieurs types de **censure à droite**:

Le premier type comprend des censures inévitables : elles sont dues à une durée d'observation inférieure à la durée effective. Dans notre étude, ce type de censure est très fréquent et représente 49 % de la population. Ces personnes ont débuté leur cumul en 2008 et au 31 décembre 2012, elles étaient toujours en cumul. Ce type de censure est tout de même informatif : on sait que la durée avant l'occurrence de l'évènement est supérieure à la durée d'observation.

Un autre type de censure correspond aux personnes « perdues » en raison de l'attrition de l'échantillon. Ce type de censure est pris en considération par les modèles de durée mais il introduit des biais, car la disparition des observations n'est pas aléatoire. On sait tout de même que la personne n'est pas sortie de l'état avant d'être perdue. Ce type de censure n'apparaît pas dans l'étude du cumul emploi-retraite.
Le dernier type de censure correspond à des retraits du champ des personnes observées. Si le fait de ne plus faire parti du champ de l'étude n'est pas lié à la survenu de l'évènement étudié, alors ces interruptions n'introduisent aucun biais. Dans notre cas, ce type de censure correspond aux 13 personnes décédées en cours de cumul.
Pour les censures à droite, même si la date de sortie du cumul est inconnue, nous connaissons une durée minimale du cumul : celle écoulée entre le début du cumul et la fin de son observation au 31 décembre 2012. Nous retenons donc cette durée dans la variable ***duree*** pour les observations censurées.

Il y a également **des censures à gauche :** des individus peuvent avoir connu l'évènement antérieurement, à une date inconnue. Les biais dans les résultats vont dépendre fortement de la proportion d'observations tronquées. SAS peut prendre en compte les censures à gauche partiellement, mais le plus souvent, la date d'origine t_0 est avancée pour éviter les censures à gauche. Dans nos données, il y a des individus en cumul emploi-retraite entre 2008 et 2012 qui ont débuté leur cumul avant 2008, à une date qui nous est inconnue. Ainsi, afin d'éviter les censures à gauche, le cumul emploi retraite est étudié à partir de 2008 et non auparavant.

Le dernier type de censure sont **les censures dans un intervalle** : l'individu a connu un évènement mais la date exacte de sa survenue est inconnue, seul un intervalle de temps est connu. Ce cas est fréquent dans la contraction de maladie. Les procédures SAS peuvent traiter ces cas de censures, mais souvent une date de survenue de l'évènement est supposée (ex : le milieu de l'intervalle).
Dans l'étude du cumul emploi-retraite, il n'y a pas de censures à gauche ou par intervalle.

En supposant que les individus censurés se seraient comportés vis-à-vis du processus étudié de la même façon que l'ensemble de la population, SAS peut produire des estimateurs de durée non biaisés. Néanmoins, dans notre étude, comme une observation sur deux est une censure, il n'est possible d'étudier le cumul emploi-retraite que sur les quatre premières années, voire cinq années selon que le cumul ait débuté en janvier ou en décembre 2008.

Au cours des premières années de cumul, les censures présentes sont principalement celles de personnes ayant interrompu leur cumul emploi-retraite, mais qui sont toujours en cumul au 31 décembre 2012. Il s'agit par exemple d'un auto-entrepreneur qui exerce son activité de mars 2008 à mars 2010, puis s'arrête. Il choisit ensuite de reprendre son activité au début de l'année 2012, et est toujours en cumul au 31 décembre 2012. 3 années de cumul sont donc observées, et sa durée totale de cumul est encore inconnue : il fait donc partie des censures observées lors de la troisième année. Ainsi, le simple calcul de la moyenne de durée de cumul, déjà observé, sous-estime la durée de cumul.
Dans notre étude, la variable indicatrice indiquant si l'observation est censurée s'appelle *censure*. **Elle vaut 0 si l'observation est censurée (par convention), et 1 sinon**.

3. ANALYSE NON PARAMETRIQUE

3.1. Présentation des modèles non paramétriques

La première étape d'une analyse de durée consiste en la réalisation d'un modèle non paramétrique. Il s'agit d'une analyse descriptive simple des durées. Il permet de donner une première série de réponses sur la durée du cumul emploi-retraite et la distribution des évènements, c'est-à-dire des sorties du cumul, en fonction du temps. De plus, il est aussi possible d'étudier la distribution des sorties de cumul pour des sous-populations. Il donne des informations sur la différenciation des comportements et une idée de l'effet relatif de certaines caractéristiques. L'analyse non paramétrique est aussi un préalable indispensable à une analyse plus poussée car elle permet de sélectionner les modèles paramétriques et semi-paramétriques les plus adaptés aux données.

Il existe deux méthodes d'analyse non paramétriques : la méthode de Kaplan-Meier (appelée aussi méthode du produit-limite) et la méthode actuarielle (appelée aussi méthode de la table de survie). Ces méthodes permettent d'obtenir la fonction de survie, ce qui correspond à la proportion de personnes encore en cumul emploi-retraite à une durée t.

Pour obtenir la fonction de survie, il faut d'abord obtenir la probabilité de connaître l'évènement, ce qui correspond dans notre cas à la probabilité de mettre fin au cumul en t. Elle est définie ainsi :

P(connaître l'évènement)= $\frac{nombre\ d'évènements\ en\ t}{population\ soumise\ au\ risque\ en\ t}$

Dans notre cas, P(mettre fin au cumul en t)= $\frac{nombre\ de\ sorties\ du\ cumul\ en\ t}{population\ soumise\ au\ risque\ de\ sortir\ du\ cumul\ en\ t}$

La fonction de survie est le produit des probabilités de ne pas connaître l'évènement, soit le produit des probabilités de ne pas mettre fin au cumul emploi-retraite. Ainsi, elle s'écrit :

$S(t)=\prod_t P(ne\ pas\ connaître\ l'évènement\ en\ t) = \prod_t \left[1 - \frac{nombre\ d'évènements\ en\ t}{population\ soumise\ au\ risque\ en\ t}\right]$

$S(t)= \prod_t P(ne\ pas\ mettre\ fin\ au\ cumul\ en\ t) = \prod_t \left[1 - \frac{nombre\ de\ sorties\ du\ cumul\ en\ t}{population\ soumise\ au\ risque\ de\ sortir\ du\ cumul\ en\ t}\right]$

Les méthodes de Kaplan-Meier et actuarielle se différencient sur deux points :
La méthode de Kaplan-Meier calcule la fonction de survie pour des durées exactes, alors que la méthode actuarielle calcule la fonction de survie pour des durées regroupées sur des intervalles de temps.
La méthode de Kaplan-Meier considère que les observations censurées sont exposées au risque jusqu'à la durée t, c'est-à-dire que la censure arrive très rapidement après la durée t. Cela revient à dire que la probabilité de connaître l'évènement étudié est la même pour les individus censurés, et ceux non censurés. Ainsi, les individus censurés font partie de la population soumise au risque. Dans la méthode actuarielle, il est supposé que la censure survient de manière uniforme sur un intervalle. Les individus censurés sont exposés au risque uniquement pendant la moitié de l'intervalle de temps.

Pour comprendre plus facilement les différences entre les deux méthodes, nous calculons la fonction de survie liée à un évènement quelconque dans une population de 7 personnes.
Individu A : l'évènement s'est produit au bout de 4 mois et demi.
Individu B : l'évènement ne s'est jamais produit. L'individu a été suivi pendant 5 mois.
Individu C : l'évènement ne s'est jamais produit. L'individu a été suivi pendant 3 mois.
Individu D : l'évènement s'est produit au bout de 5 mois.
Individu E : l'évènement ne s'est jamais produit. L'individu a été suivi pendant 3 mois.
Individu F : l'évènement s'est produit au bout de 3 mois.
Individu G : l'évènement s'est produit au bout de 2 mois.

Répertorions les évènements en fonction de la durée :

Durée en mois	Nombre d'évènements	Nombre de censures
0	0	0
1	0	0
2	1 (individu G)	0
3	1 (individu F)	2 (individus C et E)
4,5	1 (individu A)	0
5	1 (individu D)	1 (individu B)

Calculons l'estimateur de Kaplan-Meier :

Durée en mois	Nombre d'évènements	Nombre de censures	Nombre de personnes à risque	Fonction de survie S(t) de Kaplan-Meier
0	0	0	7	1
1	0	0	7	1
2	1	0	7	0.86=1-1/7
3	1	2*	6=7-1	0.71=0.86*(1-1/6)
4,5	1	0	3=6-1-2*	0.48=0.71*(1-1/3)
5	1	1	2=3-1	0.24=0.48*(1-1/2)

**les deux individus censurés lors de la durée 3 sont soumis au risque de connaître l'évènement à la date 3, car on suppose qu'ils sont censurés juste après la date 3.*

Calculons l'estimateur de la méthode actuarielle: les données sont regroupées sur des intervalles de 2 mois.

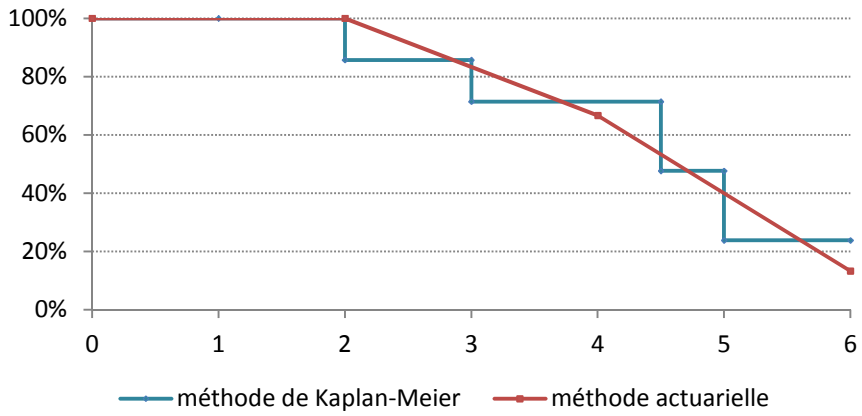
Durée en mois	Nombre d'évènements	Nombre de censures	Nombre de personnes à risque	Fonction de survie S(t) avec la méthode actuarielle
[0 ;2[	0	0	7	1
[2 ;4[	2 (indiv G et F)	2 (indiv C et E)*	6=7-0.5*2	1
[4 ;6[	2 (indiv A et D)	1 (indiv B)	2.5=7-2-2-0.5*1	0.666=1-2/6*1
[6 ;.[				0.13=(1-2/2.5)*0.666

**Les deux individus censurés sont soumis au risque de connaître l'évènement la moitié de l'intervalle.*

Ainsi, à la fin du 5ième mois, la proportion de personnes qui ont connu l'évènement s'élève à 24 % d'après la méthode de Kaplan-Meier et à 13 % d'après la méthode actuarielle (graphique 3).

Comme le calcul de la fonction de survie se fait à des dates exactes, la méthode de Kaplan-Meier est adaptée aux petits échantillons pour lesquelles les dates sont précisément mesurées. La méthode actuarielle s'utilise davantage sur des échantillons ayant un grand nombre de durées différentes ou lorsque la mesure des durées n'est pas précise.
Dans l'étude du cumul emploi-retraite, la méthode de Kaplan-Meier a été mise en œuvre dans un premier temps car les durées étaient connues précisément, et cela permettait de garder le plus d'information. Puis, la méthode actuarielle a été appliquée car, grâce au regroupement des durées par intervalle, l'information est plus compréhensible. Néanmoins, plus l'intervalle des durées est grand, plus la perte de précision de la fonction est importante. Inversement, un intervalle de durée très petit fait disparaître la différence entre les estimateurs obtenus par la méthode de Kaplan-Meier et actuarielle.

Graphique 3 : Courbes de survie obtenues avec les méthodes actuarielle et de Kaplan-Meier



Note : Dans la méthode de Kaplan-Meier, la fonction de survie ne change de valeur qu'aux temps correspondant à des évènements observés, d'où une courbe aux allures de marches d'escalier.

3.2. Mise en œuvre d'un modèle non paramétrique sous SAS

Reprenons l'analyse de la durée du cumul emploi-retraite. Nous mettons en œuvre une méthode non paramétrique pour obtenir la distribution des sorties de cumul en fonction de la durée passée en cumul.
Voici la procédure SAS des modèles non paramétriques :

Proc lifetest data=xxxx method=xxxx;
*Time variable_duree*variable_censure(valeur lorsque l'observation est censurée) ;*
Run ;

Ainsi, dans le cas de notre étude :

```
Proc lifetest data=cohorte2008 outsurv=actu method=act width=0.5
conftype=loglog
plots=(survival(atrisk cb=hw), hazard, logsurv ) graphics;
Time duree*censure(0) ;
Run ;
```

Note : la variable `duree`=(date de fin du cumul (date de cessation d'activité au RSI)-date de début de cumul(1er report postérieur à la date d'effet de la pension RG)).
Si l'observation est censurée : `duree`=date de fin d'observation (31/12/2012) - date de début de cumul (1er report postérieur à la date d'effet de la pension RG).
La variable *duree* est une variable continue, exprimée en années (exemple : 1,7 année=620 jours).

`censure(0)` : indique le nom de la variable de censure, suivie entre-parenthèses de la valeur de cette variable lorsque l'observation est censurée, 0 dans notre cas.

Options du proc lifetest

L'option `outsurv` permet de sauvegarder les résultats dans une table, ici nommée actu.

L'option `method=` permet d'indiquer la méthode à utiliser. `method=km` ou `pl` pour la méthode de Kaplan-Meier (appelée aussi produit-limite). `Method=act` ou `life` pour la méthode actuarielle. Si cette option n'est pas précisée, SAS utilise par défaut la méthode de Kaplan-Meier.

Lorsque la méthode choisie est actuarielle, il faut préciser les intervalles des durées.

- L'option `width`=x permet de faire des intervalles de taille x. Dans notre cas d'étude, les intervalles sont donc des semestres (0.5 année=semestre).
- L'option intervals permet de créer des intervalles d'amplitudes différentes. Par exemple, `intervals=1 2.5 4` crée les intervalles suivants : [0, 1[, [1, 2.5 [, [2.5, 4 [et [4 ;+inf[:
- Si les options width ou intervals ne sont pas précisées, SAS effectue automatiquement un découpage en 10 classes de son choix, avec des entiers comme bornes d'intervalles.

Il est intéressant d'obtenir l'intervalle de confiance de la fonction de survie. SAS propose de calculer cet intervalle de confiance avec la méthode des bandes de Hall-Wellner. Néanmoins, l'intervalle de confiance obtenu n'est pas forcément compris entre 0 et 1, contrairement à la fonction de survie. Ainsi, une transformation est appliquée ; plusieurs types de transformation sont possibles : log, log-log, arc-sinus de la racine carrée ou logistique. Par défaut SAS utilise la transformation log-log. L'option `conftype=` permet de préciser la transformation à appliquer.

