IRM | institut du risk management BY institut

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : Shu Louise LI

Titre : Apport des méthodes d'apprentissage supervisé à la construction d'une table de maintien en incapacité temporaire de travail pour un groupe fermé de fonctionnaires

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : Fixage
Nom : Michel PIERMAY
Signature et Cachet :

Directeur de mémoire en entreprise :
Nom : Marc DU CHOUCHET
Signature :

Invité :
Nom :
Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

TABLE DES ABREVIATIONS

Abréviation	Signification
CMO	Congé de maladie ordinaire
CLM	Congé de longue maladie
CLD	Congé de longue durée
DO	Disponibilité d'office pour raison de santé
ST	Survival Tree (Arbre de survie)
RSF	Random Survival Forest (Forêts aléatoires de survie)
SGB	Survival Gradient boosting (Gradient boosting de survie)
CART	Classification and Regression Tree (Arbre de classification et de régression)

RESUME

Le régime statutaire indemnise l'agent public en incapacité temporaire en fonction de 3 états : le congé de maladie ordinaire CMO, le congé de longue maladie CLM et le congé de longue durée CLD. Chaque état se distingue par des caractéristiques propres, notamment par une durée d'indemnisation maximale différente. Face à une population spécifique, les tables réglementaires du BCAC ne sont plus adaptées (un état d'incapacité temporaire indemnisé à 36 mois au maximum) pour estimer les provisions en arrêt de travail temporaire. La solution est alors de construire des tables d'expérience.

Pour ce faire, les méthodes classiques telles que l'estimateur non paramétrique de Kaplan-Meier ont prouvé leur efficacité. Cependant, l'essor des algorithmes d'apprentissage supervisé apporte une alternative aux approches classiques.

Les algorithmes appliqués dans ce mémoire sont l'arbre de survie, les forêts aléatoires de survie ainsi que le gradient boosting de survie. Leur point commun tient au prédicteur faible : un arbre. A partir d'un modèle dit simple qu'est l'arbre, il est théoriquement possible d'améliorer les performances de prédiction soit en moyennant les prédictions de différents arbres (les forêts de survie), soit en moyennant les prédictions corrigées de différents arbres (le gradient boosting de survie).

L'objectif étant de construire une loi de maintien par état, chaque état s'est vu appliqué les trois algorithmes. La sélection de l'algorithme optimal pour une loi donnée s'est basée sur la mesure de la performance prédictive. Pour cela, les deux indicateurs initialement retenus sont l'indicateur de concordance de Harell et le score de Brier intégré.

Suite à l'application sur le portefeuille étudié, les résultats des deux indicateurs de performance ne sont pas concordants, notamment sur les prédictions des CMO. Ce dernier présentant les plus faibles performances avec le C-index et les meilleures avec l'IBS. Cette incohérence trouve son origine dans le volume de données censurées : quand ce volume est trop faible, l'IBS est alors trop optimiste et fournit un résultat non exploitable. Par conséquent, seuls les résultats du C-index sont retenus pour sélectionner l'algorithme optimal pour chaque loi.

Enfin le C-index indique comme algorithme optimal les forêts aléatoires de survie pour les lois CLM et CLD et le gradient boosting de survie pour la loi CMO.

Pour comparer les résultats des algorithmes d'apprentissage supervisé à ceux de l'estimateur de Kaplan-Meier, un backtesting est appliqué. Cela consiste à comparer les durées de maintien estimées aux durées de maintien réelles, sur les sinistres en cours à une date d'inventaire donnée.

Les résultats du backtesting sont cohérents avec ceux de l'indicateur de performance prédictive du C-index. Notamment pour la loi CLD, les durées estimées sont plus proches des durées réelles par rapport à Kaplan-Meier.

Pour appréhender ces résultats, le pouvoir prédictif de chaque variable explicative est analysé par le C-index. Le constat est le suivant : le pouvoir prédictif d'un algorithme dépend de celui des variables explicatives. Une variable à fort pouvoir prédictif présente des modalités dont la répartition est équilibrée et dont les valeurs sont suffisamment distinctes. L'optimalité de la prédiction est donc conditionnée par la structure de la base de données.

Dans le cadre de l'analyse de survie et du portefeuille étudié, l'application d'un algorithme d'apprentissage supervisé semble pertinente, sous condition que la distribution des différentes modalités d'une variable soit équilibrée et que les modalités soient suffisamment distinctes.

NOTE DE SYNTHESE

L'incapacité temporaire de travail doit être provisionnée (loi Evin) en utilisant les lois de maintien appropriées (arrêté du 28 mars 1996) : la table réglementaire du BCAC ou les tables d'expérience.
La table de maintien en incapacité temporaire du BCAC est généralement utilisée pour des salariés du régime général de la sécurité sociale, où l'état d'incapacité temporaire dure au maximum trois ans. Mais pour la population de notre portefeuille (groupe fermé composé d'individus ayant gardé le statut de fonctionnaire), la table réglementaire n'est plus adaptée, puisque le profil des assurés est significativement différent.
Pour estimer les lois de maintien en incapacité, les méthodes classiques telles que l'estimateur non paramétrique de Kaplan-Meier ont prouvé leur efficacité. Cependant, l'essor des algorithmes d'apprentissage supervisé apporte une alternative aux approches classiques.

L'objet de ce mémoire est de comparer différentes méthodes statistiques, notamment par les algorithmes d'apprentissage supervisé, pour la construction de loi de maintien en incapacité temporaire de travail appliquée à un portefeuille spécifique (groupe fermé de fonctionnaires).

Pour ce faire, les résultats des lois estimées par Kaplan-Meier sont comparés à ceux des algorithmes d'apprentissage supervisé : l'arbre de survie, les forêts aléatoires de survie et le gradient boosting de survie. Notamment, pour appréhender les résultats de l'apprentissage supervisé, et par conséquent son apport, il sera important de comprendre dans quelle mesure ces algorithmes sont adaptés à l'analyse de survie.

Risque incapacité temporaire de travail : différences entre le régime général et le régime statutaire

Les différences notables de la couverture incapacité temporaire de travail entre le régime général et le régime statutaire permettent de comprendre l'utilité de la construction de tables d'expérience.

Les salariés du régime général de la sécurité sociale sont en incapacité pendant trois ans puis sont consolidés en invalidité.

Les fonctionnaires relèvent d'un régime différent du régime général de la sécurité sociale qui est dit statutaire et dont l'incapacité temporaire est illustrée par trois états et non plus un seul :

- le congé de maladie ordinaire CMO ;
- le congé de longue maladie CLM ;
- le congé de longue durée CLD.

Chaque état présente une durée d'indemnisation maximale différente et un fonctionnement différent.

Régime de base	Type d'arrêt	Durée maximale d'indemnisation
Régime général de la sécurité sociale	Un seul état	36 mois
Régime statutaire de la fonction publique	CMO	12 mois
	CLM	36 mois
	CLD	60 mois

L'analyse de survie avec des données incomplètes et les algorithmes d'apprentissage supervisé

Une loi de maintien en incapacité temporaire est construite avec les méthodes statistiques appartenant à l'analyse de survie.

L'analyse de survie consiste à étudier la durée de survie, soit le temps écoulé jusqu'à la survenance d'un événement d'intérêt.

La durée de survie est représentée par une variable aléatoire positive ou nulle, dont la distribution est caractérisée principalement par la fonction de survie et la fonction de hasard (ou taux de risque instantané). Ces deux fonctions étant liées.

L'étude de la durée de survie prend en compte une spécificité : l'existence de données incomplètes (partiellement observées). Pour cela, les méthodes statistiques appliquées doivent prendre en compte cette spécificité. Notamment les censures à droite et les troncatures à gauche dans le cadre de tables de maintien en arrêt de travail.

Les algorithmes d'apprentissage supervisé appliqués dans ce mémoire sont adaptés à l'analyse de survie. Les trois algorithmes étudiés sont basés sur un arbre de survie.

Un arbre de survie est un arbre de décision dont le principe consiste en un partitionnement binaire et successif de chaque variable explicative. L'arborescence est donc identique à un arbre de classification ou de régression. Cependant, un arbre de survie diffère car il prend en compte les données censurées à droite.

Concrètement, un arbre de survie se distingue d'un arbre de décision standard (classification ou régression) dans la division d'un nœud et dans le contenu d'une feuille.
Un nœud est divisé selon la statistique de test du log-rank (indice de Gini en classification, variance inter-classe en régression). Celle-ci mesure la différence de survie entre deux groupes : plus la valeur de cette statistique est élevée et meilleure est la coupure.
A la fin du partitionnement, la feuille fournit l'estimation de la distribution des individus associés à cette feuille.

Estimation des lois des maintien avec l'approche classique : l'estimateur de Kaplan-Meier

Pour construire ces lois de maintien, l'approche classique est abordée avec l'estimateur de Kaplan-Meier.

Le principe consiste à estimer à chaque pas de temps la fonction de survie (et par conséquent la fonction de hasard) : estimer la probabilité de maintien en incapacité à chaque jour d'indemnisation.

Les lois brutes calculées avec l'estimateur de Kaplan-Meier sont évaluées par un intervalle de confiance à 95%. La construction est basée sur la variance de l'estimateur de Kaplan-Meier. Un estimateur de la variance de la fonction de survie à un instant t est donné par l'estimateur de Greenwood.

Le lissage, effectué par la méthode de Whittaker-Henderson, s'est révélé inutile. Le backtesting est alors effectué sur les résultats des lois de maintien brutes.

Estimation des lois de maintien avec l'apprentissage supervisé : les arbres de survie, les forêts aléatoires de survie et le gradient boosting de survie

Le point commun aux trois algorithmes sélectionnés tient au prédicteur faible : un arbre. A partir d'un modèle dit simple qu'est l'arbre, il est théoriquement possible d'améliorer les performances de prédiction soit en moyennant les prédictions de différents arbres (les forêts de survie), soit en moyennant les prédictions corrigées de différents arbres (le gradient boosting de survie).

SYNTHESE DES CARACTERISTIQUES DE CHAQUE ALGORITHME

- **Arbre de survie :**
  - principe : partitionnement binaire et successif pour chaque variable explicative, de sorte à créer des groupes les plus homogènes possible en fonction de la variable cible ;
  - à la fin du partitionnement, les données des individus associés aux nœuds terminaux (feuilles) sont alors utilisées pour prédire la variable cible ;
  - l'arbre de survie diffère d'un arbre de classification et de régression dans la division des nœuds à chaque étape : un groupe est homogène si tous les individus dans le nœud survivent pendant une période identique de temps ;
  - la différence de survie entre les groupes (nœuds) est mesurée par le test du log-rank.

- **Forêts aléatoires de survie :**
  - méthode ensembliste ;
  - principe du bagging :
    - entraîner des prédicteurs faibles en parallèle puis agréger les prédictions qui en sont issues ;
    - le but étant d'aboutir à un prédicteur plus performant, et donc minimiser les erreurs de prédiction ;
  - les prédicteurs faibles, différents les uns des autres, sont entraînés indépendamment les uns des autres (apprentissage en parallèle).

- **Gradient boosting de survie :**
  - méthode ensembliste ;
  - cadre polyvalent dont le prédicteur faible est ici un arbre ;
  - principe du boosting :
    - entraîner des prédicteurs faibles en séquentiel puis agréger les prédictions qui en sont issues ;
    - construction récurrente : chaque nouvel arbre (prédicteur faible) est constitué en corrigeant l'erreur de prédiction du précédent arbre ;
  - la descente de gradient permet la convergence du modèle en minimisant une fonction de coût convexe.

L'objectif étant de construire une loi de maintien par état d'incapacité, chaque état s'est vu appliqué les trois algorithmes. La sélection de l'algorithme optimal pour une loi donnée s'est basée sur la mesure de la performance prédictive. Pour cela, les deux indicateurs initialement retenus sont l'indicateur de concordance de Harell (C-index) et le score de Brier intégré (IBS).

Suite à l'application sur le portefeuille étudié, les résultats des deux indicateurs de performance ne sont pas concordants, notamment sur les prédictions des CMO. Ce dernier présentant les plus faibles performances avec le C-index et les meilleures avec l'IBS. Cette incohérence trouve son origine dans le volume de données censurées : quand ce volume est trop faible, l'IBS est alors trop optimiste et fournit un résultat non exploitable. Par conséquent, seuls les résultats du C-index sont retenus pour sélectionner l'algorithme optimal pour chaque loi.

Un C-index indiquant un modèle utile doit être supérieur à 50%, et à partir de 65%, les performances prédictives sont considérées correctes. L'algorithme optimal est représenté par les forêts aléatoires de survie pour les lois CLM (72,28%) et CLD (79,86%) et par le gradient boosting de survie pour la loi CMO (67,57%).

Le lissage, effectué par la méthode de Whittaker-Henderson, s'est révélé inutile. Le backtesting est alors effectué sur les résultats des lois de maintien brutes.

Comparaison entre les lois de maintien calculées par l'estimateur de Kaplan-Meier et celles estimées par les algorithmes d'apprentissage supervisé : importance des données

Ecart entre les durées estimées et observées (estimées / observées)	CMO	CLM	CLD
Par l'estimateur de Kaplan-Meier (KM)	20,0%	14,1%	9,2%
Par l'algorithme d'apprentissage supervisé (AAS)	20,4%	14,7%	7,8%

Ecart entre l'apprentissage supervisé et Kaplan-Meier	CMO	CLM	CLD
AAS - KM	0,4%	0,6%	-1,4%

Pour valider la pertinence des tables de maintien estimées et comparer les deux approches (classique et apprentissage supervisé), un backtesting (écart entre les durées de maintien estimées et les durées réelles) sur les arrêts en cours au 31/12/2019 est appliqué. Les résultats du backtesting entre les deux approches sont quasi identiques sur les CMO et les CLM. En revanche, sur les CLD, les durées estimées par l'apprentissage supervisé sont plus proches des durées réelles que celles estimées par l'estimateur de Kaplan-Meier.

Cette différence tient à l'équilibre et à l'hétérogénéité des modalités d'une variable explicative :

- la majorité des arrêts CMO n'est pas prolongée par une DO (96%) et a un motif de sortie standard (99%). Les hommes et les femmes sont représentés à part quasi-égale (51%/49%) et ont une durée moyenne d'arrêt identique ;

- les arrêts CLM se distinguent davantage sur la prolongation par la DO (35% des arrêts) et sur le motif de sortie (14% de sortie pour retraite et 85% de sortie standard), mais peu entre les hommes et les femmes en termes de durée moyenne d'arrêt (33/32 mois) ;
- les arrêts CLD se distinguent peu sur le sexe et le motif de sortie en termes de durée moyenne d'arrêt (durée identique ou quasi-identique avec des modalités de proportions différentes). Mais les arrêts se distinguent sur la durée moyenne d'arrêt entre les arrêts prolongés par une DO (66 mois et 45% des arrêts) et ceux qui ne le sont pas (55 mois et 55% des arrêts).

Une variable explicative à fort pouvoir prédictif est une variable dont les modalités sont hétérogènes et dont la répartition des modalités est équilibrée. Les modalités doivent être suffisamment distinctes et le volume de chaque modalité doit être suffisant.
Des données composées de 55% d'arrêts sans DO avec une durée moyenne de 55 mois (contre 45% et 66 mois pour les arrêts avec DO) offrent de meilleures prédictions que des données composées à 96% d'arrêts sans DO (bien que la durée moyenne soit bien éloignée de celle des arrêts avec DO).

Dans le cadre de cette étude, l'application d'un algorithme d'apprentissage supervisé dans la construction d'une loi de maintien en incapacité semble pertinente, dans la mesure où l'estimation est plus précise, sous condition que la distribution des différentes modalités d'une variable soit équilibrée et que les modalités soient suffisamment distinctes.

Mots-clés : *machine learning, apprentissage supervisé, arrêt de travail, incapacité, tables d'expérience, tables de maintien en incapacité, Kaplan-Meier, Whittaker-Henderson, arbre de survie, forêts aléatoires de survie, gradient boosting de survie, groupe fermé, fonctionnaire, régime statutaire.*

EXECUTIVE SUMMARY

Temporary disability must be provisioned (Evin table) using the appropriate tables (decree of March 28, 1996) : the BCAC regulatory probability table or the experience probability tables.
The BCAC temporary disability probability table is generally used for employees of the general social security system, where the state of temporary disability lasts a maximum of three years. But for the population of our portfolio (closed group made up of individuals who have retained civil servant status), the regulatory table is no longer suitable, since the profile of the insured is significantly different.
To estimate disability tables, classic methods such as the nonparametric Kaplan-Meier estimator have proven their effectiveness. However, the rise of supervised learning algorithms provides an alternative to traditional approaches.

The purpose of this dissertation is to compare different statistical methods, in particular by supervised learning algorithms, for the construction of temporary disability table applied to a specific portfolio (closed group of civil servants).

To do this, the results of the probability tables estimated by Kaplan-Meier are compared to those of supervised learning algorithms: the survival tree, survival random forests and survival gradient boosting. In particular, to understand the results of supervised learning, and consequently its contribution, it will be important to understand to what extent these algorithms are adapted to survival analysis.

Risk of temporary disability for work: differences between the general regime and the statutory regime

The notable differences in temporary disability coverage between the general scheme and the statutory scheme make it possible to understand the usefulness of constructing experience tables.

Employees of the general social security system are unable to work for three years and then consolidated into disability.

Civil servants come under a regime different from the general social security regime which is called statutory and whose temporary disability is illustrated by three states and no longer just one:

- ordinary sick leave CMO ;
- long illness leave CLM ;
- long-term leave CLD.

Each state has a different maximum compensation period and how it works.

Basic social scheme	Type of work stoppage	Maximum duration of compensation
General social security system	One state	36 months
Public service statutory regime	CMO	12 months
	CLM	36 months
	CLD	60 months

Specificity of survival analysis and adaptation of supervised learning algorithms

A temporary disability table is constructed using statistical methods belonging to survival analysis.

Survival analysis consists of studying the survival time, i.e. the time elapsed until the occurrence of an event of interest.

The survival time is represented by a positive or zero random variable, the distribution of which is mainly characterized by the survival function and the hazard function (or instantaneous hazard rate). These two functions are linked.

The study of survival time takes into account a specificity : the existence of incomplete data (partially observed). To do this, the statistical methods applied must take this specificity into account. In particular the right-censored data and the left-truncated data in the context of temporary disability table.

The supervised learning algorithms applied in this dissertation are adapted to survival analysis. The three algorithms studied are based on a survival tree.

A survival tree is a decision tree whose principle consists of a binary and successive partitioning of each explanatory variable. The tree is therefore identical to a classification or regression tree. However, a survival tree differs because it takes into account right-censored data.

Concretely, a survival tree differs from a standard decision tree (classification or regression) in the division of a node and in the content of a leaf.
A node is divided according to the log-rank test statistic (Gini index in classification, inter-class variance in regression). This measures the difference in survival between two groups: the higher the value of this statistic, the better the cutoff.
At the end of the partitioning, the leaf provides the estimate of the distribution of individuals associated with this leaf.

Estimation of temporary disability tables with the classical approach: the Kaplan-Meier estimator

To construct these tables, the classical approach is approached with the Kaplan-Meier estimator.

The principle consists of estimating the survival function (and consequently the hazard function) at each time step: estimating the probability of remaining incapacitated for work on each day of compensation.

The raw distributions calculated with the Kaplan-Meier estimator are evaluated by a 95% confidence interval. The construction is based on the variance of the Kaplan-Meier estimator. An estimator of the variance of the survival function at time t is given by the Greenwood estimator.

Smoothing, carried out by the Whittaker-Henderson method, proved unnecessary. Backtesting is then carried out on the results of the raw tables.

Estimation of temporary disability tables with supervised learning: survival trees, survival random forests and survival gradient boosting

The common point between the three selected algorithms is the weak predictor : a tree. From a so-called simple tree model, it is theoretically possible to improve prediction performance either by averaging the predictions of different trees (survival forests), or by averaging the corrected predictions of different trees (survival gradient boosting).

SUMMARY OF THE CHARACTERISTICS OF EACH ALGORITHM

- **Survival tree :**
  - principle: binary and successive partitioning for each explanatory variable, so as to create the most homogeneous groups possible depending on the target variable;
  - at the end of the partitioning, the data from the individuals associated with the terminal nodes (leaves) are then used to predict the target variable;
  - the survival tree differs from a classification and regression tree in the division of nodes at each stage: a group is homogeneous if all individuals in the node survive for an identical period of time;
  - the difference in survival between groups (nodes) is measured by the log-rank test.

- **Survival Random Forests:**
  - ensemble method;
  - principle of bagging:
    - train weak predictors in parallel then aggregate the resulting predictions ;
    - the goal being to arrive at a more efficient predictor, and therefore minimize prediction errors ;
  - weak predictors, different from each other, are trained independently of each other (parallel learning).

- **Survival gradient boosting :**
  - ensemble method ;
  - versatile framework whose weak predictor here is a tree ;
  - principle of boosting:
    - train weak predictors sequentially then aggregate the resulting predictions ;
    - recurrent construction : each new tree (weak predictor) is created by correcting the prediction error of the previous tree ;
  - gradient descent allows model convergence by minimizing a convex cost function.

The objective being to construct a table per state of disability, each state had the three algorithms applied to it. The selection of the optimal algorithm for a given table was based on the measurement of predictive performance. For this, the two indicators initially selected are the Harell concordance indicator (C-index) and the integrated Brier score (IBS).

Following application to the portfolio studied, the results of the two performance indicators are not consistent, particularly on the CMO predictions. The latter presents the lowest performance with the C-index and the best with the IBS. This inconsistency finds its origin in the volume of censored data: when this volume is too low, the IBS is then too optimistic and provides an unusable result. Consequently, only the results of the C-index are retained to select the optimal algorithm for each table.

A C-index indicating a useful model must be greater than 50%, and from 65%, the predictive performance is considered correct. The optimal algorithm is represented by the survival random forests for the CLM (72.28%) and CLD (79.86%) tables and by the survival gradient boosting for the CMO table (67.57%).

Smoothing, carried out by the Whittaker-Henderson method, proved unnecessary. Backtesting is then carried out on the results of the raw tables.

Comparison between the tables calculated by the Kaplan-Meier estimator and those estimated by supervised learning algorithms : importance of the data

Difference between estimated and observed durations (estimated/observed)	CMO	CLM	CLD
By the Kaplan-Meier (KM) estimator	20.0%	14.1%	9.2%
By the supervised learning algorithm (AAS)	20.4%	14.7%	7.8%

Gap between supervised learning and Kaplan-Meier	CMO	CLM	CLD
AAS - KM	0.4%	0.6%	-1.4%

To validate the relevance of the estimated tables and compare the two approaches (classic and supervised learning), backtesting (difference between the estimated temporary disability times and the actual durations) on the outages in progress as of 12/31/2019 is applied. The backtesting results between the two approaches are almost identical on CMOs and CLMs. On the other hand, on CLDs, the durations estimated by supervised learning are closer to the real durations than those estimated by the Kaplan-Meier estimator.

This difference is due to the balance and heterogeneity of the modalities of an explanatory variable:

- the majority of CMO work stoppage are not extended by an DO (96%) and have a standard reason for discharge (99%). Men and women are represented almost equally (51%/49%) and have an identical average duration of work stoppage;

- CLM work stoppage differ more on the extension by the DO (35% of judgments) and on the reason for exit (14% of exit for retirement and 85% of standard exit), but little between men and women in terms of average duration of work stoppage (33/32 months);
- CLD work stoppage differ little on gender and reason for leaving in terms of average duration of work stoppage (identical or almost identical duration with modalities of different proportions). But the work stoppage differ on the average duration between work stoppage prolonged by an DO (66 months and 45% of stoppages) and those which are not (55 months and 55% of stoppages).

An explanatory variable with strong predictive power is a variable whose modalities are heterogeneous and whose distribution of modalities is balanced. The modalities must be sufficiently distinct and the volume of each modality must be sufficient.
Data composed of 55% of work stoppage without DO with an average duration of 55 months (compared to 45% and 66 months for work stoppage with DO) offer better predictions than data composed of 96% of work stoppage without DO (although the average duration is far removed from that of work stoppage with DO).

In the context of this study, the application of a supervised learning algorithm in the construction of a temporary disability table seems relevant, insofar as the estimation is more precise, provided that the distribution of the different modalities of a variable are balanced and that the modalities are sufficiently distinct.

Keywords: *machine learning, supervised learning, work stoppage, disability, experience tables, temporary disability tables, Kaplan-Meier, Whittaker-Henderson, survival tree, survival random forests, survival gradient boosting, group closed, civil servant, statutory regime.*

REMERCIEMENTS

En premier lieu, je tiens à remercier Monsieur Michel PIERMAY, Président de Fixage, pour la confiance accordée en m'accueillant au sein de sa structure et en m'accompagnant tout le long du parcours académique du CEA.

Je tiens tout particulièrement à remercier Marc Du Chouchet pour ses conseils avisés.

Je remercie également l'ensemble des personnes qui ont contribué à alimenter ma réflexion tout le long de ce mémoire, notamment ma collègue Claire PELTIER, mon professeur au CEA Olivier LOPEZ ainsi que mon référent académique Georges-Louis GONCALVEZ, pour leurs relectures attentives et leurs précieux conseils.

Mes remerciements s'adressent aussi à l'ensemble du corps enseignant du CEA.

Enfin, merci à mes parents, à mon petit-frère, à mon brillant mari et à mon adorable fils, pour leur soutien indéfectible.

TABLE DES MATIERES

INTRODUCTION

L'incapacité temporaire de travail doit être provisionnée (loi Evin) en utilisant les lois de maintien appropriées (arrêté du 28 mars 1996) : la table réglementaire du BCAC ou les tables d'expérience. Le point essentiel de ces lois est l'adéquation à la population assurée.
La table de maintien en incapacité temporaire du BCAC est généralement utilisée pour des salariés du régime général de la sécurité sociale, où l'état d'incapacité temporaire dure au maximum trois ans. Mais pour la population de notre portefeuille (groupe fermé composé d'individus ayant gardé le statut de fonctionnaire), la table réglementaire n'est plus adaptée, puisque le profil des assurés est significativement différent.
Les fonctionnaires relèvent d'un régime différent du régime général de la sécurité sociale qui est dit statutaire et dont l'incapacité temporaire est illustrée par trois états et non plus un seul : le congé de maladie ordinaire CMO, le congé de longue maladie CLM et le congé de longue durée CLD, chaque état présente une durée d'indemnisation maximale différente.
Une telle contrainte justifie alors le recours aux tables d'expérience qui permettent de mieux cerner le risque inhérent au portefeuille de l'assureur.

Pour ce faire, les méthodes classiques telles que l'estimateur non paramétrique de Kaplan-Meier ont prouvé leur efficacité. Cependant, l'essor des algorithmes d'apprentissage supervisé apporte une alternative aux approches classiques.

L'objet de ce mémoire est de comparer différentes méthodes statistiques, notamment par les algorithmes d'apprentissage supervisé, pour la construction de loi de maintien en incapacité temporaire de travail appliquée à un portefeuille spécifique (groupe fermé de fonctionnaires).

A cette fin, sera présentée dans un premier temps la spécificité du risque incapacité temporaire dans la fonction publique pour comprendre l'utilité des tables d'expérience pour notre portefeuille.
Dans un second temps, le cadre théorique des méthodes utilisées pour la construction des lois de maintien en incapacité temporaire sera exposé : d'une part, l'estimateur de Kaplan-Meier pour l'approche classique, d'autre part, les arbres de survies, les forêts de survie ainsi que le Gradient Boosting de survie pour l'apprentissage supervisé.
Ensuite, sera réalisée l'analyse des données pour la construction de la base de sinistres qui sera utilisée pour l'application empirique sur notre portefeuille, à savoir la comparaison de l'approche par Kaplan-Meier avec l'algorithme optimal de l'apprentissage supervisé.
Enfin cette comparaison permettra de mettre en exergue l'apport de l'apprentissage supervisé et constituera la dernière partie de ce mémoire.

1. LE RISQUE INCAPACITE TEMPORAIRE DANS LA FONCTION PUBLIQUE

Dans quelle mesure l'évaluation du risque incapacité dans la fonction publique justifie-t-elle l'utilisation de tables d'expérience ?
Après une brève présentation des fonctions publiques, ce chapitre mettra en exergue les différences notables entre la protection sociale des agents du régime statutaire et celle des salariés du régime général, notamment sur la couverture du risque incapacité temporaire.
Ces différences permettront de comprendre l'inadéquation des tables d'incapacité temporaire réglementaires du BCAC à cette population et donc l'utilité de la construction de tables d'expérience.
Ces dernières devant suivre une procédure rigoureusement établie par l'Institut des Actuaires pour pouvoir être appliquée par l'assureur, à savoir la certification, dont la présentation synthétique clôturera ce premier chapitre.

