# Etude des données

La base de donnée adhérents comporte au total 1,7 millions d’individus, ayant-droit et adhérents décédés inclus.

## Traitements initiaux

|  |  |
| --- | --- |
| On compte 13 affiliés de moins de 16 ans en droit propre | 0.0% |
| On compte 25 affiliés de plus de 120 ans sans date de décès | 0.0% |
| On compte 10 affiliés dont les ages d'affiliation, radiation, liquidation, ou décès sont >120 | 0.0% |
| On compte 70 affiliés en DP dont les ages d'affiliation, radiation, liquidation, ou décès sont <16 | 0.0% |
| On compte 744 affiliés de plus de 100 ans sans date de décès ou date de liquidation | 0.05% |
| On compte 517 affiliés décèdé sans date de décès | 0.03% |
| On compte 176 affiliés sans date de Naissance | 0.01% |
| On compte 9 149 affiliés droit-propre (DP) sans date d'Affiliation | 0.59% |
| On compte 13 affiliés décédés avant radiation | 0.0% |
| On compte 307 affiliés décédés avant liquidation | 0.02% |
| On compte 1 affiliés décédés avant affiliation | 0.0% |
| On compte 83 affiliés liquidés avant affiliation | 0.01% |
| On compte 1 affiliés radiés avant affiliation | 0.0% |
| On compte 0 affiliés affiliés avant naissance | 0.0% |
| **Au total, on exclut 11 109 affiliés** | **0.72%** |

On remplace les données manquantes par 0 là où ça a du sens.

On groupe les 2000 professions les moins courantes dans une classe AUTRE.

On groupe les 300 codes postaux les moins courantes dans une classe AUTRE.

Le taux de femmes en portefeuille (non décédés) est de **42.0%**

Le taux de PL en portefeuille (non décédés) est de **38.0%**

## Statistiques descriptives

* On compte le volume d’adhérents par catégorie et sous-catégorie.

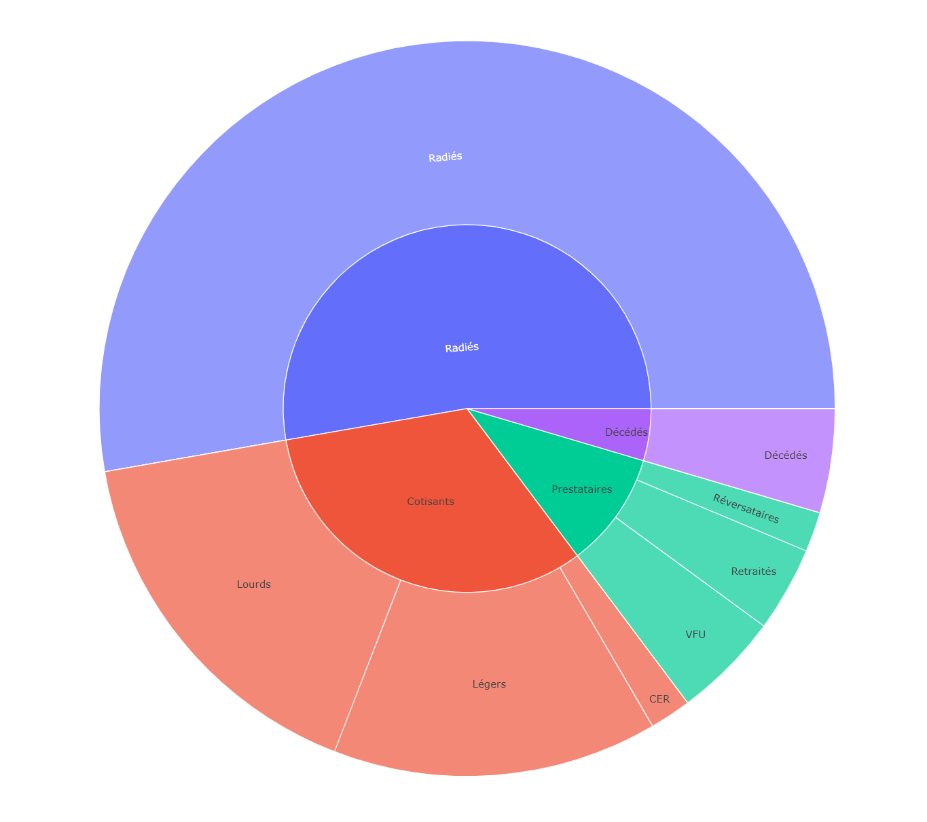
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| En milliers | | **PL** | **ME** | **Droits-dérivés** | **Total** |
| **Cotisants** | CER | 13 | 16 |  | 29 |
| Lourds | 148 | 97 |  | 245 |
| Légers | 34 | 200 |  | 234 |
| **Total** | **194** | **314** |  | **508** |
| **Radiés** | Radiés | 273 | 551 |  | 824 |
| **Prestataires** | Retraités | 55 | 4 |  | 58 |
| VFU | 46 | 27 |  | 73 |
| Réversataires |  |  | 27 | 27 |
| **Total** | **101** | **31** | **27** | **159** |
| **Décédés** | Décédés | 51 | 5 | 15 | 71 |

Les **cumuls emploi-retraite** (CER) sont considérés comme des cotisants bien qu’ils perçoivent également une pension.

On distingue les **cotisants lourds** des **cotisants légers** par leur nombre de points. Les cotisants qui, en moyenne, cotisent moins de 9 points par ans percevraient au bout de 20 ans de cotisations un versement forfaitaire unique (VFU). Ils sont considérés comme « cotisants légers ». Les autres sont des cotisants lourds.

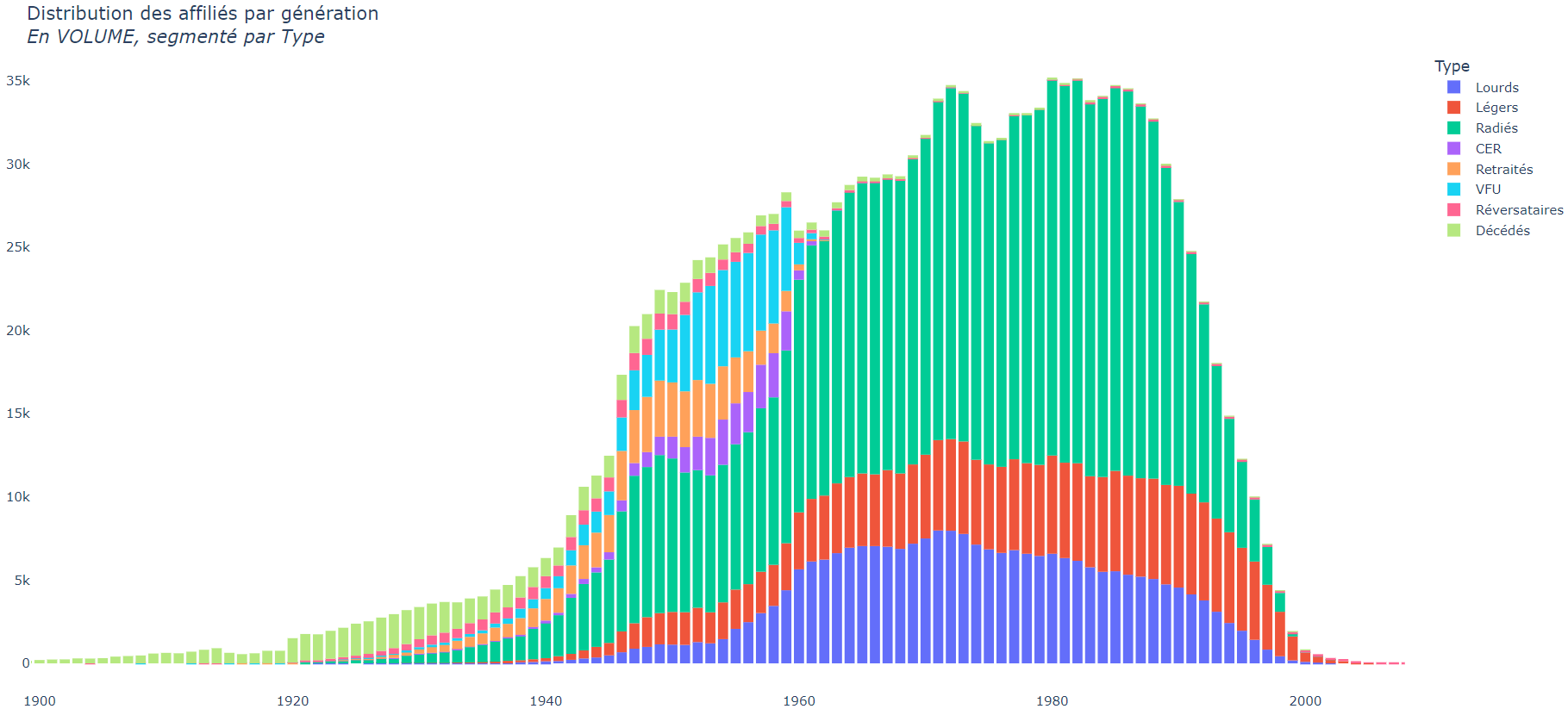
On constate que les cotisants PL sont majoritairement lourds tandis que les ME sont principalement légers. On observe plus de radiés que de cotisants, à la fois en PL et en ME. Les retraités sont pour le moment peu nombreux, et on note que les ME sont plus souvent servi en VFU. Le nombre de décès est pour le moment limité.

On peut visualiser cette répartition sur un diagramme circulaire :



On constate bien la part écrasante des radiés ici (plus de la moitié).

Il est également interessant de segmenter cette répartition par génération :



On constate que :

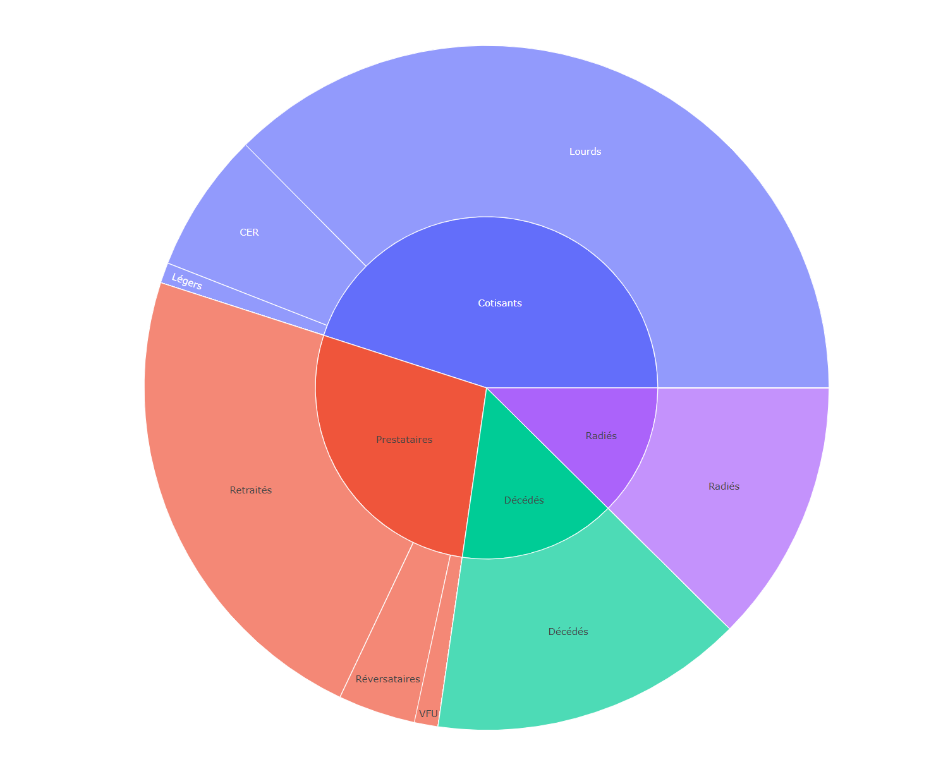
* La proportion d’adhérents lourds et légers semblent a peu près se conserver au fil des générations.
* Les radiés sont majoritairement jeunes.
* Il y a un volume important de cotisants avant 1960 (plus de 62 ans), autant que de prestataires.
* Il reste des cotisants avant 1950 (plus de 72 ans), qui ne liquideront peut-être jamais leurs droits.
* On note la mise en place du dossier de retraite en ligne depuis 2019, qui apporte de la visibilité et de la transparence concernant les droits des adhérents. L’inflexion à la baisse du nombre de radiés observée entre 1950 et 1960 en est peut-être une conséquence.
* On compte le nombre de points par catégorie et sous-catégorie.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| En millions | | **PL** | **ME** | **Droits-dérivés** | **Total** |
| **Cotisants** | CER | 25 | 6 |  | 31 |
| Lourds | 162 | 14 |  | 177 |
| Légers | 1 | 4 |  | 5 |
| **Total** | **189** | **25** |  | **213** |
| **Radiés** | Radiés | 46 | 13 |  | 59 |
| **Prestataires** | Retraités | 103 | 6 |  | 109 |
| VFU | 4 | 1 |  | 5 |
| Réversataires |  |  | 17 | 17 |
| **Total** | **107** | **7** | **17** | **131** |
| **Décédés** | Décédés | 55 | 1 | 15 | 70 |

On rappelle qu’un point cotisé correspond à une portion du revenu d’un adhérent, tendis qu’un point servi donnera une pension perçue tous les ans jusqu’au décès de l’adhérent.

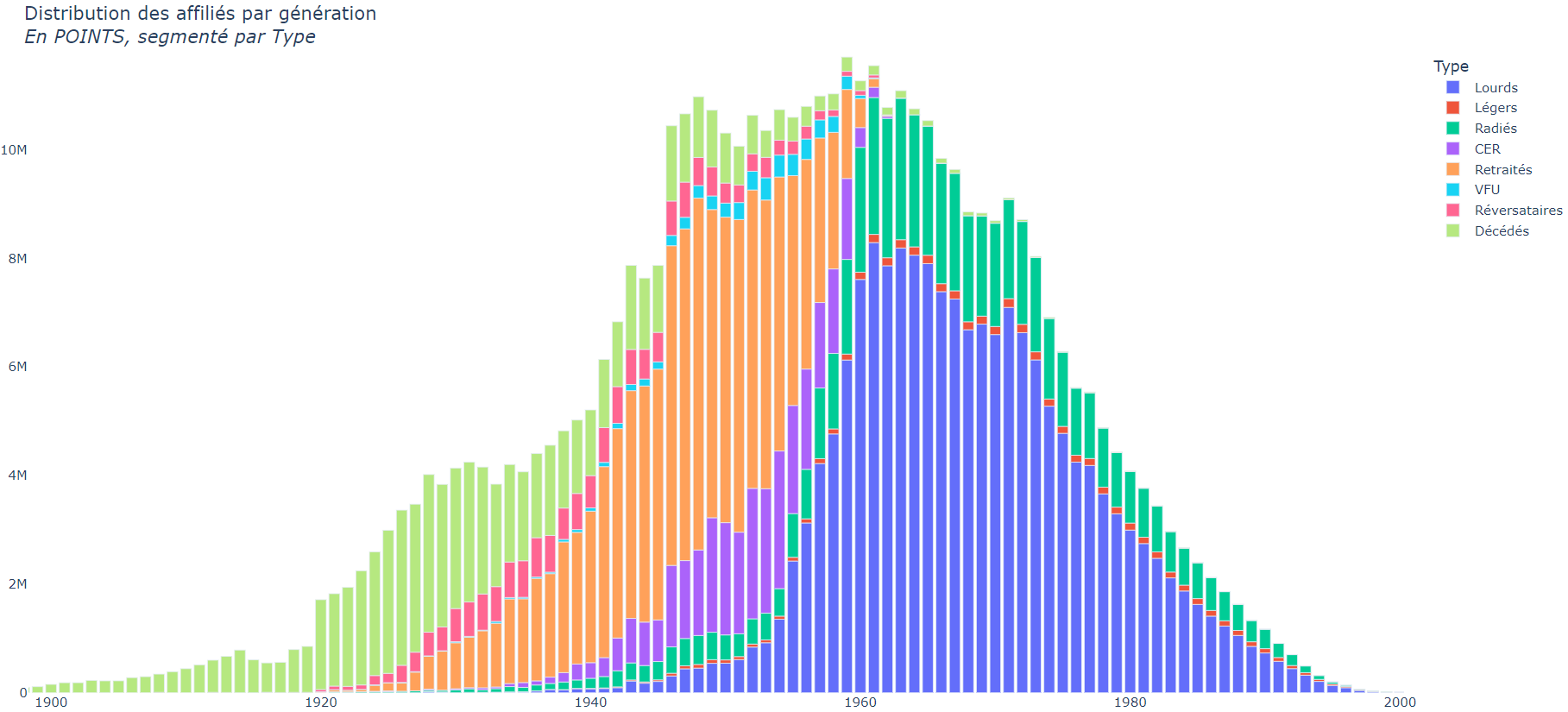
On constate que les CER, bien que peu nombreux, possèdent un nombre de points significatif. On constate que les légers possèdent très peu de points à la fois en PL et en ME. Le radiés possèdent, en proportion de leur nombre, assez peu de points, tandis que les retraités possèdent une quantité importante de points.

