

CONSEIL D'ORIENTATION DES RETRAITES

Séance plénière du 20 octobre 2010 - 9 h 30

« Elaboration de cas types pour le pilotage du système de retraite »

<i>Document N°5</i>
---------------------

Document de travail, n'engage pas le Conseil
---

**Construction de cas types représentatifs : méthode et limites.**

**Une illustration sur la comparaison des générations 1938 et 1950**

*Anais Le Gougec - DREES*

Octobre 2010

## **Construction de cas types représentatifs : méthode et limites**

*Une illustration sur la comparaison des générations 1938 et 1950*

Anaïs Le Gougec - DREES<sup>1</sup>

### **Résumé**

Dans le cadre d'une réflexion plus générale sur la mise en place d'un modèle de projection des retraites, intermédiaire entre l'approche agrégée des modèles macro et l'approche individualisée de la microsimulation, les travaux présentés ici constituent une réflexion méthodologique sur l'identification de cas types représentatifs à partir de l'Échantillon Interrégimes de Cotisants. Ces premiers travaux se limitent aux hommes nés en 1938 et 1950 qui, avant l'âge de 55 ans, ont validé l'essentiel de leurs trimestres à la CNAV.

Les classifications réalisées sur les générations 1938 et 1950 mettent en évidence, d'une part, l'existence de groupes de carrières homogènes et distincts au sein de chacune des générations, et révèlent d'autre part une apparente persistance de certains de ces groupes entre les générations, mais aussi la disparition et l'apparition d'autres groupes entre les deux générations étudiées.

Comme support à la réflexion sur la construction de cas types représentatifs, les carrières des individus les plus au centre des classes sont extraites.

Ces travaux mettent en particulier en avant plusieurs options pour construire des cas types représentatifs et facilement appréhendables et soulignent la difficulté à faire évoluer sur plusieurs générations ces cas types dont le déroulement et la représentativité se modifient.

---

<sup>1</sup> Ce travail a bénéficié des travaux préliminaires menés par Layla Ricroch (ENSAE).

## Introduction

Prévoir l'évolution du système de retraite et mesurer l'impact de réformes peut se faire à l'aide de différents outils. Trois types de méthodes sont généralement envisagés : les approches macroéconomiques, les outils de microsimulation et les études sur cas types.

Les modèles de projection macroéconomiques sont des cadres comptables qui reposent sur une conception agrégée de l'information en utilisant des individus possédant les caractéristiques moyennes de la population. Ces approches ont l'avantage de la clarté et de la légèreté de traitement. L'hétérogénéité des carrières n'est cependant pas maîtrisée, et la pension moyenne nécessite donc souvent d'être traitée comme une donnée exogène. En effet, le biais dû à l'agrégation étant décuplé par les non-linéarités du mode de calcul des droits à pension, la pension de l'individu moyen ne correspond généralement pas à la pension moyenne. L'approche agrégée diminue ainsi la pertinence des études portant sur la continuité du niveau de vie. En outre, il est impossible de traiter des questions d'équité et de redistribution par le biais de ce type d'analyse.

Les outils de microsimulation, simulant des trajectoires à l'échelle individuelle à l'aide d'une combinaison de règles déterministes, de tirages aléatoires et de fonctions de comportements, permettent en revanche d'examiner dans le détail les questions d'inégalités et de continuité du niveau de vie. Ils sont de plus largement utilisés pour leur grande précision et le haut niveau de détail qu'ils peuvent fournir, en particulier grâce à la prise en compte de toute l'hétérogénéité des carrières. En revanche, ils demandent un traitement très lourd en matière de récolte de données, sont complexes à appréhender d'un point de vue pédagogique et parfois difficiles à maîtriser techniquement en raison du nombre important de paramètres et de dimensions qu'ils font intervenir.

L'objectif des cas types est de maintenir une approche microéconomique, tout en exigeant un travail de synthétisation de l'information qui rend les résultats plus faciles à appréhender. Dans le cadre d'une étude sur le système de retraite, un cas type correspond généralement à un individu dont on connaît les éléments principaux nécessaires au calcul de sa pension, c'est-à-dire son déroulé de carrière ainsi que quelques informations sociodémographiques prises en compte dans l'évaluation des droits à retraite. Des cas types de cette forme permettent notamment de calculer aisément des indicateurs individuels et en particulier d'étudier les niveaux de revenu et leur continuité relative entre la période active et la retraite. Selon l'utilisation que l'on souhaite en faire, les cas types peuvent être théoriques, stylisés ou chercher à être représentatifs, chaque carrière type étant alors affectée d'une pondération rendant compte du nombre d'individus dans la population totale qui connaissent des trajectoires professionnelles comparables.

L'exploitation de cas types de carrière a parfois une vocation purement didactique. Des déroulés de carrières théoriques sont alors choisis pour leur simplicité, hors de réelles considérations de représentativité effective, dans le but d'illustrer clairement les mécanismes législatifs régissant le système de retraite et leurs conséquences concrètes sur les pensions. L'objectif peut être en particulier d'évaluer les conséquences individuelles de la réforme sur certaines carrières archétypales, potentiellement atypiques mais qui peuvent être importantes pour des considérations d'équité ou de justice sociale. Au-delà de leur qualité pédagogique, les cas types théoriques ont en outre l'avantage de requérir un travail de collecte de données particulièrement léger. En revanche, les cas types théoriques ne permettent pas de quantifier les impacts macroéconomiques d'une réforme, puisqu'ils ne sont pas en lien avec les données empiriques, notamment les évolutions du contexte social et économique.

Pour pouvoir réaliser une analyse macroéconomique quantitative à l'aide de cas types, il faut en fait disposer d'un ou de plusieurs cas types tels que le calcul de la moyenne de leurs pensions (en pondérant éventuellement les cas types) permette de retrouver la valeur de la pension moyenne de l'ensemble de la population.

Utiliser un seul cas type intégrant des valeurs moyennes ou médianes observées afin de se rapprocher des carrières réelles, comme par exemple une carrière complète au salaire moyen ne permet pas de retrouver la pension moyenne de la population, du fait des non-linéarités dans le mode de calcul des droits à pension. Ces non-linéarités sont induites par exemple par le dispositif du minimum contributif, les mécanismes d'indexation, le mode de calcul du salaire annuel moyen ou encore les changements de législation impactant de façon différenciée les individus en fonction de leur cohorte ou de leur année de liquidation. Avec la législation actuelle, il est donc impossible de construire un cas type unique permettant à la fois de synthétiser correctement les carrières des affiliés et d'évaluer les droits à pension correspondant à la pension de retraite moyenne.

En outre, la prise en compte des aléas de carrière affectant les droits à pension, tels que des phases de chômage ou de maladie, est complexe dans cette approche simplifiée. En effet, l'introduction de tels épisodes dans la carrière de l'individu type ne peut *a priori* se faire que de façon binaire à l'échelle trimestrielle : toute tentative d'adéquation avec des données macroéconomiques, celles-ci étant par définition des moyennes continues, est donc très délicate. Enfin, le choix de la période à laquelle l'individu est censé faire face à ces événements — en début, milieu ou fin de carrière — n'est pas dépourvu d'impact sur ses droits à pension.

Afin de résoudre ces problèmes, il est donc indispensable de disposer d'un nombre suffisant de cas types captant au mieux l'hétérogénéité des situations.

## **1. Avantages et limites des cas types représentatifs**

### **a. Qu'est-ce qu'un ensemble de cas types représentatifs satisfaisant ?**

L'objectif est d'établir un nombre limité de cas types résumant les différentes caractéristiques ayant trait aux pensions d'une génération. En outre, pouvoir passer à une analyse macroéconomique nécessite d'être capable de quantifier la représentativité réelle de chacun des cas types ainsi définis<sup>2</sup>.

Au contraire de la microsimulation qui conserve la quasi-totalité des données de carrière, l'établissement de cas types nécessite une synthèse, et donc la perte d'une partie de l'information disponible. La démarche la plus intuitive serait de focaliser l'analyse sur le niveau de pension finale. Cette approche n'est pas satisfaisante dès lors que l'on souhaite établir des cas types conservant leur pertinence au fil des réformes envisagées. Dans ce cadre, il est essentiel qu'un individu dont la pension est correctement approchée par un cas type continue d'être représenté de façon satisfaisante par ce même cas type dans le nouveau cadre législatif envisagé par la réforme. Or, si deux individus perçoivent des pensions de niveau comparable dans la législation actuelle, il est possible que ça ne soit plus le cas dans les autres modes de réglementation envisagés : rapprocher les individus en fonction de leur niveau de

---

<sup>2</sup> Des ensembles de cas types ont souvent été déterminés, mais bien souvent leur représentativité réelle n'est pas connue, ce qui entrave toute approche quantitative. Voir par exemple les cas types utilisés pour le rapport Charpin, ou encore ceux de Christel COLIN et Corinne METTE (2003 Impact des différents aléas de carrière sur les retraites : inactivité, chômage, travail à temps partiel et préretraite)

pension n'est donc pas adéquat. En revanche, deux individus connaissant un déroulé de carrière similaire, ayant le même nombre d'enfants et le même sexe, étant nés la même année et relevant en outre des mêmes régimes de retraite se verront attribuer des pensions comparables dans toutes les variantes de réforme envisageables, pourvu qu'elles soient paramétriques<sup>3</sup>. Il est donc nécessaire de bâtir les cas types en se fondant sur ces critères pour qu'ils conservent leur validité au fil des réformes étudiées, permettant ainsi d'avoir une première estimation de leur impact moyen. Cependant, il est à noter que les comportements de départ en retraite des individus évoluent avec les réformes. Or, les cas types ont des carrières rigides, ce qui limite fortement leur pertinence, à moins qu'un module de comportement soit utilisé pour adapter leur fin de carrière.

Les cas types nécessitent en outre d'être construits relativement à une génération donnée : il serait moins pertinent, par exemple, de créer des cas types censés représenter tous les individus partant en retraite une année donnée. En effet, les règles législatives tiennent largement compte de l'année de naissance des affiliés, le mode de calcul de la pension n'étant donc pas le même selon la génération considérée. De plus, les profils de carrière sont liés à un certain nombre de caractéristiques exogènes (recul de l'âge de fin d'études, variation de la conjoncture économique, etc.) qui impactent différemment les générations successives.

Au final, pour une génération donnée, un ensemble de cas types pondérés pourra être considéré comme satisfaisant si, en calculant la moyenne pondérée de leurs pensions, on estime suffisamment précisément la pension moyenne de l'ensemble de la population. Si les cas types sont robustes aux réformes paramétriques, ils devraient alors conduire à correctement prédire l'impact moyen d'une telle modification législative.

On cherche ici à raisonner sans aucun présupposé sur la forme que devraient avoir les cas types censés synthétiser les carrières réelles. Plusieurs questions se posent alors. Tout d'abord, sur quels critères objectifs se fonder pour les déterminer ? Comment ensuite quantifier précisément leur représentativité ? Enfin, si la détermination des cas types se fonde principalement sur le déroulé de carrière, comment prendre en compte les autres critères qui impactent également le niveau des pensions ?

## **b. Comment conserver la dimension temporelle du déroulé de carrière ?**

C'est par le biais du choix de la métrique – laquelle est nécessaire pour déterminer une notion de distance entre les carrières<sup>4</sup>, que l'on a cherché à répondre au mieux à l'objectif essentiel de conserver toute la richesse de l'information sur le déroulé temporel des carrières observées.

Certaines méthodes ont déjà été avancées pour établir des cas types à la représentativité quantifiable en utilisant tout le déroulé temporel de la carrière : ainsi Malik Koubi (Élaboration des carrières-types représentatives des carrières complètes des salariés du secteur privé nés en 1948 (2005)) propose de calculer la distance entre les carrières comme l'écart

---

<sup>3</sup> Actuellement, le mode de calcul des droits à retraite se fonde sur un certain nombre de paramètres (âge minimal de départ à la retraite, durée d'assurance requise...). Les réformes « paramétriques » visent à modifier la valeur de ces paramètres, à l'inverse des réformes « systémiques » qui envisagent, elles, une refonte totale du mode de calcul des pensions : l'instauration d'un régime de retraite à point est par exemple à compter parmi les réformes systémiques.

<sup>4</sup> Le nombre de carrières assimilées à un cas type du fait de leur proximité dans cette métrique mesure alors le degré de représentativité du cas type.

quadratique entre les deux chroniques de salaire, en simplifiant préalablement ces dernières par projection sur l'espace des polynômes d'un degré fixé. Cette méthode nécessite cependant de se restreindre aux carrières complètes, puisque l'intégralité de l'information concernant la carrière est assimilée à la chronique de salaire.

Karine Briard (Profils types des salariés du secteur privé : approche par une classification des carrières, *Économie et Prévision* 2007/4 n°180-181 pages 59-85), propose une technique qui pallie ce problème en permettant aussi de traiter les carrières connaissant des aléas. Ici, la chronique décrivant la carrière de l'individu est donnée année après année par une variable qualitative. Ainsi, chaque année, les individus se voient attribuer une modalité en fonction de leur situation, d'une part selon leur décile de salaire, s'ils en ont perçu un, mais aussi, d'autre part, selon qu'ils ont connu ou non des épisodes de chômage, d'invalidité, de maternité, de service militaire ou d'inactivité. La chronologie ainsi obtenue résume la carrière de l'individu, en étant moins précise quant aux revenus, décrits qualitativement et non quantitativement, mais autorisant la prise en compte des aléas de carrière. Les carrières alors représentées sous formes de séquences sont ensuite soumises à une analyse factorielle des correspondances multiples (AFCM) puis à classification (algorithme non-hiérarchique des k-moyennes avec sélection des formes stables suivi d'une classification ascendante hiérarchique).

Ici, après avoir codé les carrières selon un procédé très proche de celui employé par Karine Briard, on applique une méthode d'appariement optimal, particulièrement adaptée au traitement des séquences, afin de déterminer une distance entre les individus. On effectue ensuite une classification ascendante hiérarchique afin de laisser les carrières se regrouper spontanément en fonction de leur similitude, ce qui permet de déterminer les profils professionnels dominants. On extrait ensuite un cas type représentatif de chacun des sous-groupes révélés par la classification. Le détail de la méthodologie est exposé dans la partie 2.

### **c. Le problème du maintien de la pertinence des cas types au fil des générations**

Une fois la méthode mise en place, la représentativité associée à chaque carrière-type n'est valable que pour la cohorte considérée. La pertinence de l'analyse sera d'autant plus grande si les cas types ne sont pas seulement représentatifs d'une génération ancienne.