L'instruction `plots=` permet d'afficher les graphiques :

- De la fonction de survie « `survival` » ou « `s` ». (graphiques 4 et 5)
- De la fonction de risque instantané « `hazard` ». Le tracé de cette fonction n'est possible qu'avec la méthode actuarielle (graphique 6)
- De la fonction de risques cumulés « `logsurv` » H(t): pseudo probabilité de connaître l'évènement si un individu était constamment soumis au risque. (graphique 7, negative log of estimated survivor function=-ln(S(t))=H(t)))

Pour compléter le graphique de la fonction de survie, il est demandé :

- De tracer l'intervalle de confiance Hall-Wellner de la fonction de survie grâce à `cb=hw`.
- D'indiquer sur le graphique l'effectif de la population à risque grâce à l'option `atrisk`. En effet, avec la survenue des évènements et des censures, la population à risque diminue avec le temps. Il est préférable de s'assurer que la fonction de survie est estimée à partir d'un nombre suffisant d'individus.

Sorties SAS Kaplan-Meier

Tableau 3 : L'estimateur de la fonction de survie obtenu avec la méthode de Kaplan-Meier

Le Système SAS

Procédure LIFETEST

Product-Limit Survival Estimates

duree		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000		1.0000	0	0	0	15017
0.08219		.	.	.	1	15016
0.08219		.	.	.	2	15015
3.04932		0.5937	0.4063	0.00402	6073	8739
3.04932	*	.	.	.	6073	8738
3.05205		0.5936	0.4064	0.00402	6074	8737
3.05479		0.5935	0.4065	0.00402	6075	8736
3.05479	*	.	.	.	6075	8735
4.99726	*	.	.	.	7626	1
4.99726	*	.	.	.	7626	0

Tableau 4 : Les quantiles de la fonction de survie de Kaplan-Meier

Statistiques descriptives pour variable temps duree

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	4.28219	LOGLOG	4.10137	4.41644
25	1.66575	LOGLOG	1.58630	1.74521

Mean	Standard Error
3.39498	0.01458

Tableau 5 : Récapitulatif du nombre d'évènements observés, d'observations censurées et leur proportion

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
15017	7626	7391	49.22

Le tableau 3 donne la fonction de survie, c'est-à-dire la proportion de personnes en cumul emploi-retraite pour chaque durée. Il y a donc une ligne par individu.

Colonne 1 : durées auxquelles un évènement, une sortie de cumul ou une censure, a lieu.

Colonne 2 : valeurs de la fonction de survie.

Colonne 3 : inverse de la fonction de survie, soit la proportion d'individus ayant connu l'évènement, c'est-à-dire qui sont sortis du cumul.

Colonne 4 : écart type de la fonction de survie S(t).

Colonne 5 : effectif ayant connu l'évènement jusqu'en t, c'est-à-dire que 6073 personnes ont arrêté le cumul avant 3 ans et 2 semaines (3,049 ans).

Colonne 6 : population encore soumise au risque après t, c'est-à-dire personnes qui sont encore en cumul emploi-retraite. Après 3 ans et 2 semaines, 8739 personnes sont encore dans le cumul, et donc soumises au risque d'y mettre fin.

***** Indique que l'observation est censurée.

Le **tableau 4** donne les quantiles de S(t) avec leurs intervalles de confiance et la durée moyenne de cumul emploi-retraite.
Comme l'observation du cumul s'arrête au 31 décembre 2012, il n'est pas possible de connaître la durée des cumuls supérieurs à 5 ans. Ainsi, il n'est pas possible de calculer le dernier quantile de la fonction de survie. De même, la durée moyenne calculée n'est pas fiable. D'ailleurs, SAS indique qu'elle est sous-estimée[9].

Le **tableau 5** récapitule le nombre d'évènements et de censures, et la proportion d'observations censurées.

[9] « The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time ».

Sorties SAS de la méthode actuarielle :

Tableau 6 : Les estimateurs du modèle de durée actuariel : fonctions de survie et de risque

Procédure LIFETEST

Life Table Survival Estimates															
Interval		3	4	5	6	7	8	9	10	11	12	14 Evaluated at the Midpoint of the Interval		16	
1 [Lower,	2 Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	13 PDF	PDF Standard Error	15 Hazard	Hazard Standard Error
0	0.5	1057	4	15015.0	0.0704	0.00209	1.0000	0	0	4.2868	0.0827	0.1408	0.00418	0.145929	0.004486
0.5	1	1354	10	13951.0	0.0971	0.00251	0.9296	0.0704	0.00209	.	.	0.1804	0.00468	0.204008	0.005537
1	1.5	1042	37	12573.5	0.0829	0.00246	0.8394	0.1606	0.00300	.	.	0.1391	0.00416	0.17291	0.005352
1.5	2	1033	28	11499.0	0.0898	0.00267	0.7698	0.2302	0.00344	.	.	0.1383	0.00415	0.188117	0.005847
2	2.5	785	45	10429.5	0.0753	0.00258	0.7007	0.2993	0.00374	.	.	0.1055	0.00366	0.156421	0.005579
2.5	3	775	76	9584.0	0.0809	0.00278	0.6479	0.3521	0.00390	.	.	0.1048	0.00366	0.168542	0.006049
3	3.5	605	98	8722.0	0.0694	0.00272	0.5955	0.4045	0.00402	.	.	0.0826	0.00329	0.143714	0.005839
3.5	4	579	119	8008.5	0.0723	0.00289	0.5542	0.4458	0.00407	.	.	0.0801	0.00326	0.150019	0.00623
4	4.5	286	2822	5959.0	0.0480	0.00277	0.5142	0.4858	0.00411	.	.	0.0494	0.00287	0.098349	0.005814
4.5	5	110	4152	2186.0	0.0503	0.00468	0.4895	0.5105	0.00416	.	.	0.0493	0.00460	0.103238	0.00984
5	.	0	0	0.0	0	0	0.4648	0.5352	0.00457	.	.	.	.	.	.

Le tableau ci-dessus représente la fonction de survie, soit la proportion de personnes en cumul emploi-retraite par semestre.

Colonne 1 et 2 : début et fin des intervalles

Colonne 3 : effectif de personnes connaissant l'évènement (ici, mettre fin au cumul) entre [ti ;ti+a[

Colonne 4 : nombre de personnes censurées sur [ti ; ti+a[

Colonne 5 : nombre de personnes exposées au risque sur [ti ; ti+a[en faisant l'hypothèse de répartition uniforme des censures

Colonne 6 : colonne 3 / colonne 5=probabilité conditionnelle de connaître l'évènement « mettre fin au cumul » et son écart-type en colonne 7

Colonne 8 : fonction de survie S(t), soit la proportion de personnes en cumul en t, ou la probabilité de rester en cumul jusqu'en t (s'obtient à partir de la valeur de la date précédente par multiplication par (1-colonne 6), ex : 0,9296=1*(1-0,0704) et 0,8394=0,9296*(1-0,0971))

Colonne 9 : proportion de personnes ayant connu l'évènement « sortie du cumul »

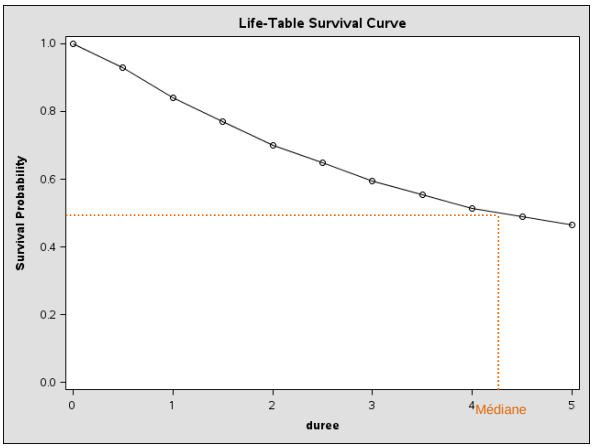
Colonne 10 : écart type de la fonction de survie S(t)

Colonne 11 : durée médiane résiduelle, soit la durée écoulée entre ti et l'instant où S(t)=S(ti)/2 et son écart-type en colonne 12.

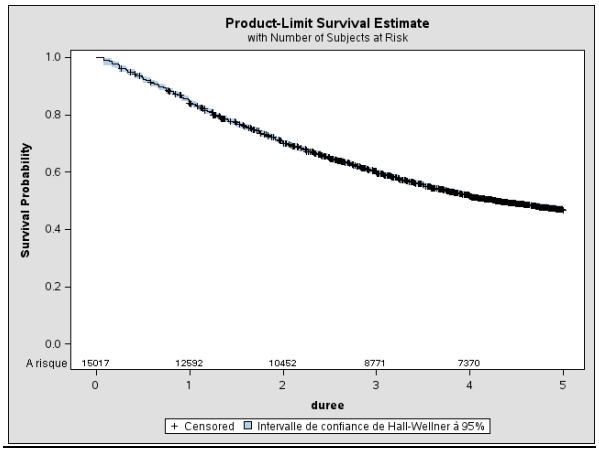
Colonne 13 : densité de probabilité f(t) et son écart-type en colonne 14

Colonne 15 : risque instantané h(t) et son écart-type

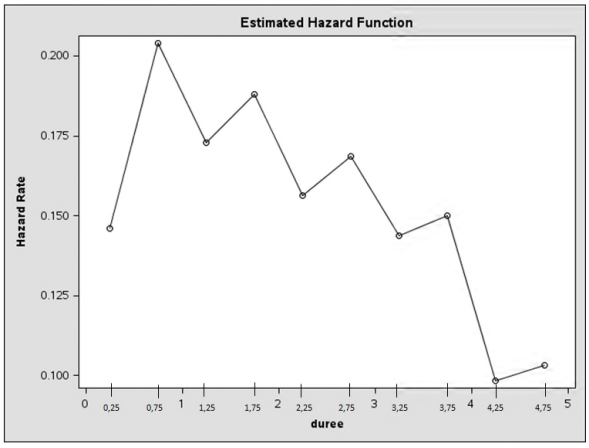
Graphique 4 : Fonction de survie avec la méthode actuarielle (intervalle de temps=semestre)



Graphique 5 : Fonction de survie avec la méthode de Kaplan-Meier

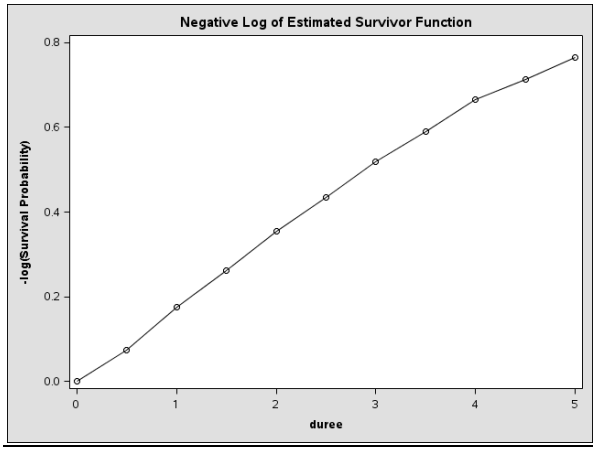


Graphique 6 : Fonction de risque instantané avec la méthode actuarielle (intervalle de temps=semestre)



. Les points sont estimés au milieu de l'intervalle. Ils peuvent être reliés entre-eux, parce que l'on suppose que les censures et les évènements se répartissent uniformément sur l'intervalle.

Graphique 7 : Fonction de risques cumulés avec la méthode actuarielle (intervalle de temps = semestre)



L'étude des résultats du modèle non paramétrique montre que le cumul d'un emploi d'indépendant et d'une retraite du régime général s'effectue sur plusieurs années. Plus de la moitié de la population est en cumul emploi-retraite pendant au moins 4 ans (voir graphique 4, ou tableau 4). Après 5 ans de cumul emploi-retraite, date de fin de la période d'observation, 47% de la cohorte est toujours en situation de cumul, pour une durée qui nous est inconnue (tableau 6, encadré vert). En parallèle, pour près d'un quart de la population le cumul emploi-retraite n'est effectué que quelques mois : 23% de la population arrête le cumul emploi-retraite avant un an et demi d'exercice (tableau 6, encadré rouge).

Il semble y avoir deux utilisations du cumul emploi-retraite : une où le dispositif serait utilisé pendant une période relativement courte, et une seconde où son usage s'étalerait davantage dans le temps. L'étude des fonctions de risque conforte cette idée. Le risque de sortir du cumul emploi-retraite est assez élevé au cours de la première année, et plus particulièrement entre le sixième et le douzième mois (graphique 6). Après 6 mois de cumul, on a presque une chance sur 10 de sortir du dispositif au cours du deuxième semestre de la

première année (tableau 6, encadré bleu). Au delà d'un an, le risque de sortir du cumul emploi-retraite décroit progressivement. Ainsi, plus la durée passée en cumul emploi-retraite augmente, et plus le risque d'y mettre fin diminue. Après 4 ans dans le dispositif, on a seulement 1 chance sur 20 de quitter le cumul emploi-retraite au cours des 6 mois suivants (tableau 6, encadré orange).

Encadré 3 : Interprétation des fonctions de risque instantané et de risque cumulé

La fonction de risque instantané permet de distinguer l'évolution générale du risque en fonction du temps. La valeur de ce risque instantané n'est pas interprétable telle quelle.

Par exemple : le risque instantané estimé entre 6 mois et un an de cumul (tableau 6 encadré vert colonne 15) est de 0,20 par unité de temps (dans notre étude l'année). Il est obtenu ainsi :

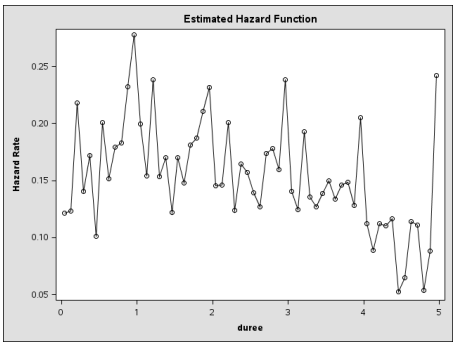
$$h(t) = \frac{f(t)}{S(t)} \text{ donc } h[0{,}5;1] = \frac{f[0{,}5;1]}{\frac{S(0{,}5)+S(1)}{2}} = \frac{2*f[0{,}5;1]}{S(0{,}5)+S(1)} = \frac{2*0{,}180443}{0{,}929604+0{,}839382} = 0{,}204008$$

Pour retrouver les valeurs de $f[0{,}5;1]$ *et S(t), voir l'encadré jaune en colonne 13 et l'encadré violet colonne 8 du tableau 6*

$h[0{,}5;1] = 0{,}20$ signifie que la sortie du cumul emploi-retraite se produira 0,2 fois dans l'année en moyenne. Le risque n'est pas toujours compris entre 0 et 1 : il est souvent supérieur à 1.

Comme l'évènement est non renouvelable, le risque peut sembler une notion absurde lorsqu'il est supérieur à 1. Le risque peut aussi être exprimé en prenant son inverse qui représente la durée moyenne avant de sortir du cumul. Si le risque instantané était identique à celui observé au cours du dernier semestre de la première année, alors la durée moyenne du cumul emploi-retraite serait de 5 ans (1/0,20). En revanche, si le risque instantané était identique à celui observé entre deux ans et demi et trois ans de cumul, la durée moyenne du cumul serait de 5 ans et 10 mois (1/0,17=5,88).