On peut visualiser cette répartition sur un diagramme circulaire :



On constante que les cotisants sont les adhérents qui ont le plus de points.

Il est également interessant de segmenter cette répartition par génération :



On constate que :

* Les cotisants légers représentent très peu de points, tout juste l’épaisseur du trait.
* Il y a un volume important de points chez les décédés (non servis aujourd’hui). Ce constat résulte de la chute progressive du rendement du point. En effet, la valeur d’achat du point était avant bien plus faible qu’aujourd’hui, et donc à revenu équivalent, les cotisants pouvaient « acheter » plus de points qu’aujourd’hui.

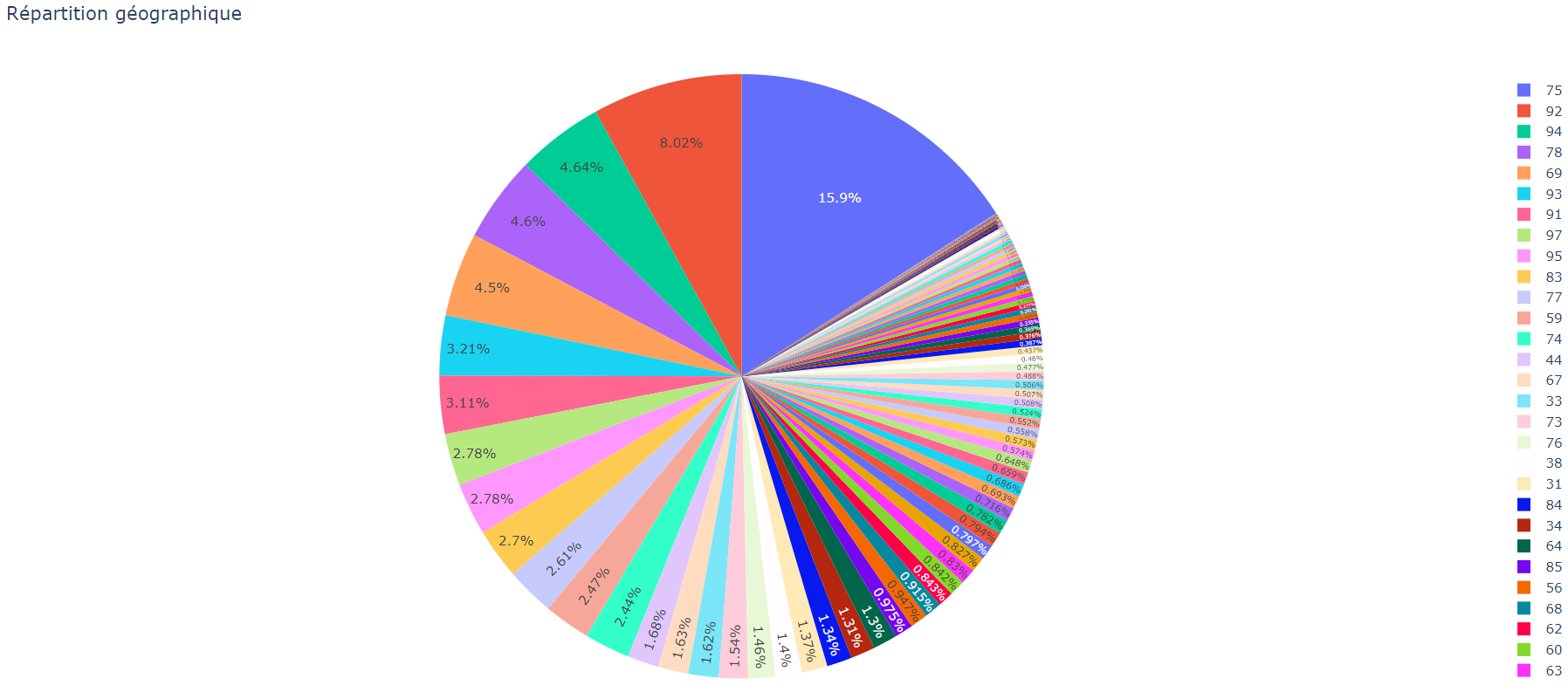
On s’intéresse ici aux éléments qui vont jouer en faveur du bénéfice technique :

* Les CER représentent une part non négligeable qui est tout à l’avantage du régime (leurs cotisations ne donnent pas lieu à des droits supplémentaires). Parmi eux, on compte des profile à forte valeur ajouter sur le marché et donc à fort revenu.
* Parmi Les radiés (et cotisants) de générations plus anciennes que 1960, certains ne liquideront jamais leurs droits malgré le fait qu’ils aient cotisés. On rappelle cependant une réforme statutaire de 2021 qui donne le droit aux débiteurs de liquider leur pension sans avoir régulariser leur dette. On s’attend donc à une réduction des adhérents qui ne liquident jamais.

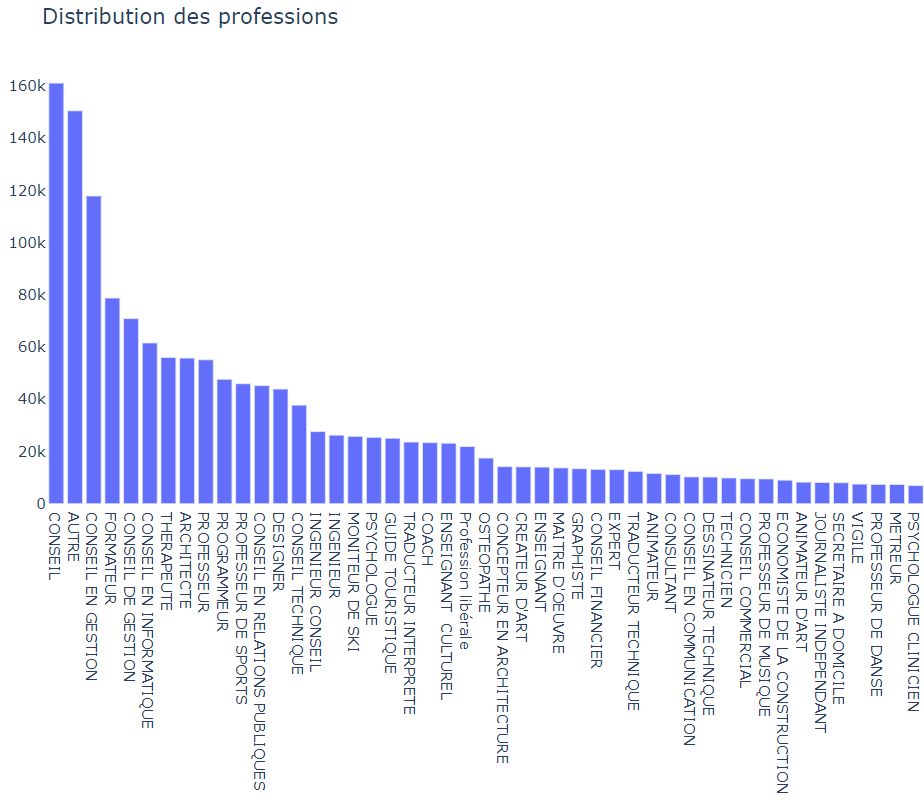
Et à l’inverse ceux qui vont jouer en défaveur :

* Les réversataires perçoivent une pension sans avoir cotisé. Ils sont peu nombre et doivent vérifier certaines conditions pour pouvoir percevoir ces droits.
* Les VFU ne représentent que peu de points. Ils ne perçoivent pas de pension, mais « rachète » leur capital cotisé à 15 fois le montant de leur pension. Cette valeur correspond à l’inverse du rendement du point. Ils récupèrent donc leur pension à la valeur d’achat actuelle. Or ils ont acheté leur point à une valeur plus faible. Ils ont donc perçu un rendement équivalent à l’évolution de la valeur d’achat. Le régime à donc perdu ce que les adhérents VFU ont gagnés.
* Répartition par zone géographique.

On dispose en base des code postaux de nos affiliés.

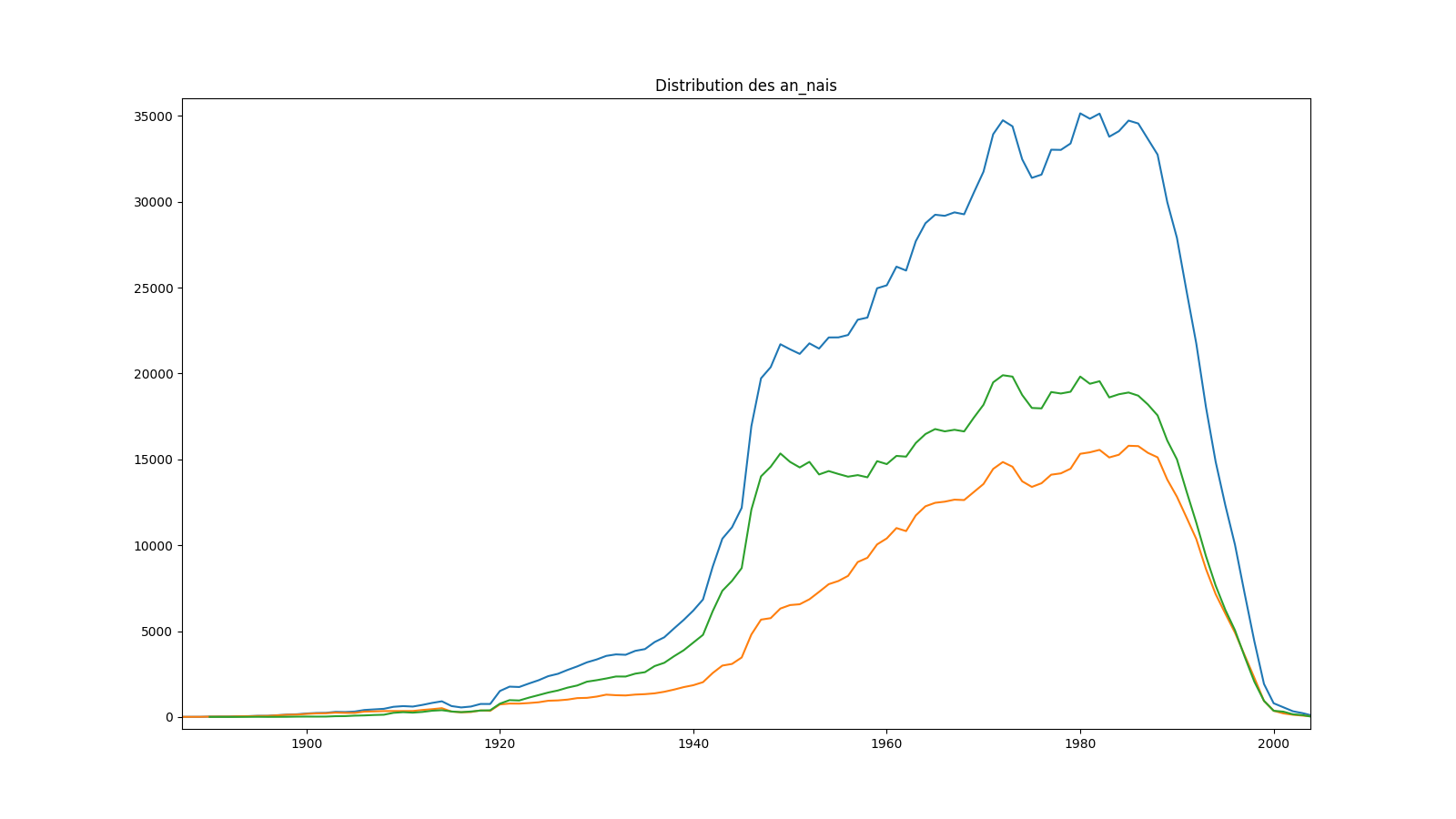
On note que Paris et sa banlieue représente une part importante de notre portefeuille (entre un quart et un tiers)

* Répartition par professions.

L’activité de conseil est, sans surprise, l’activité la plus représentée dans notre portefeuille.

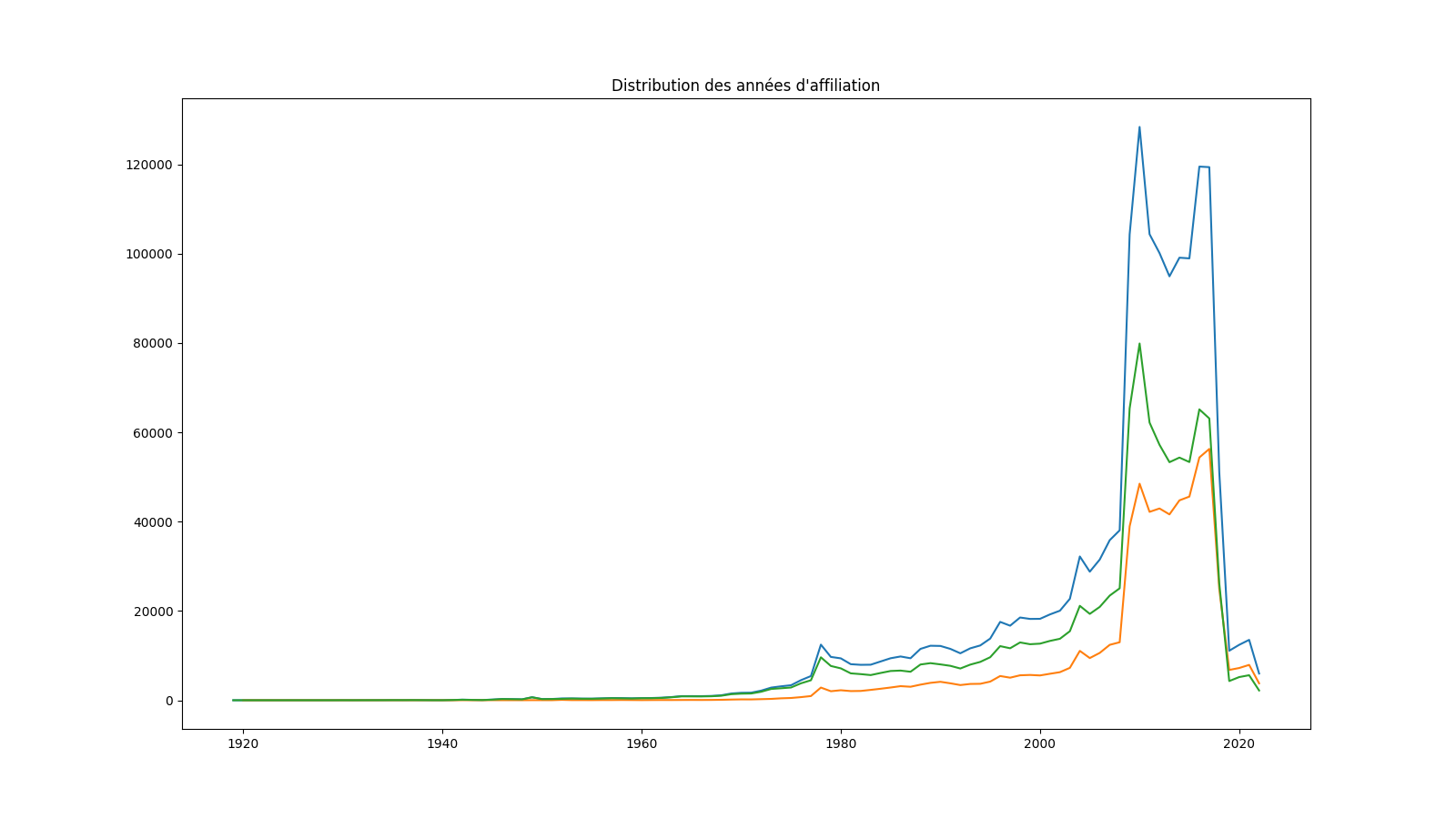
* Statistiques par génération

La majorité de nos adhérents sont nés entre les années 1940 et 1990.