Or, reprendre successivement le même exercice dans l'objectif de reconstruire un jeu de cas types adapté spécifiquement à chacune des générations soulève deux difficultés.

Tout d'abord, l'exercice est déjà complexe à effectuer pour une génération (voir infra). Répéter le même processus pour chaque cohorte serait donc extrêmement long et les avantages en termes de simplicité et de pédagogie de l'utilisation de cas types seraient très largement diminués.

À cela s'ajoute le fait que les générations jeunes n'ont effectué qu'une fraction de leur carrière : l'information est donc de plus en plus lacunaire au fil des cohortes. Il serait donc nécessaire de prolonger les carrières ou de ne mener l'exercice que sur les débuts de carrière observés (voir Briard, 2006, 2007).

Savoir si les cas types établis sur des générations antérieures peuvent être utilisés pour représenter des générations plus récentes est donc essentiel. Or, adapter le jeu de cas types au fil des générations implique de répondre à un grand nombre de questions. Si certaines configurations de trajectoire professionnelle persistent vraisemblablement au fil des générations, comment leur représentativité évolue-t-elle ? Quels profils professionnels prennent de l'ampleur ? Quels sont ceux qui se marginalisent ? Observe-t-on en outre

l'apparition de carrières types fondamentalement nouvelles, ou au contraire la disparition de certaines trajectoires pourtant emblématiques de générations antérieures ? Plutôt que de disparaître, certains cas types se contentent-ils de se déformer au fil des générations ?

Si les différences d'une génération à l'autre sont notables, comment alors faire évoluer les cas types des générations antérieures afin de pouvoir mener des analyses pertinentes en projection ?

Pour apporter des éléments de réflexions sur ces différents points, nous effectuons dans la suite le travail de construction de cas types représentatifs sur deux générations distinctes, afin d'en étudier les évolutions.

## **2. Présentation d'une méthode pour construire des cas types représentatifs sur une génération**

Afin de mettre en évidence les difficultés qui peuvent se présenter dans le cadre d'une utilisation en projection des cas types, et notamment de savoir s'il est légitime d'utiliser des cas types estimés sur la base de générations anciennes pour décrire les cohortes plus récentes, on compare les classifications établies sur la génération 1938 et la génération 1950 en utilisant exactement la même méthode dans les deux cas, afin d'observer le degré de ressemblance entre les deux systèmes de cas types retenus.

### **a. Carrière et retraite : le choix des variables d'intérêt**

L'objectif étant de rassembler les individus considérés comme proches au sens des critères pris en compte dans le calcul des droits à retraite, le déroulé de carrière – incluant notamment les épisodes de chômage ou de maladie — est une donnée essentielle. L'Échantillon Interrégimes de Cotisants<sup>5</sup> (EIC) regroupe des informations très complètes sur l'historique de la carrière de près de quatorze mille assurés de la génération 1938 et de vingt-quatre mille de la génération 1950, tous régimes confondus. Ainsi, on dispose de données concernant le montant du salaire annuel dans chacune des caisses, mais aussi le nombre de trimestres cotisés ou validés au titre du chômage ou de la maladie, l'affiliation à l'assurance vieillesse des parents au foyer, etc.. L'intérêt majeur de cette source de données réside d'une part dans son aspect multi-régime, et d'autre part dans ce qu'elle permet de retracer la chronique complète de la carrière des individus de l'échantillon.

Il est nécessaire de prendre également en compte d'autres caractéristiques socio-démographiques, comme le nombre d'enfants élevés pour les femmes, qui impactent parfois fortement les droits à pension. L'Échantillon Interrégimes de Retraités de 2004, qui est apparié à l'EIC, permet notamment de rassembler toutes les données sociodémographiques nécessaires au calcul des pensions – ainsi que le montant des pensions perçues dans les différentes caisses.

Cependant, les données de l'EIR ne sont pas disponibles pour la génération 1950<sup>6</sup>, qui ne concerne que les individus ayant déjà liquidé leur retraite. L'objectif étant ici de comparer les résultats d'une méthode appliquée de façon rigoureusement identique à deux générations, les données de l'EIR n'ont été utilisées qu'à titre descriptif, et uniquement pour la génération 1938.

---

<sup>5</sup> Voir annexe 1 pour une présentation plus détaillée de l'EIR et de l'EIC.

<sup>6</sup> Voir annexe 1 pour plus de détails sur le champ de l'EIR.

Ce même souci de comparaison a abouti à devoir tronquer les carrières de la génération 1938 à 55 ans, afin que la méthode soit appliquée sur la base de la même durée de carrière observée pour les deux générations.

## **b. Résumé de la carrière par une succession de variables qualitatives**

Le champ d'étude a été restreint aux monopensionnés de la CNAV. Les explications concernant ce choix, et le détail de la méthode de leur identification pour la génération 1950 sont exposés dans l'annexe 2. Chaque année, l'activité professionnelle des individus est résumée par une variable qualitative prenant en compte principalement le tercile de salaire, mais aussi la présence d'épisodes de chômage, d'AVPF ou de maladie.

Neuf modalités ont ainsi été retenues :

1. Inactivité : pas d'activité relevée au cours de l'année. Cela peut correspondre à une réelle inactivité, à une période où l'individu n'est pas en France ou à une activité effectuée au sein d'une caisse mal renseignée.
2. Chômage : au moins un trimestre a été validé à la CNAV au titre du chômage au cours de l'année, à l'exclusion des cas où un salaire porté au compte au titre de l'AVPF est relevé.
3. Autres trimestres gratuits : au moins un trimestre gratuit, autre qu'au titre du chômage et généralement au titre de la maladie, a été relevé au cours de l'année, à l'exclusion des cas où un salaire porté au compte au titre de l'AVPF est relevé.
4. AVPF : un salaire porté au compte au titre de l'AVPF est relevé au cours de l'année
5. Une activité salariale est détectée à la CNAV sans épisode de chômage, AVPF ou autres trimestres gratuits, mais elle permet une cotisation globale inférieure à quatre trimestres sur l'année.
6. Les quatre dernières modalités sont ensuite attribuées en fonction du décile de salaire, dans le cas où l'activité salariale a conduit à la cotisation de quatre trimestres sur l'année à la CNAV, sans épisode de chômage, AVPF ou autres trimestres gratuits :
  - Premier tercile de salaire de l'année
  - Second tercile de salaire de l'année
  - Troisième tercile de salaire hors dernier décile de l'année
  - Dernier décile de salaire de l'année

Afin d'obtenir des individus représentés par des cas types ayant des montants de pensions similaires, il est en effet apparu nécessaire de discriminer entre les différents niveaux de revenu. Pour limiter le nombre de modalités, les niveaux de salaires ont été tout d'abord découpés en tercile. En outre, le dernier décile a été isolé afin de distinguer les individus dont le niveau de retraite risque d'être fortement amplifié par les pensions complémentaires.

Les quantiles de revenus sont calculés année après année afin de contrôler l'augmentation graduelle globale des niveaux de salaire au cours de la carrière ; sachant que l'on travaille sur une génération, cela revient à fixer les quantiles pour chaque âge.

En effet, étant donné que des impératifs de lisibilité contraignent à se limiter à un nombre réduit de modalités, utiliser des quantiles calculés sur la totalité des revenus perçus par la cohorte ne permettrait pas de discriminer finement les évolutions salariales.

La carrière d'un individu est ainsi résumée par une succession de 42 modalités, représentant à chaque âge la situation de l'individu, de 14 à 55 ans.

### c. Calcul de la distance entre les individus

Une méthode d'appariement optimal est ensuite mise en œuvre pour mesurer la distance entre deux carrières. Ce type de méthode, visant à évaluer le degré de similarité entre deux séquences, est né des recherches en informatique effectuées à partir des années 50, mais a été très largement utilisé et développé récemment pour permettre le séquençage du génome. La carrière d'un individu pouvant être conçue comme une succession d'états relatifs au marché du travail, ces méthodes sont en effet particulièrement adaptées pour établir une distance entre les carrières.

Le principe général en est le suivant : pour calculer la distance entre deux individus, on compte le nombre de transformations qu'il faut effectuer pour passer de la carrière du premier individu à la carrière du second. À chaque opération envisageable (substitution, insertion, et suppression d'un caractère), on associe un coût. On détermine ensuite la suite d'opérations dont le coût total est minimal pour passer de l'une à l'autre séquence : ce coût total correspond à la distance entre les deux chroniques.

Schématiquement, admettons que l'on cherche à évaluer la distance entre ces deux séquences :

	Période 1	Période 2	Période 3	Période 4
<b>Séquence 1 :</b>	Salaire	Salaire	Salaire	Chômage
<b>Séquence 2 :</b>	Inactivité	Salaire	Salaire	Salaire

Il existe de nombreuses combinaisons d'insertions, suppressions et substitution permettant de passer de la première à la deuxième séquence. Celle comportant le moins d'opérations est une insertion d'inactivité en période 1, suivie d'une suppression du chômage en fin de séquence :

	Période 1	Période 2	Période 3	Période 4	
<b>Séquence 1 :</b>	<b>Inactivité</b>	Salaire	Salaire	Salaire	<b>Chômage</b>
<b>Séquence 2 :</b>	Inactivité	Salaire	Salaire	Salaire	

En revanche, pour comparer les deux séquences suivantes :

	Période 1	Période 2	Période 3	Période 4
<b>Séquence 1 :</b>	Salaire	Salaire	Salaire	Chômage
<b>Séquence 2 :</b>	Inactivité	Salaire	Salaire	Chômage

La méthode comportant le moins d'opération sera alors une substitution de salaire en inactivité en période 1 :

	Période 1	Période 2	Période 3	Période 4
<b>Séquence 1 :</b>	Salaire	Salaire	Salaire	Chômage
	⇕			
<b>Séquence 2 :</b>	<b>Inactivité</b>	Salaire	Salaire	Chômage

Cette méthode présente tout d'abord l'avantage de permettre de traiter l'information résumée sous une forme qualitative, et donc de ne pas avoir à exclure les carrières même très morcelées en termes de perception de salaire, ce qui est essentiel si l'on souhaite comparer les cas types de deux cohortes différentes.

Ensuite, elle permet de traiter les carrières en incluant leur dimension temporelle. Ainsi, une carrière ascendante sera estimée dissemblable d'une carrière descendante, même si ces deux carrières présentent la même fréquence d'apparition pour chaque modalité. Cette propriété est particulièrement importante pour la précision du calcul des droits à retraite, notamment du fait des indexations.

Le calcul de la distance repose largement sur les coûts associés aux différentes opérations. Ceux-ci peuvent être attribués de différentes façons<sup>7</sup>. Ici, le coût de substitution entre deux modalités est calculé en fonction de la fréquence à laquelle on observe un côtoiement entre les deux modalités dans la carrière des individus<sup>8</sup>.

#### **d. Classification : rassembler les individus les plus semblables en groupes les plus distincts possibles**

Afin de regrouper les carrières les plus semblables, on procède ensuite à une classification ascendante hiérarchique. Cette méthode permet de découper l'ensemble de la population en sous-catégories, en rassemblant les individus les plus semblables au sein de groupes les plus distincts possibles. Cela correspond bien à l'idée des cas types représentatifs : chaque cas type représentera sa classe, il faut donc qu'au sein d'un groupe les carrières soient le plus homogène possible. L'ensemble des cas type doit en revanche représenter l'hétérogénéité de la population.

Le principe de la classification est le suivant. A la première étape, on regroupe les deux individus aux carrières les plus proches, au sens de la distance définie précédemment. On itère le processus et, de proche en proche, les regroupements sont effectués de telle façon que la variance à l'intérieur des classes est minimale, et la variance entre les classes est maximale. Au fur et à mesure qu'on fusionne les groupes par proximité, on cherche à minimiser la perte d'information.

Pour choisir le nombre de classes à retenir, c'est-à-dire le moment où l'on arrête de fusionner des groupes, il faut tout d'abord déterminer une fourchette globale qui semble adaptée à l'objectif fixé. Ici par exemple, une dizaine de classes semble *a priori* suffisamment élevé pour être relativement précis, et assez restreint pour permettre une certaine lisibilité. Le nombre exact de classes autour de cet objectif est ensuite suggéré par le dendrogramme : ce type de graphique détaille précisément le niveau de perte d'information à chaque fusion entre deux groupes. Les regroupements sont donc stoppés avant que deux groupes trop dissemblables ne soient fusionnés.

Pour prendre en compte les critères autres que ceux liés directement à la carrière<sup>9</sup>, il est par ailleurs possible d'appliquer la méthode séparément sur des sous-populations, distinguées en fonction des critères sociodémographiques pris en compte dans la législation. On est ainsi assuré que les classes constituées seront totalement homogènes au vu du calcul des droits à retraite. Il est toutefois également possible de réaliser la classification sur l'ensemble de la population et de reproduire les cas types obtenus en fonction des différentes modalités de ces critères (en recalculant la pondération en fonction du nombre d'individus correspondant à

---

<sup>7</sup> Des premiers travaux ont montré que les résultats obtenus étaient peu sensibles à la détermination des coûts.

<sup>8</sup> Voir annexe 3 pour plus de détails sur le calcul des coûts de substitution.

<sup>9</sup> Ce travail n'est cependant pas effectué ici, ces données n'étant disponibles que pour la génération 1938. On rappelle que seul le sexe a pu être pris en compte.

chaque modalité du critère dans la classe<sup>10</sup>). Ces deux options (découpage de la population avant la classification ou reproduction des cas types) conduisent à retenir approximativement le même nombre de cas types. Si les carrières ne dépendent pas des critères sociodémographiques considérés, la reproduction des cas types est plus simple. Dans le cas contraire, faire la classification séparément sur chaque sous-population peut permettre de gagner en précision.

Ainsi, parmi les critères à prendre en compte on peut citer par exemple le sexe et le nombre d'enfants. En effet, ces critères ont un impact souvent important sur le déroulement de la carrière, et même un impact direct sur le calcul des droits à pension du fait de la majoration de la durée d'assurance. Il semble donc pertinent de les utiliser comme une variable de pré-découpage. Dans l'exemple qui suit, nous nous intéresserons uniquement aux hommes.

### **e. Choix d'un individu représentatif au sein de chaque classe**

Une fois construites les différentes classes, il faut définir pour chaque classe un individu type qui la représentera.