Graphique 8 : Fonction de risques cumulés avec la méthode actuarielle (intervalle de temps =mois)



Par ailleurs, les risques instantanés peuvent être très variables au cours du temps (voir la courbe de risques instantanés lorsque les durées sont regroupées par mois). Il est alors préférable d'interpréter la pente de la courbe de risques cumulés H(t) qui donne les mêmes enseignements :

-Si la courbe est linéaire, le risque est constant au cours du temps.

-Si elle est plutôt convexe, le risque diminue au cours du temps.

-Si elle est plutôt concave, le risque augmente au cours du temps.

Dans notre étude, comme la pente de la courbe de risques cumulés ne varie pas fortement au cours du temps, il n'est pas facile de l'étudier, et c'est pourquoi elle est peu utilisée dans le reste de l'étude (graphique 7).

3.3. Analyse non paramétrique par sous-population

Les modèles non paramétriques supposent que la population est homogène. Le temps de survenue de l'évènement étudié est identique pour tous les individus : ils possèdent donc la même fonction de survie, la même fonction de risque instantané, la même fonction de risque cumulée etc. Le non respect de cette hypothèse d'homogénéité des populations peut conduire à des erreurs d'interprétation des résultats.

Par exemple, prenons deux populations A et B ayant un risque de sortir du cumul emploi retraite constant dans le temps, mais différent. La fonction de risque de sortir du cumul emploi-retraite pour la population totale (A+B) est un mélange des fonctions de risque afférentes à chacune des populations.

L'estimateur de cette fonction de risque est obtenu en considérant les individus encore en cumul emploi-retraite à chaque date. Or, plus la durée du cumul augmente, plus la proportion des individus ayant un faible risque de sortir du cumul s'accroit et simultanément la proportion des individus à risque élevé diminue. Ainsi, dans la population totale, le risque estimé de sortir du cumul emploi-retraite décroit. Pour autant, il faut se garder d'appliquer cette conclusion à chaque sous-population, puisque l'on sait, par construction, que leur risque est constant.

En pratique, si l'homogénéité des individus n'est pas respectée, elle conduit à estimer des fonctions de survie, ou de risque difficilement interprétable. Il faut donc essayer de se mettre dans des conditions où elle est plutôt validée. La solution proposée est de calculer les estimateurs de durée sur des sous-populations plus homogènes.

Dans l'étude sur le cumul emploi-retraite, nous stratifions la population en fonction de caractéristiques qui semblent avoir un effet sur le cumul emploi-retraite comme le secteur d'activité, le groupe professionnel, le statut d'auto-entrepreneur et la situation au moment de la liquidation. Au niveau de la mise en œuvre sous SAS, il suffit de rajouter l'instruction `strata`.

```
Proc lifetest data=cohorte2008 method=act conftype=loglog width=0.5
plots=(survival(ch=hw), hazard, logsurv, loglogs) graphics;
Time duree*censure(0) ;
strata activite_der/ test=all;
Run ;
Note : activite_der=variable correspondant au groupe professionnel (artisans,
commerçants).
```

Les tests de comparaison de courbes de survie

Les tests de comparaison permettent de déterminer si une même distribution gouverne les évènements observés dans les différentes strates. Ils analysent si la distance entre deux courbes de survie est plus grande que ce que pourrait expliquer le hasard. Ils confrontent deux hypothèses :

H_0 : les fonctions de survie sont identiques

H_1 : les fonctions de survie sont différentes

Les tests de comparaison sont des tests de khi2 avec autant de liberté qu'il y a de sous populations à comparer moins une. Les tests les plus connus sont le test du log-rank et le test de Wilcoxon (encadré 4).

Encadré 4 : les tests statistiques de comparaison

Le principe des tests de comparaison est le suivant : à chaque durée pour laquelle se produit un évènement, on compare le nombre d'évènements observés au nombre d'évènements attendus sous l'hypothèse d'indépendance entre les groupes.

$$X^2 = \frac{\left(\sum_{j=1}^{r} w_j(d_{1j} - e_{1j})\right)^2}{var\left(\sum_{j=1}^{r} w_j(d_{1j} - e_{1j})\right)}$$

Où d_{1j} est le nombre d'évènements observés dans la strate1 au temps j,

e_{1j} est le nombre d'évènements attendus dans la strate1 au temps j, s'il n'y avait aucune relation entre les groupes. Ces effectifs correspondent aux effectifs théoriques calculés dans un test de khi2.

w_j est une variable de pondération, qui n'est pas toujours utilisée :

Le test du Log-rank est un test où $w_j = 1$.

Le test de Wilcoxon est un test où $w_j = n_j =$ population à risque à chaque temps .

Le test du logrank est fondé sur une statistique qui donne des poids égaux à toutes les observations. Le test de Wilcoxon donne plus de poids aux évènements qui surviennent en début de période d'observation qu'à ceux qui arrivent plus tardivement. En effet, la population soumise au risque n_j décroit au fur et à mesure du temps. Le test de Wilcoxon est donc plus sensible aux différences entre les groupes qui surviennent en début de période. Il est donc plus intéressant lorsque l'étude porte principalement sur les débuts d'observation.

Le nombre d'évènements observés d_j et le nombre d'évènements attendus e_j est calculé indépendamment des censures. Le test du logrank suppose donc que le mécanisme des censures est indépendant du phénomène étudié. En revanche, dans le calcul du test de Wilcoxon, on prend également en compte la population soumise au risque n_j, qui est dépendante des censures. Le test de Wilcoxon suppose donc que les censures dépendent du phénomène étudié.

D'autres tests non paramétriques existent sous SAS, ils diffèrent des 2 tests précédents par leur pondération.

Test	Weight Function
Log-rank	1.0
Wilcoxon	n_i
Tarone-Ware	$\sqrt{n_i}$
Peto-Peto	$\tilde{S}(t_i)$
Modified Peto-Peto	$\tilde{S}(t_i) * \frac{n_i}{n_i + 1}$
Harrington-Fleming	$\hat{S}(t_{i-1})^p [1 - \hat{S}(t_{i-1})]^q$

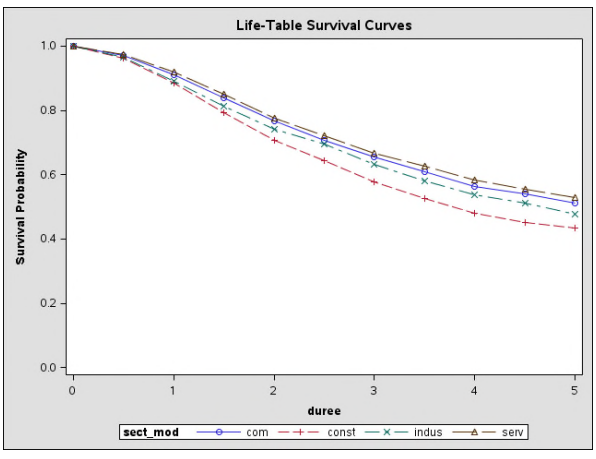
Il est également possible d'effectuer un test paramétrique : le **test du ratio de vraisemblance** -2Log(LR): qui suppose que la distribution des évènements en fonction du temps suit une loi exponentielle.

Les individus qui sont en cumul emploi-retraite peuvent exercer leur activité dans quatre secteurs d'activité : le commerce, la construction, les services et l'industrie.
Les tests de comparaison statistiques montrent que nous ne pouvons pas rejeter l'hypothèse nulle, qui suppose que les distributions des évènements sont identiques, lors de la comparaison des fonctions de survie des cumulants du secteur du commerce avec ceux de l'industrie ou des services, sinon nous aurions 38% et 10% de chances de nous tromper (tableau 7).
Entre les autres secteurs d'activité, les distributions des sorties de cumul en fonction de la durée sont significativement différentes.

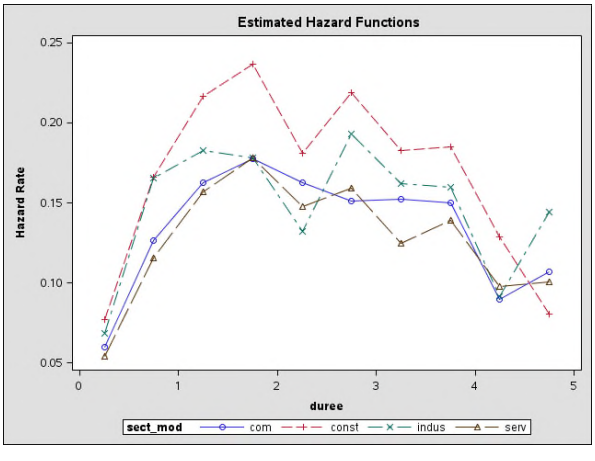
Tableau 7 : Résultats du test de comparaison du log-rank pour la variable secteur d'activité (significativité au seuil de 5%)

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
sect_mod	sect_mod	Chi-Square	Raw	Scheffe
com	const	25.6536	<.0001	<.0001
com	indus	3.0794	0.0793	0.3795
com	serv	6.3335	0.0118	0.0965
const	indus	21.9913	<.0001	<.0001
const	serv	59.4703	<.0001	<.0001
indus	serv	24.8169	<.0001	<.0001

Les personnes travaillant dans le domaine de la construction exercent le cumul emploi-retraite moins longtemps que les cumulants des secteurs de l'industrie et des services (graphiques 9 et 10). Après un an et demi de cumul emploi-retraite, 15% des cumulants du secteur des services quittent le dispositif, contre 21% de ceux du secteur de la construction. Le risque de mettre fin au cumul emploi-retraite est toujours supérieur pour les indépendants de la construction que pour les autres secteurs d'activité. Les secteurs d'activité sont à mettre en relation avec les groupes professionnels puisque les indépendants du secteur de la construction sont très souvent artisans.

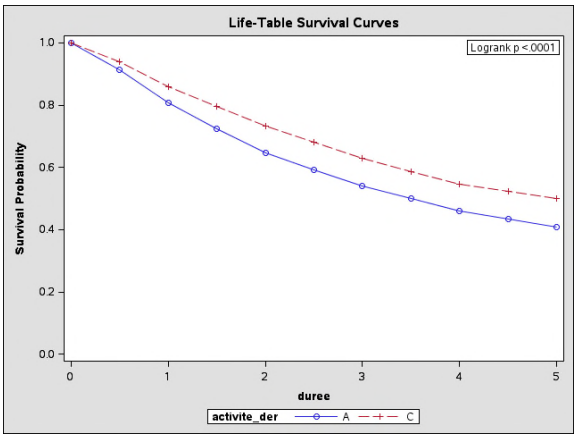


Graphique 9 : Fonction de survie avec la méthode actuarielle stratifiée par secteur d'activité (intervalle de temps= semestre)

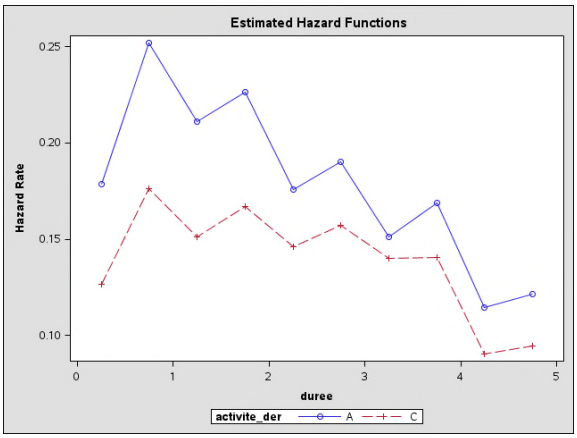


Graphique 10 : Fonction de risque instantané avec la méthode actuarielle stratifiée par secteur d'activité (intervalle de temps=semestre, hasard situé au milieu de l'intervalle)

De ce fait, les artisans restent moins longtemps en cumul emploi-retraite que les commerçants. La moitié des artisans ne sont plus en cumul emploi-retraite à 3 ans et demi, alors que pour les commerçants, c'est un an et demi plus tard, à cinq ans, que la moitié de la population a quitté le dispositif. Pour les artisans, et de façon similaire à ce qu'on avait constaté au graphique 5, le risque de sortir du cumul emploi retraite augmente très fortement au cours de la première année pour ensuite décroître au cours du temps. A contrario, pour les commerçants, le risque de sortir du cumul est presque constant sur l'ensemble de la période d'observation (graphique 12).



Graphique 11 : Fonction de survie avec la méthode actuarielle stratifiée par groupe professionnel (intervalle de temps= semestre)



Graphique 12 : Fonction de risque instantané avec la méthode actuarielle stratifiée par groupe professionnel (intervalle de temps= semestre hasard situé au milieu de l'intervalle)

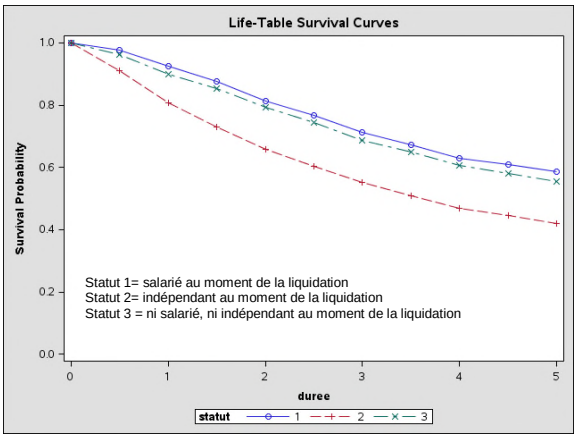
Le cumul d'un emploi d'indépendant et d'une retraite du régime général peut être le résultat de plusieurs situations :

- un indépendant, qui tout en poursuivant la même activité, décide de liquider la retraite correspondant à sa carrière de salarié du privé,

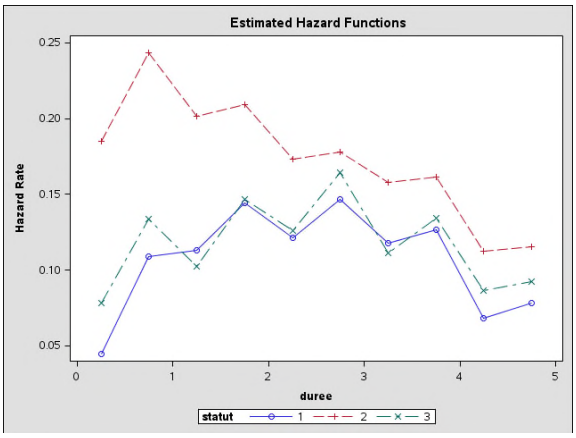
- ou bien un individu, qui suite à son départ en retraite du régime général, débute une activité d'indépendant qu'il soit salarié ou non à la date de liquidation.