1.1. Les fonctions publiques

La fonction publique française désigne l'ensemble des agents, titulaires et contractuels, occupant un poste au sein de la fonction publique de l'Etat, d'une collectivité territoriale, ou des établissements de santé.

Elle emploie près d'un salarié sur cinq (5,61 millions d'agents en 2020), dont un peu moins de la moitié par la fonction publique d'Etat (2,49 millions d'agents), le tiers par la fonction publique territoriale (1,94 millions) et enfin 21% par la fonction publique hospitalière (1,18 millions).

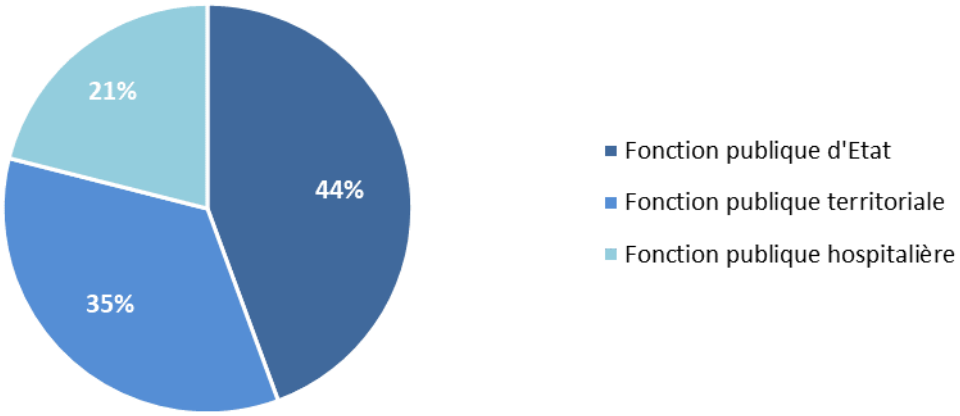


Figure 1 – Répartition des effectifs de la fonction publique en 2020
Source : https://www.fonction-publique.gouv.fr

Chaque fonction publique est régie par des dispositions particulières à caractère national. Les différentes fonctions publiques ont vu leur statut général unifié par la loi du 13 juillet 1983 (titre I - statut général) tout en précisant et préservant les spécificités de chaque secteur :

- Titre II pour les fonctionnaires de l'État
- Titre III pour les fonctionnaires territoriaux
- Titre IV pour les fonctionnaires hospitaliers

A noter que les Magistrats et militaires sont régis par un statut particulier.

Les emplois de la fonction publique sont répartis en 3 catégories hiérarchiques, chaque catégorie étant elle-même constituée de nombreux corps correspondant généralement aux diverses filières des métiers.

Les emplois de la **fonction publique d'État** se répartissent entre les administrations centrales de l'État (services centraux des ministères) et les services déconcentrés (actions de l'État au niveau de la région et du département).
On y retrouve de nombreux fonctionnaires exerçant leurs fonctions dans les établissements publics d'enseignement (universités, lycées, collèges), ainsi que dans les établissements publics administratifs rattachés aux différents ministères.

La **fonction publique territoriale** regroupe les personnels des collectivités territoriales (communes, départements, régions), des structures intercommunales (communautés d'agglomérations, communautés de communes…) ainsi que des établissements publics et des offices publics d'HLM.
Elle s'est structurée à la suite du mouvement de décentralisation des années quatre-vingts. La loi du 26 janvier 1984 pose les principes généraux définissant le cadre d'action et d'organisation de cette fonction publique.

La **fonction publique hospitalière**, telle que la définit la loi du 9 janvier 1986, concerne aujourd'hui à l'exception du personnel médical (médecins, biologistes, pharmaciens et orthodontistes), l'ensemble des emplois des établissements suivants :

- les hôpitaux publics ;
- les établissements d'hébergement pour personnes âgées ;
- les services départementaux de l'aide sociale à l'enfance ;
- les établissements publics pour mineurs ou adultes handicapés ou inadaptés ;
- les centres d'hébergement et de réadaptation sociale publics ou à caractère public.

1.2. La protection sociale des agents de la fonction publique

1.2.1. La couverture de base obligatoire (dit statutaire)

La protection sociale apporte aux agents publics des garanties pour faire face aux événements de la vie qui sont la maladie, les accidents et la maladie professionnelle, la famille, la parentalité et les aidants, le chômage ainsi que la retraite.
Les agents publics bénéficient de cette protection par leur employeur public **selon le principe de l'auto-assurance,** ce qui signifie que **les employeurs publics assurent par leurs propres moyen le financement des garanties ouvertes par les dispositions de protection sociale prévues par le droit de la fonction publique**. Seule la retraite est cofinancée à la fois par l'employeur et l'agent via des cotisations.

L'employeur public a la possibilité (et non l'obligation) d'assurer la couverture statutaire auprès d'un assureur. Ce qui n'est pas le cas dans le secteur privé, un assureur ne pouvant se substituer aux prérogatives de la sécurité sociale en matière d'assurance sociale obligatoire.

Les fonctionnaires relèvent de **régimes spéciaux de sécurité sociale.** Ces régimes garantissent des droits à prestations en espèces au moins équivalents à ceux du régime général de sécurité sociale. **Les droits apportés par le statut général de la fonction publique sont supplétifs aux droits issus des régimes spéciaux (les fonctionnaires reçoivent les meilleures prestations entre celles du régime spécial et celles prévues par le statut).** En revanche, les droits ne se cumulent pas.

A noter que les **agents contractuels de droit public relèvent quant à eux du régime général de sécurité sociale.** Les textes définissant leurs conditions d'emploi prévoient parfois des droits complémentaires à ceux qu'ils ont du régime général de sécurité sociale.

1.2.2. La couverture complémentaire

A l'instar des salariés du secteur privé soumis au régime général de la sécurité sociale, les agents de la fonction publique peuvent également bénéficier d'une couverture en complément de leur régime obligatoire.

La couverture complémentaire est **facultative** (à l'exception des frais de santé depuis l'ordonnance du 17 février 2021), la part collective est mise en place par l'employeur et la part individuelle est à l'initiative de l'agent.

Concernant l'incapacité, la couverture complémentaire prend le relais des obligations statutaires pour éviter une perte de revenu à l'agent.

Le financement de la protection complémentaire collective est défini selon le type de fonction publique et par décret :

• Le décret n°2007-1373 du 19 septembre 2007 prévoit la participation de l'Etat et de ses établissements publics au financement de la protection sociale complémentaire de leurs personnels. Chaque établissement peut souscrire une convention de participation auprès d'un organisme à travers un appel d'offre ;

• Le décret n° 2011-1474 du 8 novembre 2011 prévoit la participation des collectivités territoriales et de leurs établissements publics au financement de la protection sociale complémentaire de leurs agents.

A noter que l'ordonnance n° 2021-175 du 17 février 2021 instaure un régime de couverture complémentaire des frais de santé à adhésion obligatoire. Cette obligation de prise en charge s'appliquera dès 2024 à la fonction publique d'Etat et au plus tard en 2026 à l'ensemble des employeurs de la fonction publique.

1.3. La différence avec le régime général de la sécurité sociale : un unique état d'incapacité temporaire de trois ans maximum d'une part, trois états de durées d'indemnisation différentes de l'autre

L'arrêt de travail est une prescription du médecin attestant **que l'état de santé de l'individu ne lui permet pas d'exécuter son contrat de travail** ou de continuer son activité. Les différentes causes pouvant être une maladie, un accident du travail ou une maladie professionnelle.

L'assuré en arrêt de travail bénéficie d'une couverture de base obligatoire par son employeur dont les modalités varient en fonction de l'état de l'arrêt et peut éventuellement avoir une couverture complémentaire permettant d'éviter une perte de revenu.

Le revenu étant composé des éléments suivants :

- le traitement indiciaire brut qui correspond à la rémunération de base ;
- le supplément familial de traitement versé à partir du premier enfant à charge, le montant varie en fonction du nombre d'enfants à charge ;
- la nouvelle bonification indiciaire allouée à certains agents titulaires ;
- l'indemnité de résidence ;
- d'autres éléments de salaires.

Au sens assurantiel, l'arrêt de travail est composé par l'incapacité temporaire de travail d'une part, et par l'invalidité d'autre part. Dans le cadre de ce mémoire, il sera abordé uniquement l'incapacité temporaire suite à une maladie.

Les salariés du régime général de la sécurité sociale sont en incapacité pendant 3 ans puis sont consolidés en état d'invalidité.
L'incapacité temporaire de travail dans la fonction publique se matérialise non par un seul état (à l'instar du régime général) mais par plusieurs états avant le passage en invalidité :

- le congé de maladie ordinaire (CMO) ;
- le congé de longue maladie (CLM) ;
- le congé de longue durée (CLM).

Chaque état présente une durée d'indemnisation maximale différente : 12 mois pour le CMO, 36 mois pour le CLM et 60 mois pour le CLD. Il existe également des passages entre ces différents états qui seront détaillés par la suite.

Toutefois, il existe une possibilité d'indemnisation venant en relais de ces 3 états, après expiration des droits statutaires aux différents congés et sous certaines conditions : la disponibilité d'office pour raison de santé (DO). Elle n'est pas un état d'incapacité temporaire puisqu'elle n'ouvre pas de droit à indemnisation en tant que telle mais seulement si elle fait suite à un des trois congés. Cette spécificité est donc prise en compte dans la construction des lois de maintien, et prolonge par conséquent les durées maximales d'indemnisation.

Régime de base	Type d'arrêt	Durée maximale d'indemnisation	Complément par la DO
Régime général de la sécurité sociale	Incapacité temporaire	36 mois	na
Régime de la fonction publique	CMO	12 mois	36 mois
	CLM	36 mois	
	CLD	60 mois	

Figure 2 – L'incapacité temporaire de travail : différences entre le régime général de la sécurité sociale et celui de la fonction publique (dit régime statutaire)

1.3.1. Le congé de maladie ordinaire (CMO)

En cas de maladie attestée par un **certificat médical**, l'agent bénéficie de congés « maladie ordinaire ».

La durée totale de CMO est de **12 mois consécutifs** : 6 mois consécutifs puis le renouvellement pour 6 mois maximum est soumis à l'avis du comité médical.

L'intégralité des revenus est versé les 3 premiers mois, puis la moitié les 9 mois suivants.

A la fin du CMO:

- l'agent peut reprendre son activité dans les cas suivants :
  - si l'arrêt est inférieur à 12 mois ;
  - si l'arrêt est de 12 mois, la reprise est soumise à l'avis favorable du comité médical.
- En cas d'avis défavorable, il est soit :
  - mis en DO ;
  - reclassé dans un autre emploi ;
  - reclassé en congé de longue maladie CLM ou en congé de longue durée CLD
  - reconnu définitivement inapte à l'exercice de tout emploi et admis à la retraite pour invalidité.

A noter que le reclassement en CLM/CLD est à effet rétroactif, si l'état de l'assuré le justifie et après accord du comité médical.

Alors la date du début de CLM/CLD est celle de la constatation de la maladie pour la première fois, c'est-à-dire celle du CMO. Donc l'état CLM/CLD vient remplacer l'état CMO.

Durée maximale d'indemnisation	Prestation de base	Condition d'attribution	Situations en fin de congé
12 mois consécutifs	100% pendant 3 mois 50% pendant les 9 mois suivants	Certificat médical	- Reprise d'activité - Placement en DO - Reclassement en CLM ou CLD - Mise en retraite pour invalidité

Figure 3 – Les caractéristiques du congé de maladie ordinaire (CMO)

1.3.2. Le congé de longue maladie (CLM)

Quand la maladie présente un **caractère invalidant nécessitant un traitement et des soins prolongés**, l'assuré est alors placé en congé de longue maladie.

Le classement en CLM se produit dans les situations suivantes :

- l'affection est présente dans une liste non limitative (arrêté du 14 mars 1986 du ministère chargé de la santé) ;
- si l'affection est absente de cette liste, le CLM peut être accordé après avis du conseil médical.

Le CLM est accordé ou renouvelé par période de 3 à 6 mois, pour une durée totale de **3 ans maximum**.

Un agent ayant bénéficié d'un CLM de 3 ans ne peut en obtenir de nouveau qu'à la condition d'avoir repris ses fonctions pendant au moins 1 an.

L'intégralité des revenus est versé la première année puis la moitié les 2 années suivantes.

A la fin du congé de longue maladie, l'agent est soumis à un examen médical :

- soit il est reconnu apte à exercer ses fonctions, dans ce cas il reprend son activité à temps plein ou à temps partiel thérapeutique avec éventuellement des aménagements des conditions de travail ;
- soit il est considéré définitivement inapte à exercer ses fonctions, dans ce cas, il peut :
    - être reclassé dans un autre emploi ;
    - être placé en DO ;
    - être admis à la retraite pour invalidité

Durée maximale d'indemnisation	Prestation de base	Condition d'attribution	Situations en fin de congé
3 ans	100% pendant 1 an 50% les 2 années suivantes	Accordé par période de 3 à 6 mois renouvelable Affections de la liste de l'arrêté du 14 mars 1986 Affection hors liste par avis médical	Examen médical à la fin du CLM : - Reprise d'activité - Placement en DO - Reclassement dans un autre emploi - Mise en retraite pour invalidité

Figure 4 – Les caractéristiques du congé de longue maladie (CLM)

1.3.3. Le congé de longue durée (CLD)

Quand l'agent est atteint par l'une des maladies fixées par **l'article 3 de l'arrêté ministériel du 14 mars 1986** (tuberculose, maladie mentale, cancer, poliomyélite ou déficit immunitaire grave et acquis) et qu'il est dans l'impossibilité d'exercer ses fonctions, il peut alors bénéficier d'un congé longue durée CLD.

L'agent ne peut bénéficier que d'un seul CLD par affection au cours de sa carrière.

Le CLD est attribué à la fin de la première année de CLM rémunérée à plein traitement. Toutefois, si les droits à plein traitement de la première année en CLM sont épuisés, l'agent peut être placé directement en CLD.

La durée totale est **de 5 ans maximum**.

L'intégralité des revenus est versé pendant 3 ans puis la moitié les 2 années suivantes.

A la fin du congé de longue durée, l'agent est soumis à un examen médical :

- soit il est reconnu apte à exercer ses fonctions, dans ce cas il reprend son activité à temps plein ou à temps partiel thérapeutique avec éventuellement des aménagements des conditions de travail ;
- soit il est considéré définitivement inapte à exercer ses fonctions, dans ce cas, il peut :
    - être reclassé dans un autre emploi ;
    - être placé en DO ;
    - être admis à la retraite pour invalidité

Durée maximale d'indemnisation	Prestation de base	Condition d'attribution	Situations en fin de congé
5 ans pour la même affection	100% pendant 3 ans 50% les 2 années suivantes	Affections de la liste de l'arrêté du 14 mars 1986 (article 3) Après 1 an de CLM Droits à CLD valables une fois par affection durant la carrière	Examen médical à la fin du CLD : - Reprise d'activité - Placement en DO - Reclassement dans un autre emploi - Mise en retraite pour invalidité

Figure 5 – Les caractéristiques du congé de longue durée (CLD)

1.3.4. Le passage entre les différents états d'arrêt de travail

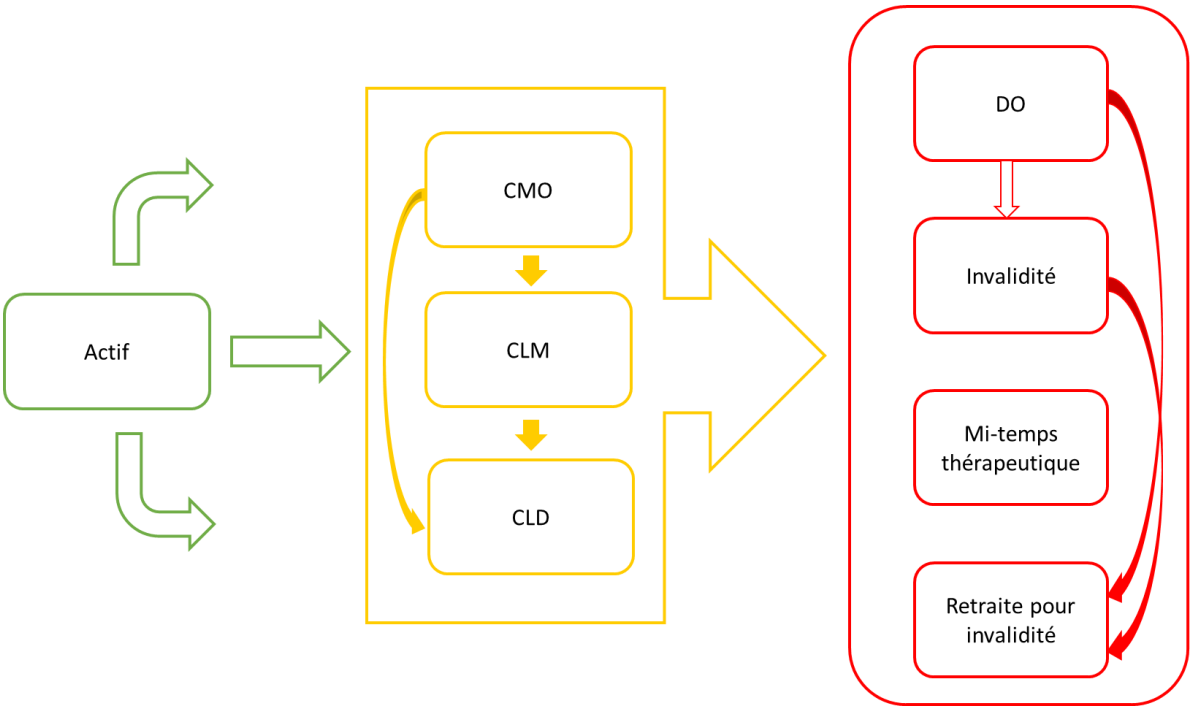


Figure 6 – Passage entre les différents états d'arrêt de travail

Le schéma de la figure 6 synthétise les passages entre les différents états d'arrêt de travail :

- chaque état CMO, CLM ou CLD :
  - peut se consolider en invalidité (à l'instar du passage de l'incapacité à l'invalidité dans le régime général de la sécurité sociale) ;
  - mais également passer en DO, à mi-temps thérapeutique ou en retraite pour invalidité ;
- ensuite l'état DO peut se consolider en invalidité ou passer en retraite pour invalidité ;
- l'invalidité pouvant se poursuivre en retraite pour invalidité.

Par ailleurs, les passages rétroactifs entre les états CMO, CLM et CLD sont également représentés par les flèches jaunes pleines : la requalification du CMO en CLM ou CLD ainsi que la requalification du CLM en CLD.
La requalification intervient quand l'état d'incapacité est modifié à titre rétroactif.

	CMO	CLM	CLD	*DO*
Durée maximale d'indemnisation	12 mois consécutifs	3 ans	5 ans	*1 an renouvelable 3 fois*
Prestation de base	100% pendant 3 mois 50% les 9 mois suivants	100% pendant 1 an 50% les 2 années suivantes	100% pendant 3 ans 50% les 2 années suivantes	*Pas de rémunération sauf conditions*
Conditions d'attribution	Certificat médical	Accordé par période de 3 à 6 mois renouvelable Affections de la liste de l'arrêté du 14 mars 1986 Affection hors liste par avis médical	Après 1 an de CLM Affections de la liste de l'arrêté du 14 mars 1986 (article 3) Droits à CLD valables une fois par affection durant la carrière	*Après épuisement des droits à CMO, CLM et CLD*
Situation en fin de congé	- Reprise d'activité - Placement en DO - Reclassement en CLM ou CLD - Mise en retraite pour invalidité	Examen médical à la fin du CLM : - Reprise d'activité - Placement en DO - Reclassement dans un autre emploi - Mise en retraite pour invalidité	Examen médical à la fin du CLM : - Reprise d'activité - Placement en DO - Reclassement dans un autre emploi - Mise en retraite pour invalidité	*- Reprise d'activité* *- Reclassement dans un autre emploi* *- Mise en retraite pour invalidité* *- Licenciement si pas de droit à la retraite*

Figure 7 – Synthèse des caractéristiques des différents états d'incapacité et de la DO

1.4. Les tables de maintien en incapacité temporaire : une table réglementaire du BCAC pour les salariés du secteur privé contre trois tables d'expérience pour les fonctionnaires

Le contexte juridique repose sur deux textes essentiels : la loi Evin et l'arrêté du 28 mars 1996.

La loi n°89-1009 du 31 décembre 1989, dite Evin, a imposé l'obligation de provisionnement des sinistres en cours pour les couvertures du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail, d'invalidité et du risque chômage.

L'arrêté du 28 mars 1996 a fixé les règles de provisionnement des garanties incapacité et invalidité :

- le risque incapacité doit être couvert par une provision incapacité en cours et une provision invalidité en attente ;
- le risque invalidité doit faire l'objet d'une provision invalidité en cours ;
- utilisation des tables réglementaires du Bureau Commun des Assurances Collectives BCAC ou d'une table d'expérience certifiée ;
- utilisation d'un taux d'actualisation inférieur à 75% du taux moyen des emprunts de l'Etat français calculé sur les 24 derniers mois (avant 2010 : base semestrielle), sans que ce taux d'actualisation ne puisse dépasser 4,5%.

1.4.1. La table réglementaire du BCAC

Initialement construites en 1993, les tables du BCAC ont été modifiées par l'arrêté du 24 décembre 2010 suite à la réforme des retraites, puis reconstruites en 2013.

Elles sont utilisées par les assureurs pour le calcul des provisions mathématiques des risques incapacité et invalidité et sont au nombre de trois :

- une table de maintien en incapacité ;
- une table de passage de l'incapacité à l'invalidité ;
- une table de maintien en invalidité.

La table de maintien en incapacité du BCAC présente l'évolution des incapables selon deux dimensions : l'âge d'entrée en l'état et l'ancienneté dans l'état en nombre de mois.

La loi de maintien est établie sur 36 mois, soit les 3 ans d'incapacité dont peuvent bénéficier les salariés du régime général de la sécurité sociale.

Au regard des spécificités de la couverture de l'incapacité temporaire des agents de la fonction publique, la table réglementaire n'est pas adaptée, ce qui induit un risque de mauvaise estimation des engagements de l'assureur.
Alors une alternative consiste à utiliser une table d'expérience certifiée et suivie par un actuaire indépendant de l'entreprise et soumis à la commission d'agrément de l'Institut des Actuaires.

1.4.2. Les tables d'expérience et la certification

De manière générale, l'utilisation d'une table d'expérience se justifie par une population couverte significativement différente de la population de la table réglementaire.

La population sous risque dans le cadre de ce mémoire est composée d'agents de la fonction publique dont l'incapacité de travail temporaire se matérialise par 3 états.

Chacun de ses états se distingue par des conditions d'indemnisation différentes et notamment par des durées d'indemnisation maximales différentes, à savoir, 1 an pour un CMO, 3 ans pour un CLM et 5 ans pour un CLD.

De plus, lorsque les droits sont épuisés, la DO permet de bénéficier (sous certaines conditions) d'un prolongement de l'indemnisation jusqu'à 3 ans.

Enfin, l'assureur du portefeuille étudié ajoute en complément une année supplémentaire d'indemnisation après épuisement des 3 ans de DO.

Par conséquent, 3 tables d'expérience seront construites : la table CMO de 60 mois, la table CLM de 84 mois et la table CLD de 108 mois (exemple pour la table CMO : 12 mois par l'état initial + 36 mois par la DO + 12 mois par la garantie complémentaire de l'assureur).

Pour utiliser une table d'expérience, celle-ci doit être certifiée et suivie annuellement par un actuaire agréé.

En pratique, la mise en place et l'autorisation d'utilisation d'une table d'expérience comportent 3 étapes :

- la construction de la table ;
- la certification initiale ;
- le suivi annuel destiné à assurer la pérennité du droit d'utilisation de la table.

En effet, en l'absence de suivi, la loi de maintien en arrêt de travail ou la loi de mortalité devient caduque au bout de 2 ans.

La durée de validité des tables de maintien en arrêt de travail est de 4 ans et des tables de mortalité est de 5 ans.

La procédure d'agrément des actuaires indépendants habilités à certifier et à suivre les tables d'expérience (mortalité et arrêt de travail) est définie par l'Institut des Actuaires après avis de l'Autorité de Contrôle Prudentiel et de Résolution.

Cette procédure comprend la mise en place d'une Commission d'Agrément, organe totalement indépendant et souverain dans ses missions d'habilitation des certificateurs de tables, dont le rôle est d'accorder le droit de certification.

Les actuaires agréés doivent répondre aux exigences suivantes :

- être membres de l'Institut des Actuaires et en activité ;
- disposer d'un niveau suffisant de formation et de qualification dans le domaine statistique appliqué à la construction des tables d'expérience ;
- disposer d'une expérience supérieure à cinq ans en tarification et calcul de provisionnement en assurances de personnes.

L'actuaire agréé ne peut à la fois construire et certifier une même table.

Le régime statutaire des fonctionnaires diffère du régime général de la sécurité sociale, et ce, de manière notable. L'incapacité de travail temporaire est identifiée par 3 états, différenciés par des durées d'indemnisation maximales différentes notamment. La construction de tables d'expérience est alors justifiée. Pour cela, les méthodes statistiques appliquées pour calculer ces lois de maintien doivent être appropriées et sont abordées dans le chapitre 2.

A retenir :

L'incapacité temporaire de travail dans le régime général de la sécurité sociale est représentée par un état avec une durée maximale d'indemnisation de 3 ans. La table de maintien en incapacité du BCAC peut être utilisée pour le provisionnement dans ce cas.

Les fonctionnaires sont soumis au régime statutaire dont la couverture incapacité temporaire est représentée par 3 états (congé de maladie ordinaire CMO, congé de longue maladie CLM et congé de longue durée CLD). Chaque état se distingue par un fonctionnement différent et une durée maximale d'indemnisation différente (12, 36 et 60 mois respectivement). La table du BCAC n'est alors pas adaptée et la construction de tables d'expérience est nécessaire pour aborder le provisionnement.

2. LES LOIS DE MAINTIEN EN INCAPACITE TEMPORAIRE DE TRAVAIL

Une loi de maintien en incapacité temporaire est construite avec les méthodes statistiques appartenant à l'analyse de survie. Cette dernière prend en compte une spécificité de l'étude de la durée de survie : l'existence de données incomplètes (partiellement observées).
Afin d'appréhender la compréhension des méthodes utilisées, sera présenté dans un premier temps le cadre théorique de l'analyse de survie.
Ensuite seront exposées les méthodes de chaque approche. D'une part, l'approche classique sera illustrée par une méthode qui a prouvé son efficacité : l'estimateur de Kaplan-Meier. D'autre part, pour l'apprentissage supervisé, 3 algorithmes seront mis en avant : l'arbre de survie, les forêts aléatoires de survie et le gradient boosting de survie.
Enfin, les tables brutes construites seront ensuite lissées par la méthode de Whittaker-Henderson, présentée dans la section clôturant ce deuxième chapitre.

2.1. Cadre théorique : l'analyse de survie

2.1.1. Distribution de la durée de survie

L'analyse de survie consiste à étudier la **durée** de survie, soit le **temps écoulé jusqu'à la survenance d'un événement d'intérêt.** Autrement dit, cela correspond à l'analyse d'un intervalle de temps avec un événement déclenchant la fin de la mesure.

La durée de survie est représentée par une variable aléatoire positive ou nulle, dont la distribution est caractérisée principalement par la **fonction de survie** et la **fonction de hasard** (ou taux de risque instantané).

Soit les notations suivantes :

$T \geq 0$: une variable aléatoire représentant la durée de survie
$\mathbb{P}$: la fonction de probabilité
$t > 0$: la durée de survie observée
dans le cas discret, T prend les valeurs t_k avec $t_1 < t_2 < \ldots . < t_K$ pour tout $k > 0$

La **fonction de survie** représente la probabilité que la durée de survie soit supérieure à un temps donné t. Ici, elle représente la probabilité que la durée de maintien en incapacité soit supérieure à t.
Elle est définie, pour tout $t > 0$, par :

$$S(t) = \mathbb{P}(T > t)$$

Si T est une variable aléatoire continue, $S(t) = \int_t^{+\infty} f(u)\, du$
où f est la densité,

Si T est une variable aléatoire discrète, $S(t)= \sum_{t_k>t} p(t_k)$

où $p(t_k) = \mathbb{P}(T= t_k)$ pour tout $t > 0$ et $k > 0$.