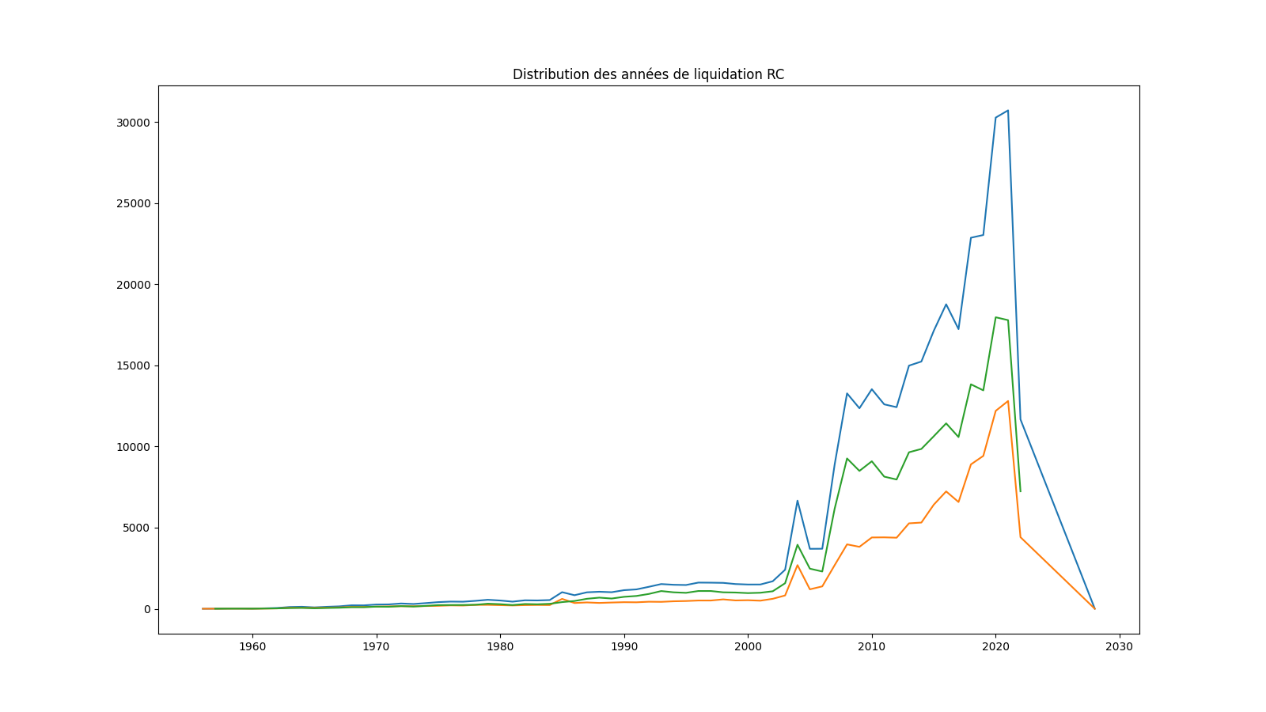
- Total - Homme - Femme

On dispose en moyenne de plus d’homme âgés en portefeuille, né avant 1960.

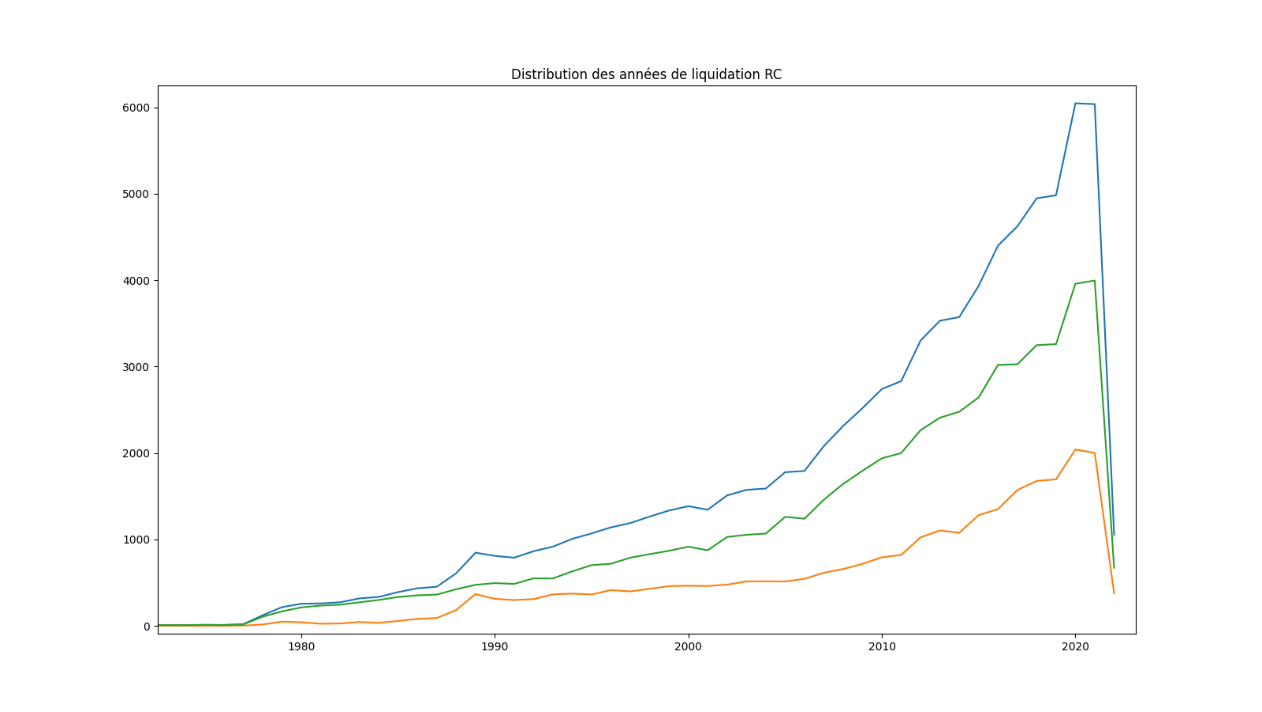


- Total - Homme - Femme

On constate une augmentation du nombre d’affiliation très marqué après 2009.

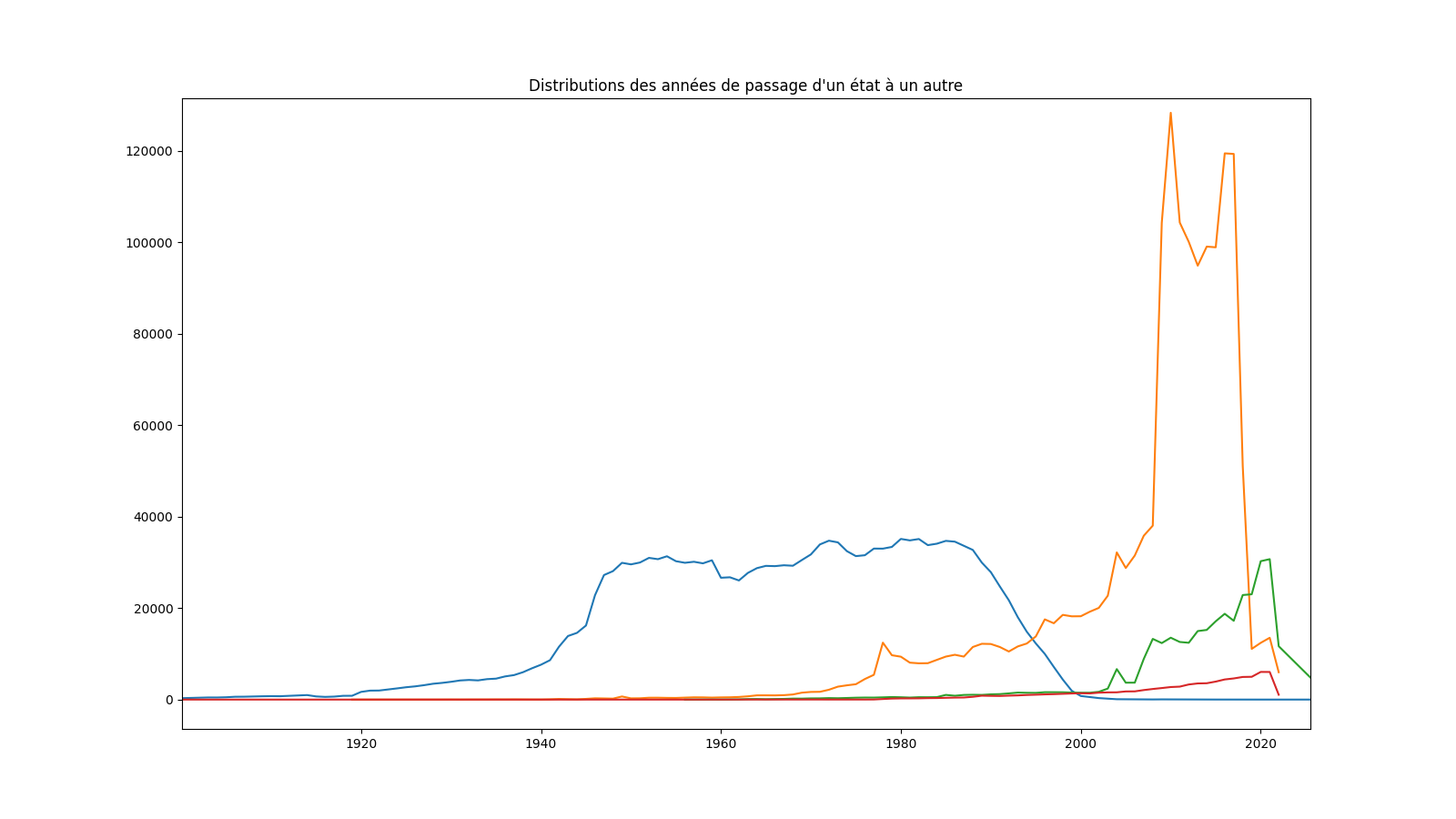


- Total - Homme - Femme



- Total - Homme - Femme

On résume sur un seul graphique :



- Naissance - Affiliation - Liquidation RC - Décès

# Projection des états

On travaille ici sur l’actualisation affilié par affilié de la chaine de Markov suivante. Cette chaine est parcourue à chaque âge de l’affilié. Les probabilités de passage sont définies par des **tables de passage** ne dépendant que de l’âge de l’affilié sauf pour le décès qui dépend aussi de la génération. Il serait théoriquement possible d’utiliser la génération pour définir les autres tables de passage, et pourquoi pas la profession, le sexe, ou la position géographique. Cependant, plus on considère de variables, plus on disperse la donnée et moins on dispose de volume statistique (par variable). Il faudrait, pour résoudre ce problème, utiliser des **modèles multivariés**.

Il serait également possible de surpondérer les données récentes dans les calculs des tables.

Les tables de passage sont lissées. On utilise un **lissage** convolutif simple (moyenne mobile). Le lissage Whittaker Henderson utilisé pour lisser la table de mortalité national n’est pas utile lorsqu’on utilise des tables unidimensionnelles.

Radiés

Radiation

Liquidation

Décès

Décès

Prestataires

Décès

Décédés

Actifs

Décès

Radiation

Liquidation

CER

Liquidation

Ayant-droit

**Les tables de passage sont définies sur les PL et ME ensemble**. On ne dispose pas de volumétrie nécessaire pour se permettre de distinguer les deux !

On rappel ici que dans les statuts, un affilié doit être **radié avant de liquider** ses droits RC. Un passage de cotisant actif à prestataire se fait donc par un rapide passage par l’état radié.

En ce qui concerne les **CER** (cumul emploi retraite), Ils peuvent liquider leurs droits sans être radié au préalable. Pendant cette phase, ils continuent d’avoir une activité rémunérée, et cotise tout en percevant une pension. On note que cette cotisation ne donne pas lieu à des droits. A l’arrêt de cette activité, ils sont radiés et passent ainsi en prestataire.

La **liquidation des droits des ayant-droit** nécessite la conjonction de deux évènements. D’une part, l’affilié ayant cotisé doit être décédé d’autre part l’ayant-droit doit atteindre ou avoir atteint 62 ans. La loi de liquidation résultante prend donc deux facteurs de nature différente et portant sur des individus différents. Elle est donc difficile à estimer. On retiendra un **taux de nuptialité** de 40% et un **taux de réversion** de 54%. Chaque nouveau décès d’un affilié pourra potentiellement engendrer un ayant-droit. L'âge de ce dernier sera tiré aléatoirement parmi la distribution des différences d'âge entre conjoints.

On prend comme loi future de **VFU** d’avoir moins de 180 points. On paie alors 15x le montant de la pension. Il est envisagé de revoir ces montants.

On ne considère pas ici les **majorations et minorations** des affiliés. Ceux-ci sont sensé être définit pour équilibrer le régime autour de l’âge pivot. En considérant que le gain perçu des cotisants liquidant plus tard compense la perte de ceux qui liquident plus tôt, on se permet ici d’omettre ce paramètre. En réalité les taux de majoration/minoration ne sont pas définit de façon à avoir un impact nul sur le régime. Ils dépendent fortement du régime de base, et sont donc exogène au régime RC. Le calcul de ces taux est donc à revoir pour affiner nos hypothèses. On note que le changement de calcul pourrait impacter le comportement de liquidation. Celui-ci sera donc à ajuster.

La méthode de cotisation choisie ici est **proportionnelle**. Le passage en proportionnelle n’impacte que faiblement les projections (impacte de 0.3% des réserves). On retient donc en hypothèse deux tranches de cotisation T1 et T2, allant respectivement de 0 à 1 PASS et de 1 à 4 PASS. Les taux respectifs sont de 9% et 22%. Cette méthodologie résulte d’une réforme statutaire de 2022.

Le taux de **rendement du point** est fixé à 5,6%. Il s’agit de la valeur anticipée sur les prochaines années. Il pourra être fait des hypothèses plus favorables sur cette valeur pour les années lointaines.

La **valeur de service du point** est fixée à 2,63. Cette valeur pourra être revalorisée, notamment en fonction de **l’inflation** future.

La **valeur d’achat du point** découle de ces deux grandeurs. Elle vaut donc 46,96.

**Les revenus** sont supposés constants. Plusieurs modèles de projections sont possibles. Certains scénarios probables seront exposés.