Les personnes déjà indépendantes au moment de la liquidation de leur retraite restent moins longtemps en cumul emploi-retraite que celles qui débutent une nouvelle activité d'indépendant (graphique 13) : 20 % des personnes qui exerçaient un emploi d'indépendant avant la liquidation de leur retraite, conservent leur activité au maximum un an. A contrario, seulement 8 % de ceux qui étaient salariés, ont cessé leur nouvelle activité d'indépendant au cours de la première année.

D'ailleurs, les risques de sortir du cumul emploi-retraite mettent en exergue deux comportements différents. Les personnes déjà indépendantes au moment de la liquidation ont de grandes chances de sortir du cumul emploi-retraite dans les premiers mois et ces chances diminuent progressivement. Les personnes qui choisissent une nouvelle activité d'indépendant au moment de leur retraite du régime général ont très peu de chances de mettre fin à cette nouvelle expérience professionnelle au cours des premiers mois. A partir de la fin de la première année de cumul, le risque de sortir du cumul emploi-retraite s'élève progressivement et reste constant sur l'ensemble de la période (graphique 14).



Graphique 13 : Fonction de survie avec la méthode actuarielle stratifiée par situation au moment de la liquidation (intervalle de temps= semestre)



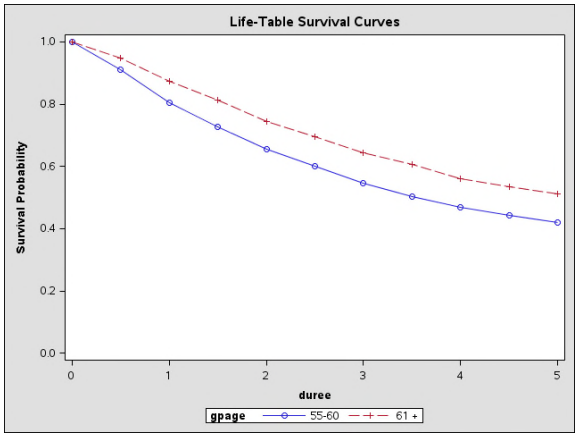
Graphique 14 : Fonction de risque instantané avec la méthode actuarielle stratifiée par situation au moment de la liquidation (intervalle de temps= semestre, hasard situé au milieu de l'intervalle)

Un autre élément qui peut être étudié est l'âge auquel un individu a débuté son cumul emploi-retraite. Un cinquième de la population a commencé son cumul entre 55 et 59 ans, 30% à 60 ans, 20% à 61 ou 62 ans, et 30% entre 63 ans et 69 ans. Les tests de comparaison statistiques montrent que les fonctions de survie ne sont différentes que pour les personnes ayant liquidé entre 55 et 60 ans inclus, et ceux ayant liquidé après 60 ans. Pour simplifier la lecture des graphiques, seulement les fonctions de survie de ces deux groupes sont représentées.

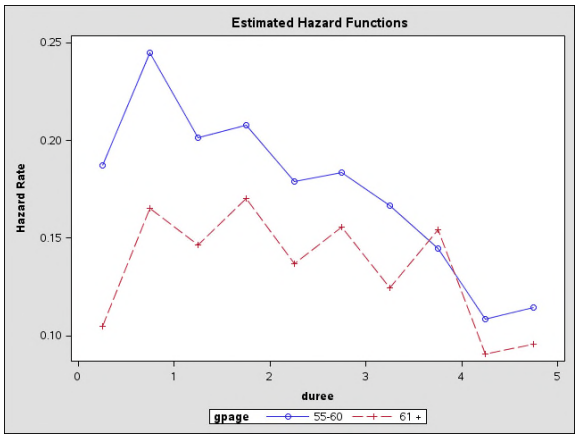
La durée passée en cumul emploi-retraite est plus courte pour les individus ayant commencé un cumul emploi-retraite entre 55 et 60 ans, dont 20 % arrêtent au cours de la première année. Elle est plus longue pour les personnes ayant débuté un cumul après 61 ans, où la moitié effectue au moins 5 ans dans le dispositif (graphique 15).

Ce résultat est un peu surprenant car on pourrait s'attendre à ce que les plus jeunes aient les cumuls emploi-retraite les plus longs. Il s'explique par une différence dans les activités exercées à la retraite. Un quart des entrants en cumul à 55-59 ans travaille dans la construction, contre seulement 10% des entrants entre 61 et 65 ans.

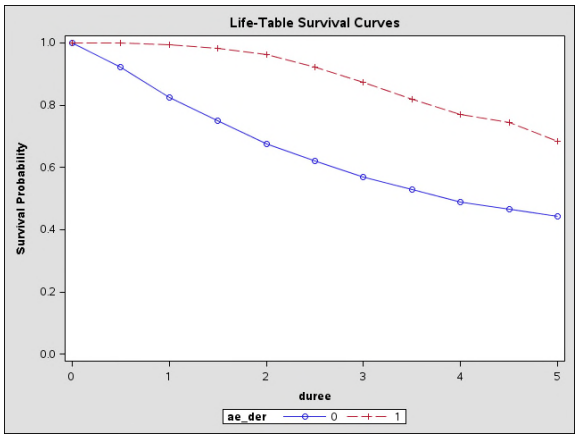
Les courbes de risques de sortir du cumul emploi-retraite laissent penser qu'il y a deux types de cumul chez les plus jeunes. Le risque de mettre fin au cumul emploi-retraite est très élevé au cours des premiers mois du cumul, mais dès la deuxième année, voire de la première, il s'effondre. Une partie des jeunes cumulants utilise le dispositif pendant un temps très court, et les autres peuvent alors rester en cumul pendant au contraire un temps long. Les personnes ayant débuté un cumul à partir de 61 ans ont un risque de quitter le cumul emploi retraite plus variable au cours du temps (graphique 16).



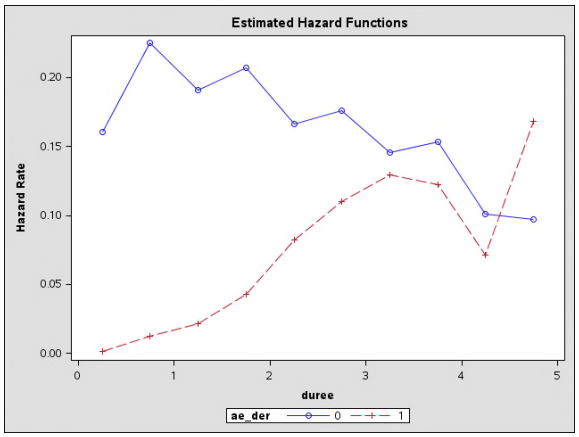
Graphique 15 : Fonction de survie avec la méthode actuarielle stratifiée par âge au début du cumul (intervalle de temps= semestre)



Graphique 16 : Fonction de risque instantané avec la méthode actuarielle stratifiée par âge au début du cumul (intervalle de temps= semestre, hasard situé au milieu de l'intervalle)



Graphique 17 : Fonction de survie avec la méthode actuarielle stratifiée par statut d'auto-entrepreneur (intervalle de temps= semestre)



Graphique 18 : Fonction de risque instantané avec la méthode actuarielle stratifiée par statut d'auto-entrepreneur (intervalle de temps= semestre, hasard situé au milieu de l'intervalle

Le statut d'auto-entrepreneur s'applique depuis le 1er janvier 2009. Ce dispositif permet à toute personne de créer une entreprise artisanale, commerciale ou libérale individuelle sous le régime fiscale de la micro-entreprise. Comme notre étude porte sur les personnes entrées en cumul emploi-retraite en 2008, elles n'ont pas pu utiliser ce dispositif dès le début du cumul. 9% d'entre elles ont cependant pris le statut d'auto-entrepreneur au cours de leur cumul. Le comportement en emploi-retraite par ces dernières est différent du reste de la population. Au cours des deux premières années de cumul, moins de 2% des auto-entrepreneurs mettent fin au cumul, alors que dans le reste de la population il s'agit de la période où le risque de sortie du cumul est le plus élevé. Le risque de sortir du cumul emploi-retraite des auto-entrepreneurs augmente progressivement jusqu'à atteindre son niveau le plus élevé pendant la troisième année. Il serait intéressant d'étudier davantage le cumul de cette population. Ces résultats sont cependant à considérer avec précaution. En effet, cette population est très peu nombreuse au sein de cette cohorte et aussi particulière car elle correspond à un changement de situation au cours du cumul emploi-retraite.

4. ANALYSE SEMI-PARAMETRIQUE : MODELE DE COX A RISQUES PROPORTIONNELS

4.1. Présentation d'un modèle de Cox à risques proportionnels

L'étude des durées de cumul par une analyse non paramétrique montre qu'il y a plusieurs utilisations possibles du cumul emploi-retraite en fonction des caractéristiques des individus. Afin d'aller plus loin dans l'analyse, un modèle semi-paramétrique de Cox est mis en œuvre. Le modèle de Cox, comme les modèles de régression standards, permet de trouver des facteurs explicatifs, et d'en mesurer leurs effets. Par ailleurs, l'objectif de l'étude est d'expliquer la durée passée en cumul emploi-retraite. Il est donc nécessaire de prendre en compte une dimension temporelle, ce qui est la première fonction du modèle de Cox.

Dans un modèle de Cox, la fonction de risque instantané est modélisée ainsi (encadré 5):
h(t)= *fonction inconnue qui correspond au risque instantané de connaître l'évènement pour l'individu de référence à la durée t (notée $h_0(t)$)* × *fonction de régression qui dépend des caractéristiques de l'individu.*

Le modèle de Cox repose sur deux règles :

la proportionnalité des risques : cela signifie que le rapport des fonctions de risques instantanés pour deux sujets dépend de leurs caractéristiques et non du temps. Cela revient aussi à supposer que les caractéristiques individuelles influencent le niveau du risque (l'intensité du phénomène) et non le profil temporel du risque.

Une forme fonctionnelle des variables continues : un changement d'une unité dans la variable continue doit avoir le même effet sur l'évènement considéré, et ce qu'elle que soit la valeur.

Le modèle de Cox est un modèle semi-paramétrique car il n'y a aucune hypothèse sur la forme de la fonction de survie mais les risques sont supposés proportionnels. C'est pourquoi dans ce modèle la fonction modélisée est celle du risque instantané et non celle de la survie. Le principal défaut du modèle de Cox est qu'il ne donne pas une estimation de l'intercepte $h_0(t)$, et donc son équation. Il n'est donc pas possible d'utiliser ce modèle pour réaliser des prévisions.

Encadré 5 : Modèle de Cox

La **fonction de risque instantané du modèle de Cox** : $h(t;z) = h_0(t) * e^{z*\beta}$,

Où : $Z_i=(Z_{i1}, Z_{i2},\ldots,Z_{ip})$ vecteur de p variables explicatives pour l'individu i. Les variables peuvent évoluer dans le temps.

$e^{z*\beta}$ est une fonction de régression explicitée paramétriquement où β est un p-vecteur de coefficients de régression inconnus.

$h_0(t)$ est une fonction inconnue de t appelée risque sous jacent. Dans l'interprétation, $h_0(t)$ sera le risque instantané (le risque) de connaître l'évènement pour l'individu de référence dont toutes les caractéristiques sont nulles, à la durée t. $h_0(t)$ est commun à tous les individus. Il est déterminé de manière non-paramétrique. Ce $h_0(t)$ correspond à l'intercepte. Au cours du calcul, cet intercepte n'est pas sauvegardé, quelles que soit les options SAS utilisées.

Les deux hypothèses du modèle de Cox :

L'hypothèse de l'existence d'une relation entre la fonction de risque instantané et les variables explicatives : un changement d'une unité dans la variable continue doit avoir le même effet sur l'évènement considéré, et ce qu'elle que soit la valeur.

L'hypothèse de proportionnalité : le rapport des fonctions de risque instantané pour deux sujets i et j de caractéristiques Z_i et Z_j ne dépend que de Z_i et Z_j et ne dépend pas du temps.

$\frac{h_i(t;Z_i)}{h_j(t;Z_j)} = \frac{e^{(\beta*Z_i)}}{e^{(\beta*Z_j)}} = \exp(\beta(Z_i - Z_j)) = K$ (K=hazard ratio)

Avant de mettre en œuvre un modèle de Cox, il faut donc s'assurer que les risques sont proportionnels et les variables continues de la forme adéquate. Néanmoins, afin d'expliquer la réalisation d'un modèle de Cox sous SAS le plus simplement possible, nous mettons en œuvre un modèle sans contrôler les hypothèses, mais nous ne tirerons pas de conclusions à partir de ces résultats.

4.2. Mise en œuvre d'un modèle de Cox sous SAS

Voici la procédure SAS du modèle de Cox :

Proc phreg data= nom_table <options> ;
Class variable_quali <options> ;
Model variable_duree*censure=variables_explicatives ;
Run ;

Ainsi, dans le cas de notre étude :

```
proc phreg data=cohorte2008 simple outest=rslCOX plots=(survival cumhaz) ;
class activite_der(ref="C") statut(ref="2") sect_mod(ref='serv');
model duree*censure(0)= activite_der statut sect_mod generation/ties=efron
risklimits ;
run;
```

La variable à expliquer est le croisement de la durée et des censures :

la variable `duree`=(date de fin du cumul (liquidation au RSI)-date de début de cumul(1er report postérieur à la liquidation RG).

- Si l'observation est censurée : `duree`=date de fin d'observation (31/12/2012)-date de début de cumul(1er report postérieur à la liquidation RG).

`censure(0)` : indique le nom de la variable de censure, suivie entre-parenthèses de la valeur de cette variable lorsque l'observation est censurée, 0 dans notre cas.

Les variables explicatives sont les suivantes :

activite_der : indique le groupe professionnel auquel appartient l'individu : artisan ou commerçant.

statut : indique la situation avant la liquidation de la retraite du régime général : indépendant, salarié, ni indépendant ni salarié.

sect_mod : indique le secteur d'activité de l'emploi relevant du RSI : construction, commerce, industrie et services.

generation : indique l'année de naissance de l'individu.

Options du proc phreg

L'instruction `class` permet de déclarer les variables qualitatives comme dans un modèle de régression.

L'option `simple` permet l'édition de statistiques descriptives des variables explicatives.

L'option `outest=nom_table` permet de sauvegarder les estimateurs de durée β dans une table. Ce tableau correspond donc à la colonne « parameter estimate » du tableau 17 de la page 26.

L'option `plot` permet d'afficher les graphiques des courbes de survie (survival), et de risques cumulés (cumhaz).

L'option `risklimits` permet d'avoir l'intervalle de confiance des risques-ratio.