La fonction de survie est le complément de la fonction de répartition F(t), c'est-à-dire :

$$S(t) = 1 - F(t)$$

avec $F(t) = \mathbb{P}(T \leq t)$

La densité vérifie $f(t) = -S'(t)$

La **fonction de hasard** ou taux de risque instantané est la probabilité que l'événement d'intérêt survienne dans l'intervalle très court [t, t+dt], sachant que l'individu a survécu au-delà de t :

Si T est une variable aléatoire continue, $h(t) = \lim_{dt \to 0} \frac{1}{dt} \mathbb{P}(t \leq T < t + dt \mid T > t) = \frac{f(t)}{S(t)}$,

Si T est une variable aléatoire discrète, $h(t) = \mathbb{P}(T = t_k \mid T \geq t_k) = \frac{\mathbb{P}(T=t_k)}{\mathbb{P}(T \geq t_k)}$

Ici, la fonction de hasard représente le taux de sortie de l'état d'incapacité, soit la probabilité que l'individu sorte de l'arrêt de travail à t+dt, sachant qu'il était en arrêt à t.

La **fonction de risque cumulé** représente la totalité des risques instantanés auxquels l'individu est exposé depuis le début de l'étude, il est défini par :

$$H(t) = \int_0^t h(u)\, du$$

H(t) vérifie les propriétés suivantes :

$H(t) = - \ln(S(t))$

$S(t) = e^{-H(t)} = e^{-\int_0^t h(u)\, du}$

Il existe un lien entre la fonction de survie et la fonction de hasard :

comme vu plus haut, $p(t_k) = \mathbb{P}(T= t_k)$ et $S(t_{k-1}) = \mathbb{P}(T \geq t_k)$,

alors $p(t_k)=S(t_{k-1}) - S(t_k)$.

Le risque instantané dans le cas discret est $h(t_k) = \mathbb{P}(T = t_k \mid T \geq t_k) = \frac{\mathbb{P}(T=t_k)}{\mathbb{P}(T \geq t_k)} = \frac{p(t_k)}{S(t_{k-1})}$

qui peut aussi s'écrire $h(t_k) = 1 - \frac{S(t_k)}{S(t_{k-1})}$

La fonction de survie dans le cas discret peut s'écrire comme le produit des probabilités conditionnelles de survie $S(t) = \prod_{t_k \leq t} \frac{S(t_k)}{S(t_{k-1})}$

Ainsi la fonction de survie est liée à la fonction de hasard par $S(t) = \prod_{t_k \leq t} (1 - h(t_k))$.

Dans le cadre de ce mémoire, l'événement d'intérêt est la sortie de l'état d'incapacité, et le temps écoulé jusqu'à la sortie de cet état (la durée de survie) correspond à la durée de maintien en incapacité. Il s'agit alors d'estimer au plus juste la probabilité de maintien en incapacité (la fonction de survie) à chaque pas de temps. Pour cela, les méthodes statistiques appliquées doivent prendre en compte une spécificité de l'analyse de survie qui est l'existence de données incomplètes : les troncatures et les censures.

2.1.2. Spécificité des données de survie : existence de données incomplètes que sont les censures et les troncatures

Une **observation complète** correspond à un individu pour lequel il est possible de calculer une durée de maintien en arrêt de travail complète : l'entrée et la sortie d'arrêt ont lieu pendant la période d'observation définie.

Cependant, certaines données ne sont observées que partiellement : ce sont les données incomplètes. Elles sont illustrées tout particulièrement par les notions de censures à droite et de troncatures à gauche dans le cadre de table de maintien en arrêt de travail.

La durée de maintien en arrêt de travail **est censurée quand cette durée est partiellement observée**. En particulier la **censure à droite** se présente quand l'individu est encore en arrêt à la fin de la période d'observation, nous perdons alors l'information concernant sa date de fin d'arrêt.

Soit les notations suivantes :

$\delta \in \{0,1\}$: un indicateur d'évènement dont la valeur est à 0 en présence de censure
$Y > 0$: la durée de survie
T : la durée non censurée
C : la durée censurée

La durée observable d'un échantillon contenant des données censurées est définie comme suit :

$$Y = \text{Min}(T,C) = T \text{ si } \delta=1$$
$$Y = \text{Min}(T,C) = C \text{ si } \delta=0$$

Les données tronquées diffèrent des données censurées en ce sens qu'elles sont manquantes : les durées ne sont pas observées car elles sont soit en-dessous d'un certain seuil soit au-dessus. En particulier la troncature à gauche se présente en présence de franchise et quand l'individu est déjà en arrêt de travail au début de la période d'observation.

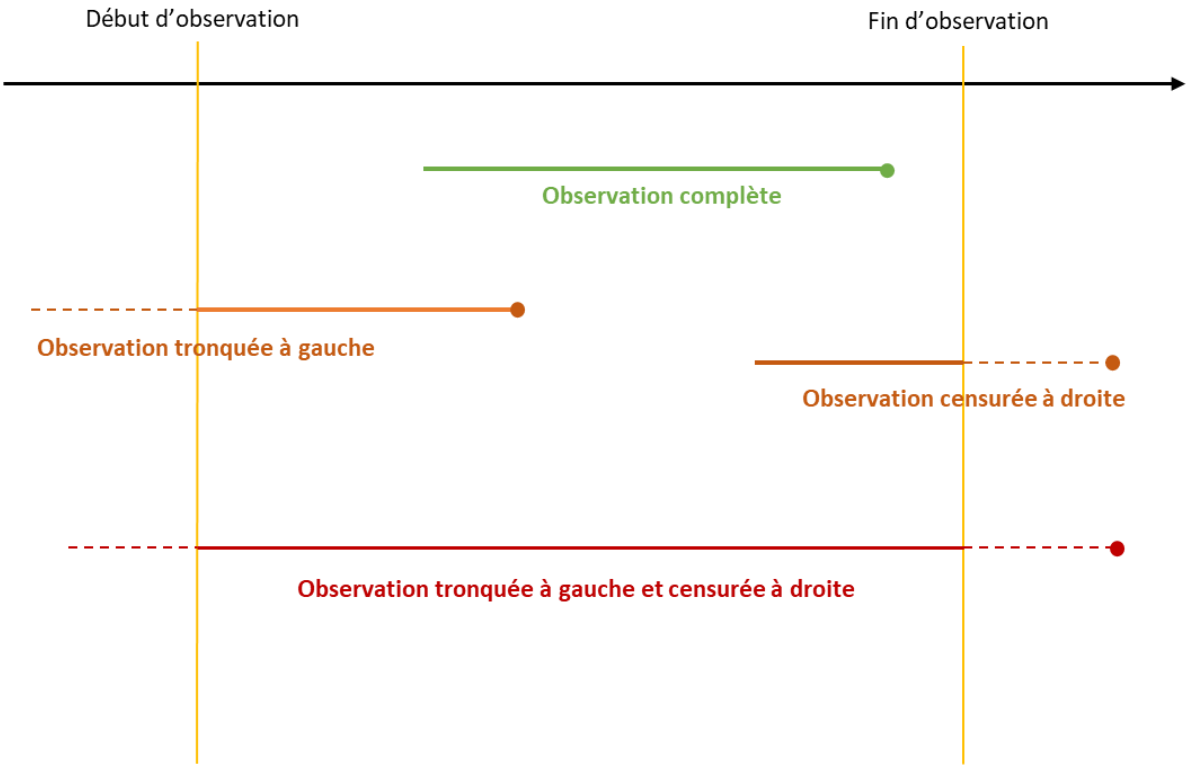


Figure 8 – Différents types de données observées sur un intervalle de temps donné

L'estimation des lois de maintien en incapacité entre dans le cadre de l'analyse de survie dont la spécificité réside dans la prise en compte de données incomplètes. Par conséquent, les méthodes statistiques employées pour calculer les fonctions de survie doivent être adaptées à ce point. Pour cela, une approche classique par l'estimateur de Kaplan-Meier sera étudiée et des méthodes d'apprentissage supervisé seront analysées dans un deuxième temps.

2.2. Approche classique : l'estimateur de Kaplan-Meier

2.2.1. Présentation de l'estimateur non paramétrique de Kaplan-Meier

L'estimateur de Kaplan-Meier est non paramétrique (pas d'hypothèse sur la forme de la loi) et prend en compte les données incomplètes (les censures et les troncatures).

Le principe de cette méthode consiste à estimer à chaque pas de temps la fonction de survie (et par conséquent la fonction de hasard) : estimer la probabilité de maintien en incapacité à chaque jour d'indemnisation.

Soit les notations suivantes :

t = le jour d'indemnisation jusqu'où les taux de sortie sont calculés
d_i = le nombre d'individus sortis de l'état d'incapacité à la date i
n_i = le nombre d'individus en incapacité à la date i
c_i = le nombre de censures observées à la date i
t_i= le nombre de troncatures observées à la date i

L'estimateur de maintien en incapacité, après t jours d'ancienneté est défini par :

$$\hat{S}(t) = \prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i}\right)$$

$\left(\frac{d_i}{n_i}\right)$ représente le taux de sortie, soit la probabilité de sortie de l'état d'incapacité entre i et i+1 jour d'ancienneté, sachant que l'individu était en incapacité au ième jour.

$\left(1 - \frac{d_i}{n_i}\right)$ représente alors le taux de maintien en arrêt de travail (taux de survie), soit la probabilité de maintien en incapacité entre i et i+1 jour d'ancienneté sachant que l'individu était en incapacité au ième jour.

n_i est le nombre d'individus en incapacité à la date i, soit la population exposée au risque à la date i et est calculé comme suit : $n_i = n_{i-1} - d_i - c_i + t_i$

La fonction de survie après t jour d'ancienneté est le produit des taux de maintien en incapacité après chaque jour d'indemnisation i, i allant de 1 à t. Elle représente la probabilité de maintien en arrêt au-delà de t jours.

2.2.2. La variance de l'estimateur de Kaplan Meier et l'intervalle de confiance

Afin d'évaluer la fiabilité et la précision des résultats obtenus, sera analysée l'intervalle de confiance à 95% des taux de maintien. Sa construction est basée sur la variance de l'estimateur de Kaplan-Meier.

Un estimateur de la variance de la fonction de survie à un instant t est donné par l'estimateur de Greenwood :

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i=1}^{t} \frac{d_i}{n_i(n_i - d_i)}$$

Soit la normalité asymptotique de l'estimateur :

$$\frac{\sqrt{n}\,|\hat{S}(t) - S(t)|}{\sqrt{\widehat{\mathrm{Var}}(\hat{S}(t))}} \rightarrow N(0,1)$$

Donc :

$$\mathbb{P}\left(\frac{\sqrt{n}\,|\hat{S}(t) - S(t)|}{\sqrt{\widehat{\mathrm{Var}}(\hat{S}(t))}} \leq u_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

L'intervalle de confiance suivant en est déduit :

$$IC_{1-\alpha}(\hat{S}(t)) = \left[\hat{S}(t) - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{Var}(\hat{S}(t))}{n}} ; \hat{S}(t) + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{Var}(\hat{S}(t))}{n}}\right]$$

Où $u_{1-\frac{\alpha}{2}}$ est la valeur critique de la loi normale centrée réduite associée à un test de niveau α.

L'avantage de l'estimateur de Kaplan-Meier tient à l'absence d'hypothèse sur la forme de la loi. Cependant, il ne prend ni en compte l'impact des variables explicatives et ni en compte les relations particulières pouvant exister entre elles. Ces deux points peuvent être corrigés par les algorithmes d'apprentissage supervisé.

2.3. Les algorithmes d'apprentissage supervisé

2.3.1. Généralités sur l'apprentissage supervisé

Selon Arthur Samuel en 1959, le machine learning ou apprentissage automatique (sans intervention humaine) est une technique d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été explicitement programmés à cet effet. Cet informaticien américain fut le pionnier du domaine en créant un programme qui jouait au Jeu de Dames et s'améliorait tout en jouant.

Le principe consiste à fournir au programme informatique des données avec lesquelles il peut s'entraîner et apprendre, et s'améliorer en conséquence. Plus particulièrement en apprentissage supervisé, les informations sur les résultats attendus sont indiquées (les valeurs de la variable d'intérêt). L'objectif est d'identifier des relations particulières dans les données (les variables explicatives), pour prédire une valeur la plus proche possible de la valeur réelle de la variable d'intérêt.

Cela revient à prédire le mieux possible une variable aléatoire Y (appelée étiquette ou label ou variable cible ou variable d'intérêt) à partir d'autres variables X (appelées prédicteurs ou variables d'entrée ou variables explicatives ou vecteur de covariables).

Les algorithmes de classification sont appliqués aux variables qualitatives et quant aux variables quantitatives, ce sont les algorithmes de régression. Certains de ces algorithmes peuvent être utilisés pour les deux types de variables. En revanche, tous ne sont pas applicables aux données de survie, en raison de la présence de données incomplètes.

2.3.2. Les algorithmes appliqués

2.3.2.1. Les arbres de survie

Les arbres de survie sont un cas particulier des arbres de décision, en ce sens, qu'ils prennent en compte la spécificité de l'analyse de survie : les données incomplètes. Afin de comprendre le processus d'un arbre de survie, il est nécessaire de présenter la méthodologie d'un arbre de décision dans un premier temps, tout particulièrement un arbre de décision CART.

Soit les notations suivantes :

Y la variable cible (quantitative)
p variables explicatives $X = (X_1,....,X_p)$
n observations
s seuils $c = (c_1,....,c_s)$ dans le cas d'une variable explicative quantitative
k modalités du caractère $m = (m_1,....,m_K)$ dans le cas d'une variable explicative qualitative

ARBRE DE DECISION CART

Un modèle d'arbre de décision CART (Classification And Regression Tree) consiste en un partitionnement de la population via des conditions binaires : l'arbre est construit de manière récursive à partir de la racine, en découpant à chaque étape (chaque nœud) la population en 2 sous-ensembles (2 nœuds fils), selon des règles de coupure, et ce, jusqu'à l'obtention d'un critère d'arrêt.

A chaque nœud, toutes les variables explicatives et toutes les découpes possibles sont ainsi testées. Le choix de la variable explicative et de la découpe est effectué en fonction de la minimisation d'un certain critère : l'indice de Gini en classification et la variance empirique des nœuds fils en régression.

A la fin du partitionnement, les nœuds terminaux (qui ne sont plus découpés), sont appelés les feuilles de l'arbre. Les données des individus associés sont alors utilisées pour prédire la variable cible. En régression, c'est la moyenne des valeurs associées aux individus de cette feuille.

La méthode consiste en une découpage binaire et successif pour chaque variable explicative, de sorte à créer des groupes les plus homogènes possible en fonction de la variable cible.

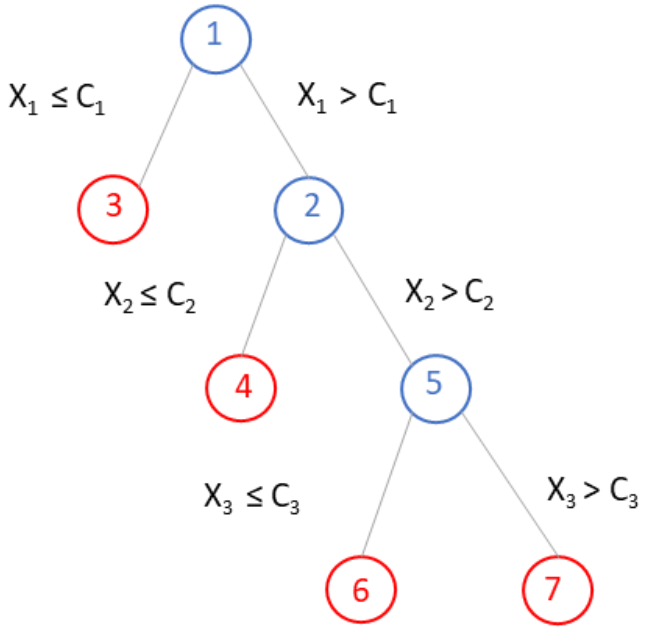


Figure 9 – Construction d'un arbre CART

La construction de l'arbre de décision CART se déroule en deux temps :

1- la construction de l'arbre maximal ;

2- l'élagage en prenant en compte un critère de performance en prédiction.

1- La construction de l'arbre maximal : les règles de coupure

Si la variable explicative est quantitative, la découpe (ou split) de chaque nœud en 2 nœuds fils s'effectue en fonction d'une variable explicative X_i et d'un seuil c_j, :

$$\{ X_i \leq c_j \} \cup \{ X_i > c_j \}$$

Les observations de la ième variable explicative X_i ayant une valeur ≤ au seuil c_j sont classées dans le nœud fils gauche, sinon elles sont classées dans le nœud fils droite.

Si la variable explicative est qualitative, $\{ X_i \leq c_j \} \cup \{ X_i > c_j \}$ devient $\{ X_i = m_k \} \cup \{ X_i \neq m_k \}$

Les individus composant ces deux groupes sont répartis dans 2 nouveaux nœuds fils, et la même procédure est répétée jusqu'à atteindre le critère d'arrêt.

Une fois l'arbre maximal construit, si le nombre de feuilles est jugé trop grand, l'arbre peut être simplifié en élaguant ses branches de bas en haut.

2- L'élagage

L'arbre pleinement développé est appelé arbre maximal : il est très précis mais présente une grande variance. Dans ce cas, il y a un risque de sur apprentissage aux données et le modèle sera difficilement généralisable. A l'inverse d'un arbre minimal constitué uniquement du nœud initial, qui présente une faible variance mais un biais élevé.

Il faut donc trouver un compromis biais-variance en recherchant l'arbre optimal parmi les admissibles, entre l'arbre maximal (le plus complexe et qui conduit au surajustement aux données) et l'arbre restreint à la racine (qui est fortement biaisé).

Les différentes tailles de l'arbre sont calibrées par des règles d'arrêts qui peuvent se combiner. Le principe consiste à quantifier un critère d'arrêt tel que la profondeur de l'arbre, le nombre minimum d'individus présents dans un nœud pour envisager une coupure ou encore le nombre minimum d'individus présents dans une feuille.

L'élagage de l'arbre se matérialise par la recherche des critères d'arrêt ou hyper paramètres optimaux.

En raison de la présence de données incomplètes, les arbres de décision CART ne peuvent être utilisés directement sous la forme présentée ci-dessus à l'analyse de survie, et nécessitent donc une adaptation : ce sont les arbres de survie.

ARBRE DE SURVIE

Les arbres de survie utilisés dans le cadre de ce mémoire sont des arbres de survie binaires, dont la méthode d'arborescence est similaire aux arbres de décision CART mais développée pour les données de survie censurées à droite. Cette méthode est basée sur la **maximisation de la différence de survie entre les nœuds** dans un arbre binaire.

A chaque nœud, pour une variable explicative X et un seuil c donnés, la division est effectuée en fonction de **la statistique de test du log-rank** . Cette statistique mesure la distance (en termes de survie) entre 2 nœuds. Plus la valeur de cette statistique est élevée, plus la différence de survie (distance entre fonctions de survie) entre les 2 nœuds fils est grande et meilleure est la coupure. La meilleure coupure pour le nœud d est déterminée en trouvant la variable X^* et la valeur de coupure c^* qui maximise la statistique de test du log-rank.

A la fin du partitionnement, **la feuille** (ou nœud terminal) d'un arbre de survie est composée de **l'estimation de la distribution** des individus associés à cette feuille (et non de la valeur moyenne, comme en régression).

Les arbres de survie diffèrent des arbres de classification et de régression d'une part dans la découpe des nœuds (découpe sur la valeur de la statistique du test de log-rank) et d'autre part dans le contenu des feuilles (distribution des individus).

LA STATISTIQUE DE TEST DU LOG-RANK

La statistique de test du log-rank est utilisée dans les arbres de survie pour maximiser les différences de survie entre les nœuds fils.

Soit les notations suivantes :

d le nœud à diviser
$(T_1,X_1,\delta_1), \ldots, (T_n,X_n,\delta_n)$ les données
où
X_i est le vecteur de covariables de l'individu i
T_i le temps observé
δ_i l'indicateur de censure pour l'individu i
X est une variable explicative
c un seuil de découpe du nœud

La coupure du nœud d avec la variable explicative X peut être :

- $X \leq c$ => nœud fils gauche L
- $X > c$ => nœud fils droit R

Soit les notations suivantes :

$t_1 < t_2 < \ldots < t_m$ les temps d'évènements distincts
Au temps t_j :
$b_{j,L}$, $b_{j,R}$ le nombre d'individus qui ont subi l'évènement (la sortie de l'arrêt de travail) à t_j
$Y_{j,L}$, $Y_{j,R}$ le nombre d'individus à risque dans les nœuds fils L et R à t_j

$$Y_{j,L} = \#\{T_i \geq t_j\ , X_i \leq c\},\ Y_{j,R} = \#\{T_i \geq t_j\ , X_i > c\}$$

Définissons :

$$Y_j = Y_{j,L} + Y_{j,R}$$
$$b_j = b_{j,L} + b_{j,R}$$

La valeur de la statistique de coupure du log-rank pour la coupure $L = \{X_i \leq c\}$ et $R = \{X_i > c\}$ est :

$$L(X,c) = \frac{\sum_{j=1}^{m}(b_{j,L} - Y_{j,L}\ \frac{b_j}{Y_j}\ )}{\sqrt{\sum_{j=1}^{m}\ \frac{Y_{j,L}}{Y_j}(\ 1 - \frac{Y_{j,L}}{Y_j}\ )(\frac{Y_j - b_j}{Y_j - 1})\ b_j}}$$

La valeur $|L(X, c)|$ est une mesure de quantité de coupure des nœuds. Plus elle est élevée, plus la différence de survie entre *L* et *R* est grande et meilleure est la coupure. La meilleure division est déterminée en trouvant la variable X^* et la valeur de coupure $c*$ telles que $|L(X*, c*)| \geq |L(X, c)|$ pour tout *X* et *c.*

Les arbres de survies sont une méthode non paramétrique (pas d'hypothèse sur la distribution de l'échantillon de données) qui présente une facilité d'interprétation des résultats et un temps de calcul raisonnable. La limite tient au fait que ce soit un prédicteur faible car instable : de minimes modifications dans les données induisent des changements dans les coupures des nœuds, et donc des changements dans l'arborescence.
Néanmoins, cette instabilité peut être corrigée par la combinaison avec d'autres arbres (méthodes ensemblistes), soit en parallèle avec les forêts d'arbres, soit en séquentiel avec le gradient boosting.

A retenir sur les arbres de survie

- La construction d'un arbre de régression binaire est définie par le choix du couple {variable explicative X_i , seuil C_j} : les groupes (nœuds) sont constitués à partir du seuil C_j pour la variable explicative X_i au nœud d.
- L'arbre de survie diffère d'un arbre de classification et de régression :
  - un nœud est pur (groupe homogène) si tous les individus dans le nœud survivent pendant une période identique de temps ;
  - un nœud est divisé en fonction de la statistique de test du log-rank : elle mesure la différence de survie entre 2 groupes. La coupure optimale est obtenue en maximisant cette statistique de test (plus la différence de survie entre les groupes est élevée, meilleure est la coupure) ;
  - à la fin du partitionnement, la feuille (ou nœud terminal) d'un arbre de survie est composée de l'estimation de la distribution des individus associés à cette feuille.

2.3.2.2. Les forêts aléatoires de survie

Une forêt d'arbres de décision est un ensemble d'arbres de décision.

Cette méthode ensembliste repose sur le principe du **bagging** qui consiste à **entraîner des prédicteurs faibles en parallèle** et ensuite d'agréger les prédictions qui en sont issues afin d'aboutir à un **prédicteur plus performant**, et donc minimiser les erreurs de prédiction. Les prédicteurs faibles, différents les uns des autres, sont entraînés indépendamment les uns des autres (apprentissage en parallèle).

Dans le cas d'une forêt d'arbres, un prédicteur faible est un arbre de décision.

Entraîner plusieurs arbres de décision en parallèle revient à entraîner les données sur des sous-échantillons de la base d'apprentissage en modifiant le nombre de variables explicatives et/ou d'individus pris en compte. Donc, les données utilisées sont légèrement différentes pour chaque arbre. Ensuite, les résultats des arbres sont agrégés afin de construire des prédictions plus robustes.

Initialement développées pour prédire une variable cible de type continue ou catégorielle, les forêts ont ensuite été étendues aux données de survie en prenant en compte les données de survie censurées à droite : ce sont les forêts aléatoires de survie ou Random Survival Forest (RSF).

Le processus se décompose en 3 grandes étapes :

Etape 1

A partir de données originelles D_n composées de n individus et p variables explicatives, générer B nouveaux jeux de données $D^1_n,\ldots,D^b_n,..,D^B_n$.
Ces nouveaux jeux de données sont construits à l'aide d'un tirage aléatoire avec remise, ou bootstrap. A noter que chaque échantillon boostrap exclut en moyenne 37% des données, appelées out-of-bag (OOB).

Etape 2

A partir de ces B nouveaux jeux de données, construire B arbres de décision $T_1,\ldots,T_b,\ldots,T_B$.
Ces arbres ont pour objectif de prédire la variable cible pour l'ensemble des individus ayant été utilisés pour les construire.

- A chaque nœud, un sous-ensemble de $m \leq p$ variables explicatives est tiré aléatoirement. On cherche la meilleure coupure uniquement en fonction des m variables sélectionnées. Pour chaque variable explicative, un ensemble de décomposition en 2 groupes est créé.

- Chacune de ces variables sélectionnées est testée dans la division du nœud, jusqu'à obtenir celle qui optimise le critère de coupure. En l'occurrence, le nœud est divisé par la coupure qui maximise la différence de survie entre les deux nœuds fils, soit la statistique de test du log-rank.

Etape 3

Agréger l'ensemble des prédictions issues des B arbres pour obtenir une unique prédiction pour les n observations (vote majoritaire en classification et moyenne en régression). Pour un arbre de survie, cela revient à calculer la fonction de risque cumulé (CHF) pour chaque arbre et ensuite de les moyenner afin d'obtenir la fonction de risque cumulé d'ensemble.

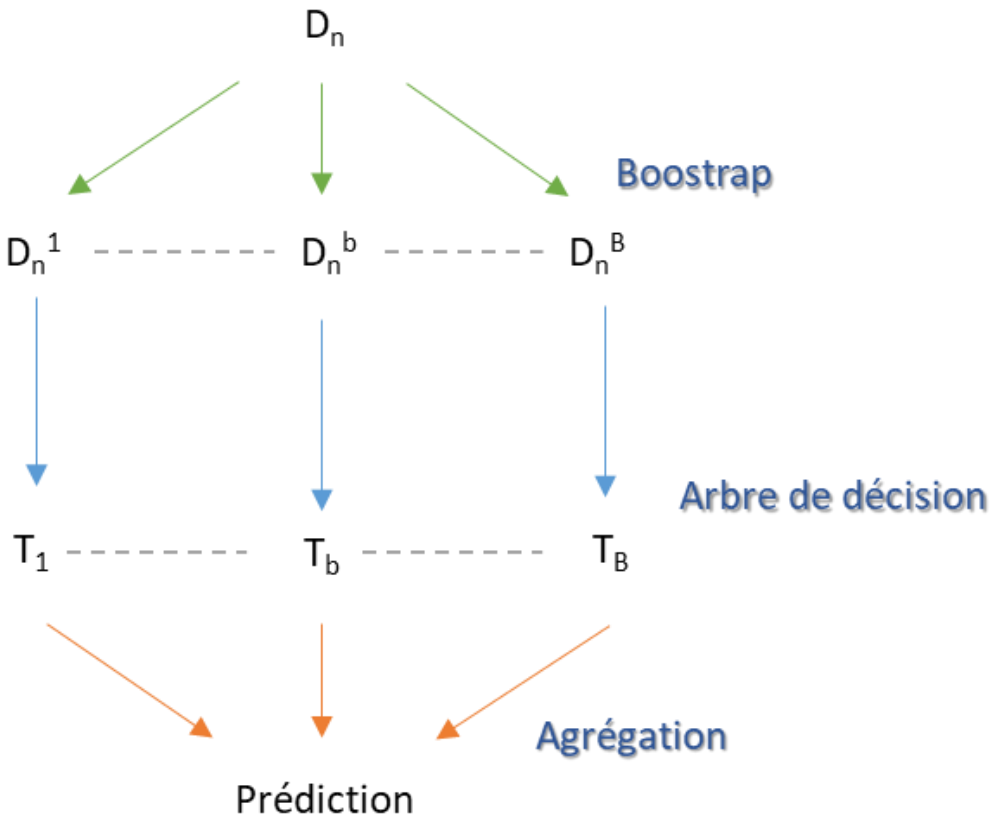


Figure 10 – Structure générale d'une forêt aléatoire

Cette méthode a l'avantage de réduire la variance des prévisions, puisque les arbres sont décorrélés. De plus, le surapprentissage est réduit par rapport à l'arbre de décision. Cependant, l'algorithme est plus lent (volume d'arbres important) et plus difficilement interprétable que les arbres.