Certains affiliés **ne liquident jamais** leurs droits. On suppose que ce phénomène se poursuivra dans le temps dans les mêmes proportions. Il est tout de même raisonnable de penser que cette proportion pourrait bien diminuer à l’avenir, notamment depuis que la possibilité de liquider à été accordée aux débiteurs. Des hypothèses prudentes pourront être prise sur ce sujet.

Pour le moment, on ne prend pas en compte les prestations invalidité, décès et la rente au conjoint.

Le **taux de recouvrement** des cotisations est fixé à 98%.

En actuariat, on introduit les notions de **troncature** et de **censure** de la donnée pour traiter les données temporelles inconnues. Par exemple, on ne dispose pas de la date de décès d’un adhérent de 90 ans toujours vivant. Néanmoins, on dispose de l’information que son âge de décès est supérieur à 90. Il est nécessaire de rester prudent lorsque l’on définit une table de passage. Il est facile de compter le nombre d’adhérents radiés à un âge par exemple, mais il n’est pas évident de définir **l’assiette** par rapport à laquelle il faut ramener ce chiffre.

On définit trois méthodologies de calcul. La première effectue un comptage des adhérents et exclue les données tronquées. La seconde utilise l’estimateur de **Kaplan Meier**, pour obtenir une table de survie. Il nous reste à inverser cette table pour obtenir notre table de passage. La dernière utilise un **réseau de neurone** pour prédire directement les âges de radiation, liquidation et décès, tirant parti de l’ensemble de la donnée.

## Table de passage définie par comptage

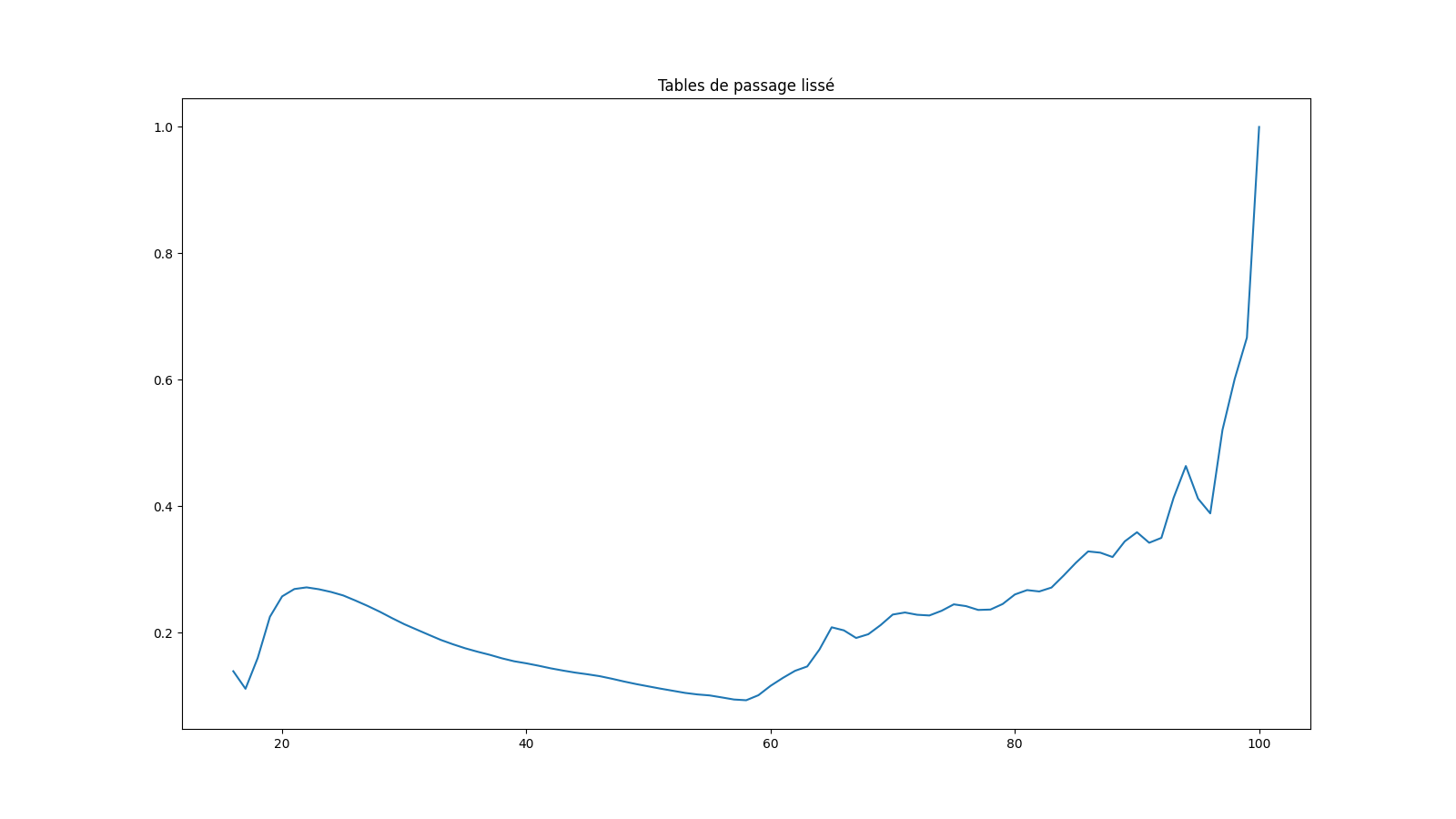
## Radiation des **actifs**

On établit la table de radiation en rapportant le nombre de radié sur une assiette de radié potentiel. On commence par exclure les ayant-droit (DD) qui ne peuvent être radié et les CER. On compte les radiés par âge. On définit l’assiette par âge comme étant le nombre d’adhérent :

* ayant déjà dépassé l’âge
* ayant été affilié avant l’âge
* n’ayant pas radié avant l’âge
* n’ayant pas liquidé avant l’âge
* n’ayant pas décédé avant l’âge
* Etant aujourd’hui radié, liquidé ou décédé

Autrement dit, on ne considère que les adhérents pouvant potentiellement être radié à un âge donné, et étant arrivé à un âge où ils ne peuvent plus être radié.

On peut aussi lisser cette courbe :



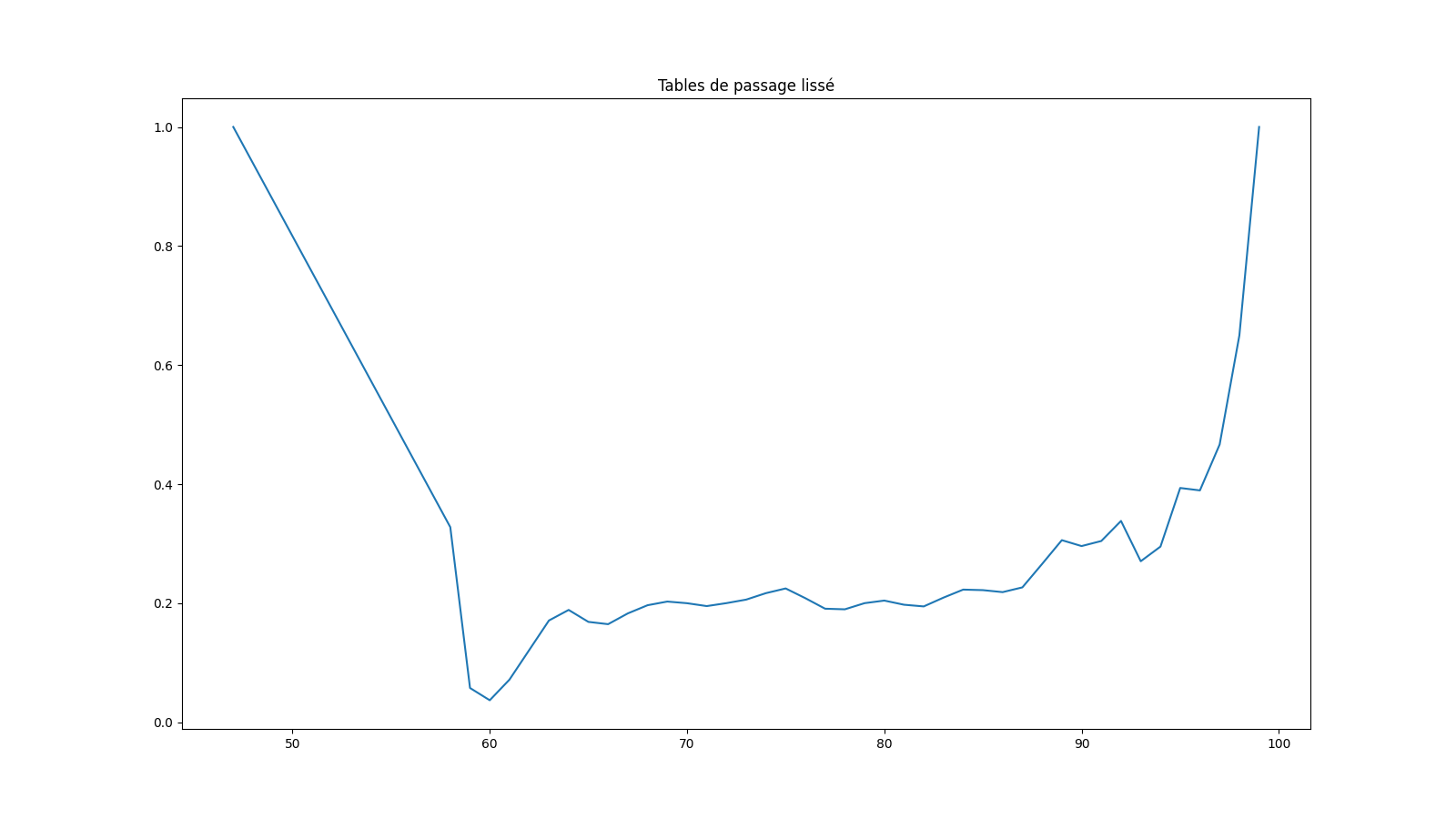
## Radiation des **CER**

On établit la table de radiation en rapportant le nombre de radié sur une assiette de radié potentiel. On commence par sélectionner les CER cotisant et prestataire. On note que les CER ont déjà liquidé mais sont encore cotisant. Après radiation ils deviendront prestataires. On compte les radiés par âge. On définit l’assiette par âge comme étant le nombre d’adhérent :

* ayant déjà dépassé l’âge
* ayant déjà liquidé avant l’âge
* ayant été affilié avant l’âge
* n’ayant pas radié avant l’âge
* n’ayant pas décédé avant l’âge
* Etant aujourd’hui radié, ou décédé

Autrement dit, on ne considère que les adhérents pouvant potentiellement être radié à un âge donné, et étant arrivé à un âge où ils ne peuvent plus être radié.

On note un taux très important à 50 ans. Cette valeur correspond probablement à une anomalie et pourrait être retraitée.

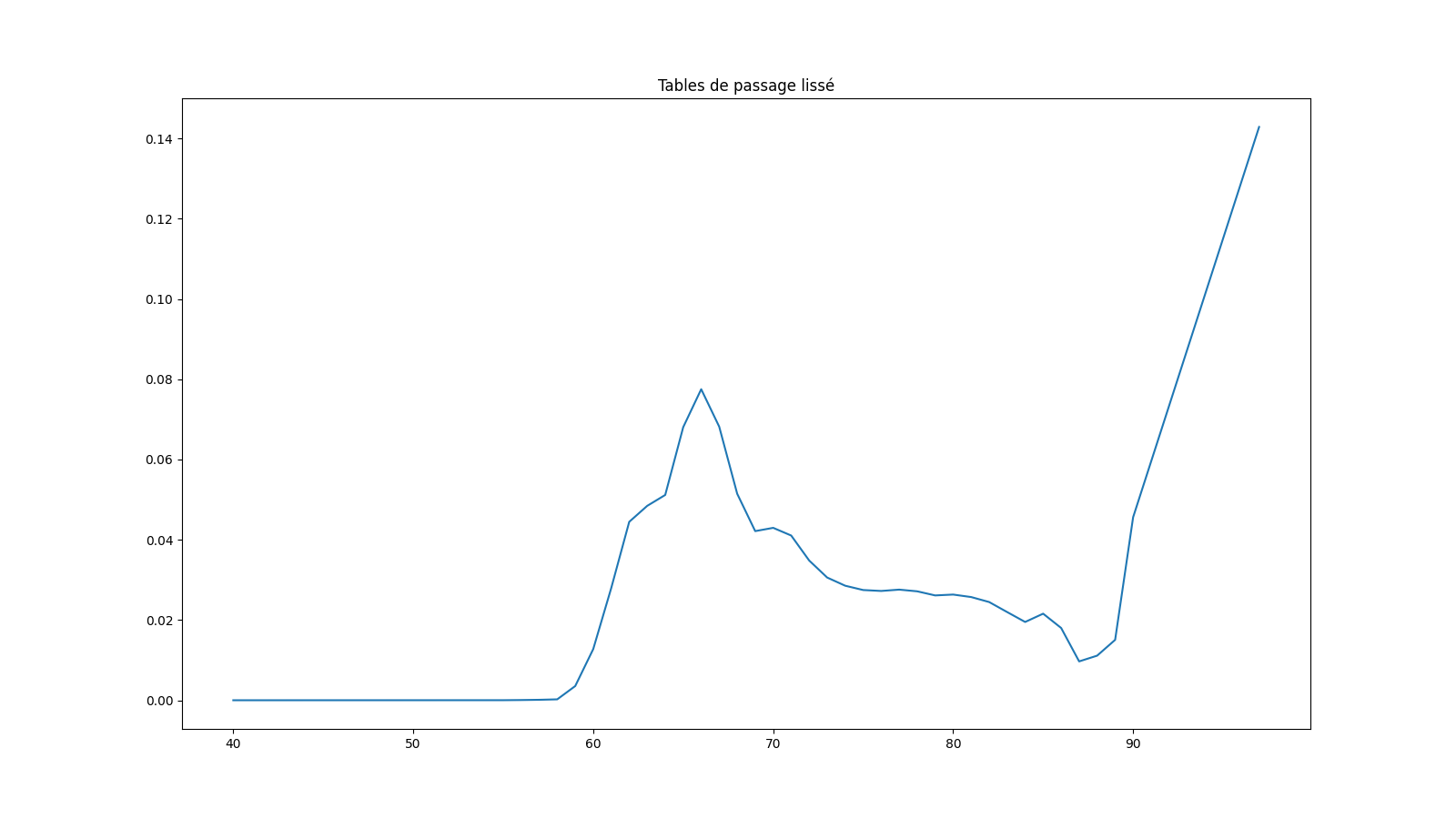


## Liquidation des **actifs** en **CER**

On établit la table de liquidation en rapportant le nombre de liquidé sur une assiette de liquidé potentiel. On commence par exclure les ayant-droit (DD) et les radiés. Par âge, parmi les actifs, on compte les liquidés en CER (on a exclu les quelques adhérents qui liquident sans avoir radié). On définit l’assiette par âge comme étant le nombre d’adhérent :

* ayant déjà dépassé l’âge
* ayant été affilié avant l’âge
* n’ayant pas radié avant l’âge
* n’ayant pas liquidé avant l’âge
* n’ayant pas décédé avant l’âge
* Etant aujourd’hui radié, liquidé ou décédé

Autrement dit, on ne considère que les adhérents pouvant potentiellement liquider à un âge donné, et étant arrivé à un âge où ils ne sont plus actifs.