L'option `ties=nom_methode` indique la méthode à appliquer pour la gestion des évènements simultanés. Il est tout à fait possible d'avoir des données où au moins deux observations ont une durée identique. Il faut dans ce cas indiquer à SAS comment traiter ces évènements simultanés :

- Méthode exacte : La méthode exacte considère que les durées sont en fait toutes différentes et que leur apparente égalité est due au manque de précision des mesures. Il faut alors considérer dans le calcul toutes les permutations possibles des égalités, ce qui peut devenir très compliqué.
- Méthode discrète : on suppose que les évènements se produisent exactement au même moment. Les durées sont mesurées sur une échelle de temps discrète. Cette méthode transforme le modèle de Cox en un modèle logistique à temps discret, car à chaque date, il calcule la probabilité pour un individu de connaître l'évènement en fonction de variables explicatives et du temps avec un modèle logit.
- La méthode de **Breslow** suppose qu'en cas d'événements simultanés, tous les événements ont le même risque que le premier d'entre eux. Il s'agit d'une approximation de la méthode exacte.
- La méthode d'**Efron** calcule un risque moyen pour tous les événements survenus simultanément. Il s'agit également d'une approximation de la méthode exacte.

Ces quatre méthodes donnent des résultats quasiment identiques. Les méthodes exacte et discrète amènent à des temps de calcul extrêmement longs. SAS utilise par défaut la méthode de Breslow. La méthode Efron donne les résultats les plus proches de la méthode exacte en un temps relativement court.

Résultats du modèle de Cox

Tableau 8 : information générale sur le modèle

Model Information	
Data Set	WORK.COHORTE2008
Dependent Variable	duree
Censoring Variable	censure
Censoring Value(s)	0
Ties Handling	EFRON

Number of Observations Read	15017
Number of Observations Used	15017

Le **tableau 8** donne les caractéristiques principales du modèle : nom de la table SAS utilisée, nom des variables de durée et de censure, méthode de gestion des évènements simultanés et nombre d'observations.

Tableau 9 : Information sur les variables qualitatives du modèle

Class Level Information					
Class	Value	Design Variables			
activite_der	A	1			
	C	0			
statut	1	1	0		
	2	0	0		
	3	0	1		
sect_mod	NR	1	0	0	0
	com	0	1	0	0
	const	0	0	1	0
	indus	0	0	0	1
	serv	0	0	0	0

Le **tableau 9** indique quelles sont les variables qualitatives, leurs modalités et les modalités de référence du modèle (vecteur nul).
Activité_der indique le groupe professionnel :
C=commerçant=modalité de référence
A=artisan
Statut indique la situation avant la liquidation
1=salarié
2=indépendant=modalité de référence
3=ni indépendant, ni salarié
Sect_mod indique le secteur d'activité
NR=secteur non renseigné
Com=commerce
Const=construction
Indus=industrie
Serv=services=modalité de référence

Tableau 10 : Tableau du nombre d'évènements observés et du nombre d'observations censurées

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
15017	7626	7391	49.22

Le **tableau 10** rappelle le nombre d'évènements observés (les sorties du cumul), le nombre d'observations censurées et leur proportion. Dans l'étude, il y a 49% d'individus pour lesquels la date de fin de cumul est inconnue.

Tableau 11 : Statistiques descriptives des variables quantitatives du modèle, ici génération

Descriptive Statistics for Continuous Explanatory Variables					
Total Sample					
Variable	N	Mean	Standard Deviation	Minimum	Maximum
generation	15017	1946	3.99761	1912	1953

Tableau 12 : Statistiques descriptives des variables qualitatives du modèle

Frequency Distribution of CLASS Variables		
Total Sample		
Class	Value	Frequency
activite_der	A	5680.0
	C	9337.0
statut	1	3368.0
	2	10734.0
	3	915.0
sect_mod	NR	1286.0
	com	4131.0
	const	2007.0
	indus	1249.0
	serv	6344.0

Tableau 13 : Statut du modèle vis-à-vis de la convergence

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Tableau 14 : Résultats des critères de qualité du modèle

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	141373.86	137305.33
AIC	141373.86	137321.33
SBC	141373.86	137376.84

Tableau 15 : Résultats des tests globaux d'hypothèse nulle

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > Khi-2
Likelihood Ratio	4068.5344	8	<.0001
Score	9277.5886	8	<.0001
Wald	5917.3090	8	<.0001

Les **tableaux 11 et 12** sont émis en raison de l'option `simple`. Ils éditent les statistiques descriptives des variables qualitatives et quantitatives.

Le **tableau 13** indique si le modèle a convergé

Le **tableau 14** permet de mesurer la qualité du modèle et sert à effectuer des comparaisons avec d'autres modèles.
Ces critères d'information permettent de choisir le meilleur modèle ayant le moins de variables afin de satisfaire le critère de parcimonie. Le meilleur modèle est celui pour lequel le critère d'information d'Akaike (AIC) ou le critère bayésien de Schwartz (SBC) est le plus faible.

Le **tableau 15** indique si au moins une variable du modèle apporte de l'information. Il s'agit de 3 tests très proches qui testent H_0= « aucune variable explicative n'apporte de l'information » contre H_1= « au moins une variable explicative apporte de l'information ». Dans notre cas, l'hypothèse nulle est rejetée : au moins une variable est utile au modèle.

Tableau 16 : Résultats des tests de significativité des variables du modèle

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > Khi-2
activite_der	1	14.0901	0.0002
statut	2	283.8912	<.0001
sect_mod	4	5538.0908	<.0001
generation	1	47.4792	<.0001

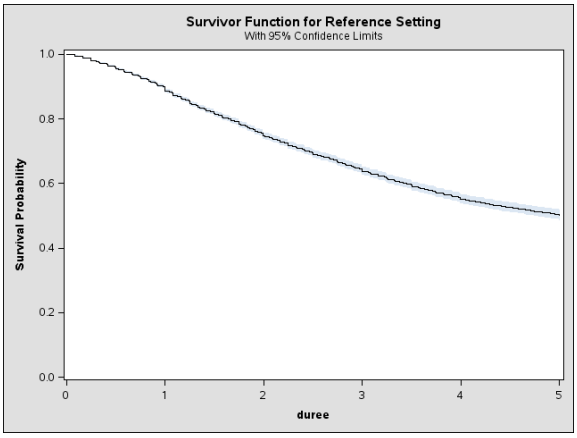
Le **tableau 16** teste la significativité de chacune des variables : on a H_0= « la variable n'est pas significative » et H_1= « la variable est significative ».
Dans notre cas, les 4 variables sont significatives.

Tableau 17 : Estimateurs du modèle de Cox et risque-ratio

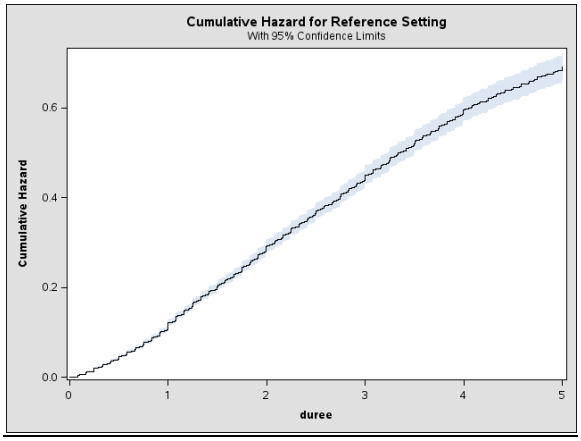
Analysis of Maximum Likelihood Estimates										
Parameter		DDL	Parameter Estimate	Standard Error	Chi-Square	Pr > Khi-2	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
activite_der	A	1	0.10854	0.02892	14.0901	0.0002	1.115	1.053	1.180	activite_der A
statut	1	1	-0.48478	0.03121	241.2227	<.0001	0.616	0.579	0.655	statut 1
statut	3	1	-0.46797	0.05419	74.5629	<.0001	0.626	0.563	0.696	statut 3
sect_mod	NR	1	2.57490	0.03702	4837.2513	<.0001	13.130	12.211	14.118	sect_mod NR
sect_mod	com	1	0.08962	0.03022	8.7927	0.0030	1.094	1.031	1.161	sect_mod com
sect_mod	const	1	0.13783	0.04065	11.4972	0.0007	1.148	1.060	1.243	sect_mod const
sect_mod	indus	1	0.03585	0.04673	0.5887	0.4429	1.037	0.946	1.136	sect_mod indus
generation		1	0.02270	0.00329	47.4792	<.0001	1.023	1.016	1.030	

Le **tableau 17** donne les résultats du modèle de Cox :
Les colonnes 1 et 2 indiquent la modalité concernée par les résultats.
La colonne 4 « parameter estimate » donne les paramètres β estimés par le modèle. Lorsque le paramètre est positif, les personnes ayant la caractéristique étudiée ont un risque *h* plus élevé que les personnes ayant la caractéristique de référence de mettre fin au cumul emploi-retraite. Ainsi, les artisans ont comparativement aux commerçants une probabilité plus grande de sortir du dispositif. Inversement, lorsque le paramètre est négatif, les personnes ayant la caractéristique étudiée ont une plus faible probabilité de mettre fin au cumul emploi-retraite que les personnes ayant la caractéristique de référence. Le paramètre estimé est négatif pour les personnes salariées avant la liquidation : elles ont une probabilité plus faible de sortir du cumul que les personnes qui étaient indépendantes avant la liquidation de leur retraite du régime général.
Pour une variable quantitative, l'interprétation n'est pas la même : une augmentation d'une unité de la variable quantitative conduit à une variation de (exp(β)-1)% de risque de survenue de l'évènement. Par exemple, si la génération augmente d'une année, le risque de mettre fin à un cumul emploi-retraite augmente de 2,3% (exp(0.02270)-1). Ce résultat, 2,3%, peut être lu directement dans la colonne hazard ratio (colonne 8).
La colonne 5 « standard-error » indique l'écart-type des paramètres β.
Les colonnes 6 et 7 correspondent aux tests de khi2 jugeant la significativité de la variable. La colonne 7 est le risque de première espèce, appelé aussi seuil de significativité. Il correspond au risque de considérer que la variable a un effet sur l'évènement étudié alors qu'elle n'en a pas (=rejeter à tort l'hypothèse nulle). Ainsi, si l'on considère que travailler dans l'industrie a un effet sur la durée passée en cumul emploi-retraite, nous aurons 44% de risques de nous tromper. En revanche, nous pouvons affirmer que toutes les autres variables de l'étude ont un effet sur le cumul emploi-retraite avec moins 1% de risques de nous tromper.
La colonne 8 donne les risques-ratio. Le risque ratio mesure le gain d'occurrence du risque chez les personnes ayant la caractéristique étudiée par rapport à ceux possédant la modalité de référence toutes choses égales par ailleurs. Il correspond aux odds-ratio des modèles de régression. Dans notre étude, le risque de sortir du cumul emploi-retraite est 1,148 fois plus

grand pour les personnes travaillant dans le secteur de la construction que pour celles travaillant dans le secteur des services, toutes choses égales par ailleurs. Le risque ratio peut être obtenu à partir des paramètres estimés : exp(0.13783)=1.148.
Les colonnes 9 et 10 sont obtenues grâce à l'option `risklimits`. Elles donnent l'intervalle de confiance du risque ratio : toutes choses égales par ailleurs, il y a 95% de chances que le risque de mettre fin au cumul emploi-retraite pour les cumulants du secteur de la construction soit entre 1,06 et 1,24 fois supérieur à celui des personnes travaillant dans le domaine des services.



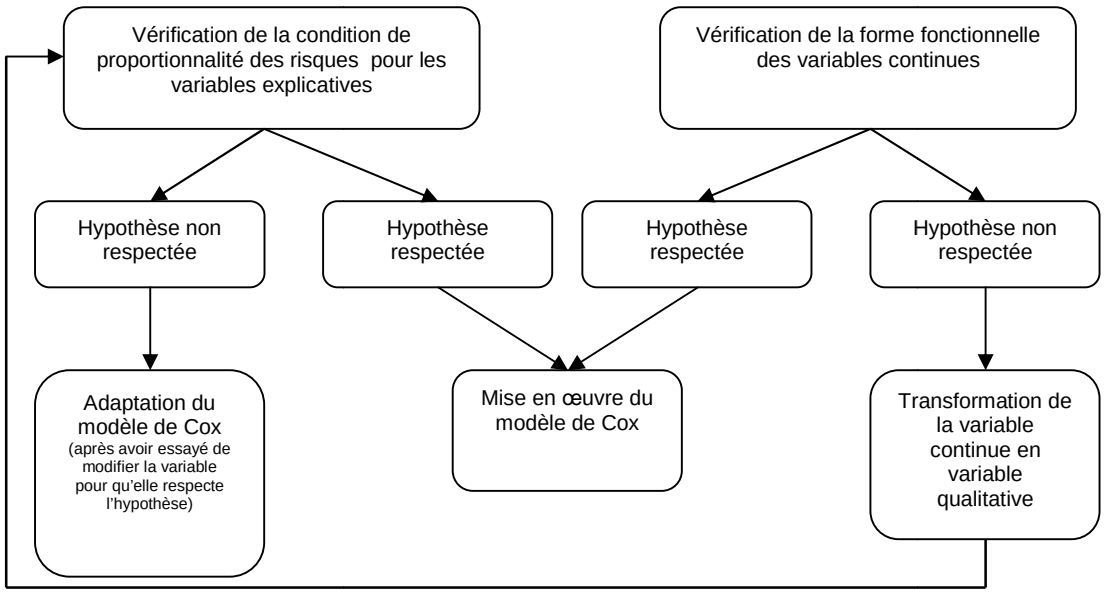
Graphique 19 : Fonction de survie obtenue avec le modèle de Cox



Graphique 20 : fonction de risques cumulés obtenue à partir du modèle de Cox

Nous n'analysons pas les sorties SAS puisque les conditions de mise en œuvre d'un modèle de Cox n'ont pas été contrôlées. Nous les vérifions maintenant.

Figure 2 : Schéma de la mise en œuvre d'un modèle de Cox



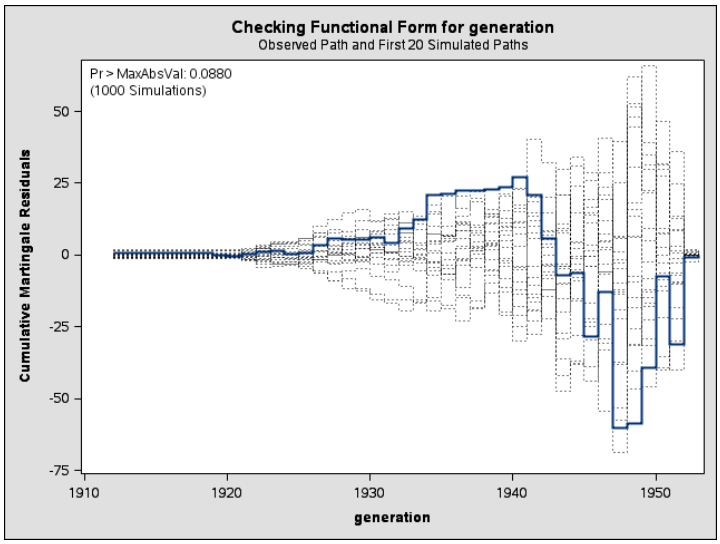
4.3. Vérification de la forme fonctionnelle des variables continues

Pour réaliser un modèle de Cox, les variables continues doivent respecter la règle suivante : un changement d'une unité dans la variable continue doit avoir le même effet sur l'évènement considéré, et ce qu'elle que soit la valeur. Afin de mieux comprendre le sens de cette règle, reprenons les paramètre estimé par le modèle de Cox précédent (tableau 17) pour la variable *generation* : lorsque la génération augmente d'une année, le risque de mettre fin à un cumul emploi-retraite augmente de 2,3% (exp(0.02270)-1). Cela signifie que lorsque l'âge à l'entrée du cumul emploi-retraite passe de 56 ans à 55 ans, le risque de sortir du cumul emploi-retraite augmente de 2,3%. Cela signifie également que lorsque l'âge à l'entrée du cumul emploi-retraite passe de 96 ans à 95 ans, le risque de sortir du cumul emploi-retraite augmente aussi de 2,3%.
Pour contrôler cette hypothèse, les résidus de Martingale sont utilisés. Ils peuvent être interprétés comme la différence au cours du temps entre le nombre d'évènements observés et le nombre d'évènements prédit par le modèle de Cox. Avec la procédure PHREG, il est plus simple d'utiliser les résidus de Martingale cumulatifs pour vérifier cette hypothèse. Une option permet de représenter un graphique des résidus de Martingale observés en fonction de la variable continue. Sur ce graphique, des simulations de résidus sont réalisées sous l'hypothèse que la variable a une forme adéquate. Il suffit alors d'observer si les résidus cumulatifs observés diffèrent des simulations. Si c'est le cas, alors la variable continue n'a pas la bonne forme.