Déterminer un individu fictif à partir des caractéristiques de la classe poserait plusieurs difficultés.

Tout d'abord, certaines classes sont caractérisées par la fréquence relativement élevée d'une modalité au contraire très rare dans le reste de la population, sans pour autant que cette modalité ne soit jamais la plus fréquente dans la classe. Construire le cas type comme un individu ayant chaque année la modalité la plus fréquente de sa classe n'est donc pas pertinent. Par ailleurs, les carrières peuvent être représentées trop grossièrement par ce procédé : ainsi, l'individu représentatif de la classe 5 de la génération 1938 (voir infra) sera considéré comme étant resté inactif pendant l'intégralité de sa carrière, alors que tous les individus qu'il est supposé représenter perçoivent des pensions non nulles, bien que faibles.

Une autre possibilité est alors de créer un individu pour lequel le nombre de chacune des modalités serait calibré sur leur fréquence totale au sein de la classe. Cependant, la question de l'ordre dans lequel devront être classées ces modalités n'est pas aisée à déterminer ; or, l'impact d'un tel choix sur le montant de la pension finale n'est pas négligeable. Il est éventuellement possible de ne pas chercher à temporaliser les informations n'ayant pas trait au niveau de salaire. Ainsi, le modèle Eclips<sup>11</sup> construit des carrières « à trous » en conservant uniquement les années où les taux de report de salaire ont été les plus élevés et la chronique des trimestres validés au titre du chômage, complétées par le nombre total de trimestres validés au titre des autres périodes telles que la maladie ou l'invalidité, sans chercher à situer temporellement ces dernières. Ces cas types reconstruits sont moins détaillés qu'un déroulé de carrière temporalisé, et la question de savoir comment les faire évoluer en projection n'est pas sans poser de questions.

Une approche différente, qui est ici retenue, est au contraire de choisir un individu existant au sein de la classe. Ainsi, l'intégralité de l'historique professionnel, incluant le détail du niveau de salaire notamment, sera disponible pour calculer tous les indicateurs envisageables.

---

<sup>10</sup> Par exemple, si les hommes et les femmes sont traités conjointement dans l'analyse, on mène ensuite une étude plus approfondie pour révéler la composition des classes en termes de sexe. Si une classe est composée à 80% de femmes et à 20% d'hommes, on duplique le cas type pondéré afin d'en créer un de sexe féminin et un de sexe masculin ; on modifie ensuite la pondération des deux cas types ainsi créés en tenant compte de leurs caractéristiques.

<sup>11</sup> Briard K. (2009), « Un modèle de carrières types dynamiques pondérées pour le Régime général d'assurance vieillesse : une application aux conséquences de la réforme de 2003 », *Économie et prévision*, n° 187.

L'individu le plus représentatif de la classe est *a priori* celui qui est situé le plus au milieu de la classe. Il est sélectionné en sommant ses distances à chacun des autres individus de sa classe : l'élément le plus « central », c'est-à-dire celui pour lequel cette somme est minimale, est ainsi retenu. Cette méthode ne donne pourtant pas toujours des résultats satisfaisants car elle conserve la singularité des individus, ce qui peut être problématique. Une autre solution, non mise en œuvre ici, serait de choisir « à la main », parmi les individus les plus centraux, celui qui ne présente pas de particularités gênantes pour pouvoir représenter sa classe de façon satisfaisante ; cette méthode incorpore cependant une part d'arbitraire dans la méthode générale. En outre, faire évoluer une carrière réelle en projection est particulièrement ardu. Ce point est détaillé plus loin.

### **3. Réflexion sur la question de l'évolution des cas types au fil des générations, à partir des générations 1938 et 1950**

On compare maintenant les résultats obtenus sur les deux générations, afin de présenter les difficultés qui peuvent apparaître.

L'interprétation des résultats de la comparaison des classes entre les deux générations doit toutefois être prudente car deux partitions d'un continuum homogène de données ne variant qu'à la marge peuvent être très différentes si les scissions ne s'effectuent pas aux mêmes endroits. Les deux classifications permettent au moins de voir en particulier s'il existe des groupes homogènes et distincts dans chacune des deux générations étudiées et si, pour certains de ces groupes, les caractéristiques sont proches entre les deux générations.

#### **a. Classifications : l'exemple des hommes monopensionnés CNAV**

Le détail des résultats de la classification, ainsi que des statistiques concernant des caractéristiques complémentaires<sup>12</sup> (statut de cadre, pays de naissance, montant de la pension, etc.), sont présentés dans l'annexe 4. La classification de la génération 1938 aboutit à la création de 9 classes, tandis que pour la génération 1950 seulement 8 classes se distinguent.<sup>13</sup>

Afin d'illustrer les caractéristiques de la carrière des individus de chaque classe, on représente la fréquence d'apparition de chacune des neuf modalités, de l'âge de 14 ans jusqu'à 55 ans. Les rapprochements entre classes se font d'abord visuellement, l'établissement d'un critère objectif déterminant la distance entre les classes des deux générations n'étant pas aisé à concevoir. On s'attache principalement à la description de l'ensemble de la classe ; les cas types sélectionnés ne sont présentés qu'à titre indicatif, la méthode retenue ne permettant pas toujours d'obtenir des résultats satisfaisants.

---

<sup>12</sup> Disponibles uniquement pour la génération 1938.

<sup>13</sup> Pour le choix exact du nombre de classes, se reporter au troisième paragraphe de la section 2.d.

On obtient les résultats suivants, en tenant compte du pic d'inactivité de la génération 1938 autour de la vingtaine, liée au service militaire<sup>14</sup> :

	Génération 1938	Génération 1950
<b>Comparables</b>	<b>Classe 1</b> : Revenus les plus élevés, début de carrière tardif <sup>15</sup> <b>Classe 2</b> : Revenus moyens, début de carrière précoce <b>Classe 3</b> : Revenus moyens et élevés, début de carrière précoce <b>Classe 4</b> : Revenu élevés, début de carrière précoce	
<b>Comparables mais déformations importantes</b>	<b>Classe 5</b> : Carrière morcelée, marquée par l'inactivité/le chômage <b>Classe 6</b> : Revenus faibles, début de carrière précoce <b>Classe 7</b> : Revenus moyens et faibles, début de carrière précoce	
<b>Disparaît</b>	<b>Classe 8</b> : Revenus faibles, début tardif, chômage <b>Classe 9</b> : Revenus disparates, chômage important	
<b>Apparaît</b>		<b>Classe 8</b> : Revenus élevés et très élevés, début tardif, chômage non négligeable

**Tableau 1 : Comparaison des classes dans les générations 1938 et 1950**

*Note de lecture* : les revenus « les plus élevés » correspondent au dernier décile, les revenus « élevés » au troisième tercile hors dernier décile, les revenus « moyens » au deuxième tercile, les revenus « faibles » au premier tercile.

#### a) Des classes *a priori* comparables mais qui se déforment

L'apparente proximité entre les carrières des classes 1 à 7 de la cohorte de 1938 et de 1950 ne permet pas de contredire en premier lieu la possibilité de correctement décrire les parcours professionnels de générations récentes à l'aide de cas types établis sur des générations plus anciennes.

<sup>14</sup> La durée du service militaire a été élevée à 30 mois durant la guerre d'Algérie. Or la génération 1938 est âgée d'une vingtaine d'années autour des années 1960, elle est donc largement concernée. En 1970 la durée du service militaire était en revanche de 1 an.

<sup>15</sup> Dans un souci de clarté, des dénominations ont été attribuées aux classes. Cependant, il s'agit d'un résumé très rapide, une analyse précise de chaque classe étant nécessaire pour mieux les qualifier.



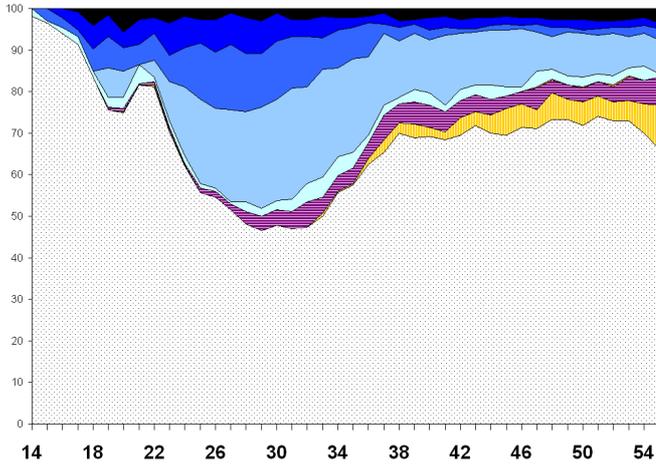




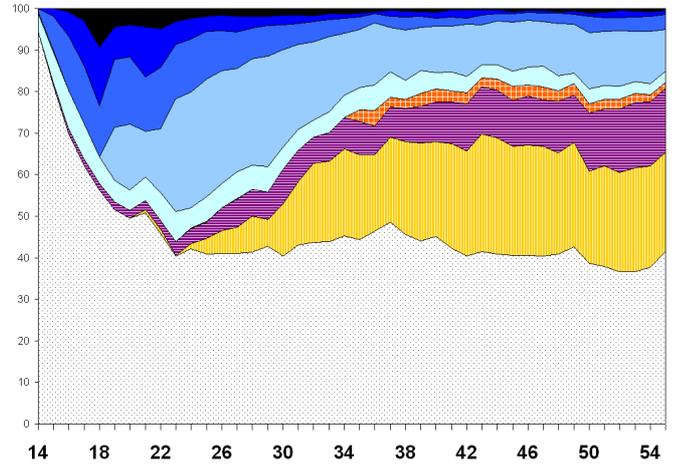


## Classe 5 Carrière morcelée, marquée par l'inactivité/le chômage

1938



1950



### Comparaison des deux individus représentatifs sélectionnés :

	Age	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55																																																
<b>CLASSE 5</b>																																																																																											
1938		□ Inactivité										■ Salaire 1er tercile										■ Salaire 2nd tercile										■ Salaire 3 tercile hors décile										■ Salaire dernier décile																																																	
1950		□ Inactivité										■ Salaire 1er tercile										■ Salaire 2nd tercile										■ Salaire 3 tercile hors décile										■ Salaire dernier décile										■ Chômage										■ AVPF au cours de l'année										■ Autres trim gratuits										□ Travail moins de 4 trim									





En revanche, la représentativité des classes évolue entre les deux générations comparées. Il est cependant à noter que la comparaison des poids entre deux classes d'une génération différente est à prendre avec précaution : les classifications ayant été réalisées de façon indépendante<sup>16</sup>.

Classe	Descriptif	Part dans les monopensionnés CNAV	
		1938	1950
1	Revenus très élevés, début tardif	6 %	6 %
2	Revenus moyens, début précoce	16 %	18 %
3	Revenus moyens et hauts, début précoce	6 %	8 %
4	Revenus hauts, début précoce	21 %	9 %
5	Carrière morcelée, chômage	9 %	13 %
6	Revenu faible, début précoce	9 %	21 %
7	Revenus moyens et élevés, début précoce	10 %	15 %

**Tableau 2 : Poids des classes 1 à 7 dans les générations 1938 et 1950**

On observe notamment une diminution importante du poids de la classe 4, qui doit être relativisée compte tenu de sa proximité avec la classe 8 pour la génération 1950 (voir *infra*). Conserver les cas types de la génération 1938 pour représenter les carrières des générations futures nécessite donc de disposer d'hypothèses d'évolution des « poids de représentativité » de chacun de ces cas types au fil des générations, ce qui est un exercice délicat.

De plus, il existe des différences parfois notables d'une génération à l'autre entre deux classes paraissant comparables à première vue. Ainsi, pour les classes 5, 6 et 7, on remarque l'apparition parfois massive d'épisodes de chômage pour la génération 1950, alors que ceux-ci étaient relativement mineurs pour la génération 1938 ; ces phases de chômage apparaissent en outre de façon plus précoce dans la carrière des individus de la génération 1950. On note aussi une diminution du niveau de revenu pour la classe 7. On remarque enfin globalement un léger recul de l'âge de début de carrière pour toutes les classes.

Utiliser les cas types déterminés à partir de la génération 1938 pour décrire les situations professionnelles les plus représentatives des générations suivantes nécessite donc non seulement de savoir faire évoluer leurs poids, mais aussi d'en modifier les profils. Des évolutions structurelles comme l'augmentation de l'activité féminine, la hausse de l'âge de fin d'études ou le développement du chômage impliquent nécessairement de faire évoluer les cas types.

Prenons l'exemple du chômage, qui affecte plus fortement les générations récentes que la génération 1938. Pour prendre en compte ces nouvelles réalités professionnelles, deux approches sont *a priori* possibles :

- Tout d'abord, il est possible d'augmenter les pondérations des cas types qui connaissent du chômage. Ceci présenterait l'avantage de conserver les mêmes cas types, mais revient à considérer que certains profils de carrières, touchés par le chômage, sont de plus en plus nombreux alors que les autres profils restent inchangés. Ceci ne correspond pas nécessairement à la réalité. De plus, la question de l'évolution des pondérations n'est pas simple à résoudre. Le plus pertinent serait d'utiliser des données macroéconomiques

<sup>16</sup> Une autre méthode de comparaison des poids aurait consisté à affecter les carrières d'une génération (1950) aux groupes identifiés sur l'autre génération (1938) par analyse discriminante, comme dans les travaux de Karine Briard (Profils types des salariés du secteur privé : approche par une classification des carrières, Économie et Prévision 2007/4).

(observées ou en projection) mais les cibles à atteindre (taux de chômage annuel par tranche d'âge) étant plus nombreuses que le nombre de cas types, il est impossible de se caler correctement. En outre, le calage des poids pourrait conduire à une déformation mal contrôlée de la population sur d'autres critères.