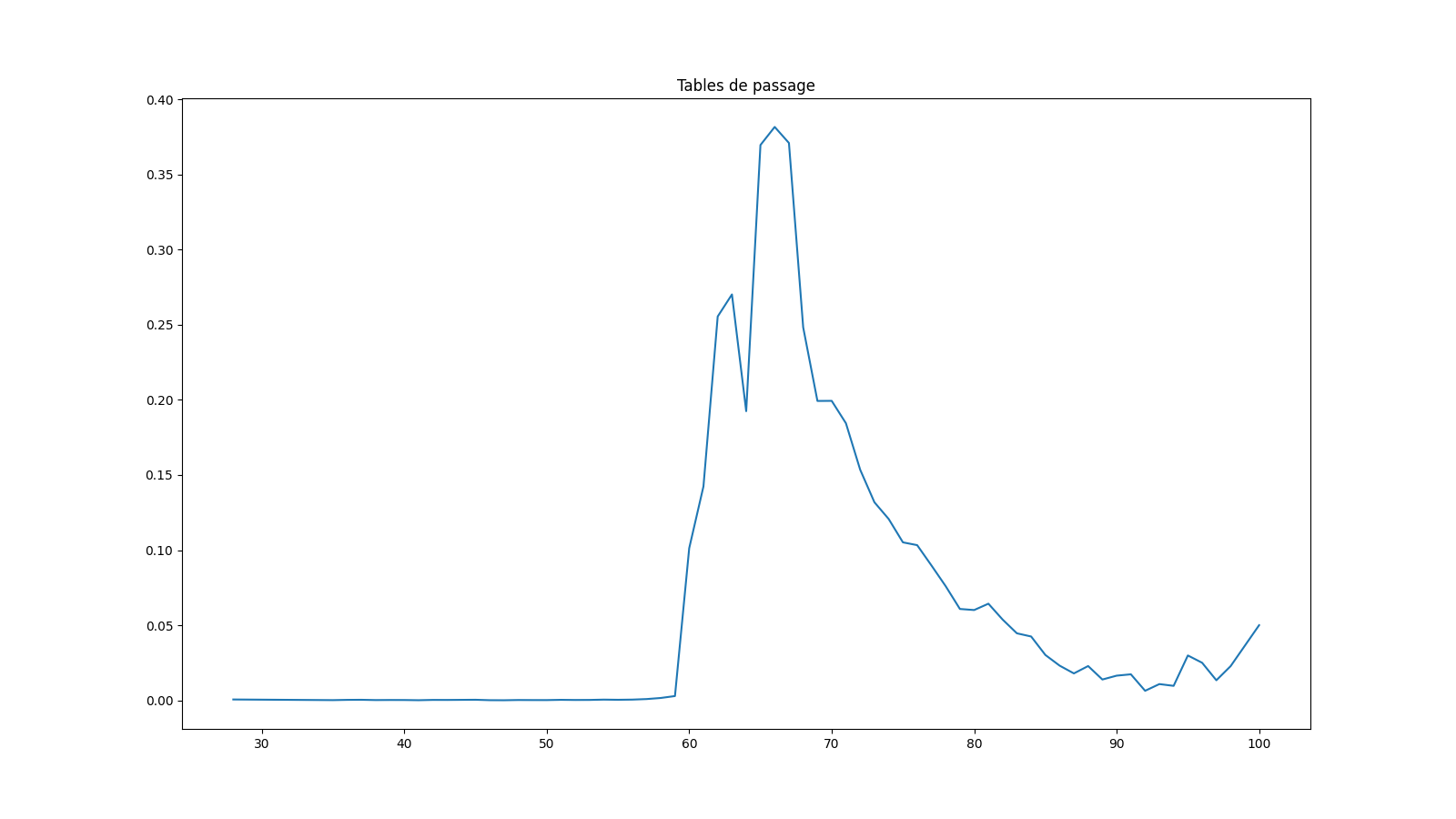


## Liquidation des **Radiés**

On établit la table de liquidation en rapportant le nombre de liquidé sur une assiette de liquidé potentiel. On commence par sélectionner radiés. On compte les liquidés par âge. On définit l’assiette par âge comme étant le nombre d’adhérent :

* ayant déjà dépassé l’âge
* ayant déjà été radié avant l’âge
* ayant été affilié avant l’âge
* n’ayant pas liquidé avant l’âge
* n’ayant pas décédé avant l’âge
* Etant aujourd’hui liquidé, ou décédé

Autrement dit, on ne considère que les adhérents pouvant potentiellement liquider à un âge donné, et étant arrivé à un âge où ils ne peuvent plus liquider.



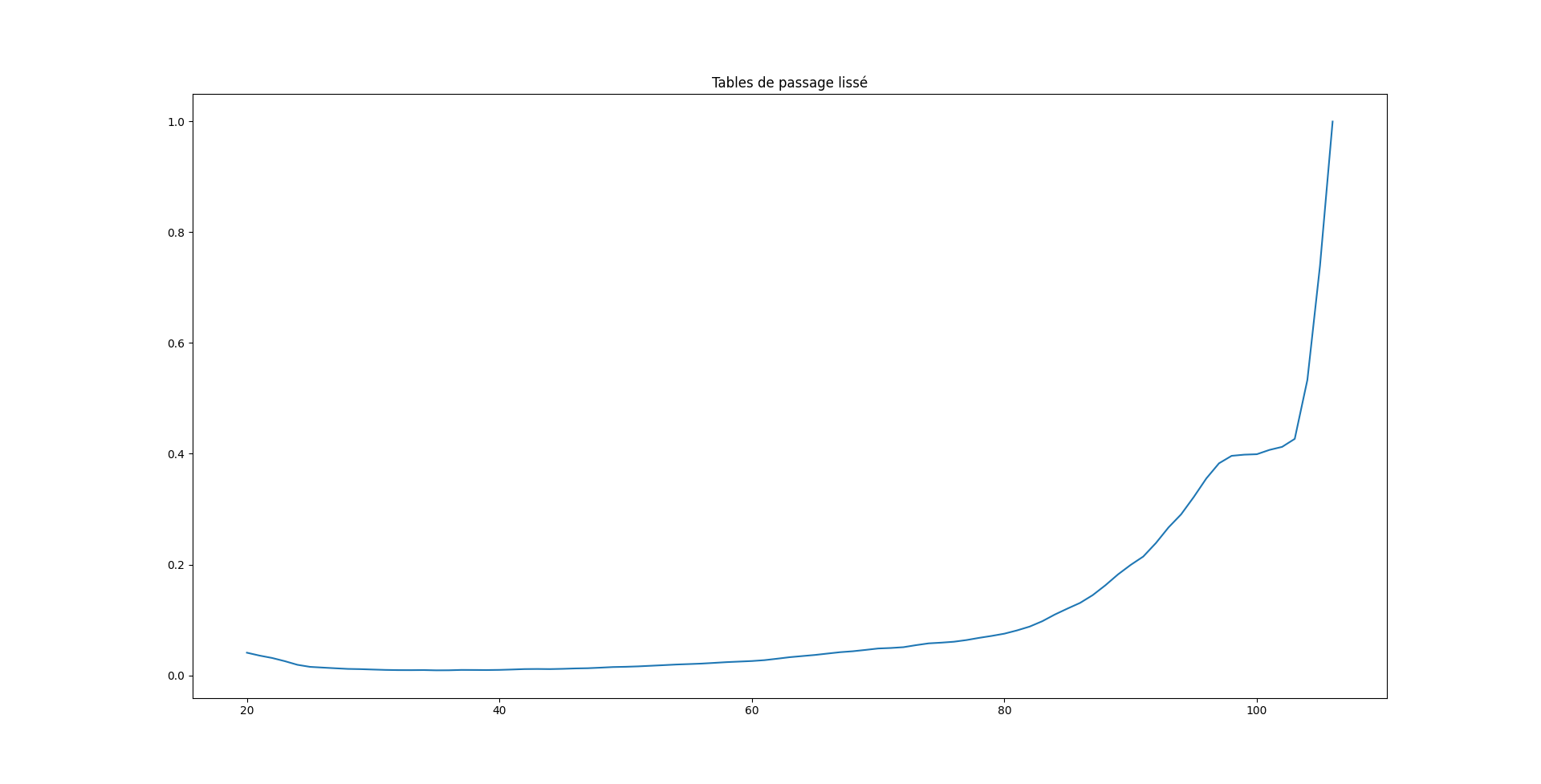
Remarque : On note que la loi n’atteint pas 100%, un nombre important d’adhérents radiés ne liquident jamais

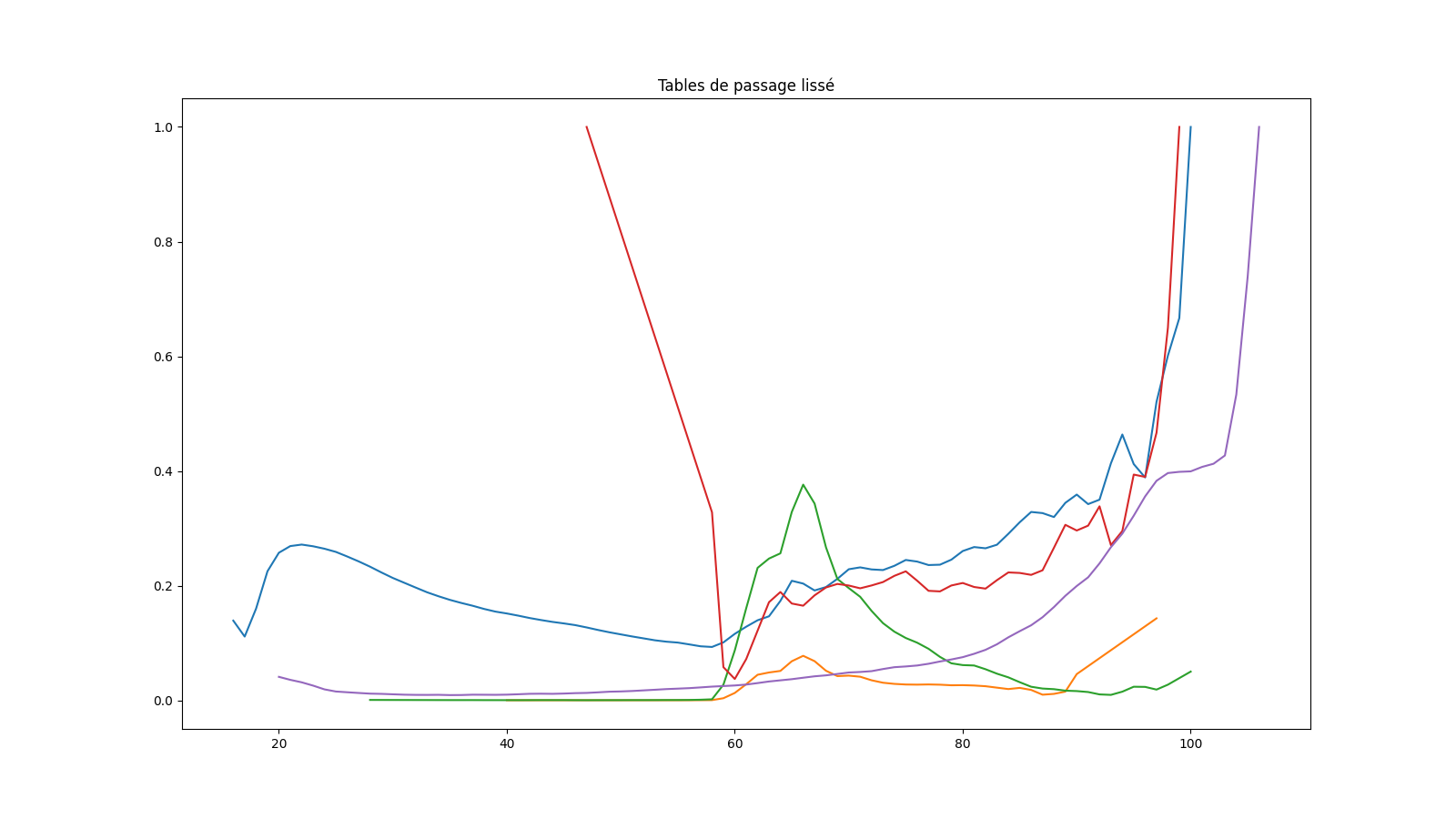
## Loi de décès

On établit la table de décès en rapportant le nombre de décès sur une assiette de décédés potentiel. On compte les décès par âge. On définit l’assiette par âge comme étant le nombre d’adhérent :

* ayant déjà dépassé l’âge
* ayant été affilié avant l’âge
* n’ayant pas décédé avant l’âge
* Etant aujourd’hui décédé

Autrement dit, on ne considère que les adhérents pouvant potentiellement liquider à un âge donné, et étant arrivé à un âge où ils ne peuvent plus liquider.





## Table de passage par Kaplan-Meier

Pour appliquer l’estimateur de **Kaplan-Meier**, il nous faut la durée d’observation, ainsi que l’information de censure. La difficulté de cette méthode est de traiter la **troncature à gauche** de la donnée. En effet, un adhérent n’est pas exposé à une potentielle radiation/liquidation/décès avant d’avoir été affilié. Il nous faut donc commencer nos observations à partir de la date d’affiliation. On n’observera donc plus l’âge de l’adhérent, mais sa **durée de cotisation**.

L’estimateur de Kaplan-Meier est l’estimation du maximum de vraisemblance non-paramétrique de S(t). C’est un produit de la forme :

* Lorsqu’il n’y aucune censure, est le nombre de survivants juste avant le temps .
* Lorsqu’il y a censure, est le nombre de survivants moins le nombre de pertes (cas censurés). Ce sont seulement ces cas survivants qui continuent à être observés (qui n’ont pas encore é censurés) qui sont « à risque » (observé).

*https://fr.wikipedia.org/wiki/Estimateur\_de\_Kaplan-Meier*

La durée d’observation représente la durée de cotisation de l’adhérent. Elle commence à partir de la date d’affiliation et se termine soit à la date minimale de radiation, liquidation, décès, soit à la date de calcul si l’affilié est toujours cotisant. La durée d’observation pour ces trois tables de passage est donc la même. En revanche, la **censure** va changer.

Passage de cotisant à radié :

Pour la radiation, un décès ou une liquidation prématurée (avant radiation) constitue une censure. La période d’observation n’est donc pas complète. Evidement un adhérent qui n’aurait été ni radié ni liquidé ni décédé sera censuré.

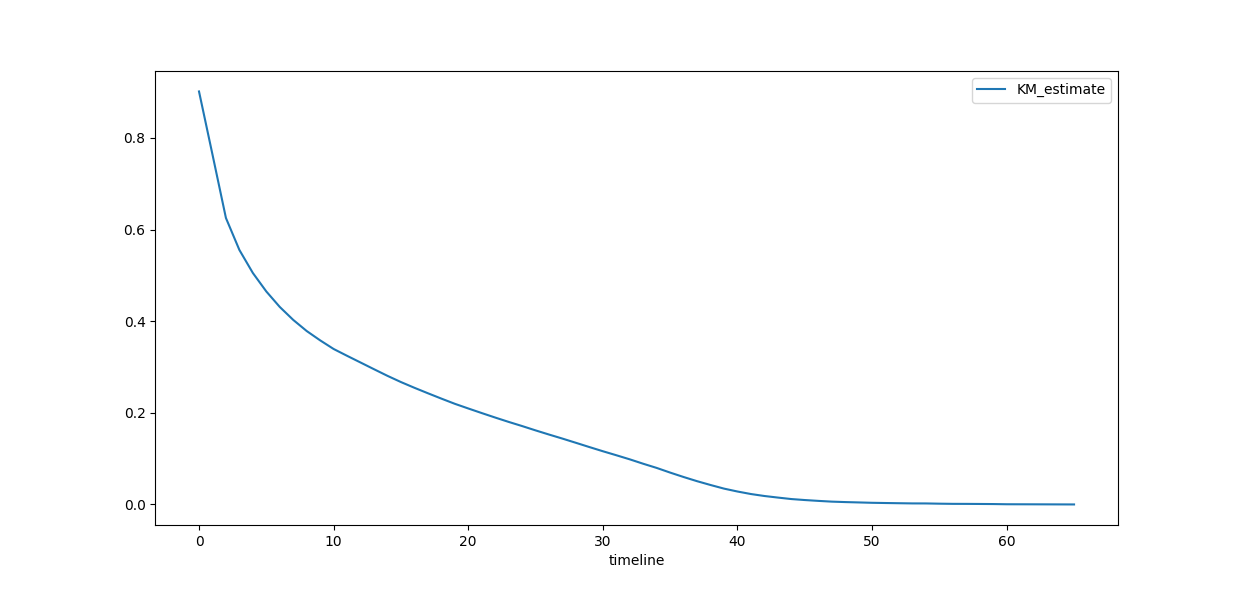


Table de survie de radiation des actifs (PL)

On constate qu’après 5 ans, 50% des cotisants sont déjà radié (s’ils n’ont pas liquidé ou décédé avant). On observe que la chute n’est pas tout à fait une décroissance exponentielle. On a en effectivement une **population hétérogène**, dont une moitié ne cotisent que peu de temps, et disparait en 5 ans, tandis que l’autre partie met encore 35 ans avant de disparaitre. Une meilleure segmentation de la donnée par des **algorithmes de classification** permettrait de fiabiliser d’avantage les projections.