Pour compléter ce test graphique, il est possible d'effectuer un test de Kolmogorov. Il permet de déterminer si deux échantillons suivent la même loi : il indique si les résidus des évènements observés suivent la même loi qu'au moins une distribution des résidus des simulations. Les deux hypothèses testées sont :
H_0=au moins une distribution des résidus des évènements simulés est identique à la distribution des résidus des évènements observés. La variable a donc la forme adéquate.
H_1= aucune distribution des résidus des évènements simulés ne suit la même loi que la distribution des résidus des évènements observés. La variable n'est donc pas de la forme adéquate.

Sous SAS, pour effectuer cette vérification, il suffit d'intégrer au modèle de Cox, créé à partir de l'instruction **`phreg`**, une instruction `assess` :
L'option `var=(nom_var)` indique que l'on souhaite tester la forme fonctionnelle de la variable dont le nom est entre-parenthèses.
L'option `resample=`xx permet d'effectuer le test de Kolmogorov sur xx simulations.
L'option `seed=`xx où xx est un nombre aléatoire entier utilisé pour effectuer les simulations de distribution des évènements.

```
proc phreg data=cohorte2008;
class activite_der(ref="C") statut(ref="2") sect_mod(ref='serv');
model duree*censure(0)= activite_der statut sect_mod generation;
assess var=(generation)  resample=50 seed=27513;
run;
```



Graphique 21 : Vérification de la forme de la variable *generation* à partir des résidus de Martingale

La distribution des résidus de Martingale des sorties de cumul emploi-retraite en fonction de la variable *generation* fait partie des distributions de résidus simulés. Par ailleurs, le test de Kolmogorov conduit à ne pas rejeter l'hypothèse nulle : la variable *generation* peut donc être considérée comme de la forme adéquate.
Si une variable continue n'est pas de la forme adéquate, il faut la transformer à l'aide des fonctions log(X), X^2 ou racine(X). Le plus souvent, la variable continue est transformée en variable qualitative, pour laquelle il est nécessaire de vérifier l'hypothèse de risques proportionnels.

4.4. Vérification de l'hypothèse des risques proportionnels

Pour mettre en œuvre un modèle de Cox, il faut que l'hypothèse des risques proportionnels soit vérifiée: autrement dit, le risque doit être constant au cours du temps, pour chaque variable explicative. Dans notre étude, il s'agit de contrôler cette hypothèse pour les variables sur le secteur d'activité, le groupe professionnel, la situation au moment de la liquidation et la génération.

Il existe 2 manières de vérifier l'hypothèse des risques proportionnels :

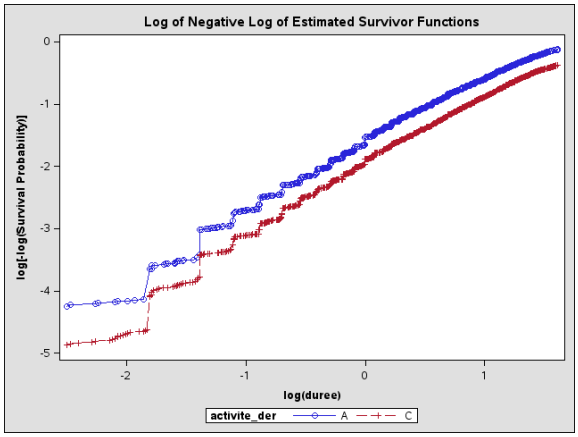
Méthode 1 : contrôler l'hypothèse de risques proportionnels revient à vérifier que les courbes Log(H(t)) selon Log(t)[10], tracées pour chaque modalité, sont parallèles entre elles. Il aurait été possible de contrôler cette hypothèse à partir de la fonction de survie, mais comme les fonctions de survie des différentes modalités se ressemblent, il est plus facile de passer par la fonction de risques cumulés.
Il faut effectuer ce test graphique pour toutes les variables. Pour les variables quantitatives, il est plus difficile à mettre en œuvre puisqu'une courbe est tracée pour chacune des valeurs de la variable.
Sous SAS pour obtenir ce graphique, il suffit de demander le tracé de la courbe Log(H(t)) avec la procédure *lifetest*. Voici le programme SAS correspondant à une vérification de cette hypothèse pour la variable *activite_der*, qui renseigne sur le groupe professionnel à partir de la méthode de Kaplan-Meier:

[10] *note : Log (-Log(S(t)))= Log(H(t))). En effet, si l'on obtient une droite y=αln(t)+μ, alors ln(-ln(S(t))= αln(t)+μ, ce qui est équivalent à : S(t)=exp(-t^α exp(μ)). On reconnait une loi de Weibull*

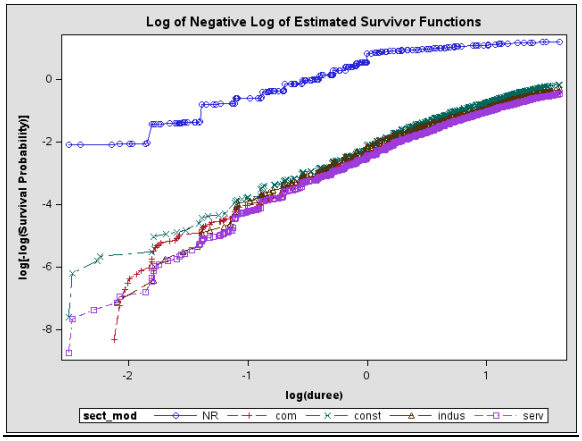
```
Proc lifetest data=cohorte2008  conftype=loglog
plots=(loglogs) graphics;
Time duree*censure(0) ;
strata activite_der;
Run ;
```

`plots=(loglogs)` est l'instruction qui permet d'obtenir la courbe Log (-Log(S(t)))= Log(H(t))).

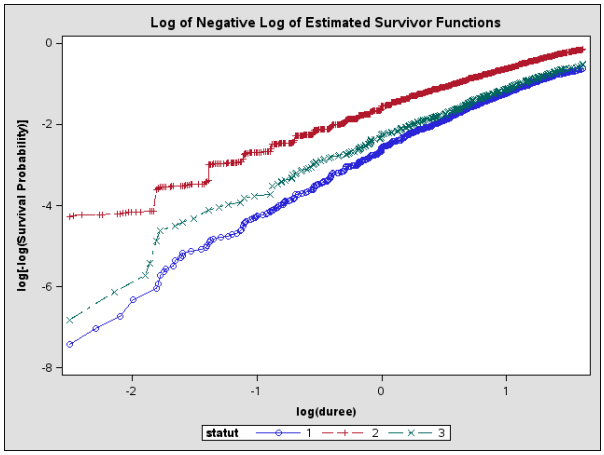
Sortie SAS :



Graphique 22 : Vérification de l'hypothèse des risques proportionnels pour la variable *activite_der* (renseignant sur le groupe professionnel) à partir des courbes du logarithme des risques proportionnels



Graphique 23 : Vérification de l'hypothèse des risques proportionnels pour la variable *sect_mod* (secteur d'activité) à partir des courbes du logarithme des risques proportionnels



Graphique 24 : Vérification de l'hypothèse des risques proportionnels pour la variable *statut* (situation avant la liquidation) à partir des courbes du logarithme des risques proportionnels

La variable *activite_der*, concernant les groupes professionnels, semble plutôt respecter l'hypothèse de risques proportionnels : la courbe représentant le logarithme de la fonction de risques cumulés des artisans est plutôt parallèle à celle des commerçants.
En revanche, les variables concernant la situation avant la liquidation (statut) et le secteur d'activité ne semblent pas respecter l'hypothèse des risques proportionnels : les courbes semblent se rejoindre.
L'inconvénient de cette méthode de vérification est de reposer sur un graphique. En général, les courbes ne sont pas strictement parallèles, et il y a donc une part de subjectivité pour considérer qu'elles le sont.

Méthode 2 : **Le processus du score** peut aussi être utilisé pour déterminer si l'hypothèse des risques proportionnels est vérifiée. Les résidus du score sont calculés pour chaque variable explicative, et chaque individu, et sont fonction de la durée. Le processus du score est obtenu en faisant la somme des résidus du score.

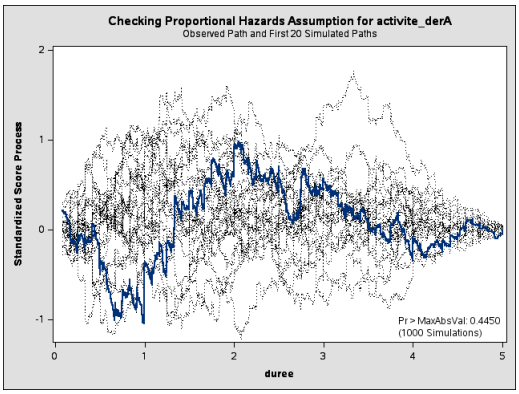
Sous SAS, pour contrôler l'hypothèse des risques proportionnels, le processus du score standardisé est représenté en fonction de la durée. Sur ce graphique, des distributions simulées de résidus réalisées sous l'hypothèse des risques proportionnels sont représentées. Il s'agit alors d'observer si le processus du score des résidus observés diffère des simulations. Si c'est le cas, alors la variable ne respecte pas la proportionnalité des risques.

En parallèle de cette vérification graphique, le **test de Kolmogorov** peut être mis en œuvre. Il indique si les résidus des évènements observés suivent la même loi qu'au moins une distribution des résidus des simulations. Les deux hypothèses testées sont :

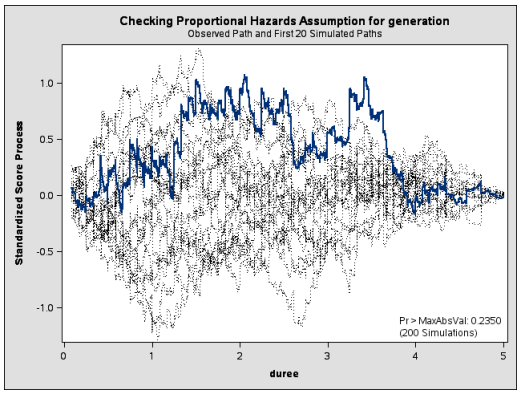
H_0=au moins une distribution des résidus des évènements simulés est identique à la distribution des résidus des évènements observés. La variable respecte donc la condition de proportionnalité des risques.
H_1= aucune distribution des résidus des évènements simulés ne suit la même loi que la distribution des résidus des évènements observés. La variable ne respecte donc pas la condition de proportionnalité des risques.

Avec SAS, pour effectuer cette vérification, il suffit d'ajouter une option de l'instruction `assess` du modèle de Cox :
L'option `ph` indique que l'on souhaite vérifier l'hypothèse des risques proportionnels (ph=proportionnal hazard), pour chaque variable explicative.

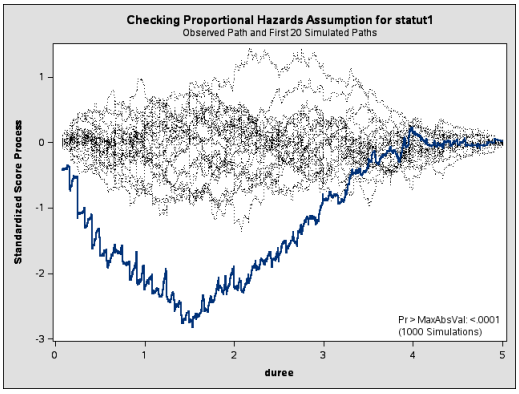
```
proc phreg data=cohorte2008;
class activite_der(ref="C") statut(ref="2") sect_mod(ref='serv');
model duree*censure(0)= activite_der statut sect_mod generation;
assess ph resample=200 seed=27513;
run;
```

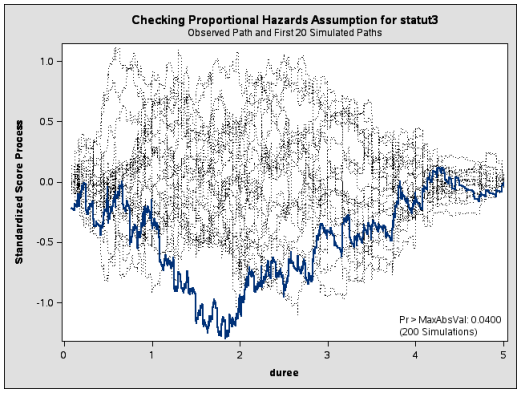


Graphique 25 : Vérification de l'hypothèse des risques proportionnels pour la variable *activite_der* (groupe professionnel) à partir des résidus du modèle

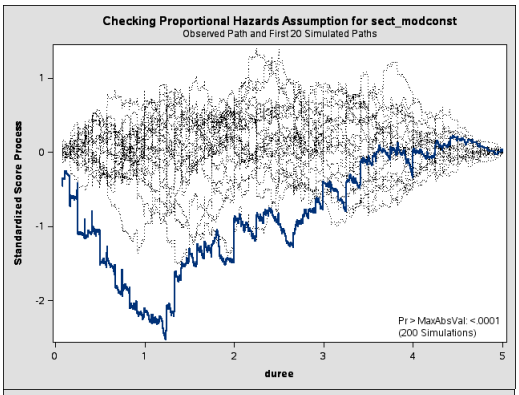


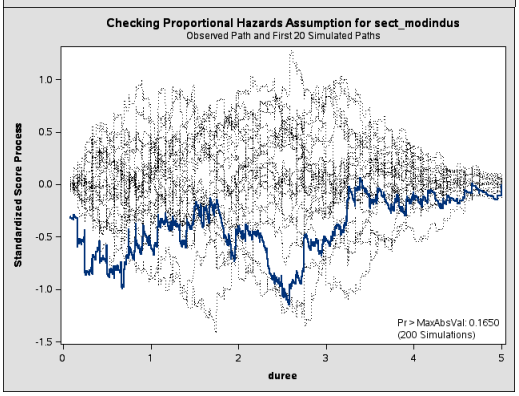
Graphique 26 : Vérification de l'hypothèse des risques proportionnels pour la variable *generation* (année de naissance) à partir des résidus du modèle