2.3.2.3. Le gradient boosting de survie

Plutôt qu'un modèle particulier, le gradient boosting est davantage un cadre polyvalent permettant également de réduire l'erreur de prédiction et donc d'optimiser la prédiction. Il s'applique aussi bien aux modèles de régression qu'aux arbres de décision ou aux réseaux de neurones.

L'algorithme de base est celui du boosting auquel est ajoutée la notion de gradient : l'algorithme du gradient boosting est le fruit de la combinaison de ces deux approches. Le gradient boosting de survie a comme prédicteur faible un arbre.

Principe

Dans le **boosting**, à l'instar du bagging, les prédicteurs faibles sont également agrégés afin de constituer un prédicteur plus robuste. Mais il diffère du bagging par la **construction récurrente**, à savoir que **chaque nouveau prédicteur est constitué en fonction du précédent**.

Notamment dans le gradient boosting de survie, **l'amélioration de l'arbre à chaque itération est réalisée en corrigeant l'erreur de prédiction** (écart entre la valeur prédite et la valeur réelle de la variable cible) **du précédent arbre**. Cette amélioration se formalise par la minimisation d'une fonction de coût convexe, matérialisée par la descente de gradient.

En minimisant cette fonction de coût, **l'objectif de la descente par le gradient est de s'assurer qu'à chaque itération il y ait une progression vers la convergence du modèle**. La fonction de coût agit alors comme un baromètre, évaluant la précision à chaque itération.

Le gradient boosting permet alors d'entraîner des prédicteurs faibles de manière séquentielle, additive et progressive avec cette fonction de coût convexe, pour construire un prédicteur plus robuste.

Processus

Les principales étapes du gradient boosting se décomposent comme suit :

Etape 1 : initialisation

La première prédiction est initialisée avec une valeur constante, par exemple la moyenne de la valeur cible ($\bar{y}$)en régression.

Etape 2 : calcul du résidu

Le résidu résulte de la différence entre les prédictions de l'arbre précédent et les valeurs réelles de la variable cible : c'est l'erreur de prédiction.

Etape 3 : entraînement de l'arbre (prédicteur ou apprenant faible)

L'arbre est entraîné pour prédire les résidus calculés à l'étape 2.

Etape 4 : pondération des résidus prédits

Les résidus prédits par l'arbre de l'étape 3 sont pondérés par un facteur inférieur à 1 : le taux d'apprentissage, ce taux correspond à la taille des étapes choisies pour atteindre le minimum. L'objectif consiste à éloigner petit à petit les prédictions de la valeur constante (ici la moyenne) pour les rapprocher de la valeur réelle de la variable cible : c'est la convergence du modèle par la descente de gradient. Donc les nouvelles prédictions sont légèrement améliorées par rapport aux précédentes.

Etape 5 : mise à jour de la prédiction (nouvelle prédiction de la variable cible)

La nouvelle prédiction est mise à jour des résidus pondérés (prédiction précédente + résidu pondéré).

Les étapes 2 à 5 sont répétées jusqu'à obtenir un résultat jugé suffisamment performant.

Etape 6 : prédiction finale

Le résultat final est obtenu en agrégeant les prédictions issues des différents arbres qui sont des prédicteurs faibles.

Soit les notations suivantes :

(x_i, y_i) un échantillon de n individus
avec i = (1,…..,n)
y_i la variable cible et x_i les variables explicatives
$f(x)$ un modèle fonction de $x=\{x^1,\ldots..,x^p\} \in \mathbb{R}^p$
$L(y, f(x))$ une fonction de coût
$\delta > 0$ le taux d'apprentissage qui est un paramètre fixé (le pas de descente de gradient)

Algorithme

- Initialisation : $\widehat{f_0}(x) = \bar{y}$ ou
- Itération : pour m= 1 à M
  1- Calculer le résidu = calculer l'opposé du gradient aux points d'observation

  $$r_{i\,m} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right] = \hat{f}_m(x) - f(x)$$

  2- Entraîner un arbre avec les données $(x_i, r_{i\,m})$ pour prédire les résidus calculés à l'étape précédente : $\hat{r}_{i\,m}$
  3- Pondération des résidus prédits : $\delta\, \hat{r}_{i\,m}$
  4- Mise à jour de la prédiction : $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \delta\, \hat{r}_{i\,m}$

L'intérêt de cette approche est l'amélioration itérative : l'erreur de prédiction est corrigée à chaque itération. Par ailleurs, il tient compte des interactions entre les variables et peut traiter de manière efficace la non-linéarité entre les variables explicatives et la variable à prédire.
Ses principales limites tiennent à la perte du caractère intelligible des arbres, et à sa gourmandise en termes de temps et de puissance machine.

Synthèse des caractéristiques de chaque algorithme

- **Un arbre de survie (ST) :**
  - principe : partitionnement binaire et successif pour chaque variable explicative, de sorte à créer des groupes (nœuds) les plus homogènes possible en fonction de la variable cible ;
  - chaque nœud est divisé en maximisant la statistique de test du log-rank. Plus la valeur de cette statistique est élevée, plus la différence de survie entre les 2 nœuds fils est grande et meilleure est la coupure.
  - à la fin du partitionnement, les données des individus associés aux nœuds terminaux (feuilles) sont alors utilisées pour prédire la variable cible ;

- **Les forêts aléatoires de survie (RSF) :**
  - méthode ensembliste ;
  - principe du bagging :
    - entraîner des prédicteurs faibles en parallèle puis agréger les prédictions qui en sont issues ;
    - le but étant d'aboutir à un prédicteur plus performant, et donc minimiser les erreurs de prédiction ;
  - les prédicteurs faibles, différents les uns des autres, sont entraînés indépendamment les uns des autres (apprentissage en parallèle).

- **Le gradient boosting de survie (SGB) :**
  - méthode ensembliste ;
  - cadre polyvalent dont le prédicteur faible est ici un arbre ;
  - principe du boosting :
    - entraîner des prédicteurs faibles en séquentiel puis agréger les prédictions qui en sont issues ;
    - construction récurrente : chaque nouvel arbre (prédicteur faible) est constitué en corrigeant l'erreur de prédiction du précédent arbre ;
  - la descente de gradient permet la convergence du modèle en minimisant une fonction de coût convexe.

2.3.3. Choix de l'algorithme optimal par les métriques de la performance de prédiction

En analyse de survie, les performances prédictives d'un modèle sont évaluées sur deux critères : la **calibration et la discrimination**. L'objectif étant de mesurer la capacité de l'outil à prédire correctement la survenue de l'événement d'intérêt et à ordonner correctement les risques.

La discrimination évalue la capacité du modèle à différencier les individus ayant connu l'événement d'intérêt de ceux ne l'ayant pas connu, c'est-à-dire à classer correctement les individus selon le niveau de risque.

La calibration consiste à mesurer l'écart entre le risque prédit et le risque observé. Par exemple, un modèle qui prédit 20% de sorties d'arrêt de travail pour un profil donné devrait amener à observer environ 20 sorties parmi les 100 individus en arrêt ayant ce profil.

Plusieurs métriques existent en analyse de survie, les plus utilisées étant l'indice de concordance de Harell (C-index) et le score de Brier intégré (IBS Integrated Brier Score). Le C-index évalue le pouvoir de discrimination d'un modèle. L'IBS évalue à la fois la calibration et la discrimination d'un modèle.

2.3.3.1. L'indice de concordance de Harell (C-index)

L'indice de concordance de Harell permet d'évaluer le critère de **discrimination** : la capacité à classer correctement les individus en termes de risque.

Il est défini comme le rapport entre des paires d'individus ordonnées et comparables, choisies aléatoirement.
2 individus sont dits « comparables » si celui ayant la durée de survie la plus faible a connu l'événement d'intérêt (l'individu n'est donc pas censuré).
2 individus comparables sont dits « concordants » si le risque estimé par le modèle est supérieur pour l'individu ayant une durée de survie plus courte. Par exemple pour la prédiction du temps de survie jusqu'à l'apparition d'une maladie : un patient présentant une durée de survie plus courte (c'est-à-dire une durée plus courte avant l'arrivée de la maladie) devrait avoir un score de risque plus élevé.

Cette métrique tient compte des données censurées et prend une valeur entre 0 et 1. Un C-index de 1 indique une parfaite concordance entre score de risque et probabilité de survie, les individus peuvent être parfaitement classés en terme de risque associé à la survie. En revanche un C-index de 0,5 correspond à une classification aléatoire, dans ce cas le modèle est inutile.

Soit les notations suivantes :

> 2 individus i et j
> leur durée de survie respective T_i et T_j
> leur risque prédit respectif η_i et η_j
> δ_j l'indicatrice de censure de l'individu j, vaut 1 si T_j n'est pas censurée et 0 sinon

Si la durée de survie jusqu'à la survenance de l'événement d'intérêt (par exemple une maladie) d'un individu i est plus petite que celle d'un individu j, un bon modèle prédira une plus grande probabilité de survie pour l'individu j et donc un risque moindre $\eta_j < \eta_i$. Dans ce cas, la paire d'individus (i,j) est dite concordante. Dans le cas contraire, à savoir la prédiction d'un risque accru pour j alors que $T_j > T_i$, la paire (i,j) sera alors discordante.

Les deux types de paires, concordantes et discordantes, constituent des paires comparables.

En reprenant les notations ci-dessus et en considérant que T_j est désormais une durée censurée. Alors une paire d'individus est comparable dans la configuration où la durée de survie de l'individu non censuré est inférieure à la durée censurée de l'individu censuré, soit $T_i < T_j$. Un individu censuré peut être comparé uniquement avec un individu antérieur non censuré, et non avec un individu (censuré ou non) postérieur puisque la durée de survie est inconnue.

Le C-index est alors défini comme le rapport entre les paires d'individus concordantes sur le nombre de paires d'individus comparables :

$$C - index = \frac{Nombre\ de\ paires\ concordantes}{Nombre\ de\ paires\ concordantes + Nombre\ de\ paires\ discordantes}$$

$$\text{C-index} = \frac{\sum_{i \neq j} \mathbb{1}\{\eta_i < \eta_j\}\, \mathbb{1}\{T_i > T_j\}\, \delta_j}{\sum_{i \neq j} \mathbb{1}\{T_i > T_j\}\, \delta_j}$$

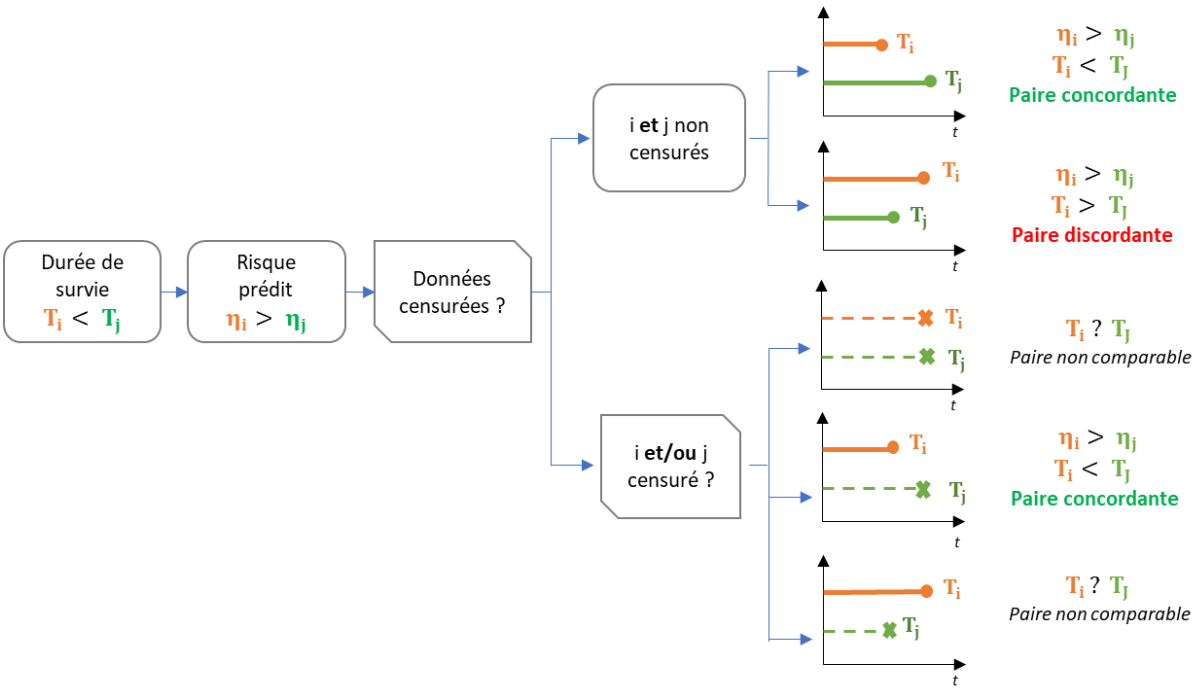


Figure 11 – C-index : définitions des paires comparables, concordantes et discordantes

2.3.3.2. Le score de Brier intégré (IBS)

Soit les notations suivantes

$Y_i(t) = 1|_{\{T_i>t\}}$: le statut de l'individu i au temps t
$\hat{S}(t|X_i)$: la probabilité de survie prédite au temps t pour l'individu i

Le score de Brier permet d'évaluer la précision de l'estimation de la fonction de survie à un instant t : il mesure la distance moyenne entre le statut réel et la fonction de survie prédite.
En l'absence de censure à un instant t > 0, le score de Brier se définit comme suit :

$$\mathrm{BS}(t, \hat{S}) = \mathbb{E}\left[(Y_i(t) - \hat{S}(t|X_i))^2\right]$$

Comme ce score correspond à une erreur quadratique, il est possible de le décomposer sous la forme d'un terme de biais et d'un terme de variance :

$$\mathrm{BS}(t, \hat{S}) = \mathbb{E}\left[(\mathbb{E}[Y_i(t)|X_i] - \hat{S}(t|X_i))^2\right] + \mathbb{E}[(Y_i(t) - \mathbb{E}[Y_i(t)|X_i])^2]$$

Le premier terme mesure la calibration, à savoir l'écart quadratique moyen entre les fonctions de survie individuelles théoriques et celles estimées. Le deuxième terme mesure la discrimination du modèle.

L'estimation du score de Brier en absence de censure est :

$$\widehat{\mathrm{BS}}(t, \hat{S}) = \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i(t) - \hat{S}(t|X_i)^2\right]$$

avec n le nombre d'individus dans l'ensemble de l'échantillon de test.

En présence de données censurées, le statut de l'individu i ne peut plus être utilisé (puisque le statut de l'individu censuré est inconnu), il est alors remplacé par le statut observé de l'individu i, soit $\widetilde{Y}_i(t) = 1|_{\{\widetilde{T}_i>t\}}$, et le score de Brier est ajusté en pondérant par la probabilité inverse des poids censurés $\widehat{W}_i(t)$:

$$\widehat{\mathrm{BS}}(t, \hat{S}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{W}_i(t)\left[(\widetilde{Y}_i(t) - \hat{S}(t|X_i)^2\right]$$

Enfin, le score de Brier intégré permet de résumer la performance prédictive estimée par le score de Brier, soit une moyenne des scores individuels de Brier :

$$\widehat{\mathrm{IBS}} = \frac{1}{\tau}\int_0^{\tau} \widehat{\mathrm{BS}}(t, \hat{S})\, dt$$

Où $\tau > 0$ qui peut être le maximum des temps observés et le score de Brier est moyenné sur l'intervalle [0,τ[

Plus la valeur du score de Brier ou de IBS est proche de 0, meilleure est la prédiction, sachant que le modèle peut être considéré utile quand le score de Brier est inférieur à 0,25.

A retenir sur les métriques de performance prédictive

En analyse de survie, les performances prédictives d'un modèle sont évaluées sur 2 critères : la discrimination et la calibration :

- la discrimination évalue la capacité du modèle à différencier les individus ayant connu l'événement d'intérêt de ceux ne l'ayant pas connu, c'est-à-dire à classer correctement les individus selon le niveau de risque ;
- la calibration consiste à mesurer l'écart entre le risque prédit et le risque observé. Par exemple, un modèle qui prédit 20% de sorties d'arrêt de travail pour un profil donné devrait amener à observer environ 20 sorties parmi les 100 individus en arrêt ayant ce profil.

Plusieurs métriques existent en analyse de survie, les plus utilisées étant l'indice de concordance de Harell (C-index) et le score de Brier intégré (IBS Integrated Brier Score) :

- le C-index évalue le pouvoir de discrimination d'un modèle. Plus sa valeur est proche de 1 et meilleures est la prédiction, sachant que le modèle peut être considéré comme utile quand le C-index est supérieur à 0,5.
- l'IBS évalue à la fois la calibration et la discrimination d'un modèle. Plus sa valeur est proche de 0 et meilleure est la prédiction, sachant que le modèle peut être considéré comme utile quand l'IBS est inférieur à 0,25.

Les différentes méthodes statistiques présentées jusqu'ici permettent de construire la loi brute de maintien en incapacité temporaire. Si les résultats présentent des variations irrégulières, mais autour d'une tendance, elles doivent être lissées. Ces irrégularités présentant un biais lors de l'application des lois. Ce point est abordé dans la prochaine section.

2.4. Lissage des taux bruts par la méthode de Whittaker-Henderson

La méthode de lissage Whittaker-Henderson est une méthode non-paramétrique reposant sur la minimisation de deux critères : un critère de fidélité (précision) et un critère de régularité en trouvant un compromis entre la fidélité aux données brutes et la régularité des données lissées.

Le critère de fidélité (fit) mesure l'écart quadratique entre le taux lissé q_x et le taux brut estimé $\hat{q}_x$, pondéré par un poids w_x :

$$F = \sum_{x=1}^{n} w_x \ (q_x - \hat{q}_x)^2$$

Le critère de régularité (smoothness) mesure l'écart entre les valeurs lissées :

$$S = \sum_{x=1}^{n-z} (\Delta^z q_x)^2$$

où z est un paramètre du modèle qui fixe le degré du polynôme utilisé pour le critère de régularité, il est généralement compris entre 2 et 4 (ce qui permet de tenir compte suffisamment d'informations consécutives)

L'objectif est de minimiser la combinaison linéaire de la fidélité et de la régularité en mettant plus ou moins l'accent au moyen du paramètre h :

$$M = F + h\,S$$

$$\text{Soit } M = \sum_{x=1}^{n} w_x \ (q_x - \hat{q}_x)^2 + h \sum_{x=1}^{n} (\Delta^z q_x)^2$$

Plus h est élevé et plus la régularité est importante. S'il est souhaité un critère de régularité plus important que celui de fidélité, il faut h>1.

La couverture incapacité temporaire du régime statutaire diffère de celle du régime général : 3 états d'incapacité de durée maximale d'indemnisation différente (12 mois en CMO, 36 mois en CLM et 60 mois en CLD) contre un seul état de 3 ans maximum. La table de maintien du BCAC n'est alors pas adaptée pour le calcul des provisions et la construction de lois d'expérience est donc justifiée.

Pour construire les tables d'expérience, les méthodes statistiques doivent prendre en compte une spécificité de l'analyse de survie : l'existence de données incomplètes, les censures et troncatures. L'approche classique est abordée avec l'estimateur de Kaplan-Meier et l'apprentissage supervisé est illustré par l'arbre de survie, les forêts aléatoires de survie et le gradient boosting de survie.

Pour appliquer ces différentes méthodes sur notre portefeuille, les données doivent être analysées et traitées : ces éléments sont abordés dans le chapitre suivant.

A retenir :

Une loi de maintien en incapacité temporaire est construite avec les méthodes statistiques appartenant à l'analyse de survie, dont la spécificité réside dans la prise en compte de données incomplètes (censures et troncatures).

L'approche classique est abordée avec l'estimateur de Kaplan-Meier :

- il est non paramétrique et prend en comptes les censures et les troncatures ;
- principe : estimer à chaque pas de temps la fonction de survie (estimer la probabilité de maintien en incapacité à chaque jour d'indemnisation) ;
- évaluation de la pertinence des estimations par un intervalle de confiance à 95% des taux de maintien. Sa construction est basée sur la variance de l'estimateur de Kaplan-Meier. Un estimateur de la variance de la fonction de survie à un instant t est donné par l'estimateur de Greenwood.

Concernant l'approche par l'apprentissage supervisé, les 3 algorithmes présentés ont un point commun : l'arbre. Notamment l'arbre de survie prend en compte les données incomplètes. A partir d'un modèle dit simple qu'est l'arbre, il est théoriquement possible d'améliorer les performances de prédiction soit en moyennant les prédictions de différents arbres (les forêts de survie), soit en moyennant les prédictions corrigées de différents arbres (le gradient boosting de survie).
La pertinence des estimations est évaluée par 2 indicateurs de performance prédictive : l'indice de concordance de Harell (C-index) et le score de Brier intégré (IBS)

3. LES DONNEES

L'analyse et le traitement des données constituent une étape cruciale et chronophage. Plus de la moitié du temps nécessaire à la construction d'une table d'expérience est imputable à cette étape, dont dépend la qualité des lois estimées.
Le chapitre 3 présentera dans un premier temps la construction de la base de sinistres : les différents traitements et contrôles nécessaires pour adapter la base de données initiale à chaque méthode exposée dans le chapitre précédent. Ensuite, de cette base de sinistre, seront mis en avant les principales caractéristiques du portefeuille étudié, via des statistiques descriptives.

3.1. Construction de la base de sinistres

Pour construire les tables d'expérience, est utilisé un fichier de données constitué d'un ligne à ligne des prestations complémentaires d'incapacité temporaire versées par l'assureur.

L'adaptation de ce fichier aux différentes méthodes utilisées constituera la base de sinistres utilisée pour l'application empirique sur le portefeuille. De cette étape, cruciale, dépend la qualité et l'efficacité des lois estimées.

Il sera nécessaire de sélectionner les variables pertinentes avec les méthodes utilisées et de réaliser les contrôles de qualité et de cohérence des données pour ensuite pouvoir implémenter les traitements spécifiques.

3.1.1. Sélection des variables pertinentes

La base de données source contient 24 variables.

Les variables retenues répondent à deux critères : fiabilité et pertinence. La fiabilité est matérialisée soit par la complétude de la variable soit par une information communiquée directement par l'assureur. La pertinence est conditionnée par les besoins de la méthode utilisée.

Le fichier source se présente comme suit :

- Date de l'exercice comptable
- Origine de l'entrepôt de données
- Identifiant 1 de l'assuré
- Identifiant 2 de l'assuré
- Identifiant 3 de l'assuré
- Sexe (F ou H)
- Date de survenance : date d'entrée en incapacité

- Type d'indemnisation (différents états de l'arrêt de travail) :
  - CMO : congé de maladie ordinaire;
  - CLM : congé de longue maladie ;
  - CLD : congé de longue durée ;
  - DI1 et DI2 : DO venant en relais d'un des états d'incapacité temporaire (CMO, CLM ou CLD).
- Date de début : date du premier jour indemnisé par l'assureur
- Date de fin : date du dernier jour indemnisé par l'assureur
- Montant unitaire d'indemnisation
- Montant total d'indemnisation
- Date de naissance de l'assuré
- Motif de sortie de l'arrêt de travail :
  - Standard : l'assuré reprend une activité ;
  - Retraité : l'assuré part à la retraite ;
  - Invalidité : l'assuré est consolidé en invalidité ;
  - Décès : l'assuré décède.
- Index
- Salaire
- Indice : correspond à un niveau de salaire
- Date de paiement
- Date de premier paiement
- Date de dernier paiement
- Date de mise en invalidité dans le cadre d'un passage en l'état d'invalidité
- Numéro de sécurité sociale
- Libellé de la structure : deux structures distinctes identifiées par A et B.

Une première sélection consiste à retenir les variables nécessaires pour les contrôles de qualité et de cohérence des données.

Les **variables non retenues**, présentées ci-dessous, ne fournissent aucune explication sur la durée de maintien en incapacité ou sont à vide pour la majorité des observations (>80%) :

- Date de l'exercice comptable : date unique correspond à la date d'arrêté des données ;
- Origine de l'entrepôt de données : source de la base infocentre de l'assureur ;
- les montants indemnisés ne sont pas pris en compte car ils n'ont pas d'effet sur la durée, mais le montant aurait pu être une variable cible à prédire ;
- Index est une variable à vide ;
- le salaire est majoritairement à 0 (contrairement à l'indice qui équivaut à un montant de salaire chez les fonctionnaires) ;
- Indice est une variable non fiabilisée (information communiquée par l'assureur) ;
- les 3 dates de paiement ne sont pas fiabilisées (information communiquée par l'assureur) ;

- Date de mise en invalidité est traduite dans la variable MOTIF par INVALIDITE ;
- le libellé de la structure est une variable incomplète à plus de 80%.

Les variables retenues sont les suivantes :

- Identifiant 1 de l'assuré
- Identifiant 2 de l'assuré
- Identifiant 3 de l'assuré
- Motif de sotie de l'arrêt de travail
- Sexe
- Numéro de sécurité sociale
- Date de naissance de l'assuré
- Date de survenance (DS) = date d'entrée en incapacité temporaire
- Date de premier paiement (DDI) = date de début d'indemnisation (présence de franchise et autre)
- Date de dernier paiement (DFI) = date de fin d'indemnisation.

3.1.2. Contrôle de la qualité et de la cohérence des données

Pour vérifier la qualité et la cohérence des données, différentes analyses sont menées et présentées dans la suite.

La recherche d'éventuelles données aberrantes :

- la complétude des données : existence de données à vide ;
- l'exactitude des données :
    - valeurs minimales et maximales des dates de naissance ;
    - valeurs minimales et maximales des dates de survenance ;
    - valeurs minimales et maximales des dates de début et de fin d'indemnisation.

La cohérence des dates de survenance et d'indemnisation :

- antériorité de la date de survenance sur la date de début d'indemnisation :
  date de début d'indemnisation >= date de survenance ;
- antériorité de la date de début d'indemnisation sur la date de fin d'indemnisation :
  date de début d'indemnisation <= date de fin d'indemnisation ;

La cohérence de l'unicité des variables :

- unicité de la date de naissance de l'assuré par identifiant assuré ;
- unicité du sexe par assuré ;

- unicité de l'état d'incapacité par période d'indemnisation : pour la même période d'indemnisation, un assuré n'est pas indemnisé au titre de deux états différents ;
- unicité du motif de sortie de l'arrêt de travail par arrêt ;
- unicité de l'indice par assuré et par arrêt.

Résultats et traitements :

- les trois variables d'identification de l'assuré : la variable *Identifiant 3* est retenue car elle ne présente aucune observation à vide et est unique par numéro de sécurité sociale. Les variables *Identifiant 1* et *Identifiant 2* sont par conséquent supprimées de l'étude ;
- la variable *Sexe* : 5% des observations sont à vides. Après récupération de l'information par le numéro de sécurité sociale (variable complète), la variable *Sexe* est complète ;
- la variable *Date de naissance de l'assuré* : un assuré présente deux dates de naissance différentes (les années sont différentes). Après récupération de l'information par le numéro de sécurité sociale, la variable est complète ;
- le numéro de sécurité sociale est supprimé pour des raisons de confidentialité des données (RGPD), l'assuré étant identifié par la variable *Identifiant 3*.

La base de données finale avant traitement spécifique se présente comme suit :

- Identifiant 3 de l'assuré
- Motif de sortie de l'arrêt de travail
- Sexe
- Date de naissance de l'assuré
- Date de survenance (DS) = date d'entrée en incapacité temporaire
- Date de premier paiement (DDI) = date de début d'indemnisation (présence de franchise et autre)
- Date de dernier paiement (DFI) = date de fin d'indemnisation.

Les tests réalisés sur les données de la base n'ont permis de déceler aucune anomalie significative pouvant remettre en cause la validité de la base de sinistres.

3.1.3. Choix de la période d'observation

Le choix de la période d'observation est important lors de l'élaboration de la loi de maintien en incapacité temporaire :

- la durée de la période d'observation doit être suffisamment longue pour disposer d'un volume de données important ;
- cette durée doit également être relativement courte pour assurer une homogénéité de l'échantillon : influence du contexte socio-économique sur la sinistralité + durée cohérente pour limiter le biais introduit par les censures à droite (les arrêts clos après la fin de la période d'observation) ;
- la période d'observation définit la proportion de censures.

La base de données étudiée est arrêtée au 31/12/2021, **la période d'observation est de 6 ans, du 01/01/2014 au 31/12/2019** :

- cette période de 6 ans permet d'avoir un échantillon homogène et représentatif ;
- la fin d'observation à N-2 permet d'avoir un recul de deux ans sur les arrêts clôturés.

3.1.4. Traitements spécifiques pour adapter les données au risque réel

Les traitements sont présentés dans l'ordre de réalisation : chaque étape dépendant de la réalisation de l'étape précédente.