Passage de cotisant à CER :

La période d’observation reste la même. En revanche la censure concernera tous les adhérents qui auront été radié avant liquidation (cas classique), tous les adhérents qui n’auront pas encore liquidé, ou tous les adhérents décédés.

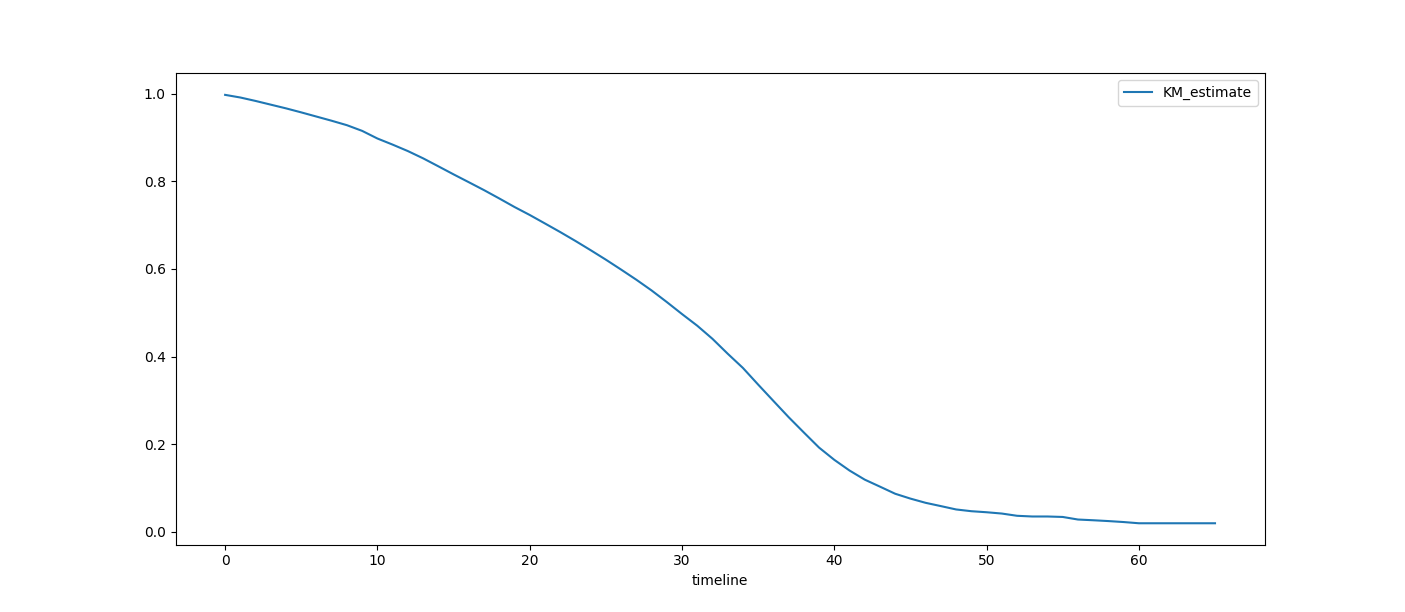


Table de survie de liquidation depuis l’état d’actif (PL)

On constate que la durée médiane de cotisation d’un CER est de 30 ans environ. Ce sont donc plutôt des profile de cotisants lourds.

Passage de cotisant à décédé :

On applique également ce calcul pour estimer la mortalité. On pourra ainsi comparer les résultats avec la table de mortalité et constater s’il apparait une dérive de mortalité. Ici on censurera tous les cotisants non décédés.

Passage de radié à prestataire :

Ici on ne considère que les adhérents qui ont déjà été radié. La censure concerne les décès et les adhérents encore radié aujourd’hui.

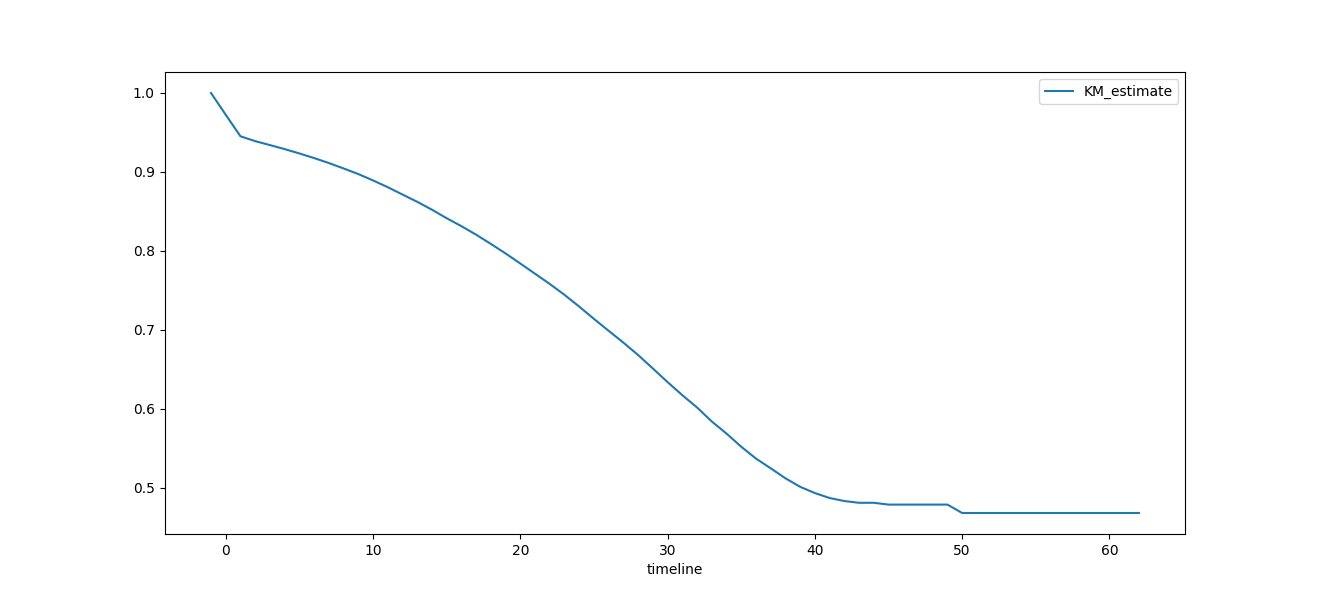


Table de survie de liquidation depuis l’état de radié (PL)

On constate que la table est moins régulière du au plus faible volume de donnée. En outre la table ne descend pas jusqu’à 0 car certains adhérents ne liquident jamais leurs droits.

Passage de radié à décédé :

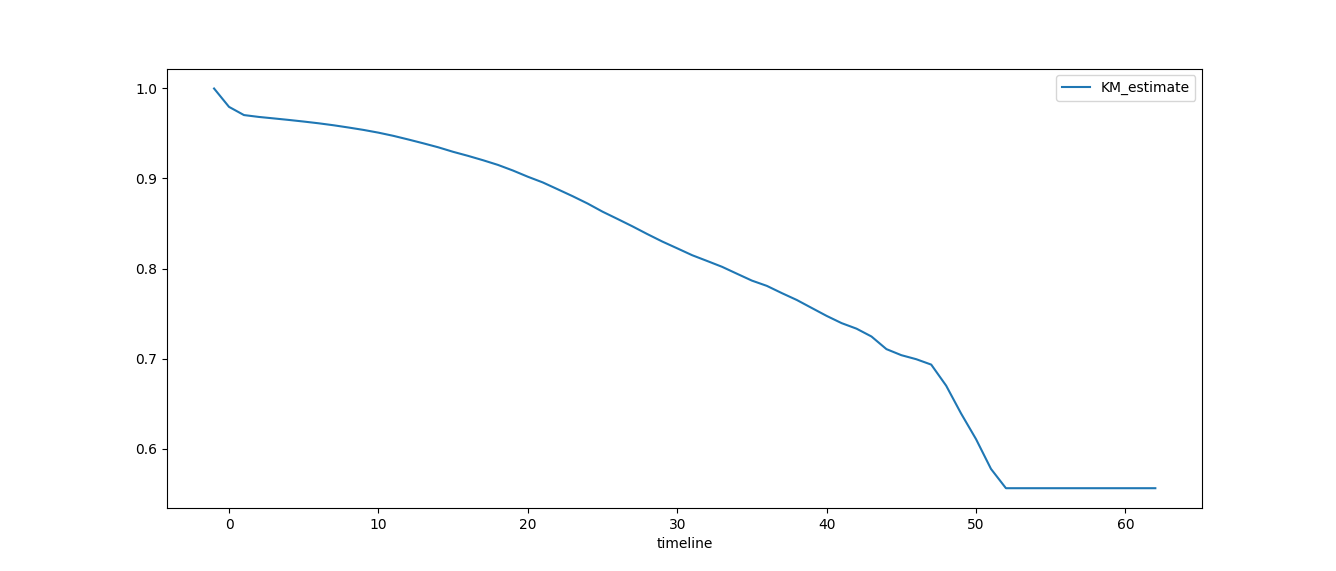


Table de survie depuis l’état de radié (PL)

On constate que les radiés ont une faible probabilité de décéder. On observe un volume de survivants encore important à 45 ans. Ce sont les adhérents qui ne liquident jamais. Ce volume tombe à 0 après 50 ans car il n’existe plus de données passé cet âge.

Passage de CER à prestataire :

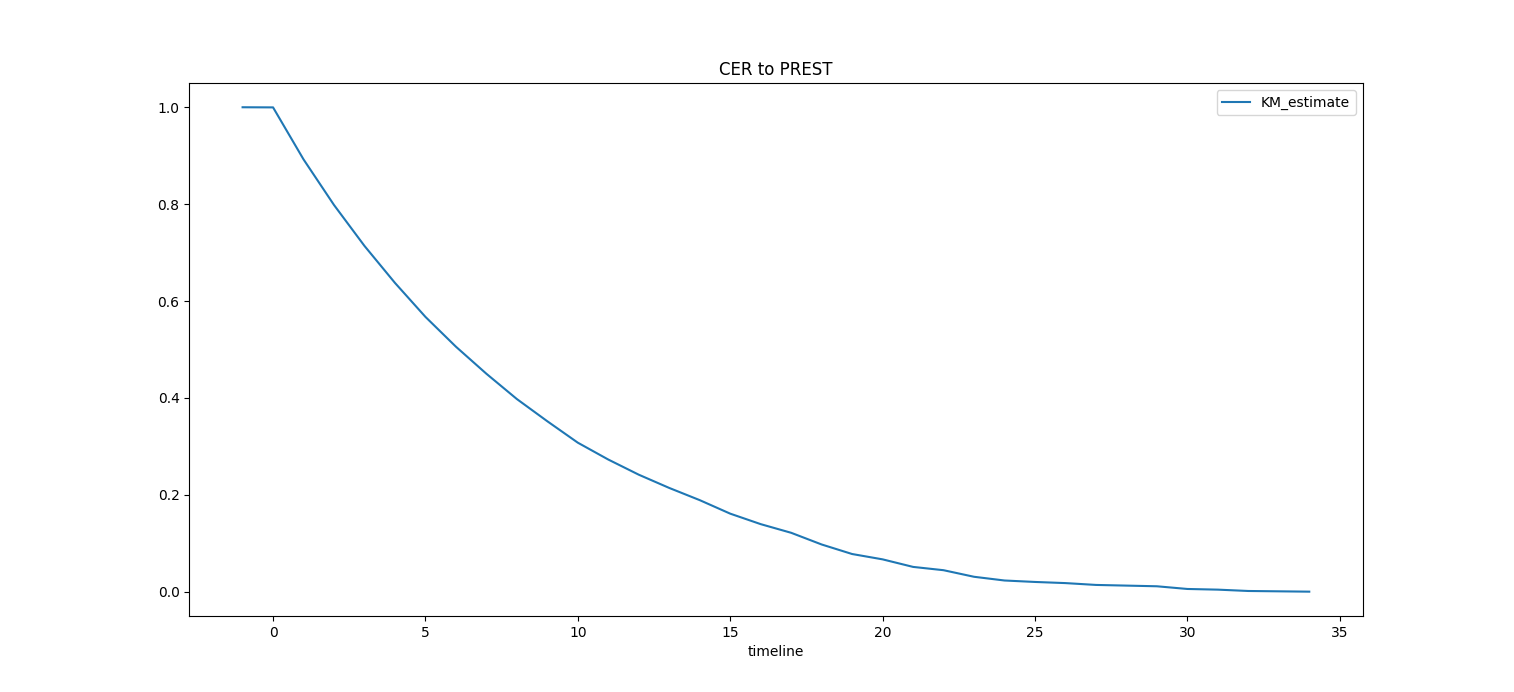


Table de survie de liquidation depuis l’état de CER (PL)

Passage de prestataire à décédé :

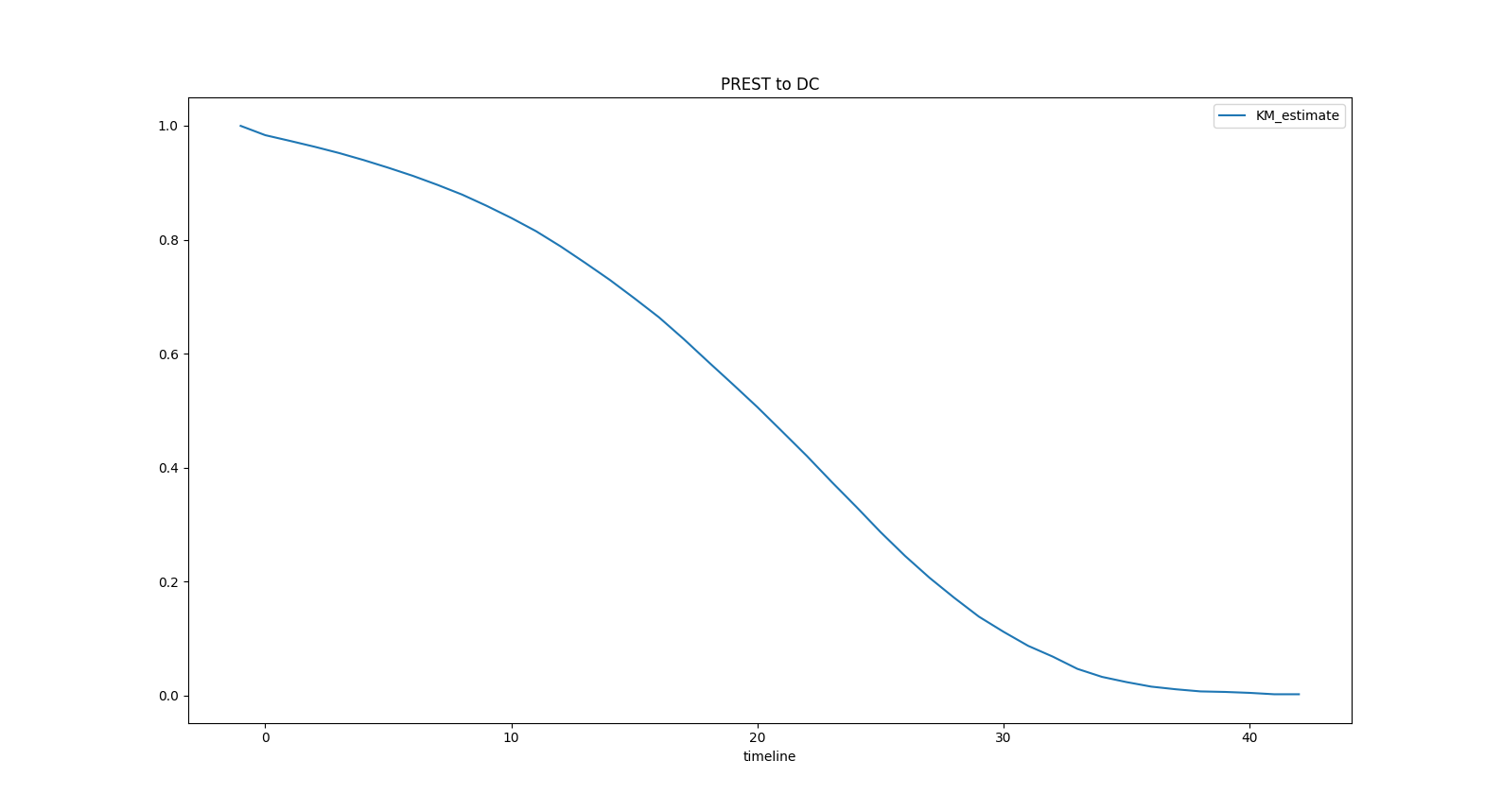


Table de survie depuis l’état de prestataire (PL)

Attention, cette table n’est pas directement comparable à une table de mortalité. On rappel qu’une troncature à gauche à été effectuée de façon à ne sélectionner que les adhérents prestataires. Les abscisses ne représentent donc pas les âges mais les années depuis liquidation des droits (ou l’arrêt d’activité dans le cas des CER).