- Il est également possible d'introduire des nouvelles périodes de chômage dans les carrières types retenues. Ceci permet de prendre en compte l'apparition du chômage en début de carrière par exemple. Cependant, si la fréquence des épisodes de chômage doit augmenter pour un cas type, dans quelles proportions cette augmentation doit-elle être prise en compte, et au détriment de quels autres types d'épisodes professionnels ? On ne dispose d'aucunes données projetées, concernant de façon isolée les sous-populations révélées par la classification, puisque celles-ci sont propres à l'exercice. Il faut donc calibrer la déformation de chacune des carrières types en se fondant uniquement sur les données macroéconomiques globales. On peut ainsi identifier des cas types plus exposés au risque de chômage et positionner les périodes de chômage en fonction de règles simples<sup>17</sup>. Il s'agit alors de choix plus ou moins arbitraires et on s'éloigne de l'idée de cas type réel.
- Enfin, il est possible de combiner les deux approches ; cependant, il n'est pas aisé de déterminer en pratique dans quelle mesure l'une est plus adaptée que l'autre pour rendre compte d'une évolution. Les problèmes liés à chaque méthode demeurent irrésolus et les problèmes de redondance entre les deux méthodes sont complexes à appréhender.

### **b) Des classes qui disparaissent ?**

Faire évoluer les cas types soulève donc de nombreuses difficultés. La comparaison des carrières types de la génération 1938 et de la génération 1950 montre de plus que faire évoluer à la marge les cas types ou en modifier le poids n'est pas suffisant. En effet, certains profils de carrière de la génération 1938 ne se retrouvent pas dans celle de 1950, tandis que de nouvelles configurations de carrières apparaissent.

Ainsi, pour les classes 8 et 9 de la génération 1938, on ne trouve pas d'équivalent évident au sein de la génération 1950. Plusieurs explications à ces apparentes disparitions sont envisageables.

Tout d'abord, les profils professionnels qui constituaient ces groupes ont pu effectivement se marginaliser au cours du temps. Ils peuvent également être désormais assimilés à diverses réalités professionnelles plus pertinentes pour la génération 1950, et donc ne plus constituer un groupe représentatif en soi. Cette explication pourrait notamment être appropriée pour la classe 9, qui rassemble des individus de tous niveaux de revenus ayant apparemment pour seule caractéristique commune de voir apparaître fréquemment des épisodes de chômage au cours de leur carrière, ce qui, par la rareté de ce type d'épisodes au sein de la génération 1938, était relativement discriminant. La génération 1950 étant en revanche plus largement touchée par le chômage, ces individus antérieurement considérés comme atypiques auront pu alors être plus aisément rapprochés d'autres individus aux évolutions de revenus comparables.

Il est en outre possible, notamment pour la classe 8, que le défaut de renseignement de début de carrière pour la génération 1938 soit responsable de la disparition artificielle de cette classe au sein de la génération 1950, qui voit, elle, ses débuts de carrière plutôt bien renseignés : elle

---

<sup>17</sup> Méthode utilisée dans le modèle Eclipss, par exemple, partant du principe que les classes comportant initialement le plus grand nombre de chômeurs sont des classes « fragiles ». Ainsi, en cas d'augmentation du chômage, ces classes sont supposées plus amplement touchées que les autres. Un coefficient, affecté à chaque classe en fonction de sa « fragilité », permet alors de répartir l'augmentation entre les classes.







## **Conclusion**

La méthode proposée permet de regrouper les individus d'une génération en fonction de leur carrière et de construire ainsi un nombre limité de cas types qui résument bien l'hétérogénéité des carrières pour cette génération. Cette analyse est très instructive car elle permet de caractériser de façon synthétique les réalités professionnelles d'une génération. Elle peut notamment permettre de décrire de façon pédagogique les impacts différenciés d'une réforme sur les différents types de carrières. Cependant, comme le montre la comparaison des générations 1938 et 1950, les cas types doivent évoluer afin de rester pertinents ; faire vieillir ces cas types est néanmoins un exercice complexe.

## **ANNEXE 1**

### **EIR, EIC**

On utilise ici les données de l'EIC de 2005 pour les deux générations, ainsi que celles de l'EIR de 2004 pour la génération 1938 – la génération 1950 étant âgée de 54 ans en 2004, l'EIR ne comporte quasiment aucune information sur cette génération.

#### **1. EIC**

L'EIC 2005 est la seconde vague de collecte de données rassemblant toutes les informations disponibles sur les droits à pension acquis au 31 décembre 2005, pour chacun des individus d'un large échantillon. La mise à jour de l'EIC se fait tous les 4 ans : la campagne de l'EIC, actualisant les données au 31 décembre 2009, a ainsi été engagée en 2010 et la collecte aura lieu en 2011. Les informations collectées lors des extractions ultérieures permettront de suivre l'évolution de la situation des cotisants.

##### **a. Sélection de l'échantillon**

L'échantillon EIC 2005 a été conçu pour représenter les personnes âgées de 31 ans à 71 ans au 31 décembre 2005. Le taux de sondage pour l'EIC est d'une génération sur quatre née entre 1934 et 1974 et de 2,68 % par génération tirée (les 10 premiers jours du mois d'octobre comptent pour 1/12,5 des naissances d'une année). Cet échantillon a été tiré dans le RNIPP – Répertoire National d'Identification des Personnes Physiques (y compris dans la section hors métropole). Cependant, pour des raisons techniques lors de la constitution de l'échantillon, les individus décédés entre le 31 décembre 2005 et 2007, soit environ 1 300 individus, ne figurent pas dans l'échantillon de l'EIC 2005.

L'EIC se présente sous la forme d'un panel : les individus tirés dans un EIC seront suivis dans les EIC suivants afin de pouvoir observer l'évolution de leurs droits acquis entre deux EIC successifs (acquisition de nouveaux droits ou mise à jour d'informations au sein des régimes). Enfin, l'EIC et l'EIR sont appariés afin de pouvoir comparer les droits acquis estimés à partir du dernier EIC et les pensions versées effectivement par chaque régime de retraite dès lors qu'un individu liquide ses droits à retraite. De même, cette articulation permet de mieux évaluer les montants de retraites perçus.

##### **b. Champ et période de référence**

Le champ de l'EIC comprend tous les individus de l'échantillon qui sont affiliés ou ont été affiliés (y compris les personnes déjà parties à la retraite) pour des droits directs, au moins une fois au cours de leur carrière à un des régimes de retraite partenaires de l'opération.

Les droits à retraite issus des régimes supplémentaires facultatifs sont exclus du champ de l'enquête, ainsi que les revenus provenant de l'épargne individuelle volontaire.

Dans son principe, l'EIC recueille auprès de l'ensemble des régimes, des données permettant le calcul des droits à pension acquis pour un échantillon de cotisants. Pour pallier l'absence de système de gestion en 2005 de la caisse de retraite des fonctionnaires de l'État, les données pour cette catégorie de population ont été reconstituées à partir d'informations provenant de l'INSEE (le panel DADS), de la Comptabilité Publique (via l'INSEE), et du Ministère de la Défense (Observatoire Social de la Défense). Par ailleurs, pour tous les individus, des données ont également été recueillies auprès de Pôle Emploi afin de disposer de certaines périodes assimilées ouvrant droits à retraite (chômage, préretraite).

La période de référence des informations sur les droits accumulés s'arrête au 31 décembre 2005 pour l'EIC 2005 et porte sur l'ensemble des droits acquis jusqu'à cette date. De la même façon, seuls les individus affiliés à un régime de retraite avant le 31 décembre 2005 figurent dans l'EIC 2005.

## **2. EIR**

### **a. Champ**

L'Échantillon Interrégimes de Retraités (EIR) est actualisé tous les quatre ans. L'EIR collecté en 2005 est la cinquième enquête menée : la première a eu lieu en 1988 ; l'expérience a ensuite été reconduite en 1993, 1997 et en 2001. L'EIR 2004 est conçu pour représenter les personnes âgées de 54 ans ou plus au 31 décembre 2004. L'échantillon a été tiré dans le RNIPP (y compris la section hors métropole) : les personnes de l'échantillon sont donc nées en France ou à l'étranger.

L'Échantillon Interrégimes se présente sous la forme d'un panel. Les personnes appartenant à l'échantillon initial sont sélectionnées à nouveau à chaque enquête (excepté les individus décédés entre deux vagues).

L'échantillon a été enrichi à chaque vague, pour mieux prendre en compte la population des retraités et répondre aux besoins d'information des acteurs sociaux.

Les échantillons de 1988 et 1993 ont été conçus de manière à observer la population des retraités de 65 ans et plus nés en France métropolitaine. L'échantillon de 1997 a été étendu à l'ensemble des retraités de 55 ans et plus nés en France métropolitaine. L'échantillon de 2001 a été étendu aux retraités nés dans les Départements d'Outre-mer. Dans la vague 2004, les personnes nées dans les TOM et à l'étranger sont aussi prises en compte, y compris parmi les générations déjà sélectionnées en 2001. Ceci permettra de couvrir l'ensemble des individus percevant des pensions des régimes de retraite français quel que soit leur lieu de naissance.

En effet, en 2001, plus d'un million de retraités nés à l'étranger recevaient un avantage vieillesse du régime général. Ils représentent près de 10 % des pensions versées par la CNAV. C'est le cas de presque la moitié des droits dérivés.

Le nombre de générations de l'échantillon a également augmenté progressivement de 4 dans l'échantillon 1988 à 21 dans l'échantillon de 2004 permettant ainsi d'améliorer la représentativité de l'échantillon.

## b. Sélection de l'échantillon

<b>EIR 2004</b> <b>Année de naissance</b>	<b>Jour de naissance</b>
1906	1 <sup>er</sup> au 10 octobre (seulement pour les personnes nées en métropole et dans les DOM)
1909	1 <sup>er</sup> au 5 octobre (seulement pour les personnes nées en métropole et dans les DOM)
1912	1 <sup>er</sup> au 10 octobre
1915	1 <sup>er</sup> au 5 octobre
1918	1 <sup>er</sup> au 10 octobre
1920	1 <sup>er</sup> au 3 octobre
1922	1 <sup>er</sup> au 10 octobre
1924	1 <sup>er</sup> au 3 octobre
1926	1 <sup>er</sup> au 10 octobre
1928	1 <sup>er</sup> au 3 octobre
1930,1932,1934,1936	1 <sup>er</sup> au 6 octobre
<b>1938</b>	1 <sup>er</sup> au 6 + <b>7 au 24 octobre</b>
<b>1940,1942</b>	1 <sup>er</sup> au 6 octobre + <b>7 au 12 octobre</b>
1944,1946	1 <sup>er</sup> au 6 octobre
<b>1948,1950</b>	<b>1<sup>er</sup> au 6 octobre</b>

Dans le tableau, les jours de naissance indiqués en gras correspondent à des « générations » entrantes par rapport à l'échantillon de 2001.

En effet, en sus des générations déjà sélectionnées en 2001, l'échantillon de 2004 comprend des nouvelles générations :

- du fait du vieillissement de la population entre 2001 et 2004, deux « jeunes » générations, celles nées en 1948 et en 1950 et âgées de 56 et 54 ans en 2004 sont entrés dans l'échantillon ;
- les générations nées en 1938, 1940 et 1942 sont surreprésentées de manière ponctuelle (le nombre de jours de tirage est multiplié par 4 pour la génération 1938 et par 2 pour les générations 1940 et 1942).

Au final, l'EIR 2004 contient 145 792 retraités.

## **ANNEXE 2 :**

### **Restriction du champ aux monopensionnés de la CNAV**

Le champ de l'étude a été restreint aux monopensionnés de la CNAV pour pallier des problèmes de données manquantes : c'est en effet sur les données fournies par la CNAV que ces difficultés sont les plus limitées. Si la détermination des monopensionnés de la CNAV est sans difficultés pour la génération 1938, pour laquelle on dispose des données de l'EIR, la question est beaucoup plus complexe pour la génération 1950, qui n'a pas encore effectué la totalité de son parcours professionnel. La carrière de chacun des individus nés en 1950 est donc prolongée à l'aide du modèle PROMESS<sup>18</sup> de la DREES. Leurs droits sont ensuite calculés sur la base de cette trajectoire professionnelle complétée, et il est ainsi possible de distinguer les monopensionnés de la CNAV. Cependant, le modèle PROMESS ne permettant pas de faire des projections distinguant le chômage des autres situations entraînant la validation de trimestres gratuits, les fins de carrière n'ont pas été incorporées à l'analyse : il a été choisi d'effectuer le choix de cas types sur la base des carrières tronquées à 55 ans pour les deux générations, afin d'utiliser dans les deux cas les données observées issues de l'EIC.

#### **1. Restriction du champ aux monopensionnés CNAV**

Certaines caisses<sup>19</sup> gèrent séparément les actifs et les retraités : après la liquidation, il n'a pas toujours été possible de récupérer les informations concernant l'historique de la carrière des individus. Or, la grande majorité des individus de la génération 1938, âgés 67 ans en 2005, ont déjà liquidé leur retraite au moment de l'enquête. Les chroniques professionnelles sont donc parfois inexistantes pour ces régimes. Ce problème est beaucoup plus marginal pour la génération 1950, puisque âgée de 55 ans en 2005 : une fraction vraisemblablement très faible de la population a déjà liquidé ses droits à pension au moment de l'enquête.

En outre, les caisses conservent avec précision les données qui leur sont utiles pour calculer les droits à pension de leurs affiliés, qui ne correspondent pas nécessairement aux exigences de l'EIC. Par exemple, nombreuses sont les caisses qui n'ont pas besoin de la chronique salariale précise : pour ces caisses, les données sur le revenu sont souvent imprécises ou partielles, ce qui pose problème pour l'exercice ici effectué.

Enfin, certaines caisses n'ont pas été en mesure de fournir à la DREES toute la chronologie des individus de l'échantillon du fait notamment de l'historique de leur système de gestion. À la SNCF, les carrières des actifs ne sont enregistrées que depuis 1987.

Pour ces raisons, il a été choisi de centrer l'analyse sur les individus ayant effectué leur carrière à la CNAV, afin de minimiser ces difficultés.

Une première possibilité aurait pu être de se limiter aux monoactifs de la CNAV : cette méthode est cependant problématique pour la génération 1938, pour laquelle certaines caisses n'apparaissent pas dans la chronique professionnelle. En revanche, si les données de l'EIC sont lacunaires, celles de l'EIR sont bien renseignées : ainsi, on sait avec précision auprès de quelles caisses chaque individu perçoit une pension<sup>20</sup>, et le montant de celle-ci. Ne conserver que les monopensionnés CNAV permet de se restreindre aux individus dont il est certain que l'essentiel de la carrière a été effectué à la CNAV, ce qui garantit la qualité des données sur le déroulé de carrière, notamment concernant le niveau de revenu.

---

<sup>18</sup> Cf. document de travail de la DREES.

<sup>19</sup> Notamment SNCF et CANCAVA.

<sup>20</sup> Hors VFU.