Recodification de la DO faisant suite à un arrêt pour incapacité

Chaque état d'incapacité temporaire (CMO, CLM ou CLD) nécessite une table d'expérience dont le nombre de mois d'indemnisation doit correspondre à la durée maximale d'indemnisation de l'état.

Toutefois, après expiration des droits aux congés d'incapacité temporaire et sous certaines conditions, la DO pour raison de santé prend le relais et permet de prolonger l'indemnisation de l'incapacité de travail temporaire.

La DO est identifiée par la variable *Type d'indemnisation* : DI1 et DI2. Quand DI1 et DI2 sont identifiées en relais des états initiaux CMO, CLM ou CLD, elles sont alors recodifiées en fonction de l'état initial.

Exemple pour un assuré en congé de maladie ordinaire

A identifiant assuré et date de survenance identiques, quand un DI1 (respectivement DI2) présente une date de début d'indemnisation venant à la suite de la date de fin d'indemnisation de la dernière période indemnisée d'un état CMO, CLM ou CLD, alors la ligne DI1 est identifiée comme une DO venant en relais de l'état initial :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	01/10/2024	CMO	01/01/2015	15/01/2015
12345678	**01/10/2024**	CMO	16/01/2015	03/02/2015
12345678	**01/10/2024**	**DI1**	04/02/2015	02/06/2015

Dans ce cas, la ligne de prestations de DI1 est modifiée de sorte que le type d'indemnisation DI1 soit remplacé par CMO et que la date de début d'indemnisation soit remplacée par celle de la dernière période indemnisée du CMO :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	01/10/2024	CMO	01/01/2015	15/01/2015
12345678	01/10/2024	CMO	16/01/2015	03/02/2015
12345678	01/10/2024	**CMO**	04/02/2015	02/06/2015

Concaténation des périodes d'indemnisation

Il faut distinguer deux types de concaténation : celle des périodes se chevauchant et celle des périodes qui se suivent.

Les périodes de chevauchement sont identifiées de la manière suivante : à identifiant assuré, type d'indemnisation, date de survenance et date de début d'indemnisation identiques, la date de fin de d'indemnisation diffère (le cas où la date de fin d'indemnisation est identique et les dates de début d'indemnisation diffèrent ne s'est pas présenté) :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	01/10/2024	CMO	01/01/2015	15/01/2015
12345678	**01/10/2024**	**CMO**	16/01/2015	03/02/2015
12345678	**01/10/2024**	**CMO**	16/01/2015	02/06/2015

Dans ce cas, la date de fin d'indemnisation la plus récente est retenue :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	01/10/2024	CMO	01/01/2015	15/01/2015
12345678	**01/10/2024**	**CMO**	**16/01/2015**	02/06/2015

La concaténation des périodes d'indemnisations successives consiste à agréger deux enregistrements qui se suivent si, à identifiant assuré, type d'indemnisation et date de survenance identiques, la date de fin de période indemnisée du premier enregistrement ± 1 jour est égale à la date de début de période indemnisée du second enregistrement :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	**01/10/2024**	**CMO**	01/01/2015	15/01/2015
12345678	**01/10/2024**	**CMO**	16/01/2015	02/06/2015

Dans ce cas, la date de début de période indemnisée de l'enregistrement agrégé est égale à la date de début la plus ancienne et la date de fin de période indemnisée de l'enregistrement agrégé est égale à la date de fin la plus récente :

Identifiant de l'assuré	Date de survenance	Type d'indemnisation	Date de début d'indemnisation	Date de fin d'indemnisation
12345678	**01/10/2024**	**CMO**	01/01/2015	02/06/2015

Application des seuils de concaténation

Un seuil de concaténation est la durée, séparant les périodes d'indemnisation de deux sinistres de même nature, en-dessous de laquelle ces deux sinistres sont considérés avoir la même origine et former un seul sinistre.

La concaténation est fonction de l'exercice de survenance et de l'état en arrêt. Ainsi, pour le risque COM, deux observations consécutives d'un même assuré et d'une même nature de risque sont agrégés si la date de début de période indemnisée de la seconde observation est inférieure à la date de fin de période indemnisée de l'observation précédent majorée de 7 jours. Pour le risque CLM, le seuil d'agrégation est de 90 jours. Pour le risque CLD, il est de 180 jours.

Suite à l'ensemble de ces traitements, la base de sinistres présente donc une ligne par arrêt.

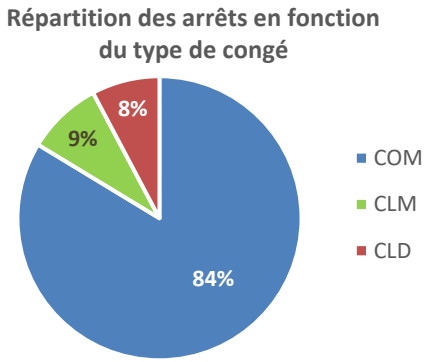
Ajout de variables

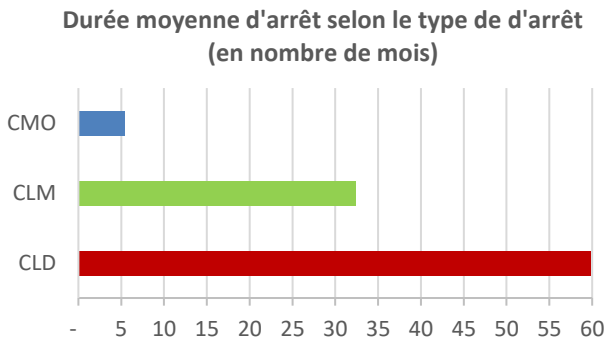
Pour les besoins des différentes méthodes qui seront appliquées pour le calcul des lois de maintien, plusieurs variables sont ajoutées :

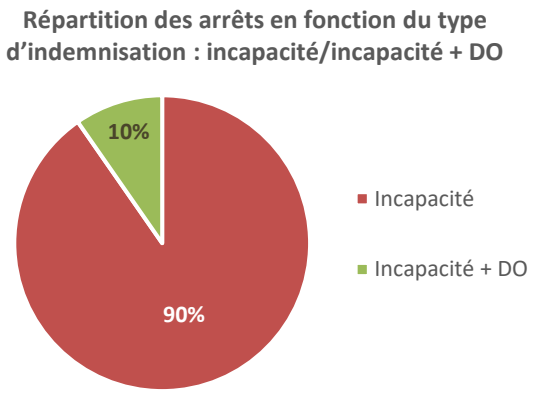
- la date de début d'observation : 01/01/2014 ;
- la date de fin d'observation : 31/12/2019 ;
- une variable permettant de distinguer les arrêts incapacité sans prolongation par la DO des arrêts incapacité avec prolongation DO ;
- une variable identifiant si l'observation est censurée ou non ;
- une variable indiquant la durée d'arrêt totale.

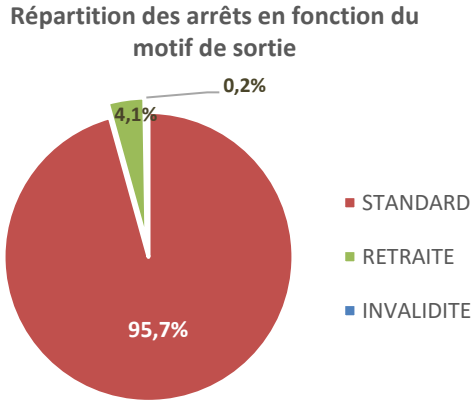
3.2. Statistiques descriptives

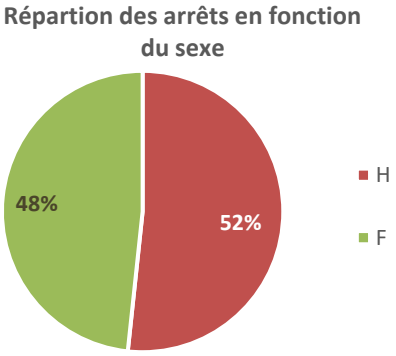
3.2.1. Ensemble du portefeuille











Le portefeuille étudié est composé majoritairement de CMO à hauteur de 84%, le reste des arrêts se partageant à part quasi égale entre le CLM pour 9% et le CLD pour 8%.

La durée moyenne d'un arrêt pour CMO est de 5,5 mois, d'un arrêt pour CLM est de 32 mois (environ 2,5 ans) et enfin d'un arrêt pour CLD est de 60 mois (5 ans).

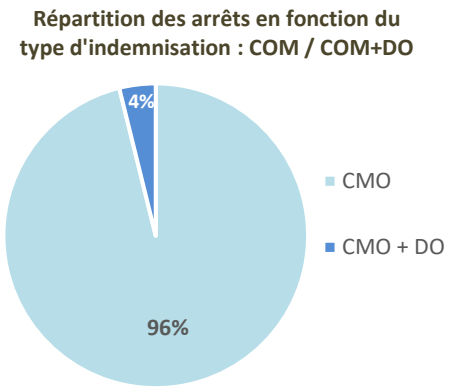
Environ 10% des arrêts sont prolongés par une DO.

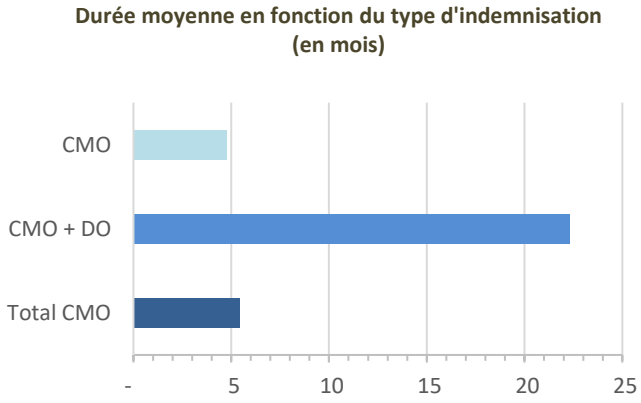
Les arrêts se répartissent quasi équitablement entre les hommes (52%) et les femmes (48%) avec un âge moyen à l'arrêt légèrement plus élevé pour les hommes à 54 ans contre 53 ans pour les femmes.

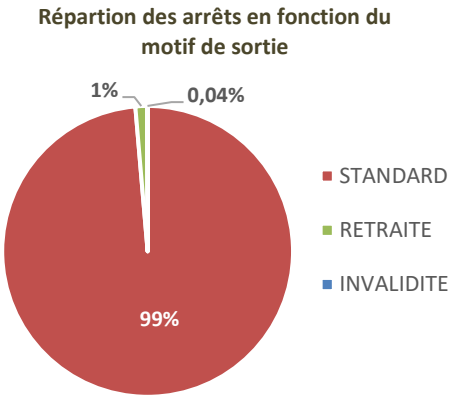
L'âge moyen à l'arrêt au global étant de 54 ans (54 ans pour les CMO et les CLM, 53 ans pour les CLD).

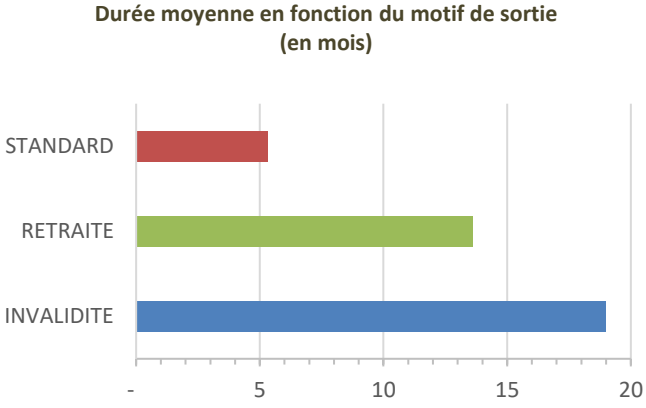
Le motif de sortie des arrêts est principalement standard (autre qu'un décès, une invalidité ou encore un départ à la retraite) pour 96% des arrêts.

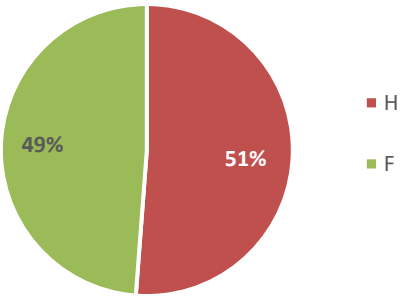
3.2.2. Arrêt pour congé de maladie ordinaire CMO

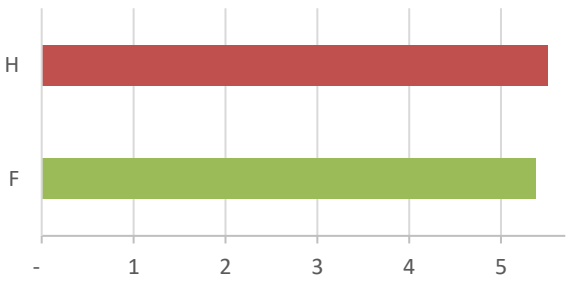


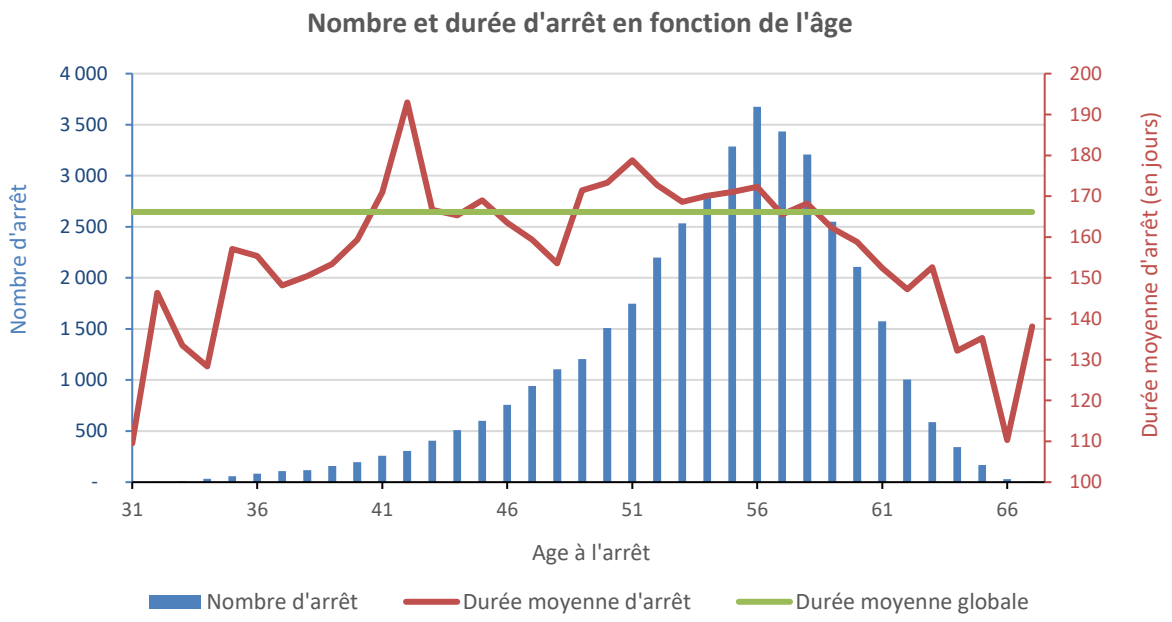










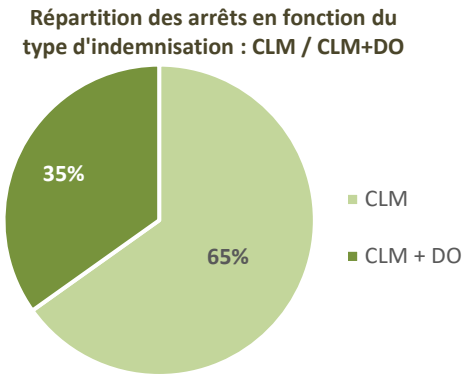


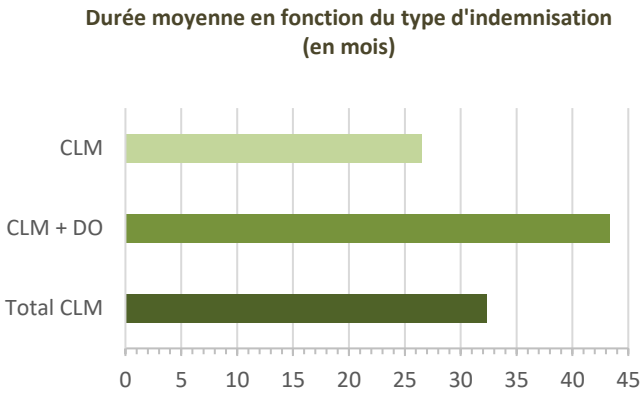
En moyenne, un arrêt pour CMO dure 5,5 mois (166 jours) avec un âge à la survenance de 54 ans, et présente un motif de sortie standard.

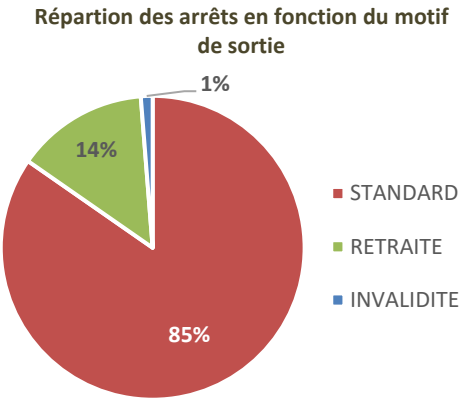
La majorité des arrêts CMO sont sans prolongation DO (96%) avec une durée moyenne de 5 mois, contre 22 mois pour les arrêts prolongés par une DO.

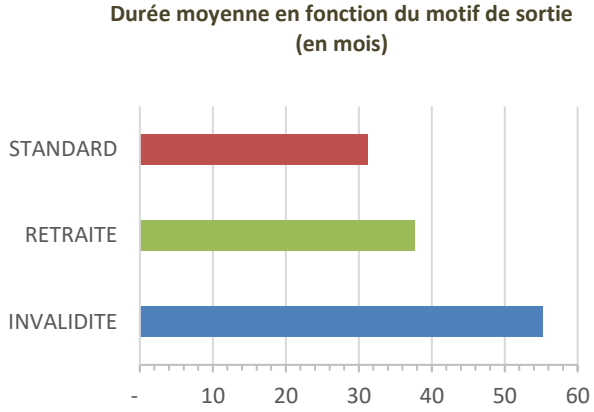
En termes de nombre d'arrêts, la tranche [48 ; 62] ans représente 85% des arrêts. Leur durée moyenne variant dans l'intervalle [-11% ; 8%] par rapport à la durée moyenne de 166 jours.

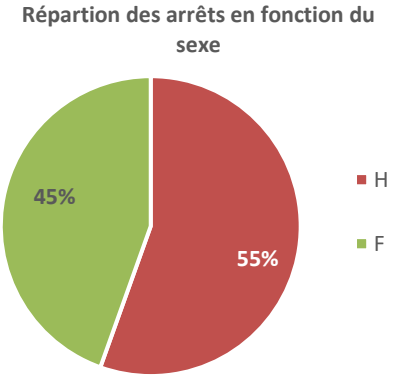
3.2.3. Arrêt pour congé de longue maladie CLM

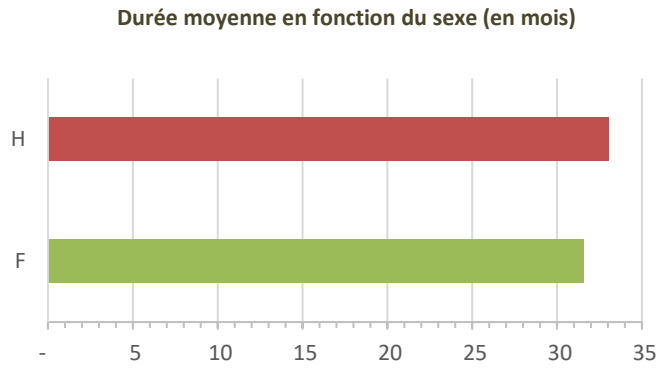


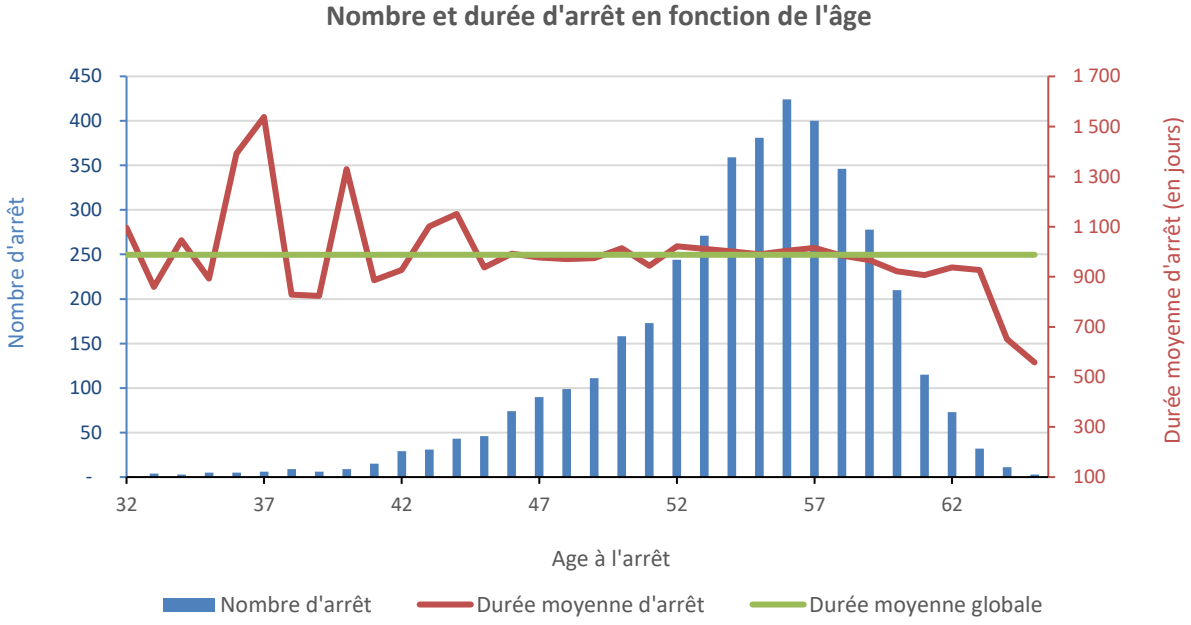



Répartion des arrêts en fonction du motif de sortie
1%
14%
85%
STANDARD
RETRAITE
INVALIDITE


Durée moyenne en fonction du motif de sortie (en mois)
STANDARD
RETRAITE
INVALIDITE
-
10
20
30
40
50
60


Répartion des arrêts en fonction du sexe
45%
55%
H
F


Durée moyenne en fonction du sexe (en mois)
H
F
-
5
10
15
20
25
30
35


Nombre et durée d'arrêt en fonction de l'âge
Nombre d'arrêt
Durée moyenne d'arrêt (en jours)
Age à l'arrêt
Nombre d'arrêt
Durée moyenne d'arrêt
Durée moyenne globale

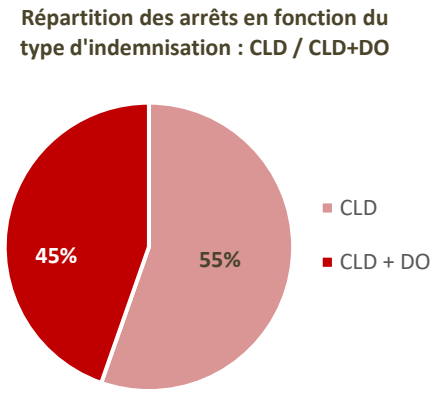
En moyenne, un arrêt en CLM dure 32 mois (988 jours) avec un âge à la survenance de 54 ans.

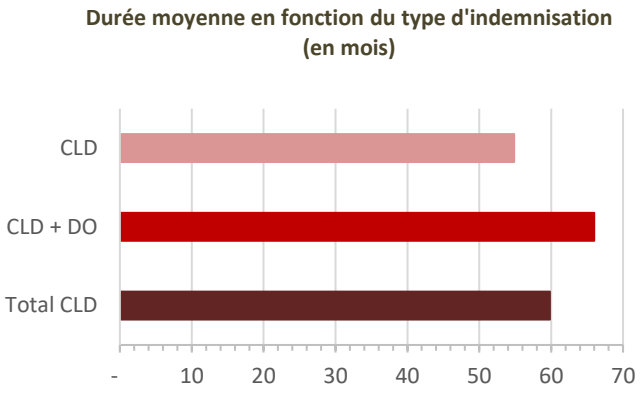
Les arrêts présentent un motif de sortie standard à hauteur de 85% et pour départ à la retraite de 14%.

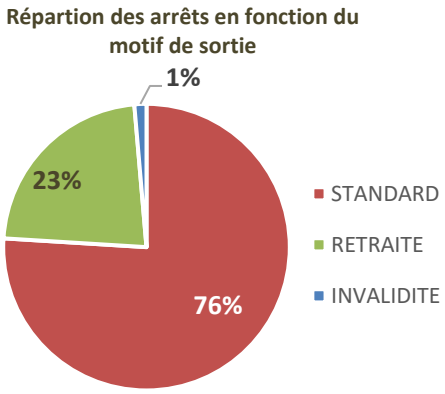
Près de 35% des arrêts CLM sont prolongés par une DO et présentent une durée moyenne d'arrêt de 43 mois, contre 27 mois pour les arrêts sans prolongation DO.

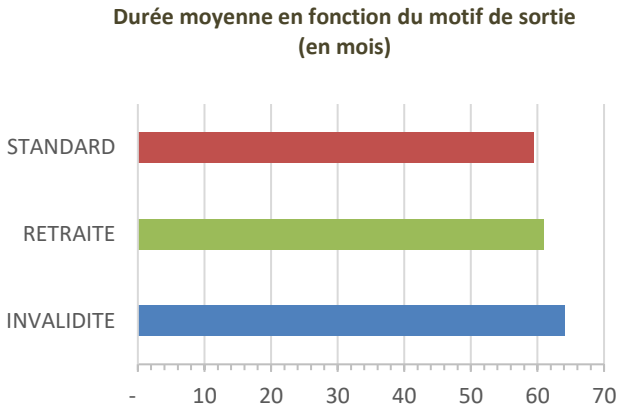
En termes de nombre d'arrêts, la tranche [48 ; 61] ans représente 88% des arrêts. Leur durée moyenne variant dans l'intervalle [-8% ; 3%] par rapport à la durée moyenne de 988 jours.

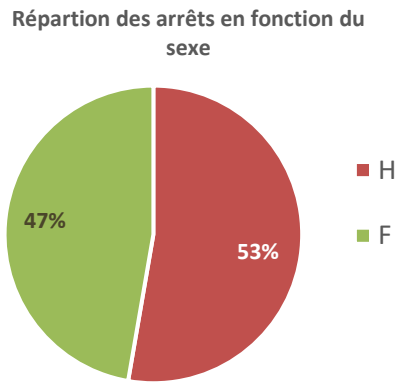
3.2.4. Arrêt pour congé de longue durée CLD

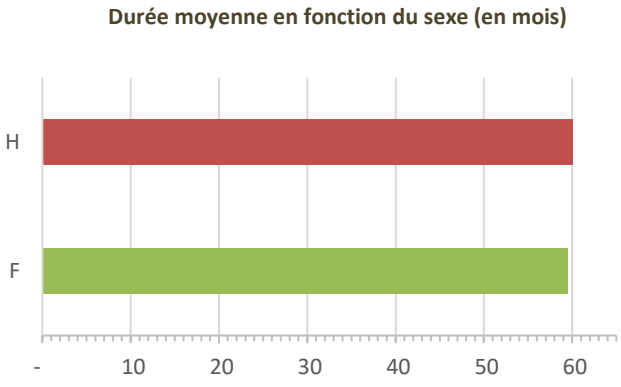


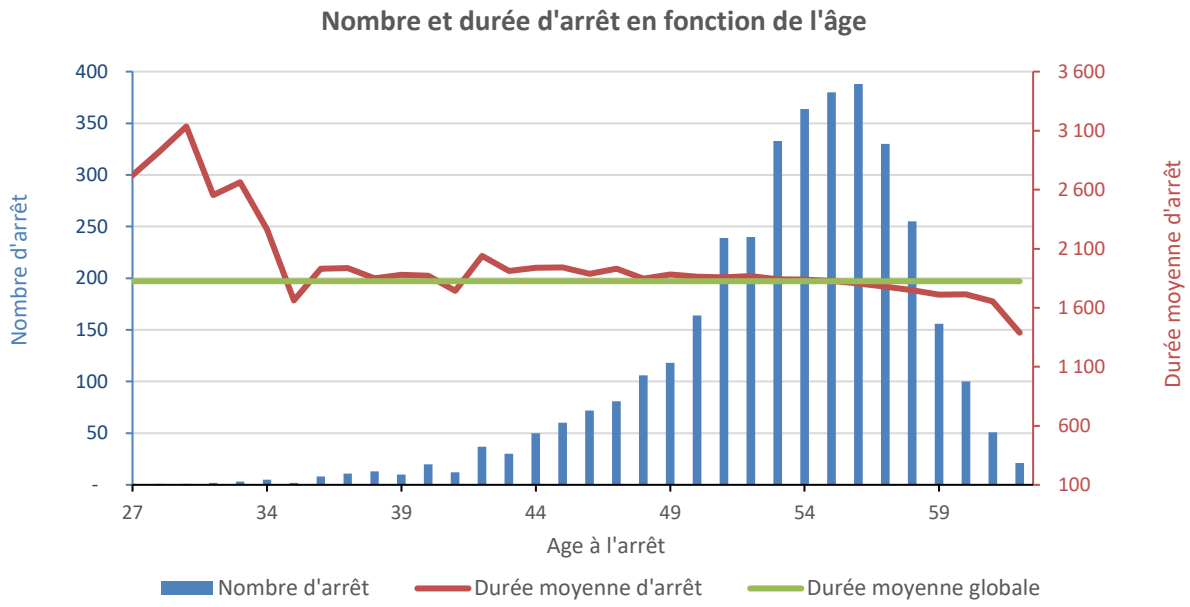












En moyenne, un arrêt pour CLD dure 60 mois (1825 jours) avec un âge à la survenance de 53 ans.