## Table de passage par modèle multivarié – Réseau de neurones

Des modèles multivariés permettent de tirer partie de la totalité de la donnée dont nous disposons. Les variables les plus discriminantes (comme l’âge, la génération ou le sexe) seront les plus explicatives, tandis que les variables redondantes seront naturellement filtrées. On utilise ainsi un modèle de réseau de neurones pour prédire simultanément l’âge de radiation, l’âge de liquidation et l’âge de décès.

Un réseau de neurones (NN) est un modèle statistique capable d’apprendre une fonction de transfert entre des variables X d’entré et une sortie Y.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **X4** | **Y** |
| 0,45 | 0,58 | 0,33 | 0,10 | 0,93 |
| 0,81 | 0,53 | 0,25 | 0,78 | 0,11 |
| 0,83 | 0,87 | 0,43 | 0,51 | 0,18 |
| 0,21 | 0,99 | 0,59 | 0,29 | 0,66 |
| 0,65 | 0,56 | 0,89 | 0,37 | 0,51 |
| 0,38 | 0,91 | 0,92 | 0,19 | 0,20 |
| 0,56 | 0,67 | 0,91 | 0,98 | 0,67 |
| 0,10 | 0,13 | 0,23 | 0,87 | 0,63 |
| … | … | … | … | … |

*Exemple de données d'entrée numériques Xi et de sortie Y*

Le NN est un modèle apprenant qui doit être entrainé lors d’une phase d’apprentissage avant d’être utilisé (phase d’inférence). Pour l’entrainement, il est nécessaire de lui fournir les variables d’entrée X ainsi que les labels Y associés.

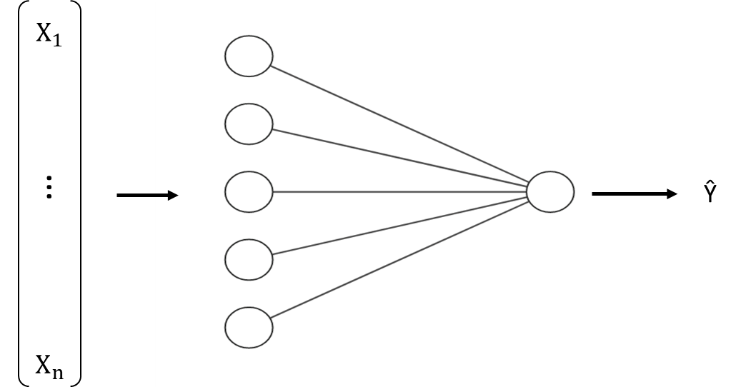
Architecture :

Un neurone est une combinaison linéaire des variables d’entrée :

Avec :

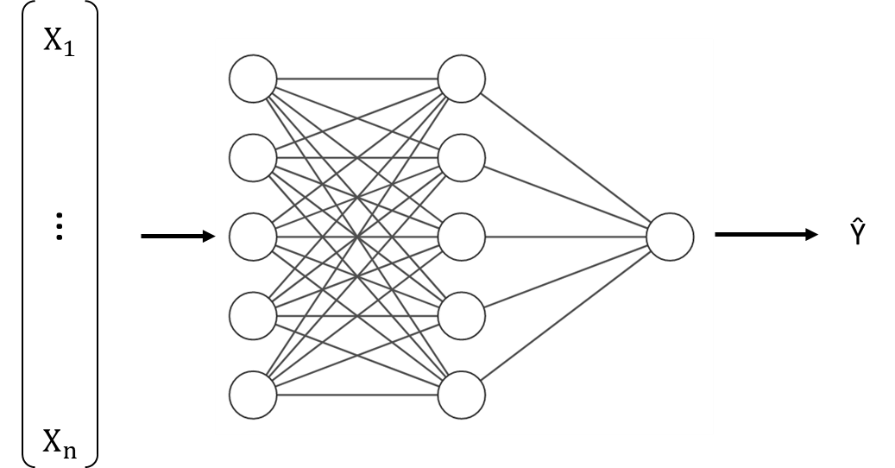
Un neurone seul est capable de réaliser une régression linéaire. Elle consiste à déterminer les poids et le biais *b* de façon à minimiser l’erreur entre la sortie et la donnée *Y.*

Un réseau de neurones utilise plusieurs neurones disposés en série et en parallèle :



*Schéma d'un réseau avec 5 neurones et une sortie*

Un réseau est profond lorsqu’il possède plusieurs couches (dites couches cachées) dont les neurones prennent en entrée l’intégralité des sorties de la couche précédente.



*Schéma d'un réseau avec 2 couches de 5 neurones et une sortie*

La fonction d’activation (souvent sigmoid, ReLu ou Tanh) est une fonction non linéaire dont le but est de casser la linéarité du réseau et de lui permettre d’apprendre des fonctions complexes.

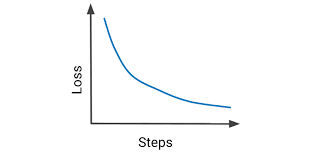
Entrainement :

Comme pour la régression linéaire, entrainer un réseau de neurones consiste à déterminer l’ensembles des poids et des biais de chaque neurone de façon à minimiser l’erreur globale.

L’apprentissage est réalisé par la méthode de la descente de gradient comme suit :

* Le réseau est initialisé avec des poids aléatoires
* La donnée est fournie au réseau qui donne une prédiction Ŷ
* On calcule une erreur quadratique moyenne avec la donnée réelle L = mean( ( Ŷ – Y )² )
* On dérive l’erreur par rapport à chaque poids
* On met à jour les poids :
  + Avec : le taux d’apprentissage (constante)

Ces étapes sont répétées sur plusieurs itérations (ou epoch) jusqu’à atteindre une erreur minimale.

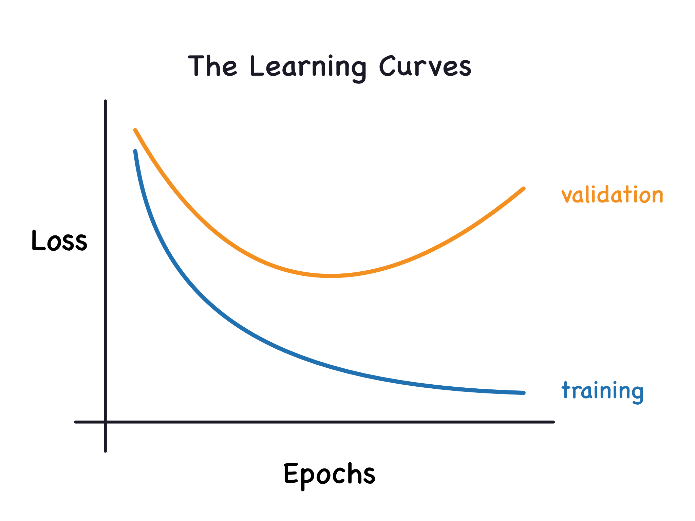


*Diminution de l’erreur au cours des itérations*

Optimisation :

Pour évaluer un modèle, on divise la donnée disponible en deux sets de données. Un set d’entrainement et un set de validation. L’entrainement est réalisé uniquement sur le set d’entrainement tandis que le set de validation permet d’évaluer le modèle au cours de son apprentissage.

Un réseau de neurones est capable de sur-apprendre la donnée d’apprentissage. C’est-à-dire qu’il apprend « par cœur » la donnée plutôt que de retrouver un modèle qui l’explique. Il aura alors une faible capacité de généralisation à de la donnée inconnue. Son erreur de validation sera alors élevée par rapport à son erreur d’entrainement.



*Courbes d’apprentissage caractéristiques d’un sur-apprentissage.*

Pour s’assurer d’une capacité de généralisation du modèle, il est nécessaire de choisir un réseau et une méthode d’apprentissage qui permettent de minimiser l’erreur de validation. Pour cela on réalise une optimisation des hyperparamètres. Ce sont les paramètres qui définissent le réseau et la méthode d’apprentissage, tels que le nombre de couches, de neurones, le taux d’apprentissage, etc…

L’optimisation consiste à essayer plusieurs configurations de paramètres jusqu’à trouver celle qui donnera la meilleure performance de validation. Pour ce faire, on utilise en pratique l’algorithme d’optimisation bayésienne.

Traitement de la donnée :

On ne s’intéresse ici qu’au droit propre (incluant les CER). Il est nécessaire d’effectuer quelques traitements préalables pour obtenir une base adaptée au réseau de neurones. Tout d’abord, le réseau ne traite que des **données numériques**. Il est donc nécessaire d’encoder les **variables catégorielles** pour obtenir des variables numériques. Une option simple serait de simplement remplacer les catégories par des nombres. Cependant, pour des variables comme la profession par exemple, on dispose de 90 valeurs différents (après avoir regrouper les professions rares dans ‘AUTRE’). Avec un tel encodage, le réseau risque de surpondérer l’importance de la 90e valeur comparée à la première (celle-ci vaux 90x plus, et rend donc négligeable la première). L’encodage choisi est donc **l’encodage ‘1 HOT’** qui scinde la colonne en 90 colonnes booléennes différentes qui prennent la valeur 1 si la profession est bien exercée par l’adhérent, et 0 pour sur toutes les autres colonnes. En ce qui concerne les **variables numériques**, certaines peuvent être supérieur à d’autres de plusieurs multiples, comme le nombre de points comparé à l’âge d’affiliation. Ces variables sont donc toutes **centrés et réduites**.

Méthodologie :

Le réseau va prédire simultanément l’âge de radiation, l’âge de liquidation et l’âge de décès. Ces variables sont donc des **variables cibles** (des Y). Or ces données sont incomplètes (on veut justement les compléter). Un réseau de neurone ne peut pas traiter des données incomplètes. Il faut donc les compléter sans biaiser la donnée.

On commence par enregistrer l’information de présence de la donnée dans des **variables booléennes**. On complète ensuite les **données manquantes** :

* Pour l’âge de décès, on complète les valeurs manquantes par l’âge actuel à la date de calcul.
* Pour l’âge de radiation, on complète la donnée manquante par l’âge de décès si celui-ci est présent, sinon par l’âge actuel.
* Pour l’âge de liquidation, on complète la donnée manquante par l’âge de décès si celui-ci est présent, sinon par l’âge actuel.

Il nous reste à définir une **fonction de cout cohérente** pour pénaliser correctement nos projections et **traiter les censures**.

Notre fonction de cout est la somme des fonctions de cout des trois variables cibles. Partout où la donnée est présente, le cout sera la **somme des écarts au carré avec la donnée réelle** (moindre carrés). Lorsque la donnée est manquante, remplacée par l’âge actuel (ou l’âge de décès), la fonction de cout **ne pénalisera que lorsque la prédiction sera inférieure à la valeur remplacée**, en carré également. Ainsi, pour un adhérent qui a aujourd’hui 62 ans et n’a pas liquidé (valeur manquante et remplacé par 62), si la prédiction est de 60, elle sera pénalisée (de ), mais si elle est de 65, elle ne le sera pas. La fonction de cout sera donc nulle.

Le modèle se rapprochera donc des valeurs réelles lorsqu’elles sont présentes, et pénalisera les prédictions impossibles là où la valeur est manquante. On prend ainsi en compte les **données censurées**. Tout comme pour Kaplan-Meier, cette prise en compte à pour effet de rehausser la prédiction.

Amélioration possible :

Cependant, l’âge de radiation, si elle est déjà présente en base, donne beaucoup d’information sur l’âge de liquidation. De même, l’âge de liquidation donne beaucoup d’information sur l’âge de décès. Il est donc nécessaire que ces variables soient également des données d’entré (des X) pour tirer parti d’un maximum d’information. Une nouvelle fonction de cout doit alors être déterminée.

Validation du modèle :

Pour s’assurer de la **performance du modèle** et éviter le **surapprentissage**, il est nécessaire d’effectuer une validation. On scinde donc nos données en un dataset d’apprentissage et un dataset de validation. Le modèle sera entrainé sur le dataset d’apprentissage et testé sur le dataset de validation. Si l’écart observé est faible, le modèle est efficace. Le dataset de validation nécessite que les trois âges soient déjà présents (pour comparer la prédiction à des données réelles). Or les adhérents dont toutes les dates sont renseignées sont ceux qui détiennent le plus d’information, ils sont donc importants pour l’entrainement du modèle. On sélectionne 550 adhérents pour notre base de validation.

On calcul l’écart moyen en valeur absolue (norme 1) à la donnée de validation. Pour ce qui est de **l’âge de radiation, cet écart est de 3,5 ans**. Pour **l’âge de liquidation il est de 3 ans**. Naturellement, la liquidation est plus précise car elle tire parti de l’information disponible sur l’âge de radiation. En revanche l’âge de décès est naturellement plus difficile à prédire, les causes sont multiples et la volatilité de cette donnée est plus forte. **L’écart est de 5 ans**. Il est possible de comparer cet écart avec cellui que l’on aurait eu avec la table de mortalité national. Cependant, il ne serait pas pertinent de mesurer la date de décès de ces adhérents à la naissance, celle-ci serait naturellement sous-évaluée. Il nous faut donc définir à partir de quel âge on choisit de prédire l’âge de décès. Une première option est de choisir l’âge moyen des adhérents non décédés en portefeuille. L’écart est alors de 13,5 ans. Une option plus favorable est de choisir un âge 5 ans inferieur à l’âge de décès réel de l’adhérent. L’écart est alors de 12.5 ans. On constate que le réseau de neurone est bien plus performant que la table de mortalité pour notre portefeuille.