Le renseignement des débuts de carrière présente tout de même des lacunes, y compris à la CNAV ; elles sont certes assez marginales pour la génération 1950, mais encore relativement importantes pour la génération 1938. Par ailleurs, les périodes de travail à l'étranger ne sont pas forcément connues. Les périodes de service militaire ne sont connues généralement qu'à la liquidation sauf pour la plupart des régimes spéciaux. Il en va de même des majorations de durées d'assurance pour enfants qui ne sont souvent comptabilisées qu'à la liquidation.

## **2. Détermination des monopensionnés CNAV**

La détermination des monopensionnés de la CNAV pour la génération 1938 ne pose *a priori* aucune difficulté. Le champ d'étude doit également être restreint aux futurs monopensionnés CNAV de la génération 1950. Ceci est plus complexe, puisque la fin de la carrière n'est pas connue. Une simulation des fins de carrière a donc été effectuée à l'aide du modèle PROMESS de la DREES ; les droits à pension ont ensuite été calculés sur cette base<sup>21</sup>.

### **a. Traitement de la génération 1950**

#### **i. Prolongation de la carrière à l'aide de PROMESS**

Le modèle PROMESS, qui s'appuie sur les données collectées dans l'EIR et l'EIC, permet de simuler les probabilités de sortie vers la retraite à chaque âge sur la base des distributions de trimestres accumulés pour chaque génération.

#### **ii. Détermination des monopensionnés CNAV**

Le montant des droits à pension est ensuite calculé sur la base des carrières complétées à l'aide du modèle PROMESS.

On retient alors les individus qui perçoivent une seule pension pérenne, versée par la CNAV, en prenant en compte les éléments suivants :

- Seules les pensions supérieures au montant minimal au-dessous duquel un versement forfaitaire unitaire est automatiquement proposé sont considérées comme des pensions pérennes. Par souci de simplicité, la valeur de ces montants minimaux est égale à celle déclarées par la CNAV jusqu'en 2010 ; elle est ensuite considérée comme constante. Il est en particulier à noter que ces montants ne correspondent pas nécessairement à ceux pratiqués réellement par les caisses concernées.
- Étant donné que pour les petites caisses les données salariales sont très lacunaires, tous les individus y ayant validés des droits ont été préalablement éliminés du champ, afin d'écartier tous ceux qui pourraient potentiellement percevoir une pension de ces caisses.
- De la même façon, les salaires ne sont pas indiqués pour la CNRACL. On élimine donc d'office de l'ensemble des monopensionnés CNAV tous les individus ayant cotisé plus de 15 ans à la CNRACL, car ils sont supposés percevoir une pension pérenne de cette caisse.

### **b. Traitement de la génération 1938**

---

<sup>21</sup> Certaines informations usuellement récupérées au moment de la liquidation, telles que le nombre d'enfants, les trimestres rachetés, les déclarations de trimestres validés au titre du service militaire ou encore les régularisations, ne sont pas disponibles pour la génération 1950. De fait, les montants de pensions pour la génération 1950 sont donc sous-estimés par rapport à ceux de la génération 1938.

Par souci de comparaison, on élimine de la même façon du champ de l'étude les monopensionnés CNAV de la génération 1938 dont la pension de base aurait suscité la proposition d'un versement forfaitaire unitaire. Cependant, les individus percevant, en plus de leur pension CNAV, des pensions d'autres caisses que la CNAV mais d'un montant inférieur au seuil de versement forfaitaire unitaire de la CNAV ne sont pas réintroduits dans le champ des monopensionnés CNAV.

Le champ est donc plus restreint pour les individus de la génération 1938.

### **c. Représentativité du champ retenu**

Les monopensionnés de la CNAV tels qu'ils sont déterminés dans l'analyse représentent 47 % des affiliés pour la génération 1938, et 57% pour la génération 1950.

Le champ des monopensionnés CNAV est en effet plus restreint pour la génération 1938. En outre, la CNAV a connu une importance croissante entre les deux générations. En effet, pour 50 % des individus nés en 1950, les reports relevés dans l'EIC 2004 ont été entièrement portés au compte de la CNAV. Cette proportion était de trois points inférieure pour la génération 1938, alors même que les reports aux comptes des autres caisses sont d'une moins grande fiabilité pour cette génération, la monoactivité CNAV étant donc surreprésentée. Cette augmentation de la monoactivité CNAV est probablement liée à la décroissance d'autres caisses<sup>22</sup>.

---

<sup>22</sup> Par exemple, la monoactivité dans une des caisses de la MSA concernait 2,57 % des individus nés en 1938, et seulement 0,37 % des individus nés en 1950

### ANNEXE 3

#### Calcul des coûts de substitution dans la Méthode d'Appariement Optimal

Le coût d'une substitution entre deux modalités A et B est défini comme suit.

En notant  $N_t(A)$  le nombre de séquences ayant la modalité A à l'instant t, et  $N_{t,t+1}(A,B)$  le nombre de séquences étant passées de la modalité A à t à la modalité B à t+1, on obtient la probabilité de passer de la modalité A à la modalité B, notée  $p(A,B)$ , selon la formule suivante :

$$p(A,B) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(A,B)}{\sum_{t=1}^{T-1} N_t(A)}$$

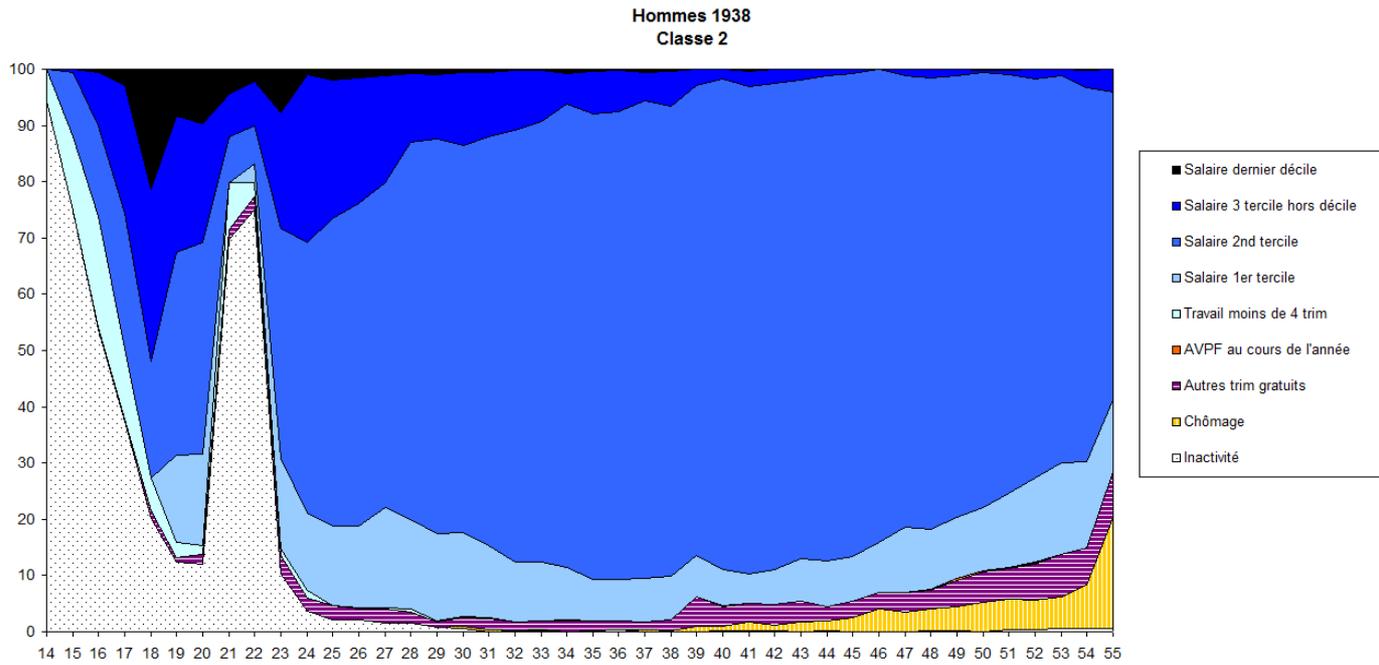
Le coût de substitution entre A et B, noté  $c(A,B)$ , est ensuite défini par :

$$c(A,B) = 2 - p(A,B) - p(B,A)$$

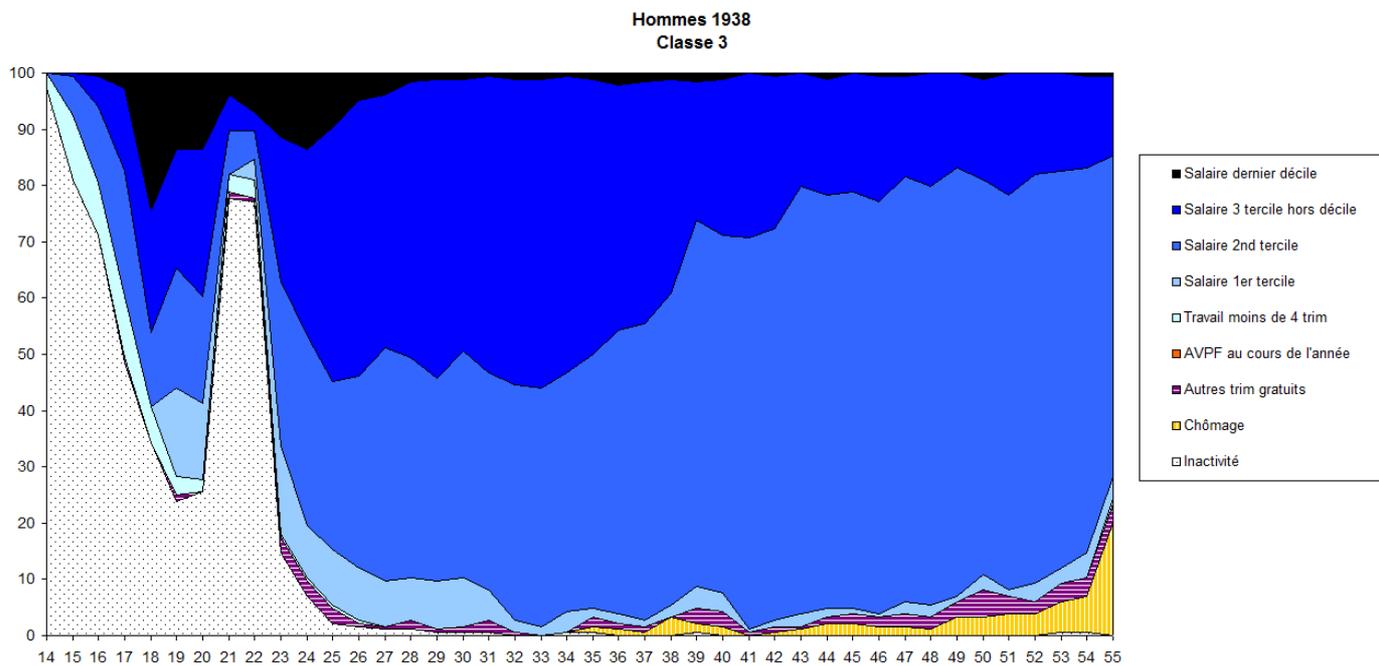
Ainsi, le coût sera nul si A et B se succèdent systématiquement l'un l'autre. À l'inverse, si A et B ne se côtoient jamais immédiatement au sein d'une même séquence, le coût sera de 2, c'est-à-dire équivalent à une insertion et une suppression.



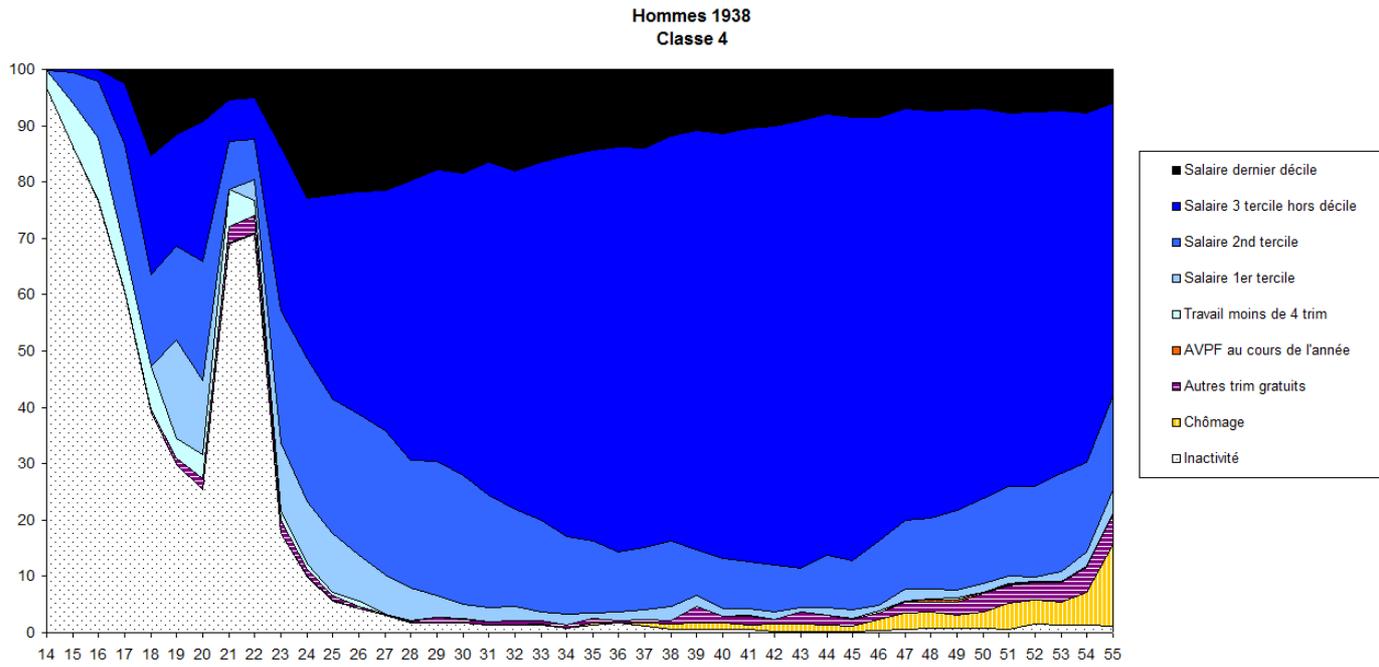
**Classe 2 : Revenus moyens, début de carrière précoce. 16 % des monopensionnés CNAV**