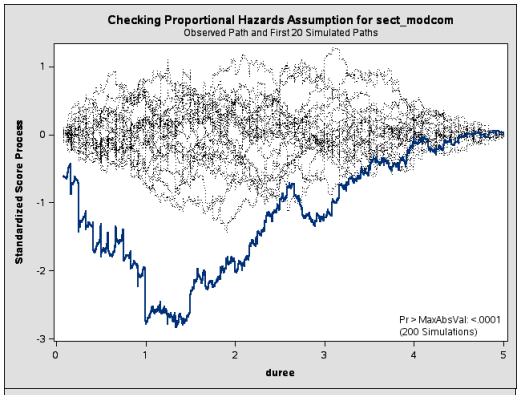


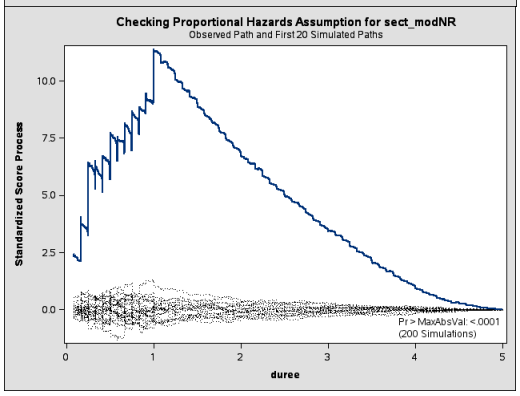


Graphiques 27 : Vérification de l'hypothèse des risques proportionnels pour la variable *statut* (situation avant la liquidation) à partir des résidus du modèle









Graphiques 28 : Vérification de l'hypothèse des risques proportionnels pour la variable *sect_mod* (secteur d'activité) à partir des résidus du modèle

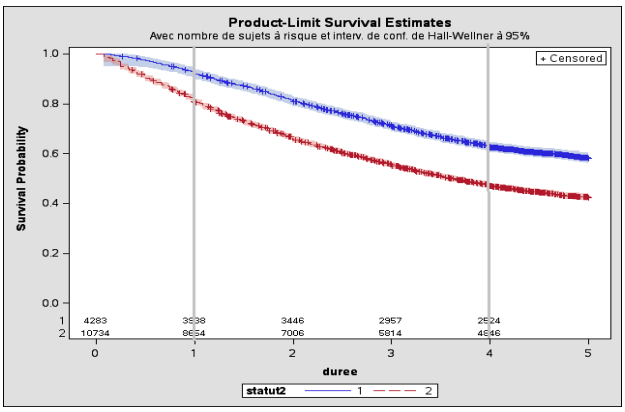
Les variables concernant les groupes professionnels (*activite_der*) et l'année de naissance (*generation*) vérifient l'hypothèse de risques proportionnels : la distribution des évènements fait partie des distributions possibles, et les tests de Kolmogorov ne conduisent pas au rejet de H_0.
En revanche, les variables sur la situation avant la liquidation (*statut*) et sur le secteur d'activité (*sect_mod*) ne respectent pas la condition de risques proportionnels. Il n'est donc pas possible de mettre en œuvre le modèle de Cox tel quel : il sera nécessaire de le transformer pour lever la condition de risques proportionnels.

Néanmoins, nous essayons, dans un premier temps, de transformer les variables qui ne respectent pas la condition de risques proportionnels, afin de diminuer au maximum leur nombre et faciliter la transformation du modèle de Cox :
Dans notre étude, la variable *statut* comporte 3 modalités : salarié, indépendant et ni indépendant, ni salarié. Les modalités « salarié » et « ni indépendant, ni salarié » sont regroupées en une seule modalité « non indépendant ». Cette nouvelle variable est nommée *statut2*.
Les personnes pour lesquelles le secteur d'activité est inconnu sont supprimées de l'analyse. Ces personnes représentent 8,6% de la base de données.

La condition de risques proportionnels est de nouveau vérifiée pour ce nouveau modèle : elle est désormais respectée pour la variable sur le secteur d'activité. En revanche, malgré le regroupement, la variable informant de la situation avant la liquidation ne respecte pas cette condition. Un modèle de Cox avec des risques non proportionnels est alors mis en œuvre.

5. MODELE DE COX A RISQUES NON PROPORTIONNELS

Lorsque la condition de risques proportionnels ne peut pas être respectée, le modèle de Cox est modifié de manière à lever cette condition. Dans notre étude sur le cumul emploi-retraite, comme nous l'avons observé ci-dessus, le risque de sortir du cumul emploi-retraite n'évolue pas de manière proportionnelle à la durée écoulée lorsque la population est stratifiée en fonction de la situation à la liquidation.
Nous incluons donc dans le modèle de Cox initial une variable renseignant la situation avant la liquidation qui dépend du temps. Dans le cas de l'ajout d'une variable dépendante du temps, l'hypothèse de proportionnalité n'a besoin d'être validée que dans les intervalles de temps où les variables dynamiques sont constantes. Cela revient à supposer une proportionnalité par morceaux (encadré 6).
Pour créer cette variable dépendante du temps, il faut déterminer les intervalles de temps sur lesquels le risque de sortir du cumul emploi-retraite semble évoluer de la même façon. En général, ces intervalles sont choisis à partir de l'allure des courbes de survie (graphique 29). Nous considérons que la pente des courbes de survie est identique entre 0 et 1 an, puis entre 1 an et 4 ans, et entre 4 ans et 5 ans. Ce découpage de la variable en fonction du temps à partir d'une représentation graphique est subjectif. Il s'obtient en partie par tâtonnement.



Graphique 29 Fonctions de survie stratifiées en fonction de la situation au moment de la liquidation, créées à partir de la méthode de Kaplan-Meier

Encadré 6 : Modèle de Cox à risques non proportionnels

Lorsqu'une variable ne répond pas à l'exigence de risques proportionnels, il est possible de modifier le modèle de Cox de 3 manières :

- Une modélisation introduisant une variable explicative dépendante du temps
- Une modélisation par partie : réalisation d'un modèle de Cox différent pour chaque intervalle de temps.
- Une stratification du modèle de Cox en fonction de la valeur de la variable pour laquelle l'hypothèse n'est pas vérifiée.

Le modèle de Cox stratifié

La stratification consiste à calculer un modèle de Cox en attribuant une valeur différente au risque de base $h_0(t)$ à chaque catégorie de la variable de stratification. En revanche, les coefficients des variables explicatives (les β) et donc l'influence des variables explicatives sont identiques dans chacune des strates. Il est possible d'effectuer une stratification en fonction de plusieurs variables.

$Soit\ s\ une\ strate, h(t;z)_s = \mathrm{h}(0)_s * e^{\beta 1 z 1 + \beta 2 z 2 + \cdots \ldots + \beta k z k}$,

Par ce moyen la condition de risques proportionnels est levée :

Soit deux individus i et j n'appartenant pas à la même strate:

$$\frac{h_i(t)}{h_j(t)} = \exp\left(h_{0;1}(t) - h_{0;2}(t)\right) * exp\left[\sum_{m=1}^{k} \beta_m(Z_{im} - Z_{jm})\right]$$

$$et\ h_{0;1}(t) \neq h_{0;2}(t)\ le\ risque - ratio\ se\ modifie\ avec\ t$$

En revanche, l'hypothèse de risques proportionnels continue d'être valide pour les individus appartenant à la même strate.

Dans un modèle stratifié, une fonction de vraisemblance partielle est construite séparément pour chacune des variables explicatives, la fonction que l'on maximise pour trouver les estimateurs des paramètres étant le produit de ces vraisemblances partielles. Soit L_s la fonction de vraisemblance partielle de la strate s : $L=L_1*L_2* \ldots *L_s$

Un modèle de Cox stratifié repose sur l'hypothèse de non-interaction entre la variable de stratification et les variables explicatives du modèle.

Pour mettre en œuvre un modèle de Cox stratifié sous SAS, il suffit de rajouter l'instruction strata :

```
proc phreg data=cohorte2008;
class activite_der(ref="C") sect_mod(ref='serv');
model duree*censure(0)= activite_der sect_mod generation;
strata statut;
run;
```

Sous SAS, la variable dépendante du temps est créée au sein de la procédure **phreg**, et non dans une étape data. Cette variable est le résultat de l'interaction entre la variable dépendant du temps et le temps : var_dependant_temps*duree. Elle est une indicatrice et donc une variable numérique.

```
proc phreg data=cohorte2008;
class activite_der(ref="C") sect_mod(ref='serv');
model duree*censure(0)= activite_der sect_mod statutdur1 statutdur2
statutdur3 generation/ ties=efron risklimits;
statutdur1=statut2*(duree <1);
statutdur2=statut2*(1<=duree<4);
statutdur3=statut2*(duree>=4);
run;
```

Rappel : la variable *activite_der* renseigne le groupe professionnel (artisan ou commerçant). La variable *sect_mod* correspond au secteur d'activité (services, commerce, industrie, construction). La variable *generation* est l'année de naissance des individus.
La variable *statut2* correspondant à la situation au moment de la liquidation est croisée avec la durée : le résultat est stocké dans les variables *statutdur* qui sont les variables dépendantes du temps du modèle de Cox.
Voici un extrait des résultats du modèle de Cox :

Tableau 18 : Estimateurs du modèle de Cox à risque non proportionnel

Analysis of Maximum Likelihood Estimates									
Parameter		DDL	Parameter Estimate	Standard Error	Chi-Square	Pr > Khi-2	Hazard Ratio	95% Hazard Ratio Confidence Limits	Label
activite_der	A	1	0.10256	0.03340	9.4260	0.0021	1.108	1.038 1.183	activite_der A
sect_mod2	com	1	0.08276	0.03038	7.4230	0.0064	1.086	1.023 1.153	sect_mod2 com
sect_mod2	const	1	0.15353	0.04219	13.2412	0.0003	1.166	1.073 1.266	sect_mod2 const
sect_mod2	indus	1	0.04788	0.04746	1.0175	0.3131	1.049	0.956 1.151	sect_mod2 indus
statutdur1		1	0.53089	0.07188	54.5535	<.0001	1.700	1.477 1.958	
statutdur2		1	0.29354	0.03526	69.3026	<.0001	1.341	1.252 1.437	
statutdur3		1	-0.01836	0.09242	0.0395	0.8425	0.982	0.819 1.177	
generation		1	0.02439	0.00354	47.5086	<.0001	1.025	1.018 1.032	

Exemple d'interprétation des variables dépendantes du temps : Le paramètre associé à la variable statutdur1 est positif. En conséquence, lorsqu'un individu était indépendant avant la liquidation, et non salarié (c'est-à-dire que la variable statutdur1 augmente d'une unité), le risque de sortir du cumul emploi-retraite est 1,7 fois plus élevé. On a moins d'1% de risques de se tromper en l'affirmant.
La variable statutdur3 n'est pas significative : on a 84% de risque d'avoir une estimation erronée pour cette variable.

6. LE MODELE DE COX FINAL

Après avoir étudié différents modèles de Cox, nous avons choisi de ne pas retenir la variable *generation* dans le modèle car elle ne semble pas être la plus pertinente. Nous préférons retenir la variable *gpage* qui correspond à l'âge au moment de la liquidation : 55-59 ans, 60 ans, 61 ans et plus. Cette nouvelle variable donne indirectement des informations sur les conditions de départ en retraite du régime général : en bénéficiant du dispositif de retraite anticipée, ou en partant à l'âge légal.
La variable *activite_der* renseignant sur le groupe professionnel est retirée du modèle car le secteur d'activité est lié au groupe professionnel : les personnes travaillant dans le secteur de la construction sont principalement des artisans. Les variables *sect_mod2* et *activite_der* apportent donc une information similaire.

Dans ce nouveau modèle, les variables *sect_mod2* et *gpage* respectent la condition de risques proportionnels, et non la variable *statut2* (annexe 1). Le nouveau modèle de Cox contient donc comme variables explicatives : *sect_mod2* (secteur d'activité), *gpage* (âge à la liquidation), et *statutdur1-statutdur3* (situation à la liquidation en fonction de la durée de cumul).

Programme SAS du dernier modèle de Cox :

```
proc phreg data=cohorte2008 ;
class sect_mod2(ref='serv') gpage(ref='61+');
model duree*censure(0)= sect_mod2 statutdur1-statutdur3 gpage / ties=efron
risklimits;
statutdur1=statut2*(duree <1);
statutdur2=statut2*(1<=duree<4);
statutdur3=statut2*(duree>=4);
run;
```

Sorties SAS

Tableau 19 : information générale sur le modèle

Model Information	
Data Set	WORK.COHORTE2008
Dependent Variable	duree
Censoring Variable	censure
Censoring Value(s)	0
Ties Handling	EFRON

Number of Observations Read	15017
Number of Observations Used	13731

Tableau 20 : Information sur les variables qualitatives du modèle

Class Level Information				
Class	Value	Design Variables		
sect_mod2	com	1	0	0
	const	0	1	0
	indus	0	0	1
	serv	0	0	0
gpage	55-59	1	0	
	60	0	1	
	61+	0	0	

Tableau 21 : Tableau du nombre d'évènements observés et du nombre d'observations censurées

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
13731	6387	7344	53.48

Tableau 22 : Statut du modèle vis-à-vis de la convergence

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Tableau 23 : Résultats des critères de qualité du modèle

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	117591.91	117269.49
AIC	117591.91	117285.49
SBC	117591.91	117339.58

Tableau 24 : Résultats des tests globaux d'hypothèse nulle

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > Khi-2
Likelihood Ratio	322.4294	8	<.0001
Score	316.3487	8	<.0001
Wald	311.1779	8	<.0001

Tableau 25 : Résultats des tests de significativité des variables du modèle

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > Khi-2
sect_mod2	3	35.0594	<.0001
statutdur1	1	57.3818	<.0001
statutdur2	1	77.5367	<.0001
statutdur3	1	11.4596	0.0007
gpage	2	54.4525	<.0001

Tableau 26 : Estimateurs du modèle de Cox et risque-ratio

Analysis of Maximum Likelihood Estimates									
Parameter		DDL	Parameter Estimate	Standard Error	Chi-Square	Pr > Khi-2	Hazard Ratio	95% Hazard Ratio Confidence Limits	Label
sect_mod2	com	1	0.06727	0.02977	5.1056	0.0238	1.070	1.009 1.134	sect_mod2 com
sect_mod2	const	1	0.21621	0.03682	34.4866	<.0001	1.241	1.155 1.334	sect_mod2 const
sect_mod2	indus	1	0.09279	0.04485	4.2798	0.0386	1.097	1.005 1.198	sect_mod2 indus
statutdur1		1	0.54370	0.07177	57.3818	<.0001	1.722	1.496 1.983	
statutdur2		1	0.29900	0.03396	77.5367	<.0001	1.349	1.262 1.441	
statutdur3		1	0.39127	0.11558	11.4596	0.0007	1.479	1.179 1.855	
gpage	55-59	1	0.23471	0.03266	51.6345	<.0001	1.265	1.186 1.348	gpage 55-59
gpage	60	1	0.12929	0.03008	18.4727	<.0001	1.138	1.073 1.207	gpage 60

Nous essayons de comprendre les différents usages du cumul emploi-retraite à partir de variables sur le secteur d'activité de l'emploi exercé en cumul emploi-retraite, l'âge au début du cumul, et la situation de l'individu au moment de la liquidation de sa retraite au régime général : indépendant ou salarié. Le sexe ne fait pas partie des variables retenues pour expliquer les différentes durées de cumul, car les statistiques descriptives ont montré que les hommes et les femmes avaient des caractéristiques proches.