Le motif de sortie standard est présent à hauteur de 76% des arrêts et le départ à la retraite pour 23%.

Environ 45% des arrêts CLD sont prolongés par une DO et présentent une durée moyenne d'arrêt de 66 mois, contre 55 mois pour les arrêts CLD sans prolongation DO.

La tranche [48 ; 60] ans représente 86% des arrêts. Leur durée moyenne variant dans l'intervalle [-6% ; 6%] par rapport à la durée moyenne de 1825 jours.

La couverture incapacité temporaire du régime statutaire diffère de celle du régime général : 3 états d'incapacité de durée maximale d'indemnisation différente (12 mois en CMO, 36 mois en CLM et 60 mois en CLD) contre un seul état de 3 ans maximum. La table de maintien du BCAC n'est alors pas adaptée pour le calcul des provisions et la construction de lois d'expérience est donc justifiée.

Pour construire les tables d'expérience, les méthodes statistiques doivent prendre en compte une spécificité de l'analyse de survie : l'existence de données incomplètes, les censures et troncatures. L'approche classique est abordée avec l'estimateur de Kaplan-Meier et l'apprentissage supervisé est illustré par l'arbre de survie, les forêts aléatoires de survie et le gradient boosting de survie.

Pour appliquer ces différentes méthodes sur notre portefeuille, les données sont analysées et traitées. Les données du portefeuille étudié mettent en avant des caractéristiques différentes pour chaque état. Ce constat confirme la nécessité de construire une loi de maintien par état. Les différents traitements et contrôles appliqués à la base de données initiale ont abouti à la base de sinistre, qui sera alors utilisée pour appliquer les méthodes statistiques présentées dans le chapitre 2.

A retenir :

La construction de la base de sinistres est une étape cruciale : la qualité et l'efficacité des lois estimées en dépendent. Cette construction consiste à nettoyer les données (sélection des variables pertinentes et divers traitements) et contrôler la qualité et la pertinence de ces données.

Ensuite, de cette base nettoyée, sont mises en avant les caractéristiques du portefeuille, qui se distinguent pour chaque état d'incapacité.

Les arrêts pour congé de maladie ordinaire sont relativement homogènes : la majorité des CMO sont sans prolongation de DO (96%) et présentent un motif de sortie standard (99%). Les hommes et les femmes sont représentés à part quasi-égale (51%/49%) et ont une durée moyenne d'arrêt identiques.

Les arrêts pour congé de longue maladie se distinguent davantage : 35% des arrêts sont prolongés par une DO, 85% des arrêts ont un motif de sortie standard, le reste étant des sorties pour départ à la retraite.

Les arrêts pour congé de longue durée se distinguent également : 45% des arrêts sont prolongés par une DO, 76% des arrêts ont un motif de sortie standard et 23% des arrêts sortent pour un départ à la retraite.

4. APPLICATION DES DEUX APPROCHES : COMPARAISON DE L'APPROCHE KAPLAN-MEIER AVEC L'APPROCHE APPRENTISSAGE SUPERVISE POUR NOTRE PORTEFEUILLE

Chaque état d'incapacité est spécifié par des caractéristiques propres, notamment une durée maximale d'indemnisation différente. Une loi est alors construite par état :

- une loi de maintien en incapacité temporaire pour le CMO ;
- une loi pour le CLM ;
- une loi pour le CLD.

Les lois de maintien en incapacité temporaire sont construites de sorte à prendre en compte l'ensemble de la durée d'indemnisation d'un état d'incapacité, complétée par une éventuelle prolongation par la DO et par la garantie complémentaire de l'assureur.

Le tableau suivant récapitule l'ensemble de ces différentes durées d'indemnisation.

Etat d'incapacité	Durée maximale d'indemnisation de l'état d'incapacité (a)	Dont durée de franchise	Durée maximale d'indemnisation par la DO (b)	Durée maximale d'indemnisation par la garantie complémentaire de l'assureur (c)	Durée de la loi de maintien en incapacité y compris la période de franchise (a)+(b)+(c)
CMO	12 mois	3 mois (sauf exception)	36 mois	12 mois	60 mois
CLM	36 mois	12 mois			84 mois
CLD	60 mois	36 mois			108 mois

Par ailleurs, au regard de la spécificité du portefeuille (groupe fermé) et par souci de simplification, la loi construite n'est pas discriminée par âge. La table de maintien en incapacité se présente donc sous la forme d'un tableau de deux colonnes : le nombre d'incapacité par mois d'ancienneté sur une base de 100 000 incapables au mois 0.

Ce chapitre exposera dans un premier temps les résultats de la construction des lois par l'approche classique : l'estimation des lois brutes par Kaplan-Meier, l'évaluation de la qualité de ces estimations par un intervalle de confiance à 95% ainsi que le lissage par la méthode de Whittaker-Henderson.
Dans un deuxième temps, seront mis en avant les résultats des algorithmes d'apprentissage supervisé. Une explication synthétique du déroulé de la procédure de calcul d'un algorithme précèdera l'analyse des résultats des performances prédictives. La dernière partie de cette section présentera les résultats des lois brutes. Sachant que chaque état d'incapacité voit sa loi construite avec l'algorithme ayant les meilleures performances prédictives en termes de C-index et/ou d'IBS.
Enfin, un backtesting illustrera la comparaison entre la méthode classique et l'algorithme d'apprentissage supervisé et permettra de conclure son apport.

4.1. Construction des lois de maintien en incapacité avec l'approche classique

Pour chacune des sous-sections :

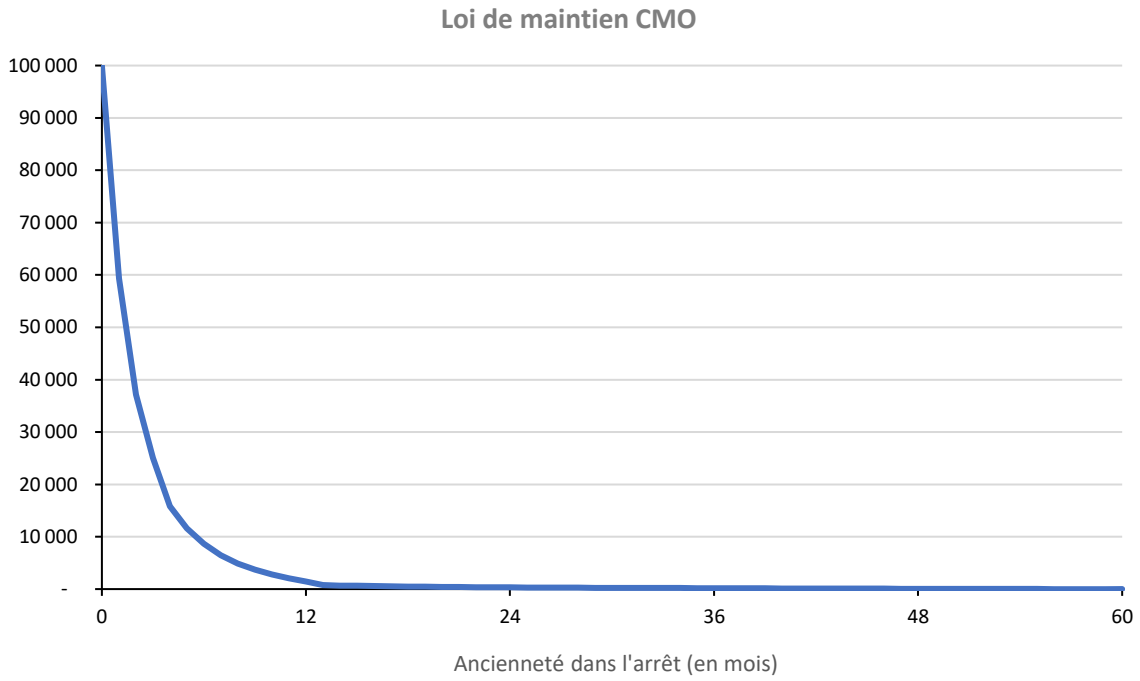
- le calcul des lois brutes par l'estimateur de Kaplan-Meier ;
- la validation des calculs avec un intervalle de confiance à 95% ;
- le lissage par la méthode de Whittaker-Henderson ;

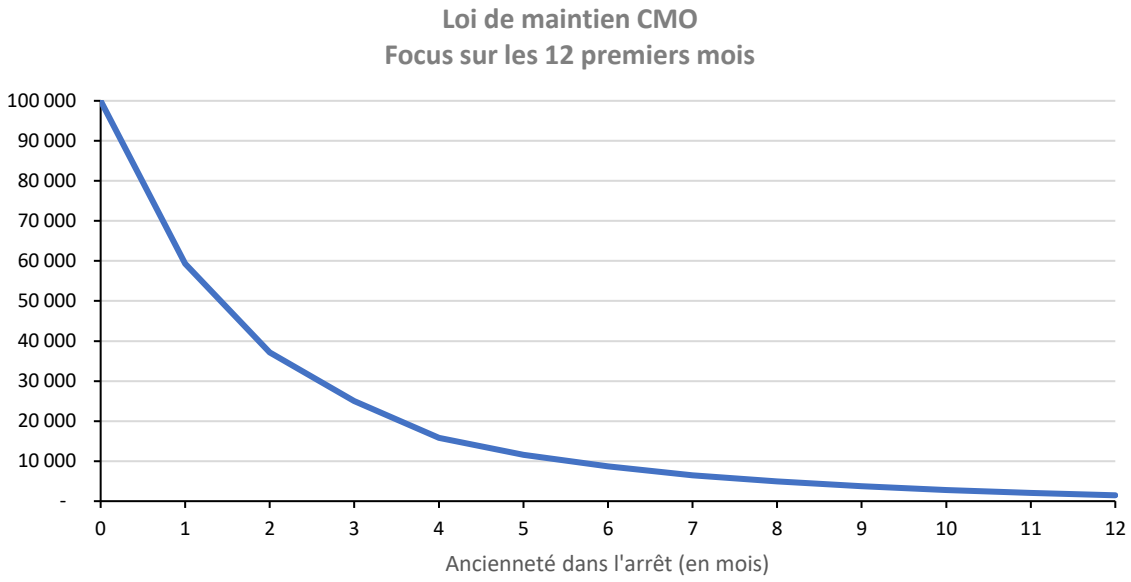
les résultats sont présentés par état d'incapacité : CMO, CLM et CLD.

Comme la majorité des arrêts CMO sort au bout de 12 mois (96%), en complément sera exposé un focus sur les 12 premiers mois.

4.1.1. Calcul des lois brutes avec Kaplan-Meier

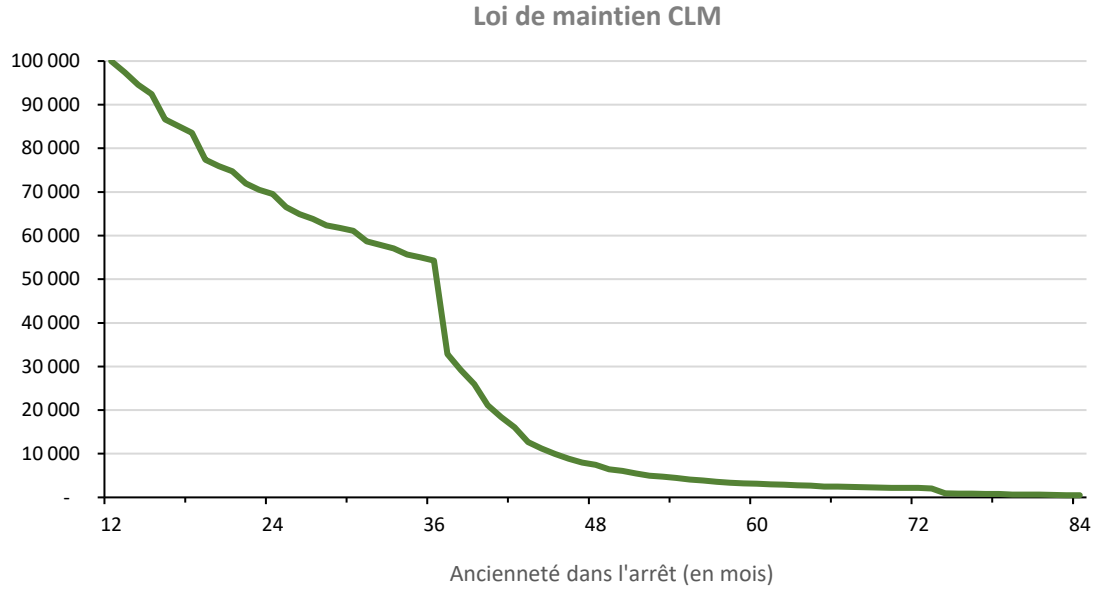
Le congé de maladie ordinaire CMO





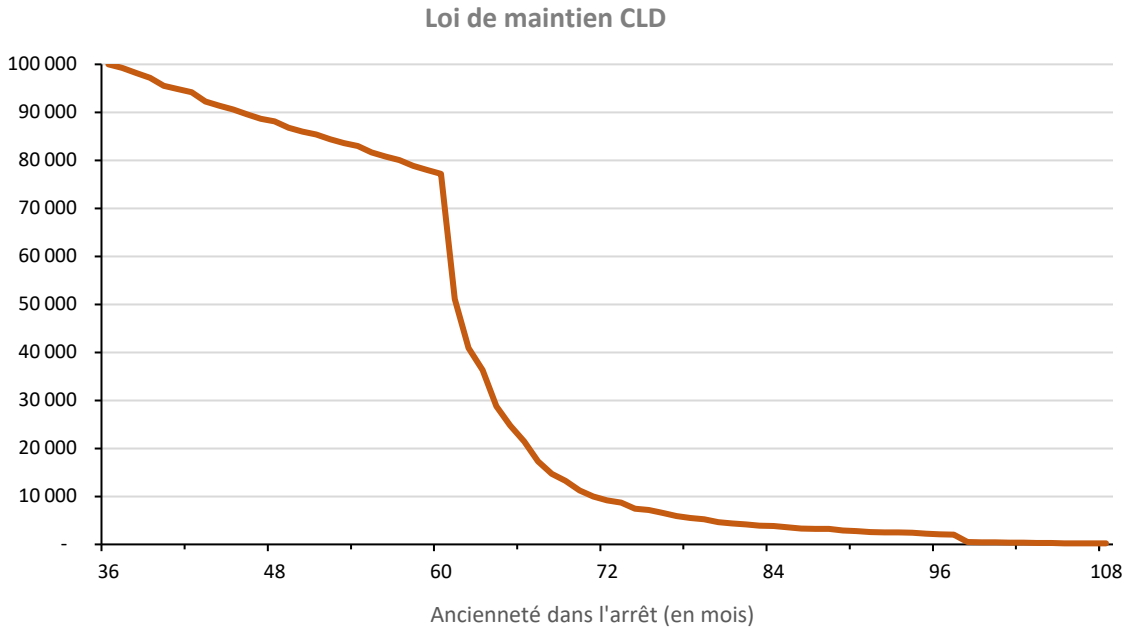
En cas d'incapacité pour maladie ordinaire, l'intégralité du salaire est maintenu par l'employeur pendant une période de 3 mois, cette période est assimilée à de la franchise dans les calculs. Cependant, si le fonctionnaire a été arrêté plus de 3 mois durant les 12 derniers mois, il sera indemnisé dès le premier jour. C'est la raison pour laquelle la loi de maintien brute en CMO démarre dès le premier jour, en l'occurrence dès le premier mois.
La loi de maintien en CMO est régulière et décroissante.

Le congé de longue maladie CLM



En cas d'incapacité pour longue maladie, l'intégralité du salaire est maintenu par l'employeur pendant une période de 12 mois, soit un an. Cette année est alors assimilée à une période de franchise dans les calculs. C'est pourquoi la loi de maintien brute en longue maladie démarre à 12 mois d'ancienneté dans l'arrêt.
Le pic observé au 36ème mois, soit après une durée de maintien en arrêt de travail de 3 ans, correspond à la fin de la durée maximale d'indemnisation pour le congé de longue maladie.

Le congé de longue durée CLD



En cas d'incapacité pour longue durée, le salaire est maintenu intégralement par l'employeur pendant une période de 36 mois, soit 3 ans. Ces 3 années sont alors assimilées à une période de franchise dans les calculs. C'est pourquoi la loi de maintien brute en longue durée démarre à 36 mois d'ancienneté dans l'arrêt.
Le pic observé au 60ème mois, soit après une durée de maintien en arrêt de travail de 5 ans, correspond à la fin de la durée maximale d'indemnisation pour le congé de longue durée.

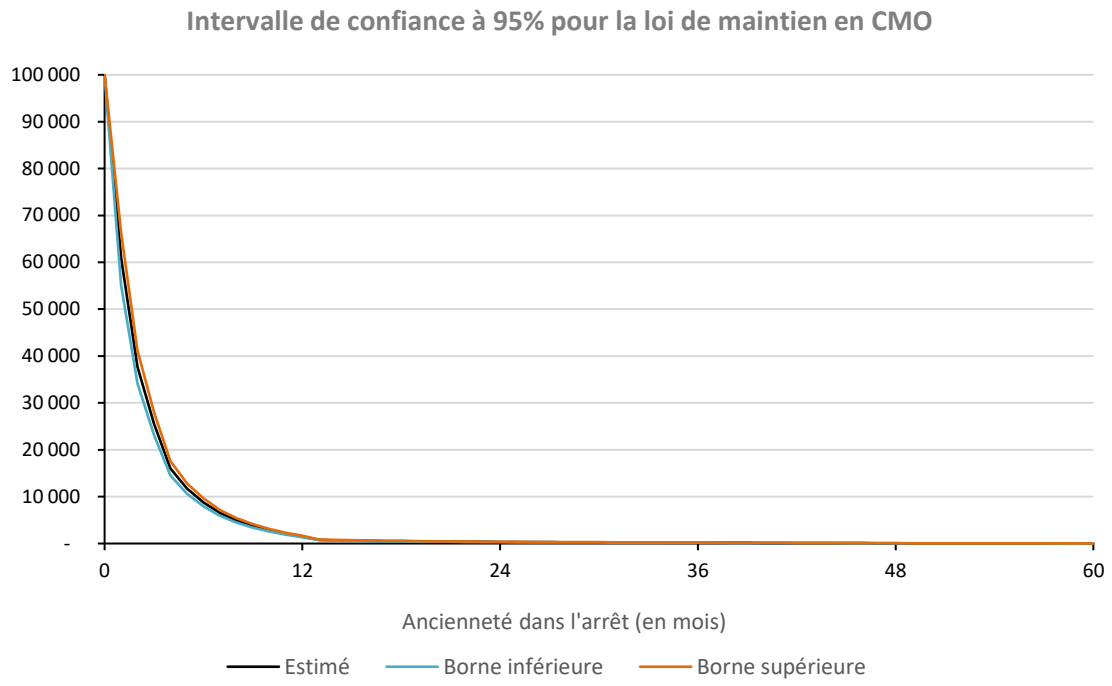
4.1.2. Contrôle de la qualité des estimations : les intervalles de confiance

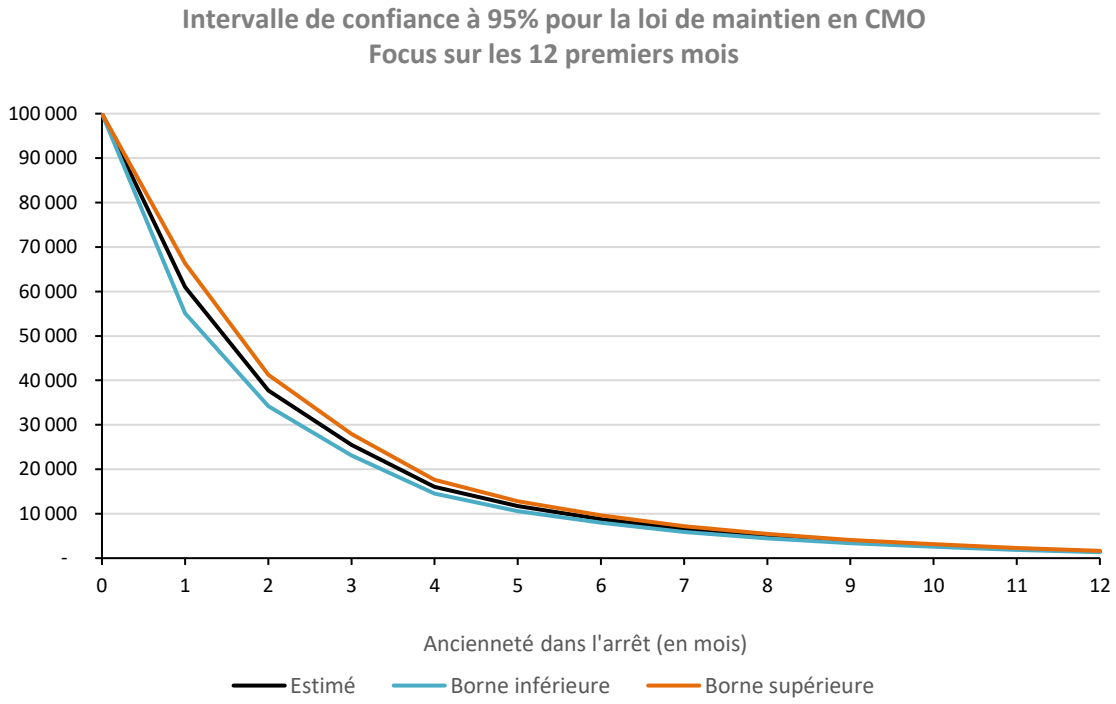
La fiabilité et la précision de l'estimation des lois brutes sont évaluées par un intervalle de confiance à 95%. La construction est basée sur la variance de l'estimateur de Kaplan-Meier.

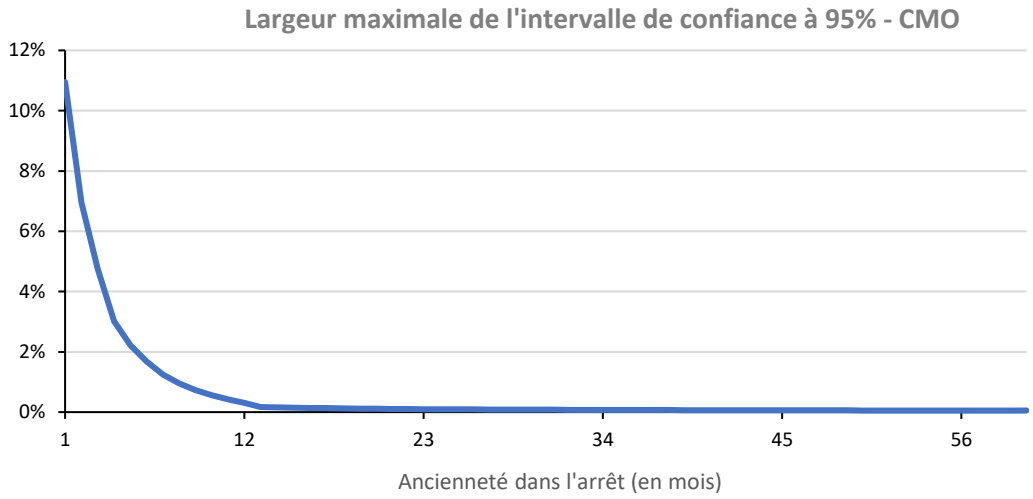
Chaque état est illustré par 2 graphiques :

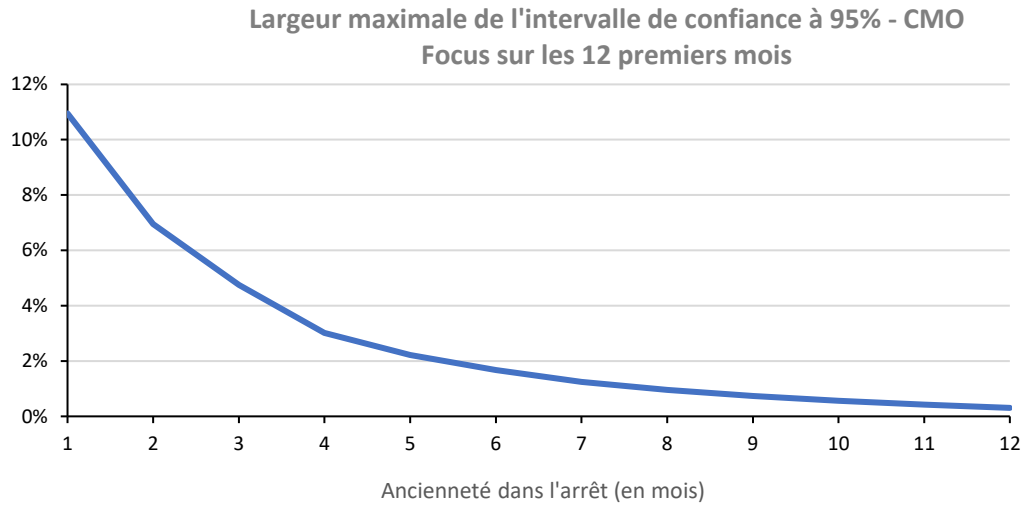
- le premier présente l'intervalle de confiance à 95% : l'estimé ainsi que la borne inférieure et la borne supérieure ;
- le deuxième expose la largeur maximale de l'intervalle de confiance : l'écart entre la borne inférieure et la borne supérieure, et ce, à chaque pas de temps.