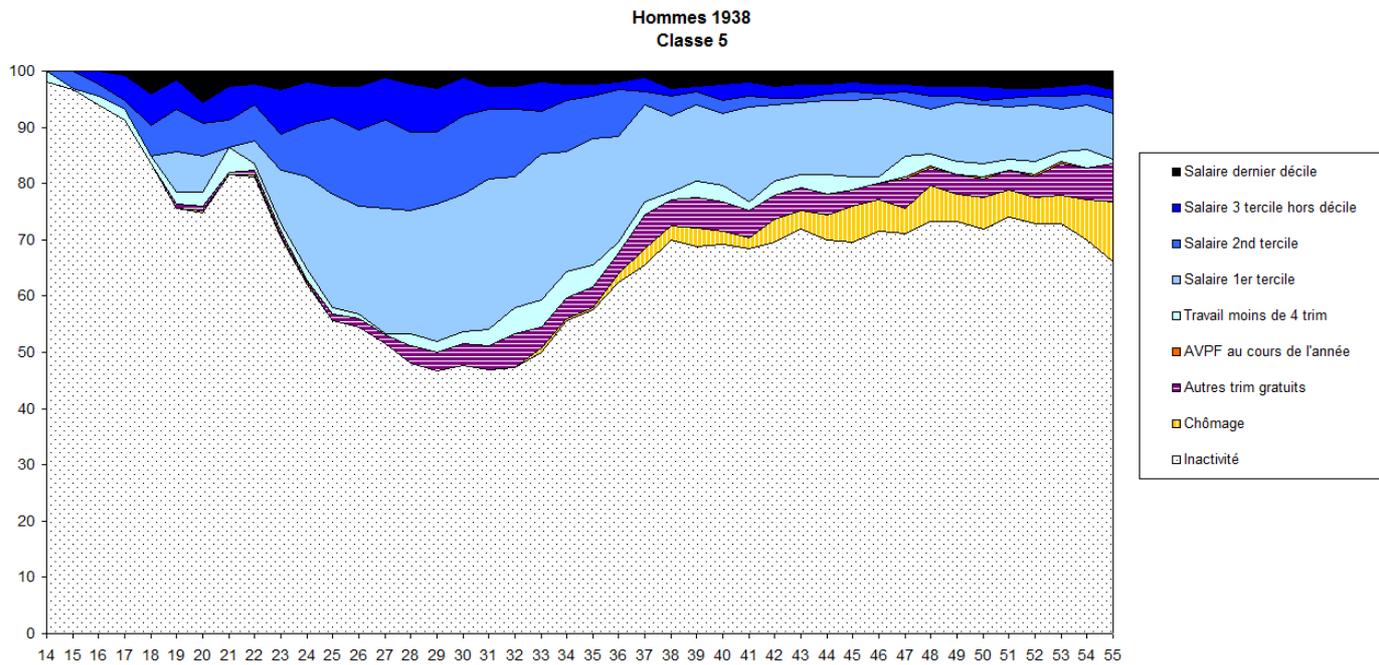
**Classe 3 : Revenus moyens et élevés, début de carrière précoce. 6 % des monopensionnés CNAV**



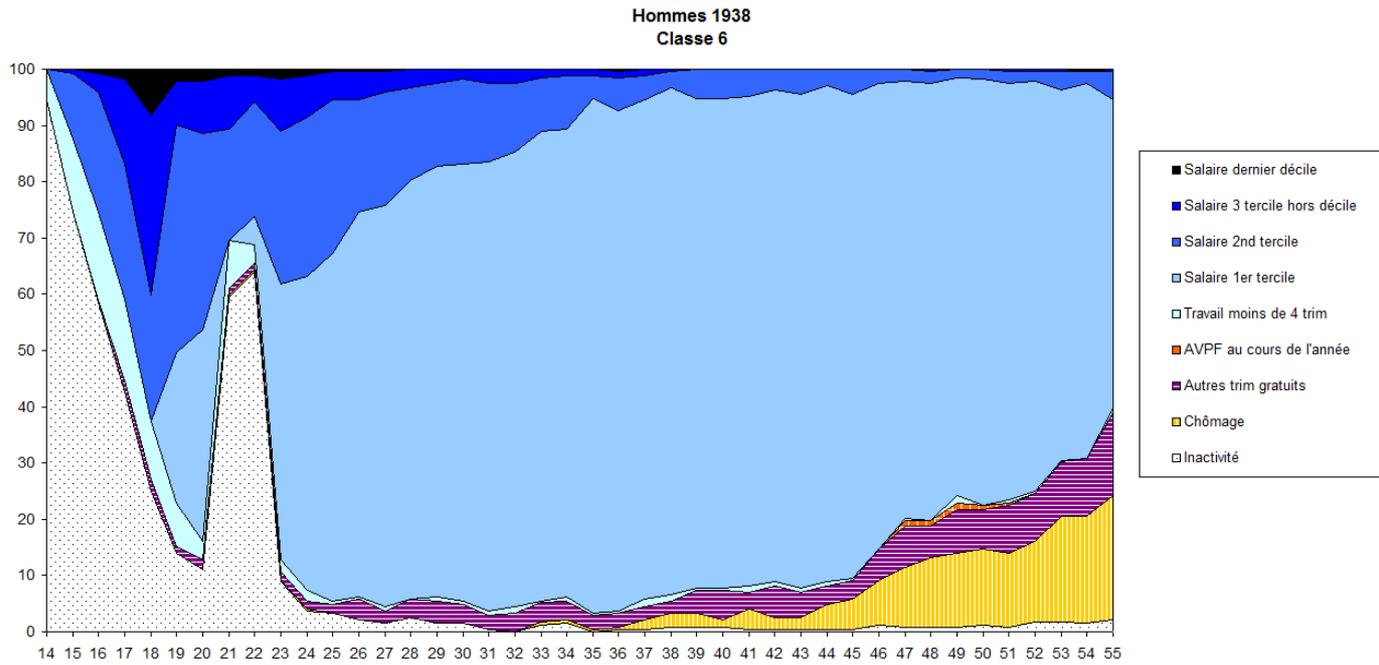
**Classe 4 : Revenu élevés, début de carrière précoce. 21 % des monopensionnés CNAV**



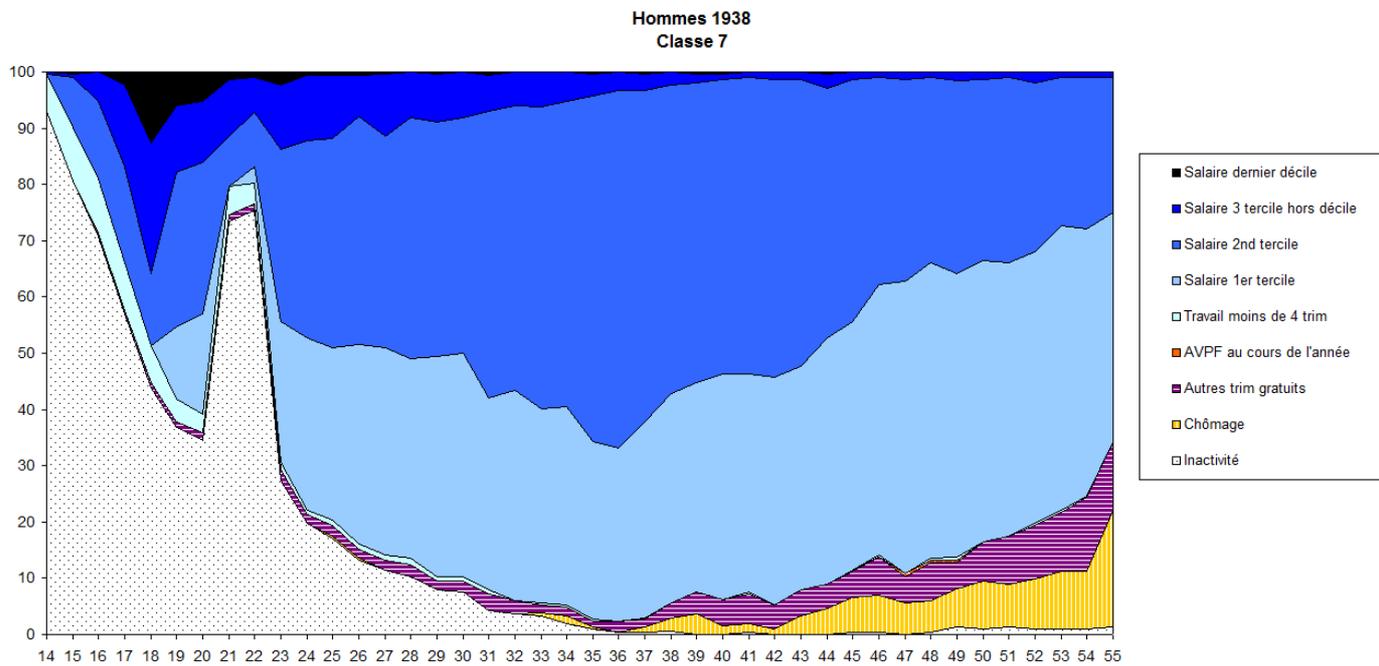
**Classe 5 : Carrière morcelée, beaucoup d'inactivité. 9 % des monopensionnés CNAV**



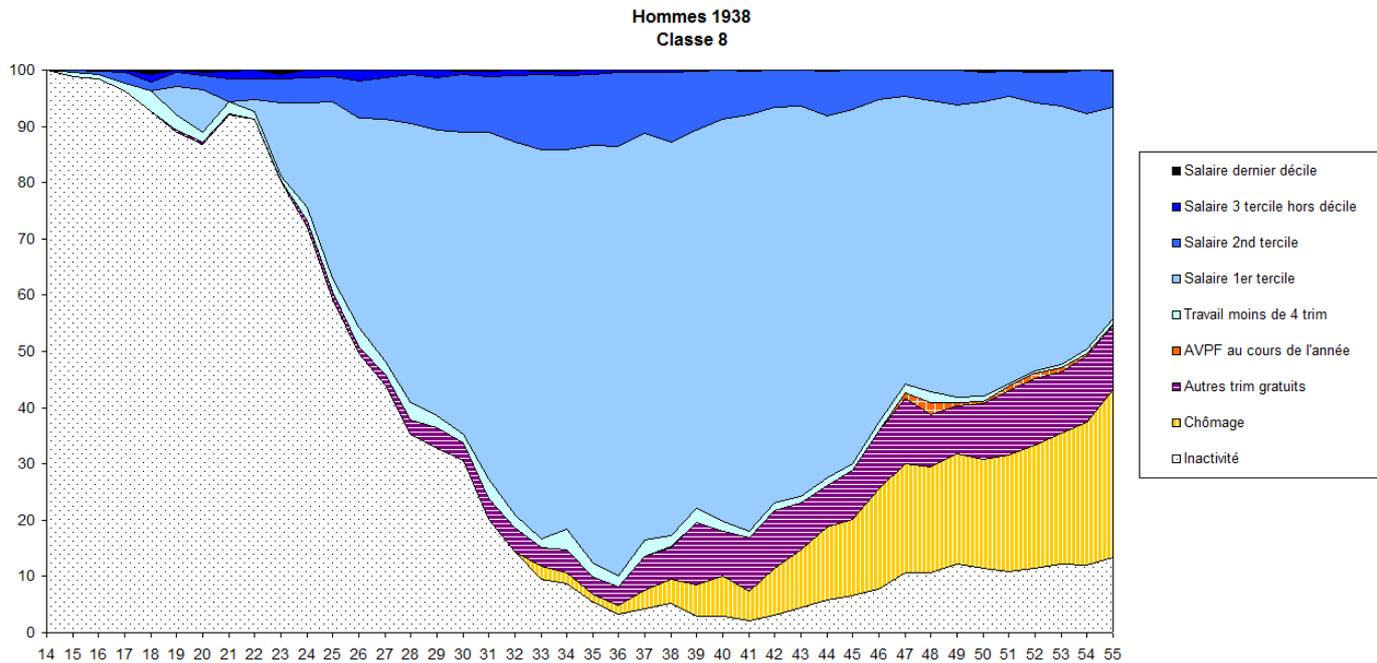
**Classe 6 : Revenus faibles, début de carrière précoce. 9 % des monopensionnés CNAV**



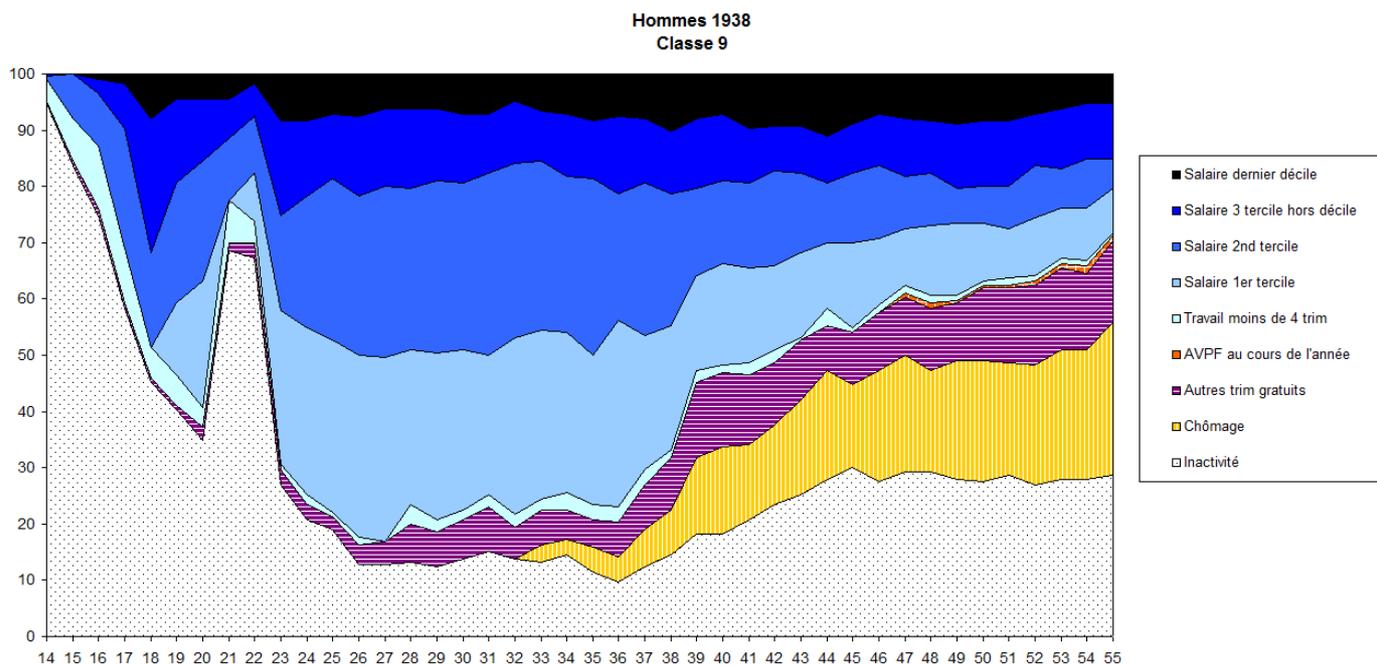
**Classe 7 : Revenus moyens et faibles, début de carrière précoce. 10 % des monopensionnés CNAV**



**Classe 8 : Revenus faibles, début de carrière tardif, chômage important. 15 % des monopensionnés CNAV**



**Classe 9 : Revenus disparates, chômage important. 8 % des monopensionnés CNAV**



**b. Statistiques descriptives**

Pour la génération 1938 uniquement, on dispose grâce à l'EIR d'informations complémentaires portant sur le pays de naissance, la durée d'assurance prise en compte dans le calcul de la pension, l'âge de liquidation, le salaire annuel moyen, le type de pension

perçue, le montant de la pension, le statut de cadre ou le nombre d'enfants<sup>23</sup>. Ces données permettent de mieux caractériser les regroupements établis par la classification par le biais d'un prisme autre que celui du pur déroulé de carrière.