Seuillage de la donnée :

Il faut garder en tête que le réseau de neurones est un modèle statistique. Il donne une « valeur moyenne » du résultat avec une certaine confiance. Il n’est donc pas impossible que certaines projections soient improbables (faible probabilité) voir impossible compte tenu des condition réelles. Par exemple, l’âge de radiation peut être projeté avant l’âge d’affiliation. Il nous faut donc seuiller les valeurs prédites de façon qu’elles ne soient pas incohérentes avec la donnée.

## Validation des hypothèses

## Hypothèse d’affiliations nouvelles

On relève 10 000 nouveaux affiliés par ans en 2019 – 2020. Notons que le périmètre futur des professions est incertain (PLFSS2018). Nous avons retenu une hypothèse prudente de 6000 nouveaux entrants par ans, pour les PL et les ME.

Pour créer de nouveaux affiliés tout en conservant la typologie du portefeuille adhérents (notamment les âges relatifs d’affiliation, de liquidation et de décès), nous utilisons une méthode d’échantillonnage. On tire ainsi aléatoirement des adhérents dans notre portefeuille, et on translate les dates de façon à obtenir la date d’affiliation voulue. L’uniformité du tirage garantit la conservation des proportions (Homme/Femme, PL/ME, Statut, …).

On précise que ces nouveaux entrants n’incluent pas l’entrée de nouveaux réversataires, projetés par ailleurs.

## Projections

Préciser que l’on procède par projection des ages

Préciser que le passé aussi est projeté : points par ans,

Attention extension des table de mortalité après 2005 en étandant la colonne 2005

On souhaite projeter les dates de radiation, liquidation et décès de tous les adhérents en portefeuille, ainsi que les réversataires. On projettera également les nouveaux adhérents.

On commence par **exclure les adhérents décédés** qui ont atteint leur état stationnaire. Comme la base contient plus d’un million de lignes, **on échantillonne 10% de la donnée**, pour permettre des calculs plus rapides. Il suffira de corriger les résultats d’un facteur 10 pour obtenir un estimateur satisfaisant.

Ainsi, **on parcourt la base d’adhérents ligne à ligne** (tête par tête) et pour chaque individu, on projette les âges de passage qui n’ont pas encore survenu. On utilise pour la projection des décès les **tables de mortalités générationnelles nationales** (TGH-05, TGF-05).

On va balayer les âges un à un depuis l’âge actuel de l’adhérent à l’âge terminal possible, soit ici 120 ans. Pour chaque âge, on effectue un **tirage aléatoire exclusif** entre les passages possibles.

Autrement dit, pour chaque âge futur de l’adhérent, il pourra, en fonction de son état actuel, soit se maintenir dans cet état, soit changer d’état. Pour ce tirage, le choix est exclusif, seul un passage est possible à la fois. Ainsi, à l’inverse des projections effectuées précédemment, on tire les probabilités de passage simultanément, et non successivement. L’avantage est de ne pas induire un biais dû à l’ordre de tirage.

Dans nos projections, on ne retient que les âges comme des valeurs entières. Ainsi, à l’âge actuel réel de l’adhérent, on ne sait pas s’il vient de célébrer son anniversaire, ou bien s’il est sur le point de le faire. Ainsi, si l’adhérent à une probabilité p de décéder à son âge, on ne sait pas s’il a déjà « tiré sa probabilité » (s’il est en fin d’âge) ou s’il doit encore la tirer. En supposant l’âge des adhérents équirépartis sur l’année, pour la première année de projection (l’âge actuel), on tire une pièce pour savoir si on effectue les projections pour cet âge ou si l’on passe à l’âge suivant.

On note cependant qu’il est possible qu’un adhérent soit radié et liquide ses droits la même année. Également, **tous autres passages d’état peuvent s’opérer la même année**. On définit donc des **taux intra-annuels de passage** sur l’historique. Après chaque tirage de passage, on effectue un second tirage (bernoulli dont la probabilité est le taux intra-annuel) pour potentiellement rechanger d’état. Les triples changements d’état intra-annuel, bien que possible, sont suffisamment rare pour être négligé.

Enfin, lors du décès d’un prestataire, on effectue un tirage selon le **taux de nuptialité**. Si celui-ci est positif, on créer en base un **nouvel ayant-droit**. Les seules variables qu’il est nécessaire de définir pour cet ayant droit seront celle nécessaire à calculer son âge de décès. Ainsi, son sexe sera simplement l’opposé du sexe de l’adhérent décédé (les couples homosexuels sont ici exclus afin de simplifier les calculs). L’âge du réversataire (et donc son année de naissance) est tiré selon une **distribution d’écart d’âge au conjoint** historique.

Ecart d’âge entre un adhérent et son réversataire

On trace donc la distribution historique d’écart d’âge. On constate qu’en moyenne, un adhérent (majoritairement masculin) a 2-3 ans de plus que son réversataire (majoritairement féminin). On note une queue de distribution longue à droite qui représente les orphelins.

Enfin, les **points attribués** au réversataire sont ceux de son conjoint, multiplié par un **taux de réversion** définit dans les statuts.

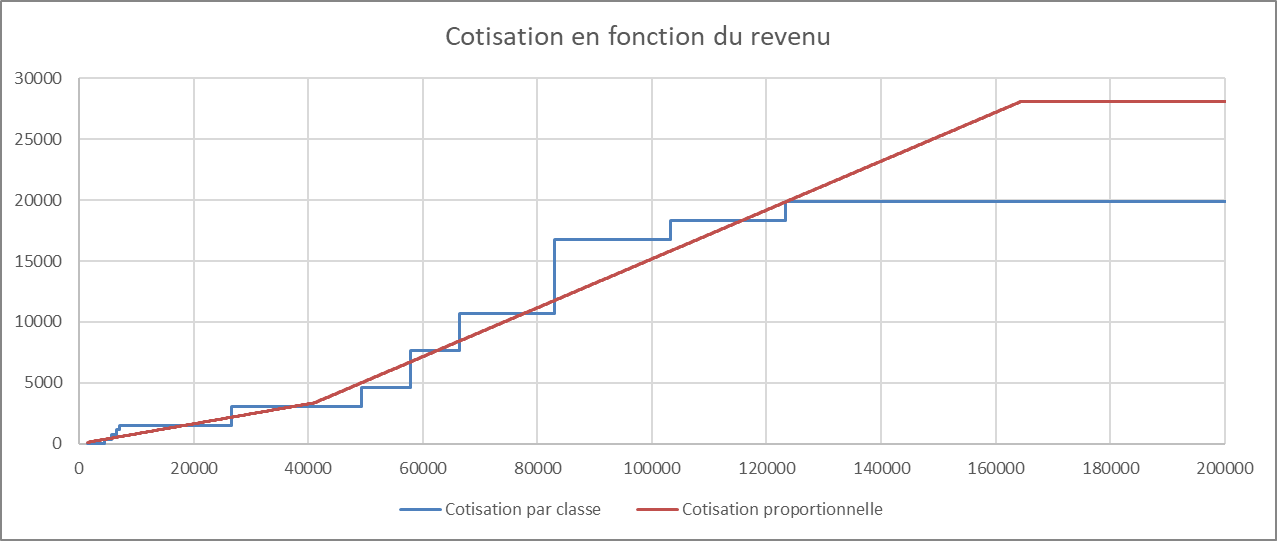
On projette alors le décès de ces ayants-droits avec la table de mortalité nationale.

## Projection des points

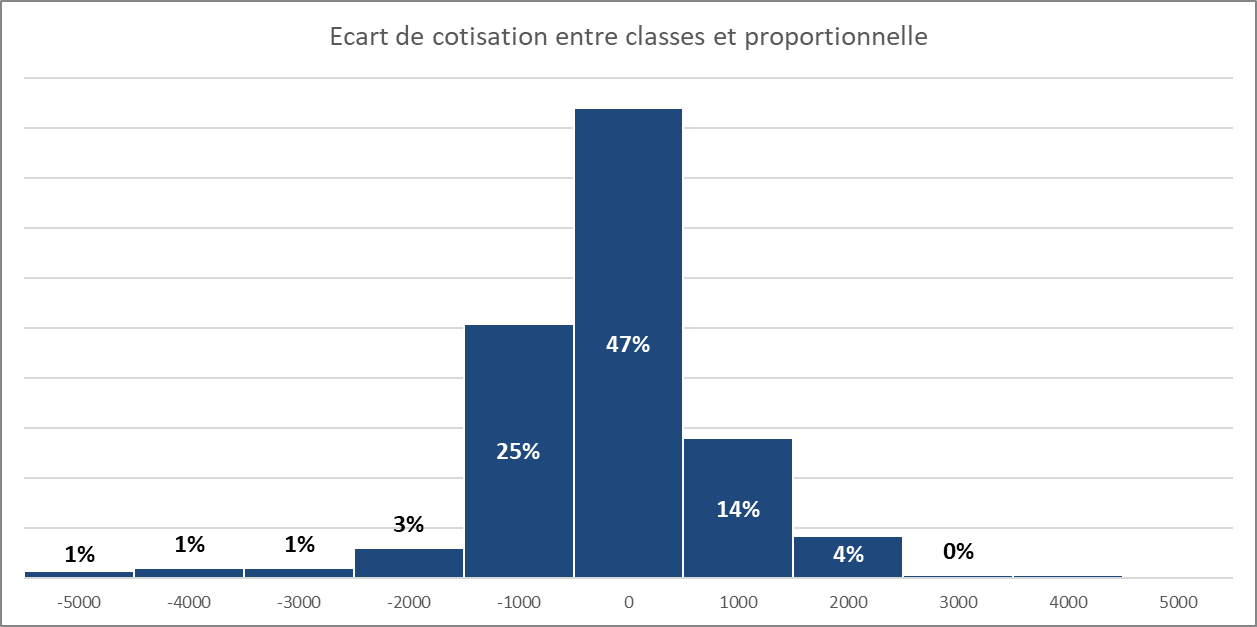
Le régime fonctionne par répartition, en distribuant les droits par points. Pour anticiper les cotisations ou prestations futurs, il nous faut pouvoir anticiper les points par adhérents. Ceux-ci dépendent historiquement du revenu de façon forfaitaire.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Revenus** | **Montant Cotisé** | **Points** |
| **Classe A** | Revenus < 26 580 € | 1 527 € | 36 |
| **Classe B** | de 26 581 € à 49 280 € | 3 055 € | 72 |
| **Classe C** | de 49 281 € à 57 850 € | 4 582 € | 108 |
| **Classe D** | de 57 851 € à 66 400 € | 7 637 € | 180 |
| **Classe E** | de 66 401 € à 83 060 € | 10 692 € | 252 |
| **Classe F** | de 83 061 € à 103 180 € | 16 802 € | 396 |
| **Classe G** | de 103 181 € à 123 300 € | 18 329 € | 432 |
| **Classe H** | Revenus > 123 300 € | 19 857 € | 468 |

Aujourd’hui, cette relation est proportionnelle, ou plus précisément proportionnelle par morceau. On dispose ainsi de deux taux de cotisations pour deux tranches de revenus, ainsi que d’un plafond de revenu à 4xPASS.



Pour le régime, ce changement de cotisation n’impacte que très peu le volume de cotisation percu.



On constate une distribution plutôt centrée sur 0. Le graphique n’affiche cependant pas les cotisants les plus fortunés qui dépassent le plafond des 4xPASS. Leur cotisation augmente significativement, de tel sorte que l’écart moyen est positif :

|  |  |
| --- | --- |
| **Moyenne des écarts** | 113 |
| **Médiane des écarts** | -61 |
| **Minimum des écarts** | -5025 |
| **Maximum des écarts** | 8263 |
| **Taux d'adhérents qui cotisent plus** | 47% |

Par la suite on ne s’intéressera plus au revenu mais uniquement aux points. En particulier, on s’intéresse au nombre de **points moyen cotisé par an**. Pour le calculer, on ramène le nombre de points acquis sur la durée de cotisation, à savoir, l’âge de dernière cotisation (minimum entre l’âge actuel et les âges de radiation, liquidation, et décès) moins l’âge d’affiliation. On suppose ici qu’il n’y a **pas d’évolution du revenu** avec le temps. Historiquement, celui-ci ne s’est révélé que légèrement supérieur à l’inflation. A cet effet, il faut ajouter la hausse historique de la valeur d’achat du point. Les adhérents pouvaient dans le temps acheter leur point moins cher et donc en acheter plus. On considère pour le moment que ces deux effets s’annulent. Un modèle de projection pourrait affiner la modélisation. Les nouveaux adhérents ne sont que des anciens adhérents échantillonnés dont on a translaté les âges. Leurs points sont donc similaires à ceux constaté historiquement. Si on voulait prendre en compte l’évolution des volumes de points dans le temps, il faudrait modifier leur nombre de points. Les points des ayants-droits sont définis par les points des adhérents et par un taux de réversion.

On note que jusqu’en 2012 les valeurs d’achat et de service étaient nettement supérieurs (resp. 289 et 25,47). En revanche le rendement du point lui a conservé sa tendance. Lors du changement de valeur en 2013, le nombre de points a été ajusté de façon que les prestataires conservent la même pension. Le changement n’est donc que cosmétique, mais nécessite de mettre en « as-if » la donnée historique pour assurer une continuité cohérente.

Enfin, certain adhérent ayant peu cotiser liquident en **versement forfaitaire unique** (VFU). Historiquement, les causes de versement en VFU étaient différentes. Aujourd’hui il s’agit simplement des adhérents ayant cotisé moins de 180 points au moment de la liquidation de leurs droits. Ces adhérents perçoivent donc un versement ponctuel et non une rente.

Avec :

* VS la valeur de service du point
* 40 représente la valeur d’achat du point au moment de calibration des VFU

## Volumes annuels à partir d’une base projetée

On considère une **base projetée** sur une **période future définie**. Autrement dit, on choisit une période, d’aujourd’hui à un horizon long terme, par exemple dans 40 ans, on projette les adhérents actuels jusqu’à leur décès, et on définit des nouveaux entrants jusqu’à la date de terme. Dans cette base, tous les adhérents sont décédés (potentiellement après la date de terme sélectionnées), et le dernier affilié s’est affilié à la date de terme. On note qu’il est nécessaire d’ajouter les adhérents décédés avant la projection (que l’on avait retiré car ils avaient déjà atteint leur état stationnaire). Pour toutes les années à venir jusqu’à la date de terme, on va compter le nombre d’adhérents par état (cotisant, CER, radié, …). Ainsi, les cotisants regroupent :

Les radiés regroupent :

Les prestataires regroupent :

Les VFU regroupent :

Les CER regroupent :

Les décédés regroupent :

# TODo gérer entré DD - Les adh dans la baseDD sont tous liquidés, soit rb, rc ou id

Regime ID

On suppose de la classe A tout adh qui acquiert moins de 200 points.

Ne concerne ici que l’invalidité.

# TODo majo/mino

# TODo revoirs les DD

# TODO décalage de l'age de retraite

# TODo projeter le point

# TODO affiner la mortalité

# TODo régime ID (et gérer les DD id)

# TODo DC CER sous-estimé

# TODo def loi sur pl puis me

# TODo vfu des DD (cf doc des reversation sur tempo)

# TODO hyp sur le revenu

# TODO hyp sur l'inflation

# TODO hyp de réduction des ne-liq\_jms

# TODO choc d'hypothèse de passage

# TODO hyp de rendement fi

# TODO integrer la mise en asif de l'age min historique -> nouvelle hyp de passage

# TODO découpage plfss

# évaluer à d’autre taux de cotis et d’autre seuil 5PASS etc

Choix de liq :

* Doit on consierer les ne liq jms
* Considérer que l’age
* Génération ? pblm de volume
* Lisser
* Sur-ponderer les données récentes
* Traitement des données manquantes
* Loi de mortalité