Le congé de maladie ordinaire CMO



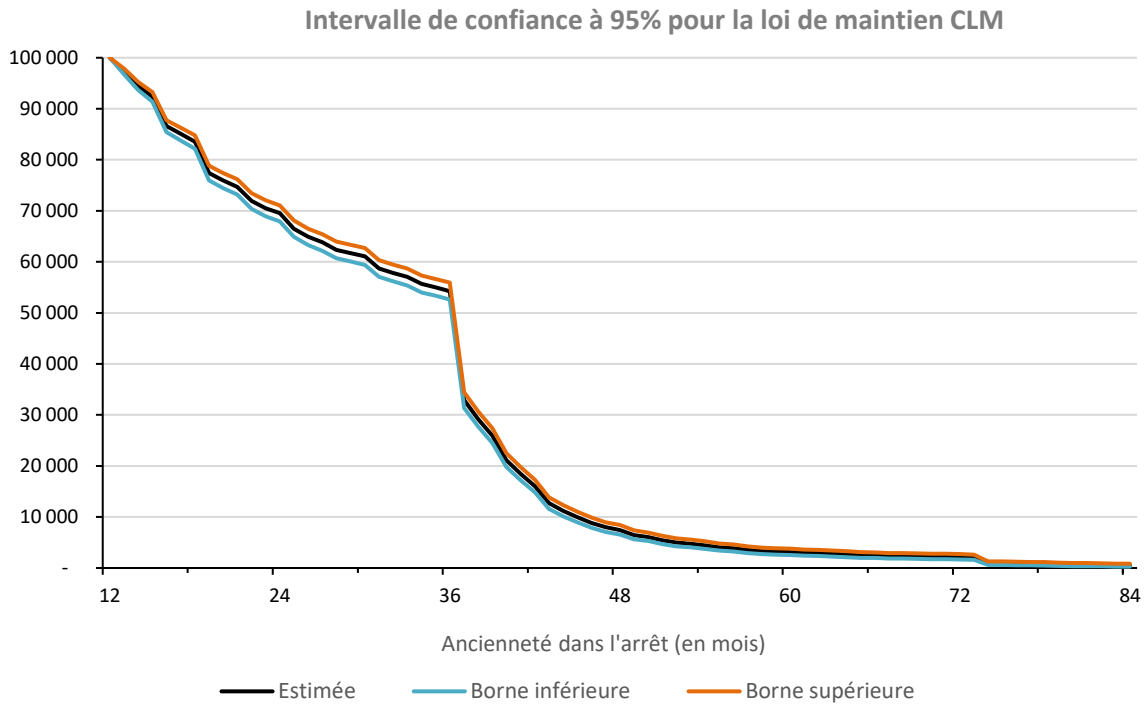


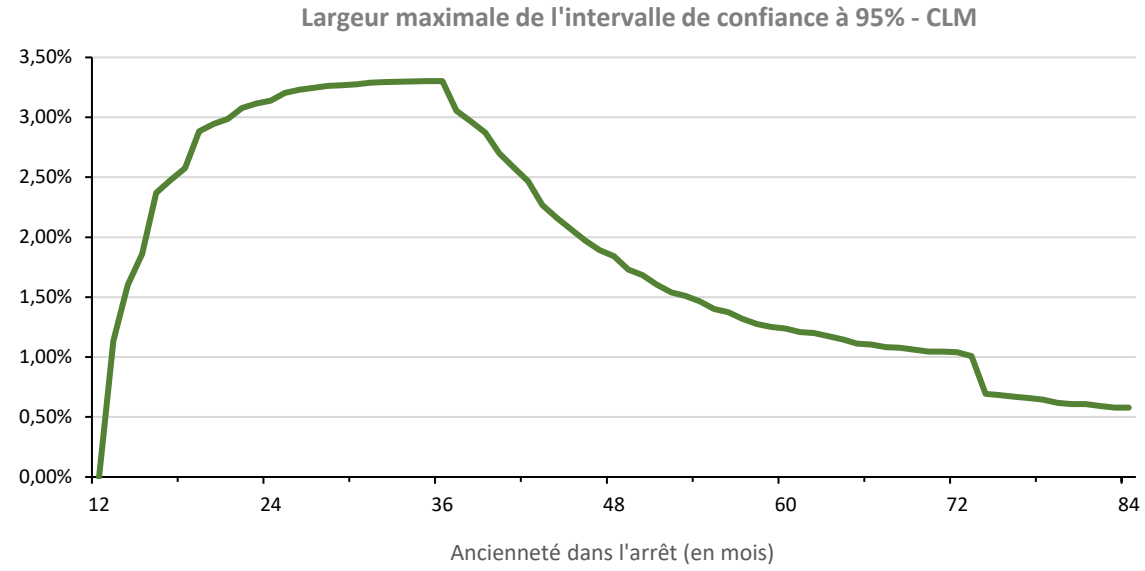




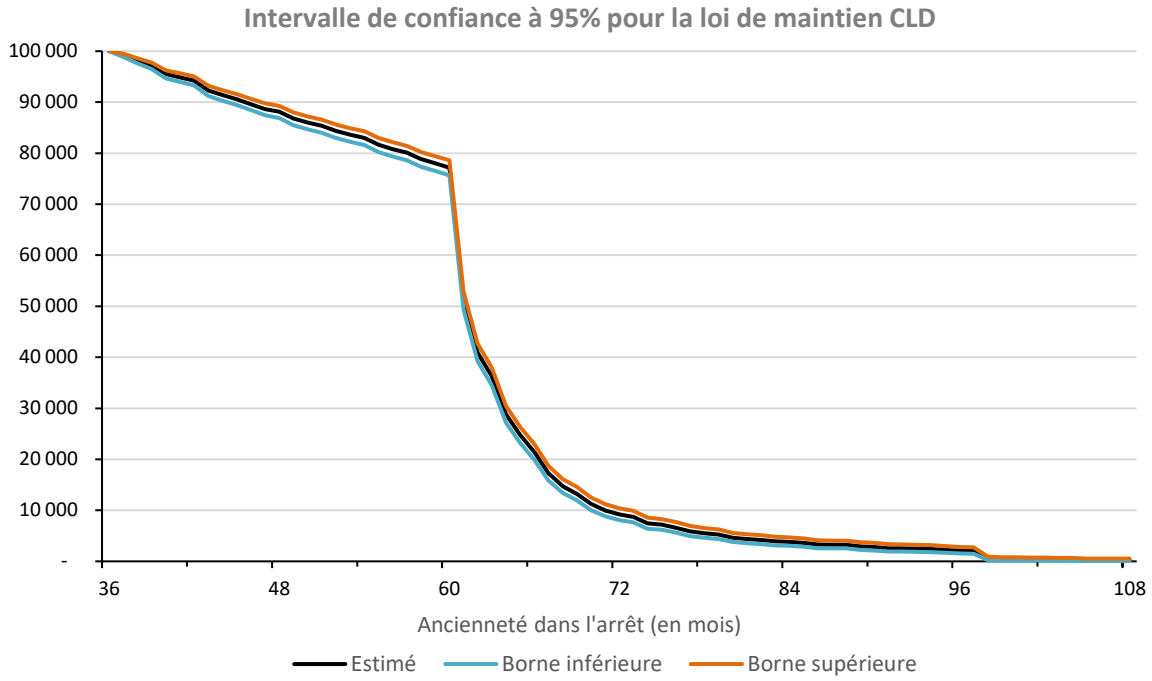
Pour les CMO, les largeurs d'intervalle de confiance à 95% les plus importantes concernent les arrêts ayant une ancienneté comprise entre 0 et 6 mois et sont décroissantes en fonction de l'ancienneté dans l'arrêt. L'intervalle variant de 0 à 11%.

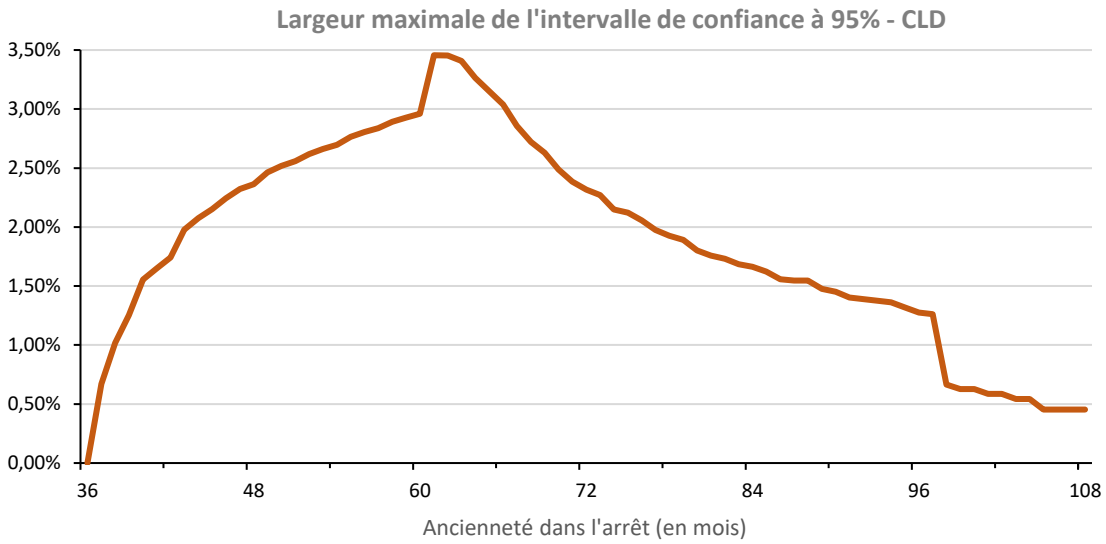
Le congé de longue maladie CLM





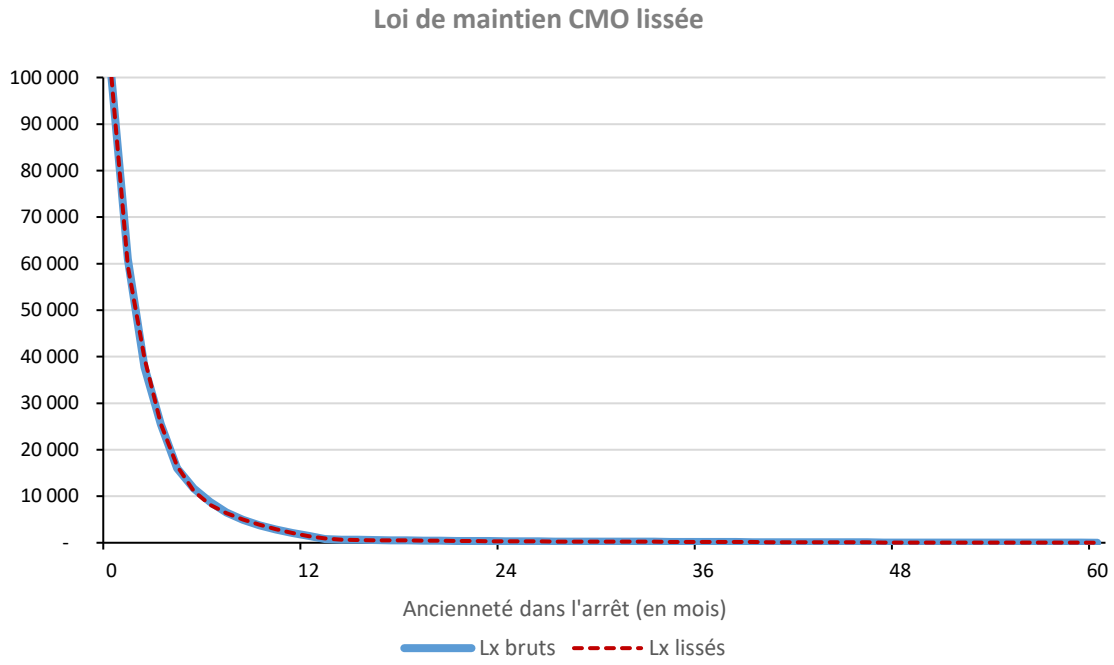
Le congé de longue durée CLD

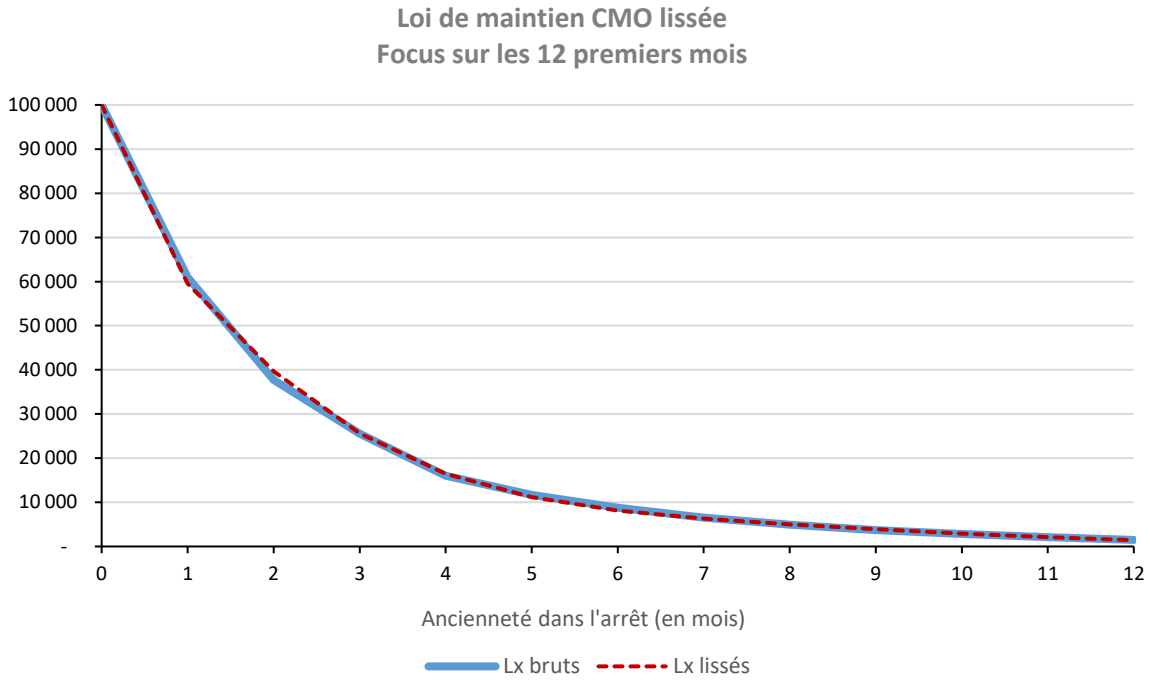


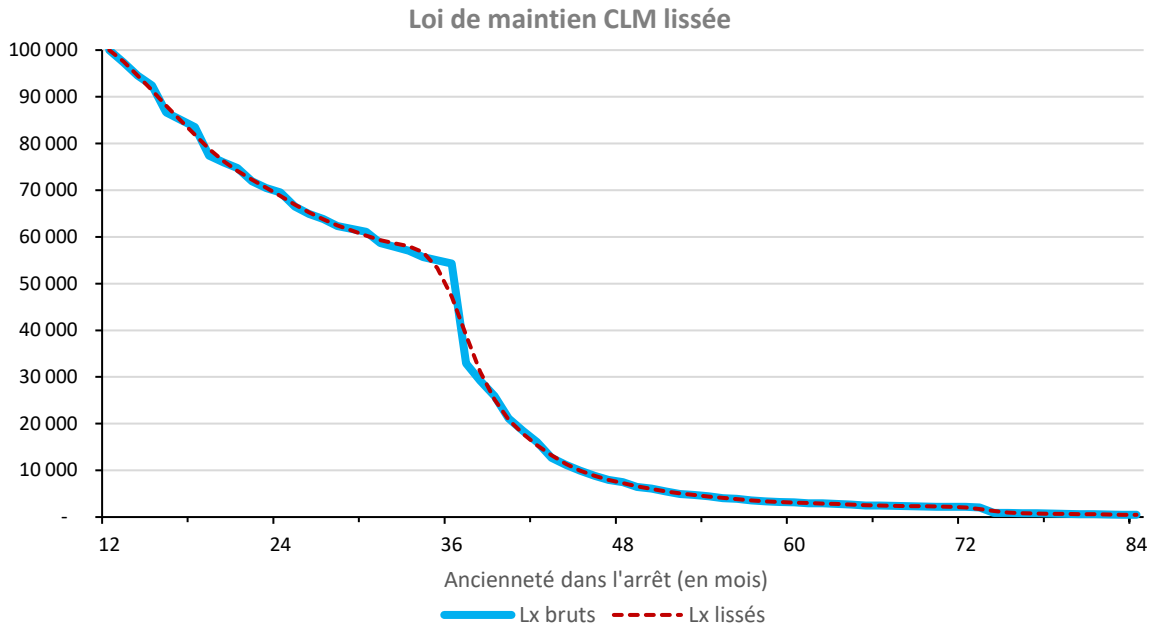


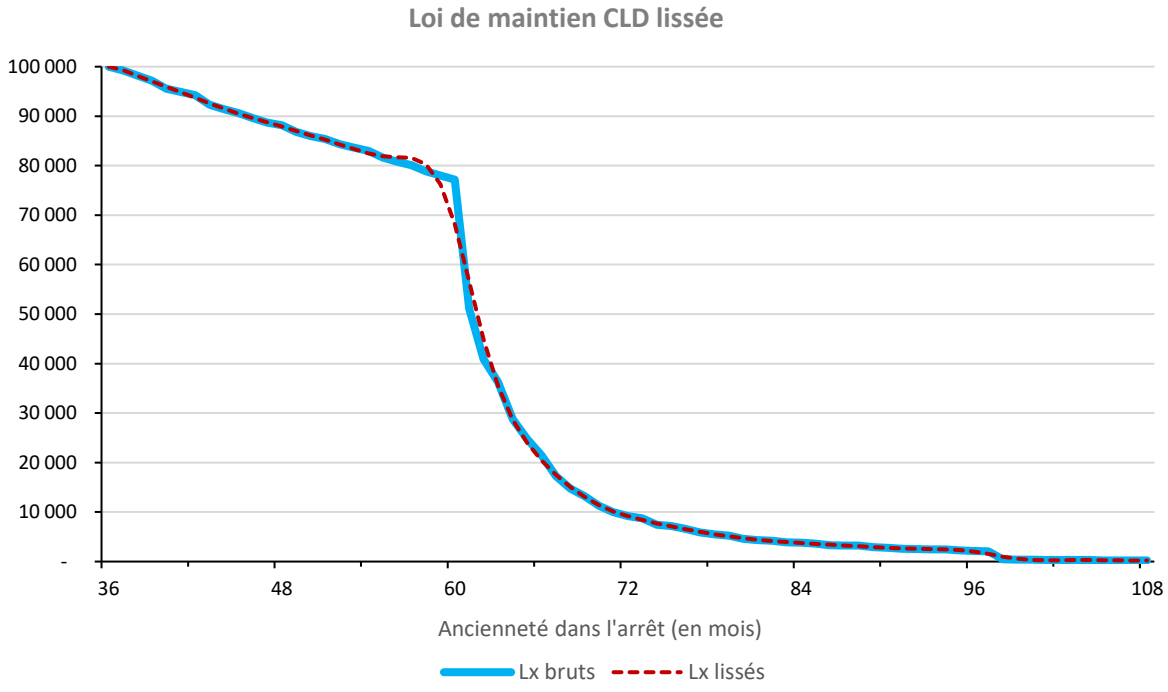
Concernant les arrêts CLM et CLD, les largeurs maximales des intervalles de confiance sont croissantes jusqu'à l'ancienneté maximale dans l'arrêt (36 mois pour la longue maladie, 60 mois pour la longue durée) puis baissent ensuite. L'intervalle de confiance à 95% variant de 0% à 3,50%.

4.1.3. Lissage des taux bruts avec Whittaker-Henderson









Au regard des résultats, le lissage est inutile pour les tables construites avec l'estimateur de Kaplan-Meier. Le backtesting sera effectué sur les lois de maintien bruts.

4.2. Construction des lois de maintien en incapacité avec l'apprentissage supervisé

Motivation

L'arbre de survie présente l'avantage d'être facilement interprétable (dans la mesure où l'arbre est peu complexe) mais est peu performant : c'est un prédicteur faible. Cet inconvénient pouvant être corrigé par la combinaison avec d'autres arbres de survie : soit en parallèle avec les forêts de survie, soit en séquentiel avec le gradient boosting de survie.

Donc, en partant d'un modèle dit simple qu'est l'arbre, il est théoriquement possible d'améliorer les performances de prédiction : soit en agrégeant les prédictions de différents arbres (forêts aléatoires) soit en agrégeant les prédictions corrigées de différents arbres (gradient boosting).

Le lien entre ces 3 algorithmes (un arbre qui est un prédicteur faible) a motivé leur sélection pour calculer les lois de maintien en incapacité avec l'apprentissage supervisé.

Cadre de calcul

Le calcul est effectué en utilisant *scikit-survival* (*SKSURV*), développé par Sebastian PÖLSTER, chercheur en intelligence artificielle dans le secteur médical.

SKSURV est un programme open-source de Python, développé pour l'analyse de survie. Il prend en compte les données censurées à droite et est compatible avec *scikit-learn* (programme open-source pour la classification et la régression).

Base de données

Les variables explicatives sélectionnées sont les suivantes :

- la variable distinguant les arrêts sans DO des arrêts avec DO ;
- l'âge d'entrée en arrêt ;
- le sexe ;
- le motif de sortie de l'arrêt.

La variable cible à prédire est caractérisée par le couple (censure ou non – durée d'arrêt) :

- une variable identifiant si l'observation est censurée ou non ;
- une variable indiquant la durée d'arrêt.

Le tableau suivant fournit le nombre d'observations ainsi que les proportions de censures pour chaque état :

Etat d'incapacité	Nombre d'observations/lignes	Proportion d'observations censurées
CMO	39 673	3%
CLM	4 064	16%
CLD	3 672	23%

Dans le cadre de l'apprentissage supervisé, la fiabilité et la précision des résultats dépendent de la base de données mais également du réglage des hyper paramètres. Il est donc important d'analyser les paramètres adaptés à chaque type d'algorithme. Ce point sera développé dans une première partie.
Ensuite seront présentés les résultats des performances prédictives. En fonction de ces résultats et pour chaque état, est sélectionné l'algorithme adapté à l'estimation de la loi de maintien.
Enfin, les lois de maintien brutes seront exposées dans la troisième partie, clôturant le chapitre 4.

4.2.1. Déroulé de calcul d'un algorithme

Etape 1 : division de la base de données en une base d'entrainement (d'apprentissage) et une base de test dans les proportions respectives de 75% - 25%.

Les proportions peuvent être différentes. Le choix étant motivé par la taille de la base de données : il faut suffisamment de données pour entraîner correctement le modèle afin que ce dernier puisse aboutir à des performances prédictives optimales. Au regard de la taille des bases de données pour les CLM et les CLD, une proportion plus importante dans la base de test aurait pénalisé l'entrainement des données.

Par ailleurs, il faut maintenir les proportions des différentes variables explicatives ainsi que la distribution des variables quantitatives. Par exemple s'il est observé 5% de censures et une durée moyenne d'arrêt de 3 mois dans la base initiale, les bases d'entraînement et de test doivent présenter des caractéristiques identiques.

La base de test est utilisée pour prédire la valeur cible et pour mesurer la performance prédictive de l'algorithme : elle évalue la capacité de généralisation du modèle. C'est pourquoi la base d'entrainement ne doit pas être utilisée pour mesurer la performance prédictive.

Etape 2 : optimisation des hyper paramètres par validation croisée.

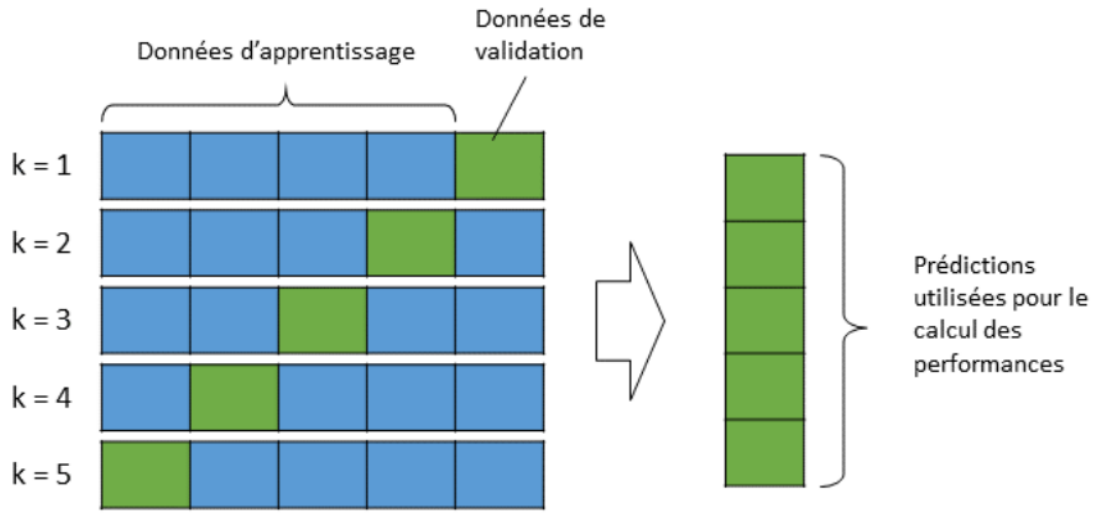
L'objectif de cette étape consiste à rechercher les critères optimaux permettant d'aboutir à la meilleure performance prédictive. Le principe consiste à **pénaliser le modèle afin d'éviter le surapprentissage aux données** (c'est-à-dire de coller parfaitement aux données de la base d'entrainement mais peu à celles de la base de test, dans ce cas le modèle n'est pas généralisable).

La validation croisée est appliquée sur **la base d'entrainement** qui est séparée en une base d'apprentissage et une base de validation.

La validation croisée K_blocs (« cross validation » en anglais) permet d'évaluer et de comparer les algorithmes d'apprentissage supervisé en divisant aléatoirement les données d'entrainement en K blocs. L'objectif consiste à minimiser l'erreur de prédiction du modèle.
Pour chaque bloc k (k=1,…,K), la méthode est la suivante :

- entraînement à partir des données d'apprentissage issu des (K-1) blocs excluant le bloc k ;
- validation à partir des données de validation provenant du k-ème bloc.

Les probabilités individuelles de l'ensemble des individus sont obtenues par l'agrégation des probabilités calculées pour chacun des K blocs. En pratique, le nombre de blocs est couramment fixé pour K ∈ [5 ; 10].



Le réglage des hyper paramètres doit tenir compte de plusieurs facteurs :

- le nombre de paramètres à régler dépend :
  - des contraintes matérielles : le temps machine et la puissance de l'ordinateur nécessaires pour les calculs sont croissants avec le volume de paramètres ;
  - ainsi que du volume de variables explicatives : plus il y a de variables explicatives, plus le nombre de paramètres à régler est important ;
- de l'intervalle de valeurs testé : chaque paramètre doit être simulé sur un intervalle de valeurs. Par exemple, un nombre de prédicteurs de base pour les forêts compris entre 100 et 500.

Pour le portefeuille étudié, en général 2 à 3 paramètres sont optimisés.

Les fonction suivantes de *scikit-learn* sont utilisées :

- *Kfold* pour la validation croisée avec un choix de K coupures (ici choix de K=5) ;
- *GridSearchCV* pour la recherche des paramètres optimaux.

La procédure consiste à indiquer à la fonction *GridSearchCV* le ou les paramètres à optimiser ainsi que l'intervalle de valeurs sur laquelle le paramètre est calculé. Le temps de calcul est croissant avec le nombre de paramètres et le nombre de valeurs de chaque paramètre, puisque la fonction *GridSearchCV* va chercher pour chaque combinaison paramètre-valeur et coupure de la validation croisée un score.

Etape 3 : évaluation de la performance prédictive de l'algorithme

Deux indicateurs présentés précédemment dans le chapitre 2 sont utilisés pour évaluer la performance prédictive :

- l'indice de concordance de Harrel ou C-Index qui évalue le critère de discrimination (capacité à différencier les individus en arrêt de ceux qui sont sortis de l'arrêt) ;
- le score de Brier intégré ou IBS (la moyenne des scores individuels de Brier) qui évalue à la fois la discrimination et la calibration. Ce dernier critère mesure l'écart entre le risque prédit et le risque observé.

Pour chaque état d'incapacité, l'algorithme optimal sera sélectionné en fonction de ces 2 métriques de performance prédictives.

4.2.2. Résultats des performances prédictives : sélection de l'algorithme optimal pour chaque état d'incapacité

Cette section présente les résultats des étapes 2 et 3 exposées précédemment.

Résultats du réglage des hyper paramètres

Pour chaque algorithme, la fonction *GridSearchCV* propose une valeur de paramètre par défaut, celle-ci est utilisée comme valeur minimale de l'intervalle de valeurs testées. La valeur maximale est déterminée en fonction des variables explicatives : nombre de variables explicatives et nombre de modalités existant dans chaque variable.

Le paramètre *Max_depth* détermine la profondeur de l'arbre (le nombre d'embranchements) : plus un arbre est profond, plus le modèle est complexe.
Min_samples_leaf fixe le nombre minimum d'individus d'une feuille (nœud final d'un arbre) : plus sa valeur est petite et plus l'arbre est complexe.
Min_samples_split détermine le nombre minimum d'individus pour créer une nouvelle coupure (pour diviser un nœud) : plus sa valeur est petite et plus l'arbre est complexe.

N_estimators est spécifique aux forêts (RSF) et au gradient boostig (SGB) puisque ce paramètre fixe le nombre d'arbres (nombre de prédicteurs) à construire.
Il est calibré pour les RSF et pas pour le SGB (valeur par défaut à 100) car les performances de l'ordinateur utilisé pour son calcul sont insuffisantes.

Certaines variables, considérées comme importantes dans la littérature actuarielle, ne sont pas prises en compte ici principalement pour des limitations en termes de puissance de la machine :

- *Max_features* qui permet de sélectionner le nombre de variables explicatives à chaque nœud ;
- *Learning_rate* (pour le SGB) : permet de réduire la contribution de chaque arbre, par défaut sa valeur est à 10% ;
- *Subsample* (pour le SGB) : correspond à la fraction d'échantillons (la quantité de sous-échantillons) à utiliser pour ajuster les prédicteurs de base individuels, par défaut sa valeur est à 1.