Tout d'abord, les individus ayant le statut de cadre sont repérés par leurs cotisations à l'AGIRC ; ils sont largement majoritaires dans les classes 1 et 4, et constituent une part importante de la classe 3, ces trois classes regroupant les plus hauts salaires.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9
Part de cadres	<b>93 %</b>	27 %	<b>59 %</b>	<b>89 %</b>	11 %	3 %	9 %	1 %	33 %

Par ailleurs, les données concernant le type de pension suggèrent que les classes 5, 8 et 9 rassemblent des individus percevant plus souvent que la moyenne des pensions au titre de l'invalidité ou en tant qu'ex-invalides. On observe en particulier que la classe 5 comporte une grande proportion d'inaptes au travail, ce qui explique peut-être la présence dans cette classe au poids important (10 %) de carrières fortement marquées par l'inactivité, phénomène *a priori* peu commun pour la génération 1938.

Type de pension	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9
Normale	98 %	91 %	96 %	95 %	<b>71 %</b>	83 %	84 %	<b>73 %</b>	<b>72 %</b>
Inapte au travail	1 %	1 %	2 %	1 %	<b>25 %</b>	3 %	4 %	<b>13 %</b>	<b>13 %</b>
Ex-invalides	1 %	8 %	2 %	4 %	5 %	<b>13 %</b>	<b>12 %</b>	<b>14 %</b>	<b>15 %</b>

La nationalité des individus pourrait également expliquer leur profil de carrière, les étrangers ayant vraisemblablement une plus grande probabilité de n'effectuer qu'une fraction de leur carrière en France. L'EIR ne recense cependant que le pays de naissance des affiliés, ce qui n'est qu'un proxy de leur nationalité. La proportion importante d'individus nés à l'étranger dans les classes 5 et 8 semble cependant évoquer une piste d'explication, avec l'invalidité, quant au caractère disparate et morcelé des carrières des affiliés qu'elles rassemblent.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7	Classe 8	Classe 9
Nés en France	90 %	88 %	89 %	91 %	25 %	86 %	67 %	16 %	66 %
Nés à l'étranger	10 %	12 %	11 %	9 %	<b>75 %</b>	14 %	33 %	<b>84 %</b>	34 %

Quant à l'âge de liquidation, l'âge de 60 ans revient de façon très dominante dans les classes 2 (97 % des affiliés liquident à 60 ans), 3 (91 %), 4 (88 %), 6 (96 %) et 7 (84 %). En revanche, les individus des classes 5, 8, ainsi que, dans une moindre mesure, la classe 9, semblent souvent attendre l'âge du taux plein à 65 ans, et la liquidation à 60 ans y concerne respectivement 41 %, 46 % et 64 % des individus. En effet, le pourcentage d'individus liquidant à 65 ans s'élève respectivement à 45 %, 35 % et 19 %.

Dans la classe 1, celle des plus hauts salaires débutant tard, le départ à 60 ans n'est pas non plus systématique, puisque seulement 48 % des affiliés sont concernés. Cependant, l'absence de mode à 65 ans et l'étalement des départs entre 61 et 64 ans suggèrent plutôt une prolongation de la carrière pour cause de trimestres manquants pour l'obtention du taux plein, vraisemblablement due à un âge de fin d'études plus tardif que la moyenne.

<sup>23</sup> Celui-ci n'étant en général pertinent que pour les femmes, les résultats ne sont pas retranscrits ici.

Par ailleurs, les données relatives au salaire annuel moyen (SAM) révèlent notamment une dispersion importante des niveaux de revenu pour les classes 5, 8 et 9.

<b>SAM</b>	<b>Classe 1</b>	<b>Classe 2</b>	<b>Classe 3</b>	<b>Classe 4</b>	<b>Classe 5</b>	<b>Classe 6</b>	<b>Classe 7</b>	<b>Classe 8</b>	<b>Classe 9</b>
Moyenne	23 359	20 921	22 036	22 240	12 566	15 179	19 288	14 546	17 640
Écart-type	1 725	1 480	1 740	1 408	<b>5 281</b>	2 256	2 509	<b>3 858</b>	<b>4 970</b>

Les durées d'assurance prises en compte dans le calcul de la pension sont également très variables dans les classes 5, 8 et 9.

<b>Durée d'assurance</b>	<b>Classe 1</b>	<b>Classe 2</b>	<b>Classe 3</b>	<b>Classe 4</b>	<b>Classe 5</b>	<b>Classe 6</b>	<b>Classe 7</b>	<b>Classe 8</b>	<b>Classe 9</b>
Moyenne	160	172	169	167	104	169	166	141	150
Écart-type	10	8	9	10	<b>49</b>	10	14	<b>24</b>	<b>32</b>

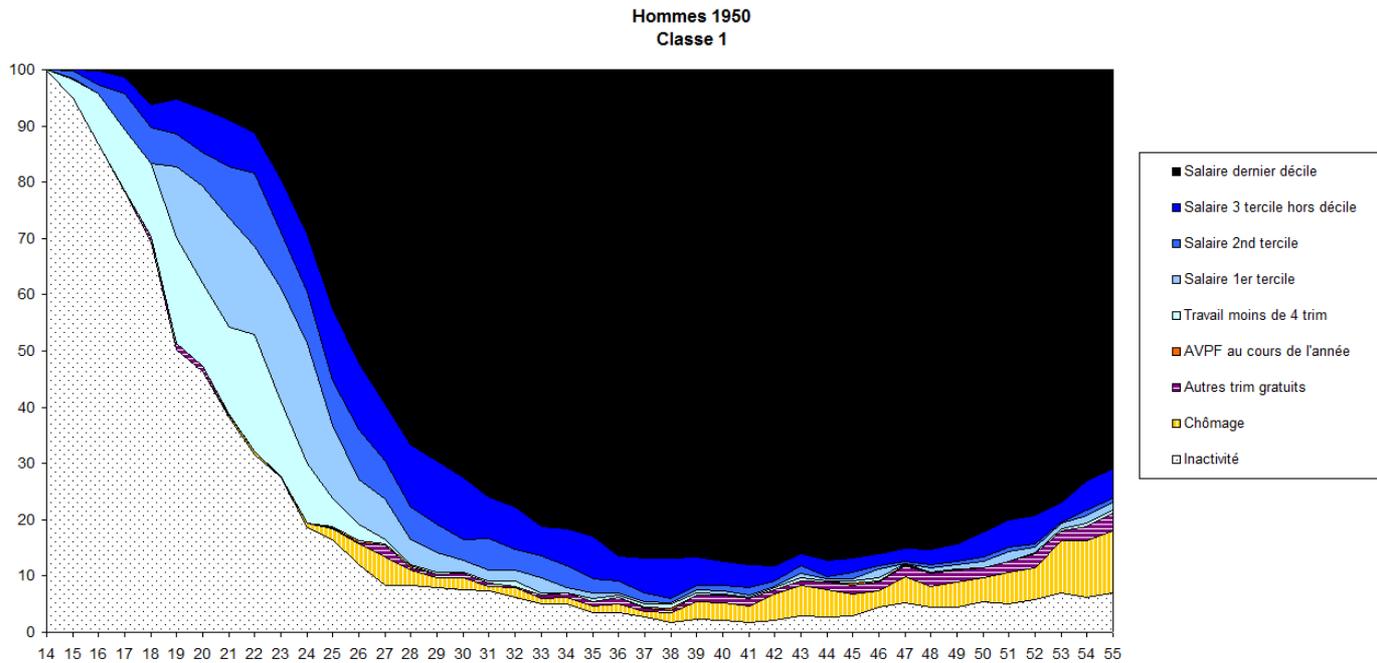
Cette hétérogénéité des classes 8, 9 et surtout 5 se retrouve *in fine* dans le montant des pensions de droit direct : cependant, la classe 5 est largement concernée par le minimum contributif, ce qui tronque la dispersion des montants de pension.

<b>Pension</b>	<b>Classe 1</b>	<b>Classe 2</b>	<b>Classe 3</b>	<b>Classe 4</b>	<b>Classe 5</b>	<b>Classe 6</b>	<b>Classe 7</b>	<b>Classe 8</b>	<b>Classe 9</b>
Moyenne	1 036	952	997	1 003	287	695	852	592	699
Écart-type	85	68	66	74	<b>181</b>	96	98	<b>177</b>	<b>265</b>

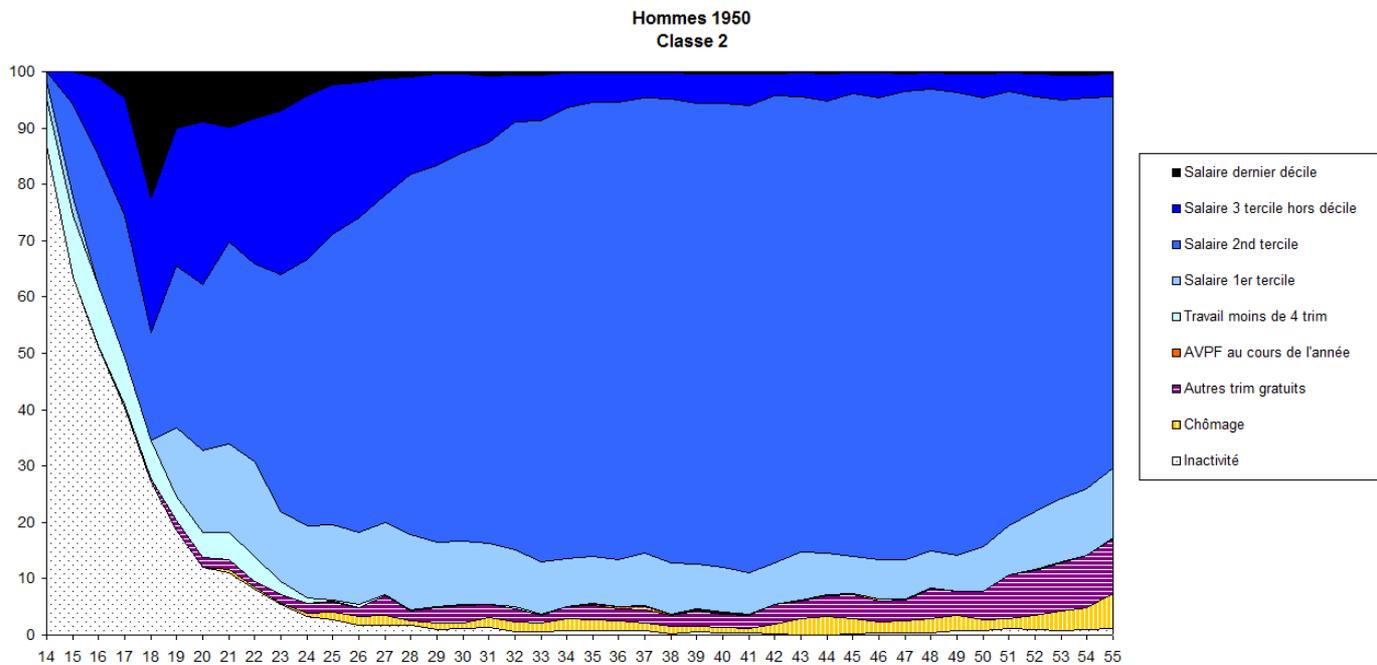
## 2. Génération 1950

En comparaison, la classification sur les hommes de la génération 1950 suggère un découpage en 8 classes :