L'élément qui a le plus d'influence sur la durée d'un cumul emploi-retraite est la situation de l'individu au moment de sa liquidation. Comme le mettait en avant l'analyse non paramétrique, les personnes qui étaient déjà indépendantes avant la liquidation de leur retraite restent moins longtemps en cumul que celles qui étaient salariées. Le modèle de Cox précise ce résultat : au cours de la première année, le risque de sortir du cumul emploi-retraite est 1,7 fois plus élevé pour les indépendants que pour les personnes qui n'étaient ni salariées, ni indépendantes, toutes choses égales par ailleurs (encadré orange tableau 26). Cette surexposition au risque de sortir du cumul emploi-retraite pour les indépendants est présente sur toute la période d'observation, même si elle diminue légèrement après la première année. Entre la première et la quatrième année de cumul, les indépendants ont une probabilité

de mettre fin au cumul de 35% supérieure à celle des salariés (encadré rouge tableau 26). Pour les personnes déjà indépendantes avant la retraite, le cumul emploi-retraite semble être un complément de revenu accompagnant leur cessation d'activité. En revanche, les personnes qui exerçaient une autre activité avant leur retraite du régime général ont investi dans un nouveau projet professionnel, et prolongent ainsi leur cumul.

Le deuxième élément qui joue un rôle sur le cumul emploi-retraite est l'âge auquel un individu débute une activité d'indépendant après avoir liquidé sa retraite au régime général. D'après le modèle non paramétrique, les personnes débutant un cumul emploi-retraite entre 55 ans et 60 ans ont des durées de cumul emploi-retraite plus courtes que celles commençant un cumul à partir de 61 ans. L'analyse descriptive montrait que les jeunes cumulants travaillaient plus souvent dans le secteur de la construction, ce qui expliquait peut-être leur plus courte durée de cumul. D'après le modèle de Cox, à situation avant la liquidation et secteur d'activité similaires, les entrants en cumul à 55-59 ans ont une probabilité plus grande de mettre fin au cumul que celles entrées à 61 ans et plus, et dans une moindre mesure à celles entrées à 60 ans (encadrés bleu tableau 26). Le modèle de Cox montre donc que l'âge d'entrée en cumul a un effet propre, et influence le cumul d'une façon qui pourrait paraître surprenante : les personnes les plus jeunes ont des cumuls plus courts.

L'âge à l'entrée en cumul dépend de l'âge auquel un individu liquide sa pension de retraite. C'est peut-être à partir des conditions d'accès à la retraite du régime général que nous pourrions expliquer ce résultat. Les individus ayant liquidé une retraite au régime général avant 60 ans ont bénéficié des dispositifs de retraite anticipée. Ils ont sûrement eu une longue carrière ou des emplois exigeants physiquement les amenant plus rapidement à une retraite totale que le reste de la population. Par ailleurs, comme les jeunes cumulants travaillent plus souvent dans le domaine de la construction, nous pouvons supposer qu'ils exerçaient avant leur retraite une activité dans ce domaine, ce qui confirme l'idée qu'ils ont connu des emplois physiques.

L'activité exercée pendant le cumul emploi-retraite a également un effet sur la durée de ce cumul. Ainsi, le modèle de Cox confirme que les indépendants du secteur de la construction dont les activités sont souvent plus physiques, exercent moins longtemps leur activité que ceux dans les services. Ils ont une probabilité de mettre fin au cumul emploi-retraite de 24% supérieure, à âge à l'entrée en cumul et situation au moment de la liquidation identiques (encadrés violet tableau 26). Par ailleurs, les personnes travaillant dans le commerce ou l'industrie ont une chance un peu plus grande de mettre fin au cumul emploi-retraite que celles exerçant une activité dans les services.

7. CONCLUSION

Les modèles de durée ont permis d'apporter un nouvel éclairage sur le cumul d'une activité d'indépendant avec une retraite du régime général en mettant en exergue les facteurs influençant la durée passée en cumul emploi-retraite. La durée d'un cumul emploi-retraite semble s'expliquer principalement par la carrière des individus. Elle dépend en partie de l'âge auquel les cumulants liquident leur retraite, qui est lui-même la conséquence des dispositifs de retraite anticipée, et donc de la carrière. Elle résulte aussi de la dernière activité effectuée avant le cumul emploi-retraite.

Les modèles de durée donnent des informations sur la durée du cumul emploi-retraite alors que près d'une personne sur deux n'a pas encore liquidé sa retraite d'indépendant. Afin de confirmer et de mieux comprendre l'utilisation de ce dispositif, une étude comparable sera menée lorsque l'intégralité du cumul sera observable.

Annexe 1 : Modèle de durée paramétrique

Les modèles paramétriques, comme le modèle de Cox, permettent de trouver des facteurs explicatifs, et d'en mesurer leurs effets, tout en prenant en compte une dimension temporelle. Ils peuvent également être utilisés pour réaliser des projections car l'équation du modèle est connue. Néanmoins, leur utilisation est moins fréquente car les modèles paramétriques imposent de nombreuses contraintes :
une distribution statistique connue à la distribution des évènements observés, c'est-à-dire qu'il faut trouver une loi statistique qui jouera le rôle de fonction de survie. SAS ne propose pas de fonction en escalier, qui est souvent utilisée.
suppose l'influence attribuée aux caractéristiques individuelles, c'est-à-dire qu'ils imposent la forme à la fonction de risques.
Il existe ainsi trois types de modèles paramétriques selon les hypothèses émises :

Modèles à risques proportionnels

On fait l'hypothèse **que les caractéristiques agissent de façon multiplicative sur les risques,** cela se traduit par le fait que les risques sont proportionnels entre-eux.
Dans un modèle à risques proportionnels, le quotient (le risque) pour un individu ayant les caractéristiques z est de la forme : $h(t;z) = h_0(t) * e^{z*\beta}$
Où $h_0(t) = h(t,0)$représente le risque instantané pour un individu dont le vecteur de caractéristiques est nul. Ce risque peut prendre toutes les formes paramétriques connues. Par exemple, Si $h_0(t)$ est constant, le modèle est exponentiel etc.
Noms de modèle de ce type : Weibull, exponentielle, Gompertz

Modèles à sorties accélérées

On fait l'hypothèse que **les caractéristiques agissent de façon multiplicative sur les fonctions de séjour**. Ceci signifie que si $S_i()$ et $S_j()$ sont les temps de survie afférents à deux individus i et j, alors il existe une constante Φ_{ij} telle que $S_i(t)=S_j(\Phi_{ij}*t)$ pour tout t. Comme exemple de cette configuration, on peut citer la relation selon laquelle une année de la vie d'un chien équivaut à 7 années de la vie d'un homme.
Soit $S_0(t)$ est la fonction de séjour pour l'individu de référence (pour lequel l'ensemble des caractéristiques est nul), la fonction de séjour pour les individus aux caractéristiques z est de la forme :

$$S(t;z) = S_0(t * e^{z\beta})$$

Ces modèles ont la particularité de comporter un résidu aléatoire indépendant des caractéristiques individuelles de z.
Noms des modèles de ce type : modèle log-normaux, gamma, log-logistique, et exponentiel et de Weibull

Modèles à risques proportionnels et à sorties accélérées

Certains modèles peuvent avoir des **caractéristiques qui agissent de façon multiplicative à la fois sur les risques et à la fois sur les fonctions de séjour.** La relation suivant doit alors être vérifiée :

$$\underbrace{h_0(t) * e^{z*\beta}}_{\text{Modèle à risques proportionnels}} = \underbrace{S_0(t * e^{z\beta})}_{\text{Modèle à sorties accélérées}}$$

Seules les distributions de Weibull et exponentielle, sous certaines conditions, vérifient ces deux propriétés.

Mise en œuvre d'un modèle paramétrique

Pour mettre en œuvre un modèle paramétrique, il faut auparavant trouver la loi statistique qui correspond à la distribution des évènements supposés. Pour trouver cette loi statistique, il est possible d'utiliser des tests de rapport de vraisemblance. Il s'agit dans ce cas de réaliser deux modèles paramétriques dont l'un des modèles est un cas particulier de l'autre. Par exemple, le modèle exponentiel est un cas particulier du modèle de Weibull. Il faut ensuite, à partir des valeurs de la log-vraisemblance de chacun des modèles, effectuer un test de ratio de vraisemblance qui indiquera si un modèle est préférable. Il existe également des tests graphiques représentant la fonction du risque ou les résidus des modèles qui aident dans le choix de la meilleure distribution statistique.
Une fois le modèle statistique choisi, SAS l'exécute grâce à la procédure suivante :

```
Proc lifereg data=nom_table ;
class nom_var_qualitative;
Model var_duree*var_censure(0)=nom_var_explicative/d=nom_distribution;
run;
```

Annexe 2: Stockage des résultats d'un modèle de Cox

Les résultats peuvent être stockés à 3 niveaux :

Option outest dans la procédure proc phreg. Une observation est stockée avec les estimateurs des coefficients de régression (page 18).

Une instruction output donne une sauvegarde de statistique (fonction de survie etc.) au niveau individuel

Une instruction baseline permet une sauvegarde de statistiques à chaque instant pour un ensemble de valeurs choisies de variables explicatives. La liste de ces valeurs doit se trouver dans une autre table dont le nom est spécifié dans l'option cov= de l'instruction baseline. De plus, les variables doivent être référencées sous le même nom.

Exemple de programme SAS :

```
data b;
input  activite_der $ statut $ sect_mod $ generation;
cards;
C 1  serv 1948
A 1 serv 1948;
run;
proc phreg data=cohorte2008 outest=rslCOX ;
class activite_der(ref="C") statut(ref="2") sect_mod(ref='serv');
model duree*censure(0)= activite_der statut sect_mod generation/ties=efron
risklimits ;
baseline out=c covariates=b survival=s logsurv=ls/nomean;run;
```

Annexe 3 : Déterminer les interactions à ajouter dans un modèle

Il est possible que l'effet d'une variable varie en fonction des valeurs prises par une des autres variables introduite dans le modèle. Le croisement entre variables aura davantage d'effet que chaque variable séparément.
Pour détecter ces interactions, la méthode la plus fréquemment utilisée est de tester plusieurs interactions dans un modèle et d'analyser les résultats des tests de significativité, les tests de vraisemblance partielle, ou encore à partir du critère d'information d'Akaike et du critère bayesien de Schwartz.
Une fois la ou les variables d'interaction trouvées, il suffit de les inclure dans le modèle avec les variables explicatives : `nom_var1*nom_var2`.

```
proc phreg data=cohorte2008 ;
class sect_mod2(ref='serv') gpage(ref='61+');
model duree*censure(0)= sect_mod2 statutdur1-statutdur3 gpage gpage*sect_mod2
/ ties=efron risklimits;
statutdur1=statut2*(duree <1);
statutdur2=statut2*(1<=duree<4);
statutdur3=statut2*(duree>=4);
hazardratio gpage / at(sect_mod2=all) cl=pl;
hazardratio sect_mod2 / at(gpage=all) cl=pl; run;
```

SAS ne calcule pas automatiquement les risques-ratio des variables qui sont en interaction avec d'autres variables. Il faut donc rajouter l'instruction `hazardratio` qui calcule les risques-ratio pour une variable, en fonction des caractéristiques d'une autre variable, spécifiée dans l'instruction `at(nom_var=all)`. L'option `cl` permet de demander le calcul d'un intervalle de confiance. L'intervalle de confiance peut être celui de Wald (`cl=Wald`) ou celui obtenu à partir du calcul de la vraisemblance (`cl=pl`).

Bibliographie

Afsa C., 1999, «L'allocation de parent isolé : une prestation sous influences. Une analyse de la durée de perception » in Economie & prévision, n°137, p13-31.

Alberti C., Chevret S, Timsit J-F, 2005, « Le modèle de Cox » in SPLF (société de pneumologie de langue française), n°22

Bac C., Gaudemer C., 2010 « Actif au RSI et retraité au régime général», *Zoom sur,* n°41*,* RSI.

Bac C., Gaudemer C., 2012 « Actif au RSI et retraité au Régime général - évolution de cette situation de cumul entre 2008 et 2010 », *Zoom sur,* n°64*,* RSI.

Bringé A., Carré A., Support de cours, analyse biographique, 12-13 décembre 2013

Bringé A., Lelièvre E., Manuel pratique pour l'analyse statistique des biographies : présentation des modèles de durée et utilisation des logiciels SAS, TDA, STATA, Paris, INED, 1998, 189 p. (coll. Méthodes et Savoirs)

Colletaz G., 2012, « Modèles de survie, notes de cours, master 2 ESA »

Courgeau D., Lelièvre E., 2001, « Analyse des biographies », in Casseli G. (dir), Vallin J. (dir), Wunsch G. (dir), Démographie : analyse et synthèse I, La dynamique des populations, Paris, INED, p503-518

Cucherat M., « Interprétation des essais cliniques pour la pratique médicale, Les courbes de survie», http://www.spc.univ-lyon1.fr/polycop/courbes%20de%20survie.htm, août 2009

Dardier A., Gaudemer C., 2014 « Actif au RSI et retraité au général à la fin 2012 », *Zoom sur,* n°82*,* RSI.

Inconnu, « L'approche semi-paramétrique : le modèle de Cox », http://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Survie%20semi_parametrique.pdf

Le Goff JM., cours d'Analyse des biographies, IDUP, 2012.

Lollivier S., « Modèles univariés et modèles de durée sur données individuelles », Méthodologie Statistique, INSEE

Pailhé A., Solaz A., 2012, « Durée et conditions de retour à l'emploi des mères après une naissance », Retraite et société, n°63, p53-75

Paletta M., Survival analysis using the proportional hazards model, SAS Institute, 2009.

Magnac T., Rapoport B., Roger M., 2006, « Fins de carrière et départs à la retraite : l'apport des modèles de durée », Solidarité et santé, n°3, p101-117

Rapoport B., 2009, « En début de carrière, moins d'acquisitions de droits à la retraite pour les jeunes générations », Solidarité et société n°10, p24-41