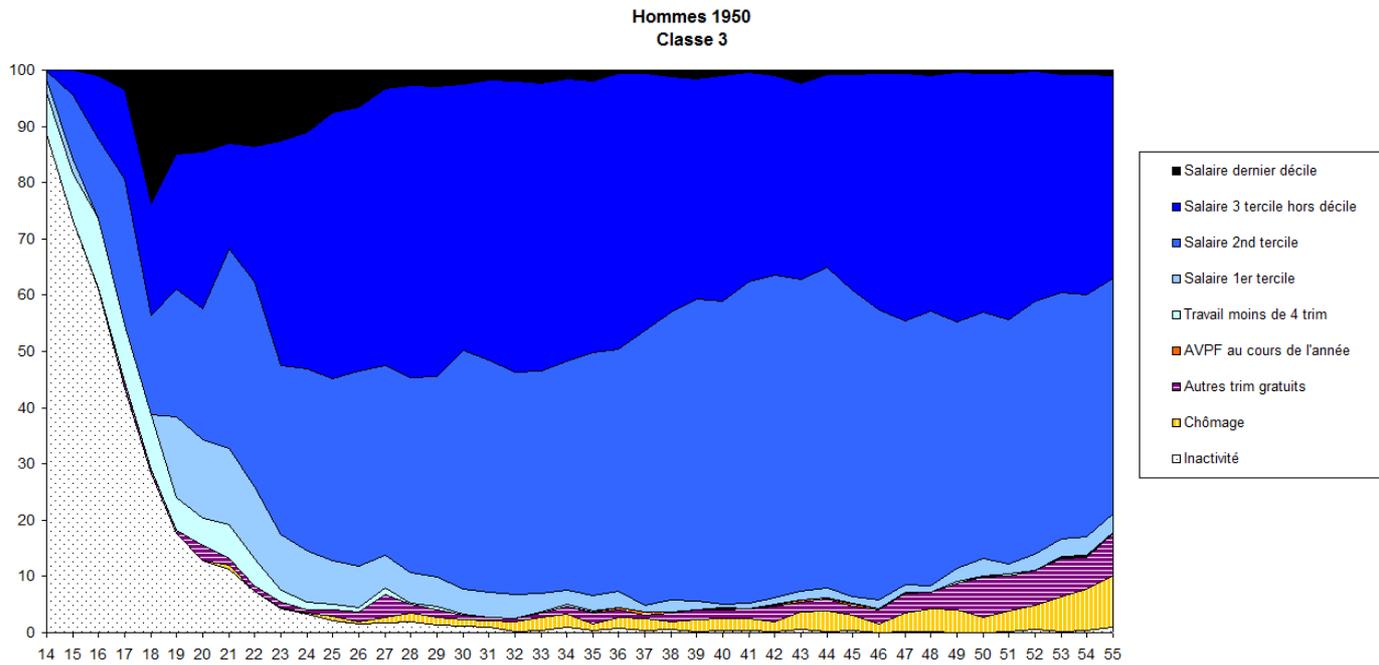
**Classe 1** : salaires les plus élevés, début de carrière tardif. **6 % des monopensionnés CNAV**



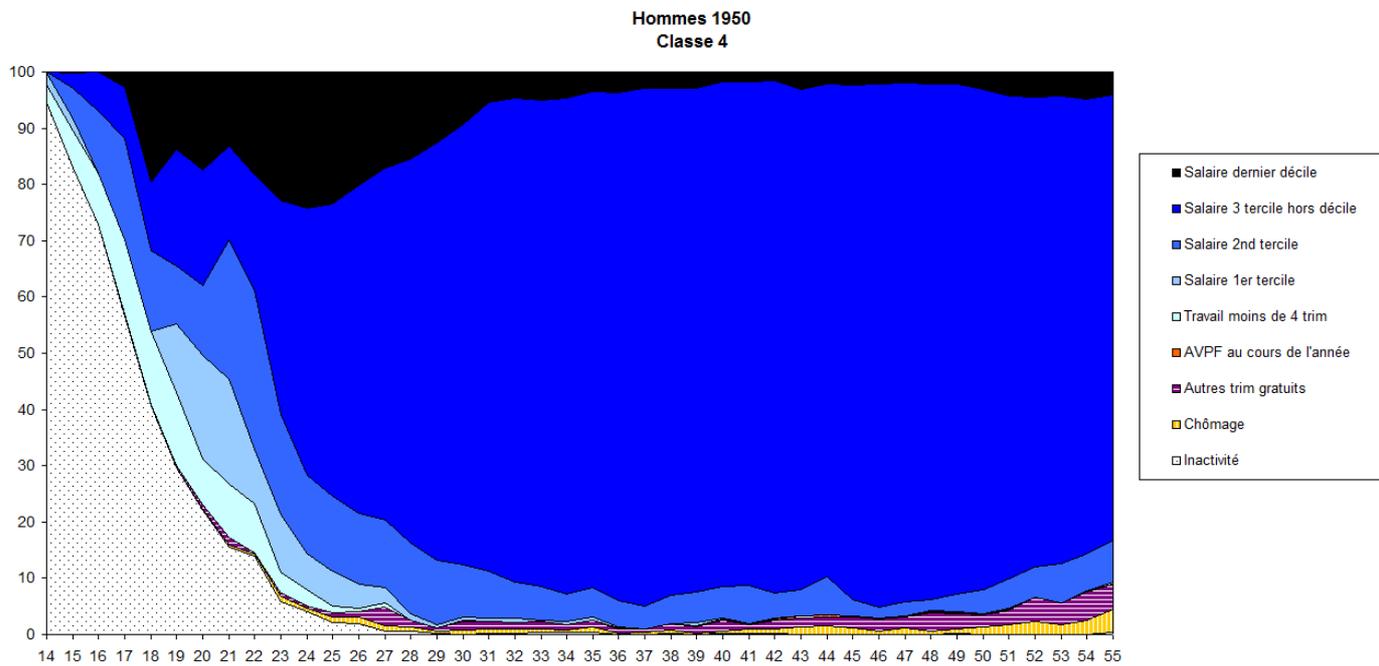
**Classe 2** : Salaires moyens, début de carrière précoce. **18 % des monopensionnés CNAV**



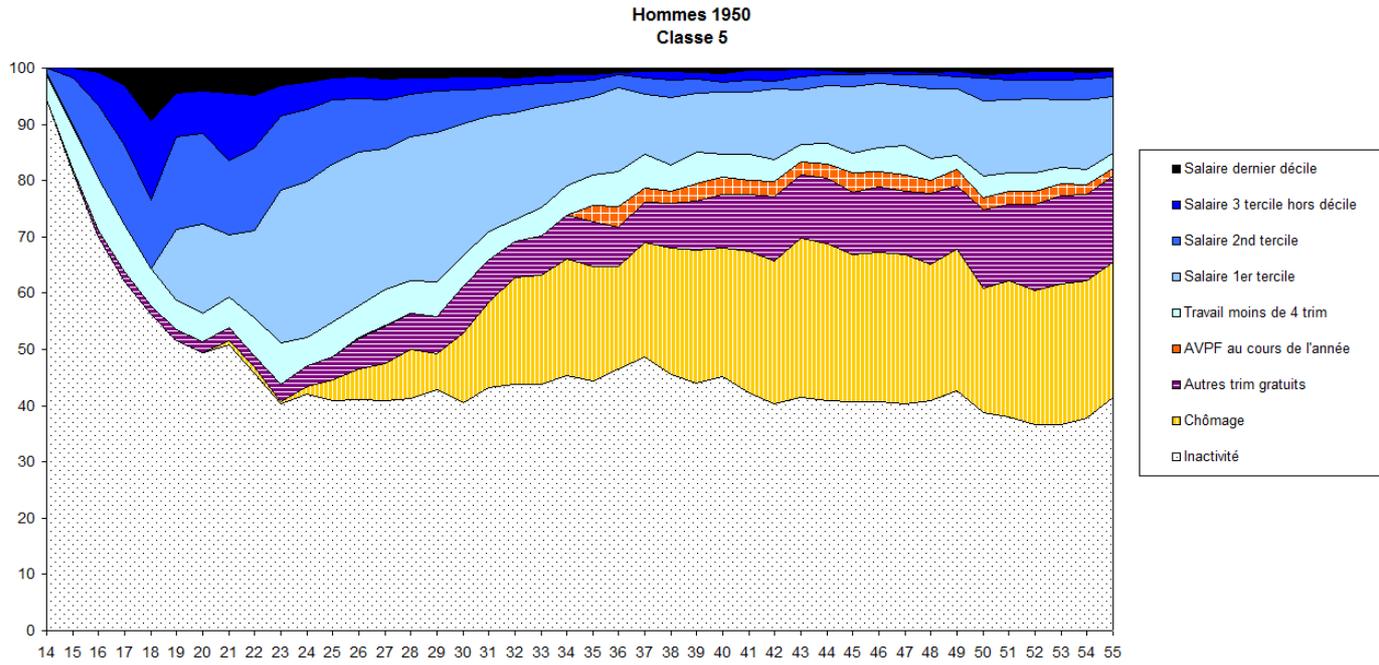
**Classe 3 : Salaires moyens et élevés, début de carrière précoce. 8 % des monopensionnés CNAV**



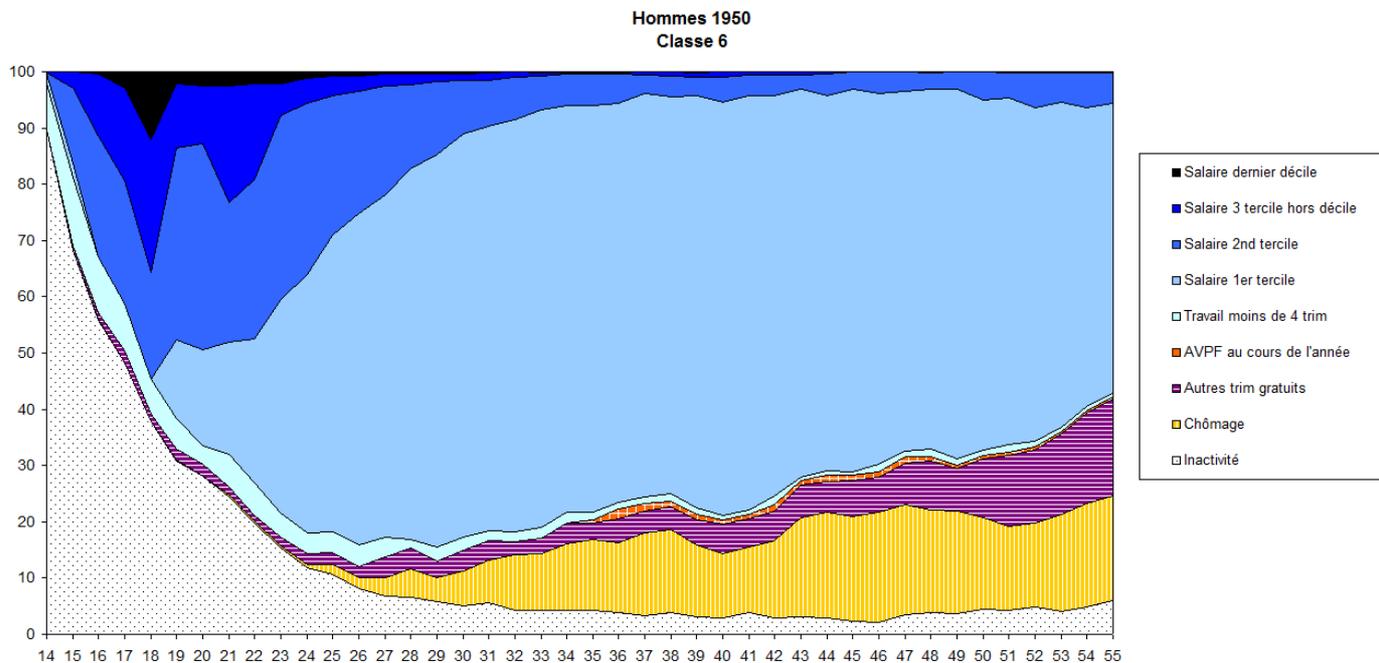
**Classe 4 : Salaire élevés, début de carrière précoce. 9 % des monopensionnés CNAV**



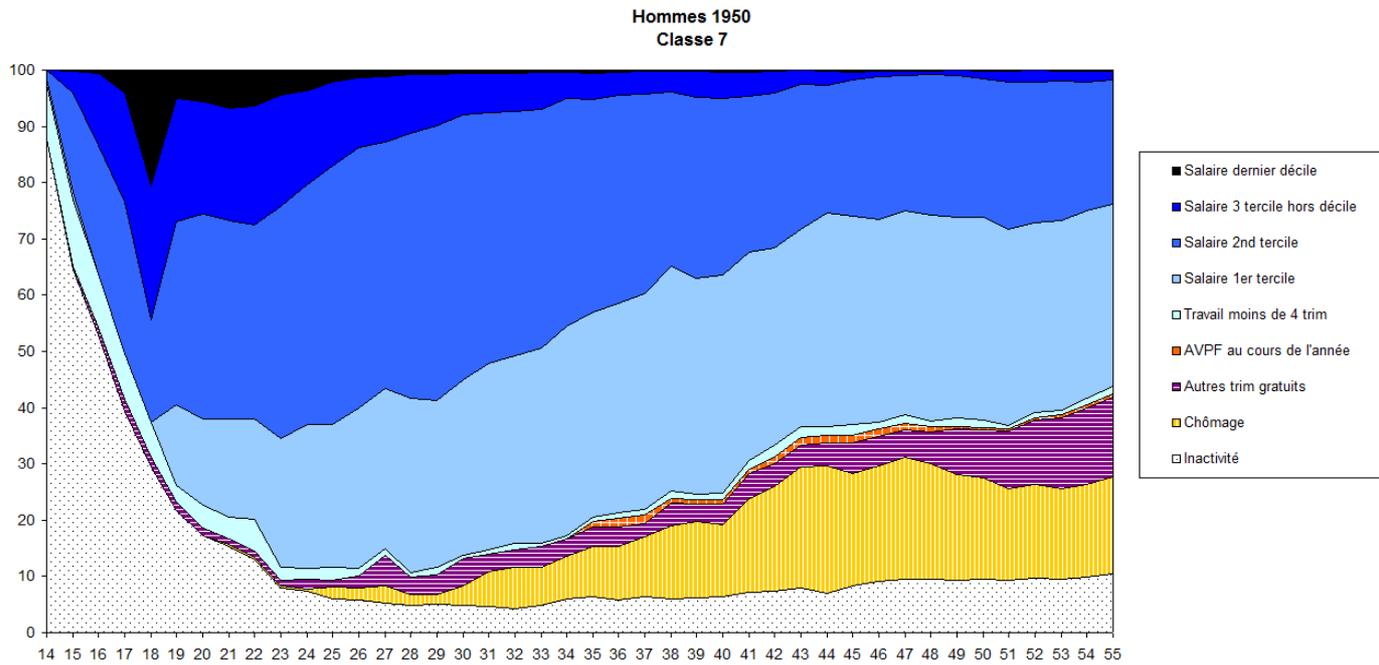
**Classe 5** : Carrière morcelée, beaucoup d'inactivité, chômage important dès le début de carrière. **13 % des monopensionnés CNAV**



**Classe 6** : Salaires faibles, début de carrière assez précoce, chômage important dès le début de carrière. **21 % des monopensionnés CNAV**



**Classe 7** : Salaires moyens et faibles, début de carrière précoce, chômage important dès le début de carrière. **15 % des monopensionnés CNAV**



**Classe 8** : Salaires élevés et très élevés, début de carrière tardif, chômage non négligeable dès le début de carrière. **10 % des monopensionnés CNAV**