Congé de maladie ordinaire

Hyper paramètre optimisé	Caractéristique de l'hyper paramètre	Valeurs testées	Arbre de survie	Forêts de survie	Gradient boosting de survie
Max_depth	Profondeur de l'arbre	1 à 10 (ST et RSF) 3 à 10 (SGB)	3	4	3
Min_samples_leaf	Nombre minimum d'individus dans une feuille	3 à 10	3	3	5
Min_samples_split	Nombre minimum d'individus pour créer un nouveau nœud	6 à 10	6	6	
N_estimators	Nombre d'estimateurs (d'arbres)	100 à 500 (par tranche de 100)		300	

Congé de longue maladie

Hyper paramètre optimisé	Caractéristique de l'hyper paramètre	Valeurs testées	Arbre de survie	Forêts de survie	Gradient boosting de survie
Max_depth	Profondeur de l'arbre	1 à 10 (ST et RSF) 3 à 10 (SGB)	4	5	5
Min_samples_leaf	Nombre minimum d'individus dans une feuille	3 à 10	3	3	6
Min_samples_split	Nombre minimum d'individus pour créer un nouveau nœud	6 à 10	6	6	
N_estimators	Nombre d'estimateurs (d'arbres)	100 à 500 (par tranche de 100)		300	

Congé de longue durée

Hyper paramètre optimisé	Caractéristique de l'hyper paramètre	Valeurs testées	Arbre de survie	Forêts de survie	Gradient boosting de survie
Max_depth	Profondeur de l'arbre	1 à 10 (ST et RSF) 3 à 10 (SGB)	8	5	3
Min_samples_leaf	Nombre minimum d'individus dans une feuille	3 à 10	8	4	6
Min_samples_split	Nombre minimum d'individus pour créer un nouveau nœud	6 à 10	6	6	
N_estimators	Nombre d'estimateurs (d'arbres)	100 à 500 (par tranche de 100)		400	

Résultats des performances prédictives pour chaque état d'incapacité

Pour rappel, 2 métriques sont utilisées pour évaluer la performance prédictive : le C-index de Harell (évalue la discrimination) et l'IBS (évalue la discrimination et la calibration).
Plus le C-index est élevé (supérieur à 50%), plus le critère de discrimination est bon. Plus l'IBS est bas (inférieur à 25%) et meilleure est la performance prédictive.

Initialement, chaque état d'incapacité voyait sa loi construite avec l'algorithme présentant le couple {C-index le plus élevé ; IBS le plus bas}. Cependant, l'IBS s'est révélé inadapté aux données du portefeuille étudié. Par conséquent, seul le C-index sera appliqué pour évaluer les performances prédictives.

Dans la suite, est présenté pour chaque état d'incapacité, un tableau récapitulatif des indices de performance permettant de sélectionner l'algorithme optimal.

Chaque indice de performance est calculé sur la base de test pour mesurer la capacité de généralisation du modèle. Le C-index calculé sur la base d'apprentissage a pour objectif de contrôler un éventuel risque de sur ajustement aux données, et ce, en le comparant au C-index de la base de test.

Congé de maladie ordinaire CMO	Arbre de survie	Forêts de survie	Gradient boosting de survie
C-index (base de test)	67,28%	67,31%	**67,57%**
IBS	2,40%	2,73%	2,32%
C-index (base d'apprentissage)	*67,75%*	*68,03%*	*68,80%*

Congé de longue maladie CLM	Arbre de survie	Forêts de survie	Gradient boosting de survie
C-index (base de test)	71,25%	**72,28%**	71,56%
IBS	7,78%	7,77%	8,06%
C-index (base d'apprentissage)	*72,87%*	*73,94%*	*73,51%*

Congé de longue durée CLD	Arbre de survie	Forêts de survie	Gradient boosting de survie
C-index (base de test)	77,89%	**79,86%**	75,96%
IBS	7,57%	7,29%	7,91%
C-index (base d'apprentissage)	*79,06%*	*80,33%*	*76,49%*

Les meilleurs performances prédictives, tant sur le C-index que sur l'IBS, sont obtenues avec :

- le gradient boosting de survie pour les CMO ;
- les forêts de survie pour les CLM et les CLD.

Selon la littérature, à partir d'un C-index de Harell de 65%, l'algorithme présente des performances prédictives correctes. Au regard des résultats, les performances prédictives de chaque algorithme sélectionné sont correctes.

Les IBS sont bien en-deçà des 25% et indiquent également des algorithmes performants. Cependant, sur le portefeuille étudié, les IBS sont incohérents avec le C-index (ce constat sera confirmé par les résultats du backtesting, présenté dans la section suivante).

Le C-index est croissant du CMO au CLD, quel que soit l'algorithme (plus le C-index est élevé et plus l'algorithme est performant) :

C-index (CMO) < C-index (CLM) < C-index (CLD).

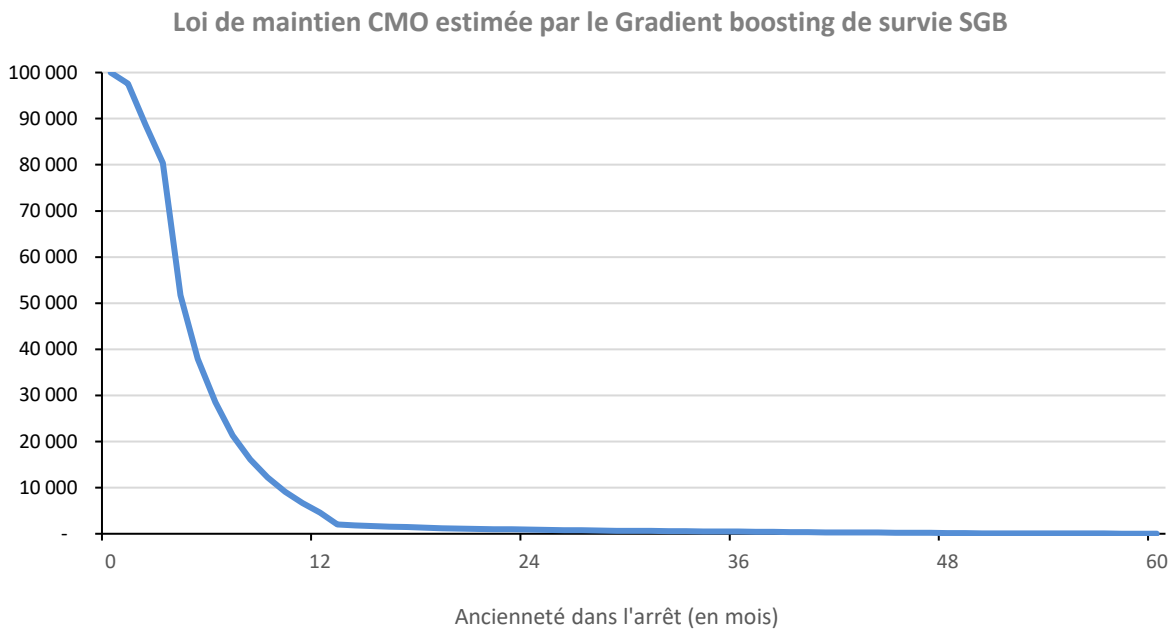
Dans ce cas, les résultats de l'IBS devraient être décroissants du CMO au CLD :

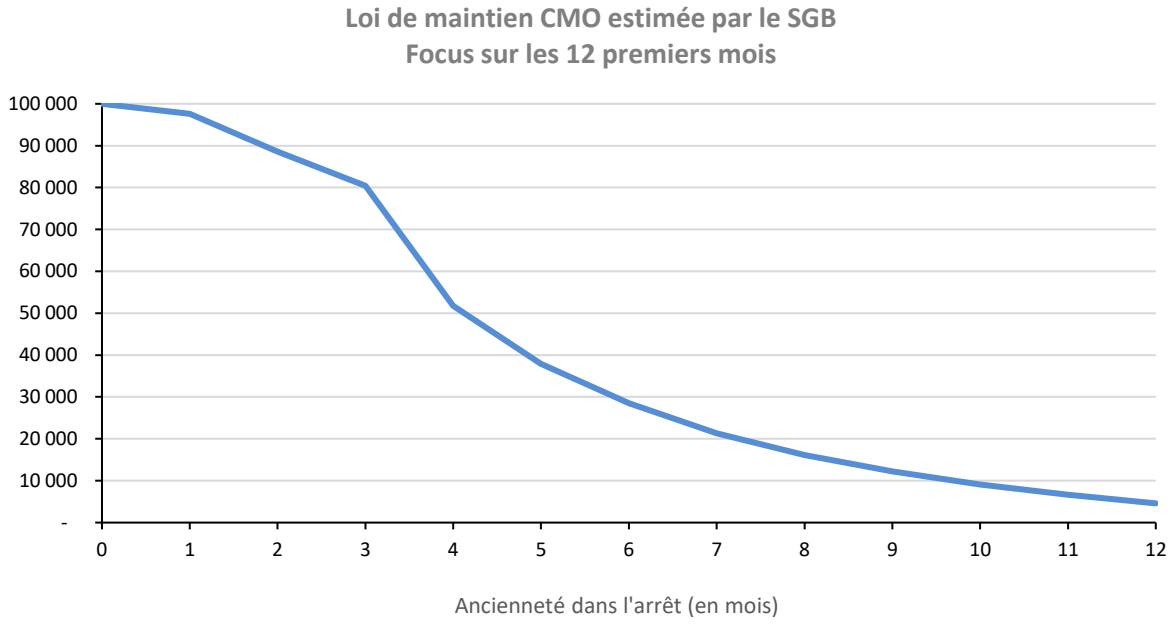
IBS (CLD) < IBS (CLM) < IBS (CMO).

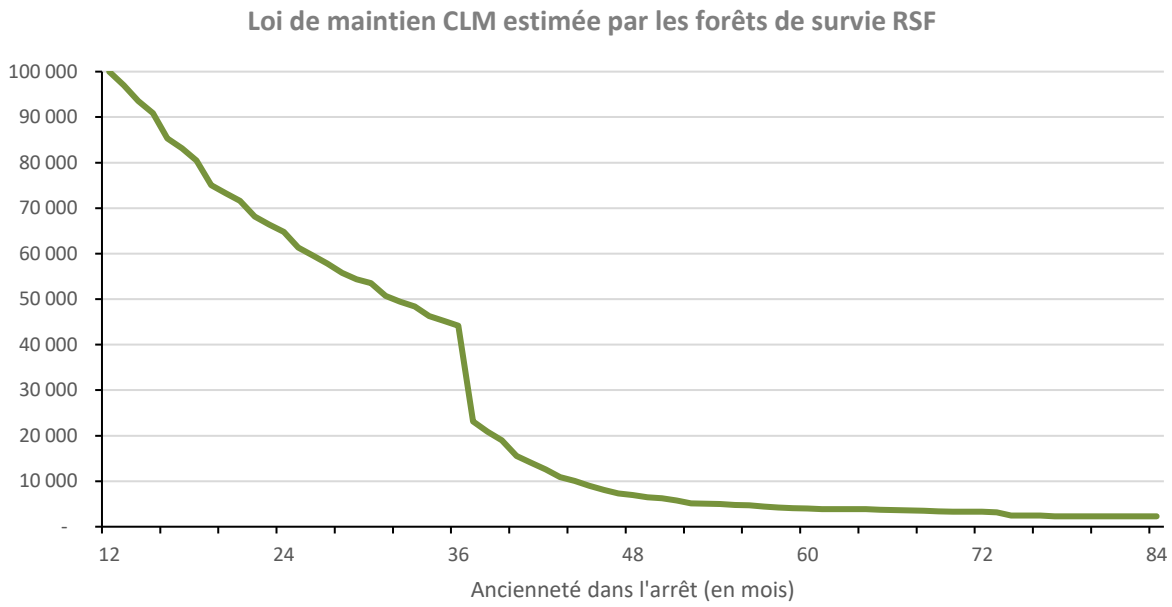
Or, il est constaté que :
IBS (CMO) < IBS (CLD) < IBS (CLM).

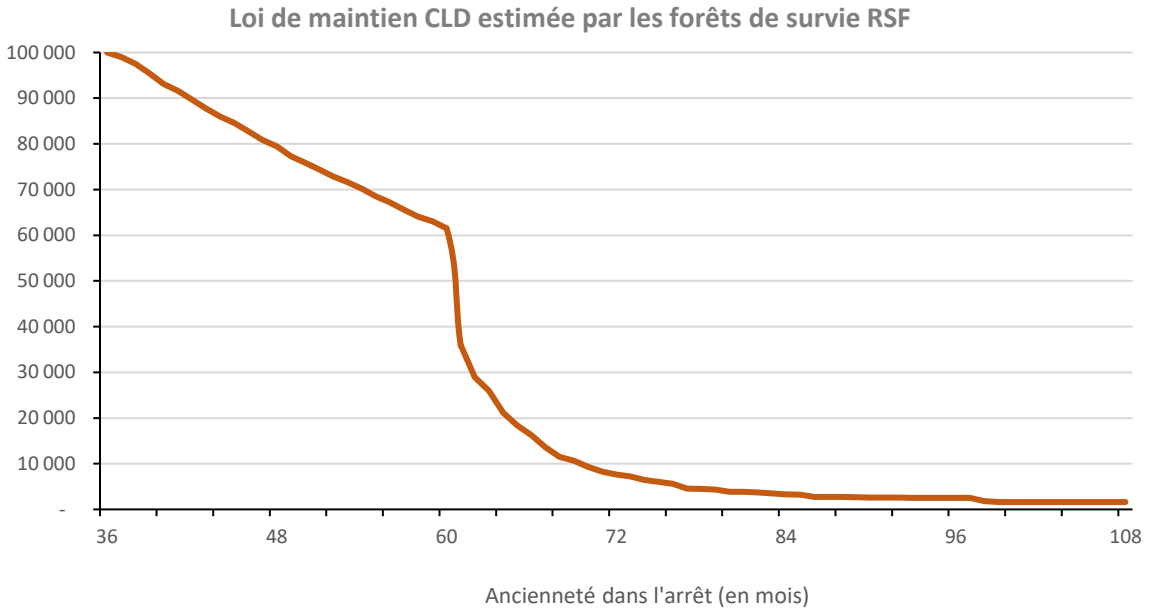
Cette incohérence trouve son origine dans la proportion de données censurées : les données CMO présentent 3% de censures contre 16% et 23% pour les CLM et CLD respectivement. L'IBS est un indicateur sensible à la proportion des données censurées et non censurées. Dans le cas où il est observé un déséquilibre des proportions de données censurées et non censurées, à l'instar des données CMO, l'IBS est alors trop optimiste et fournit un résultat non exploitable (cf. article de Assel et al.).

4.2.3. Calcul des lois avec l'algorithme optimal









Les lois estimées par les algorithmes d'apprentissage supervisé ne présentent aucun intérêt au lissage (les résultats du lissage par la méthode de Whittaker Henderson sont disponibles en annexe de ce mémoire).

En sus de la validation des estimations par le C-index de Harell, un intervalle de confiance à 95% est appliqué sur les fonctions de survie estimées. Ce complément permettant de valider la suffisance de prudence (les résultats sont présentés en annexe de ce mémoire).

4.3. Comparaison entre les deux approches

4.3.1. Comparaison des durées de maintien estimées aux durées réelles : backtesting

La pertinence des lois construites doit être vérifiée. Pour cela, les durées de maintien en incapacité temporaire estimées sont comparées aux durées réelles : c'est le backtesting.

Le déroulé du calcul est présenté par la suite.

Etape 1

A partir de la loi construite, calculer les durées résiduelles tronquées (DRT) en fonction de l'ancienneté dans l'état, en nombre de mois.

$$DRT_t = \frac{\sum_{i=t}^{max(t)} L_i}{L_t} + \frac{1}{2}$$

Avec L_t le nombre d'individus en incapacité à l'ancienneté t

Etape 2

Pour chaque sinistre en cours au 31/12/2019 (i.e. ayant une date de fin d'indemnisation > 31/12/2019), calculer l'ancienneté dans l'état à la date d'inventaire (31/12/2019), en nombre de mois :

$$anc = \left(\frac{date\ d'inventaire\ -\ date\ de\ survenance}{365{,}5} \right) x\ 12$$

Etape 3

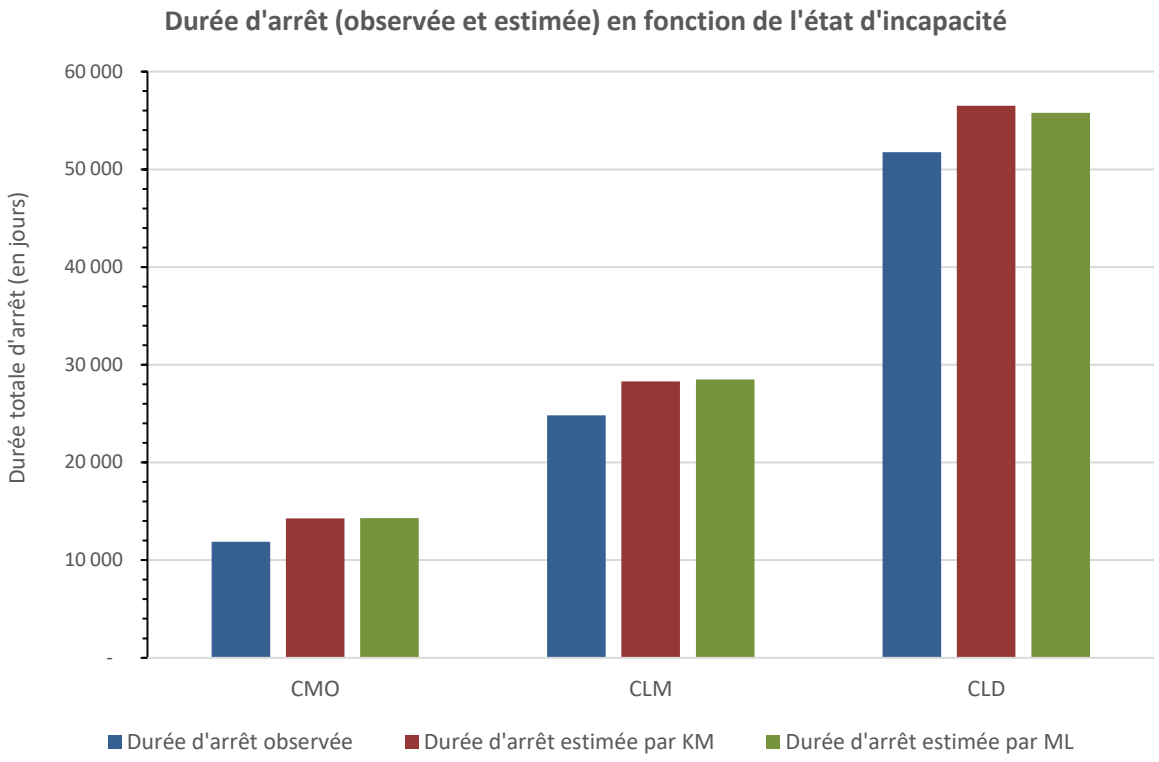
La durée totale d'incapacité estimée pour un sinistre est la somme de DRTt + anc.
Cette durée est sommée par survenance.

Etape 4

Ensuite, une comparaison est faite entre les durées estimées et réelles par survenance.

Le backtesting est effectué avec les deux approches pour chaque état d'incapacité. Cela permet également de comparer l'approche classique par Kaplan-Meier avec l'algorithme d'apprentissage supervisé :

- le CMO avec Kaplan-Meier et le gradient boosting de survie (SGB) ;
- le CLM et le CLD avec Kaplan-Meier et les forêts aléatoires de survie (RSF).



Ecart entre les durées estimées et observées (estimées / observées)	CMO	CLM	CLD
Par l'estimateur de Kaplan-Meier (KM)	20,0%	14,1%	9,2%
Par l'algorithme d'apprentissage supervisé (AAS)	20,4%	14,7%	7,8%

Ecart entre l'apprentissage supervisé et Kaplan-Meier	CMO	CLM	CLD
AAS - KM	0,4%	0,6%	-1,4%

Par rapport à l'estimateur de Kaplan-Meier :

- la loi CMO estimée par le SGB est plus prudente de 0,4% : 20,4% contre 20% pour Kaplan-Meier ;
- la loi CLM estimée par les RSF est plus prudente de 0,6% : 14,7% contre 14,1% pour Kaplan-Meier ;
- la loi CLD estimée par les RSF est moins prudente de -1,4% mais présente un écart moindre avec l'observé : 7,8% contre 9,2% pour Kaplan-Meier.

Au regard des résultats, chaque loi présente suffisamment de prudence et est donc pertinente pour envisager son utilisation dans une logique de provisionnement. Tant par l'approche classique que par l'algorithme d'apprentissage supervisé.

4.3.2. Conclusion sur l'apport de l'apprentissage supervisé

Les variables explicatives : hétérogénéité des modalités

Variable explicative		CMO	Pouvoir prédictif	CLM	Pouvoir prédictif	CLD	Pouvoir prédictif
Type d'indemnisation : sans DO/avec DO	Répartition	96% / 4%	67,9%	65% / 35%	69,2%	55% / 45%	76,6%
	Durée moyenne d'arrêt	5 mois / 22 mois		27 mois / 43 mois		55 mois / 66 mois	
Motif de sortie : standard/retraite/invalidité	Répartition	98,6% / 1,3% / 0,04%	50,8%	85% / 14% / 1%	55,4%	76% / 23% / 1%	47,2%
	Durée moyenne d'arrêt	5 mois / 14 mois / 19 mois		31 mois / 38 mois / 55 mois		59 mois / 61 mois / 64 mois	
Sexe : H/F	Répartition	51% / 49%	50,4%	55% / 45%	48,5%	53% / 47%	49,8%
	Durée moyenne d'arrêt	5,5 mois		33 mois / 32 mois		60 mois	
Age	Répartition	(1)	51,6%	(2)	52,5%	(3)	51%
	Durée moyenne d'arrêt						

Durée moyenne d'arrêt sur l'ensemble de l'état	5,5 mois (166 jours)	32 mois (988 jours)	60 mois (1825 jours)

(1) : la tranche [48 ; 62] ans représente 85% des arrêts avec une durée moyenne d'arrêt de 166 jours (5,5 mois).
(2) : la tranche [48 ; 61] ans représente 88% des arrêts avec une durée moyenne d'arrêt de 980 jours (32 mois).
(3) : la tranche [48 ;60] ans représente 86% des arrêts avec une durée moyenne de 1814 jours (60 mois).

Le tableau ci-dessus récapitule les statistiques descriptives de chaque état d'incapacité, présentés dans la section 3.2 du chapitre 3. Est ajouté en complément, le pouvoir prédictif (de chaque variable explicative) évalué par le C-index de Harrel : plus il est élevé et plus le pouvoir prédictif est important (supérieur à 50%).

La variable *Sexe* présente un C-index proche de 50%, notamment sur les CMO :

- les arrêts CMO se répartissent à part quasi égale entre les hommes et les femmes (51%/49%). De plus, la durée moyenne d'arrêt est identique pour chaque modalité ;
- bien que la répartition par sexe soit moins homogène sur les CLM et CLD (55%/45% et 53%/47% respectivement), la durée moyenne d'arrêt est cependant identique.

La variable *Motif de sortie* a également un faible pouvoir prédictif pour les CMO et les CLD :

- la majorité des arrêts CMO sont du standard (99%) ;
- bien que les arrêts CLD se répartissent entre les motifs standard et de départ à la retraite pour 76% et 23% respectivement, les durées moyennes entre les deux modalités sont cependant relativement proches (59 et 61 mois respectivement).

La variable *Type d'indemnisation* a le plus fort pouvoir prédictif selon le C-index :

- les 2 modalités de cette variable (sans prolongation DO / avec prolongation DO) présentent des durées moyennes d'arrêt suffisamment distinctes, et ce, pour chaque état d'incapacité ;
- la répartition des 2 modalités est suffisamment équilibrée pour les arrêts CLM et CLD, mais pas pour les CMO (96% d'arrêts sans DO).

Les variables à faible pouvoir prédictif ne sont pas retirés des simulations car elles apportent une information complémentaire et par conséquent, améliorent la prédiction. Notamment, l'algorithme du SGB modélise des relations non linéaires entre les variables explicatives. Des simulations sans ces variables ont abouti à une détérioration des prédictions.

Une variable explicative à fort pouvoir prédictif est une variable dont les modalités sont hétérogènes et dont la distribution est équilibrée. Les modalités doivent être suffisamment distinctes et le volume de chaque modalité doit être suffisant.
Des données composées de 55% d'arrêts sans DO avec une durée moyenne de 55 mois (contre 45% et 66 mois pour les arrêts avec DO) offrent de meilleures prédictions que des données composées à 96% d'arrêts sans DO (bien que la durée moyenne soit suffisamment éloignée de celle des arrêts avec DO).

Dans le cadre de l'analyse de survie et du portefeuille étudié, l'application d'un algorithme d'apprentissage supervisé est intéressante sous condition que les données soient diversifiées et équilibrées. L'optimalité de la prédiction est donc conditionnée par la structure de la base de données.

CONCLUSION

L'incapacité temporaire de travail doit être provisionnée en utilisant les lois de maintien appropriées. Quand la population couverte est significativement différente de celle de la table réglementaire, comme celle du portefeuille étudié, le recours à une table d'expérience se révèle alors indispensable.

Pour ce faire, les méthodes classiques telles que l'estimateur non paramétrique de Kaplan-Meier ont prouvé leur efficacité. Cependant, l'essor des algorithmes d'apprentissage supervisé apporte une alternative aux approches classiques.

La sélection des 3 algorithmes appliqués est motivée par leur point commun : l'arbre. L'arbre est un prédicteur faible (faible pouvoir prédictif) mais simple à interpréter. Il peut être amélioré en entrainant plusieurs arbres en parallèle : ce sont les forêts aléatoires. L'amélioration des prédictions peut également être réalisée en corrigeant les erreurs des arbres au fur et à mesure : c'est le gradient boosting.

Pour valider la pertinence des tables de maintien estimées et comparer les deux approches, un backtesting (écart entre les durées de maintien estimées et les durées réelles) est appliqué. Les résultats du backtesting entre les deux approches sont quasi identiques sur les CMO et les CLM. En revanche, sur les CLD, les durées estimées par l'apprentissage supervisé sont plus précises (plus proches des durées réelles que celles estimées par l'estimateur de Kaplan-Meier). Cette différence tient à l'équilibre et à l'hétérogénéité des modalités d'une variable explicative :

- les différentes modalités ou classes d'une variable explicative doivent être diversifiées : une durée moyenne d'arrêt de 2 mois pour les femmes et de 8 mois pour les hommes sera mieux prédite que 2 mois et 2,5 mois respectivement ;
- ces modalités doivent être équilibrées : 45% de femmes et 55% d'hommes aboutiront à une meilleure prédiction que 3% et 97% respectivement.

Dans le cadre de ce mémoire, l'application d'un algorithme d'apprentissage supervisé dans la construction d'une loi de maintien en incapacité apporte une meilleure précision, sous condition que la distribution des différentes modalités d'une variable soit équilibrée et que les modalités soient suffisamment distinctes.

Dans la prolongation de ce mémoire, deux axes d'amélioration semblent pertinents à analyser.

Le premier axe concerne l'amélioration du pouvoir prédictif des algorithmes d'apprentissage supervisé :

- d'une part, par de meilleures ressources matérielles, à savoir une machine plus puissante (certains algorithmes tels que les forêts et le gradient boosting, particulièrement gourmands, pourraient-ils être davantage optimisés par des ordinateurs plus puissants ?) ;
- d'autre part, par le volume de données (avec davantage de données, l'algorithme disposant par conséquent de plus de données pour s'entraîner, fournirait-il de meilleures prédictions ?).

Le deuxième axe concerne la stabilité dans le temps des algorithmes étudiés. L'étude menée est une photographie à une période donnée. Alors quid des résultats si le portefeuille est analysé dans un ou deux ans ?

BIBLIOGRAPHIE

Articles

PÖLSTERL S., [2020]. *scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn*, Journal of Machine Learning Research, vol. 21, no. 212, pp. 1–6.

WANG P., LI Y., REDDY CK., [2017]. *Machine learning for survival analysis : a survey.*

ISHWARAN et al., [2008]. *Random survival forests*. The Annals of Applied Statistics.

GENUER R., POGGI JM., [2017]. *Arbres CART et Forêts aléatoires, Importance et sélection de variables*. hal-01387654v2

HUCHON M., [2021]. *Sélection de variables à l'aide de forêts aléatoires pour données de survie de grande dimension*. Santé publique et épidémiologie. dumas-03377763

HARREL et al., [1996]. *Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in Medicine, 15, 361–387.

BRIER et al., [1950]. *Monthly Weather Review*, vol. 78, issue 1, p. 1

ASSEL et al., [2017]. *Diagnostic and Prognostic Research.* DOI 10.1186/s41512-017-0020-3

Sites internet

https://www.fonction-publique.gouv.fr
PÖLSTERL S., *scikit-survival* (Package website)

Cours

LOPEZ O., [2023]. Science des données / Analyse de survie. Cours CEA
PLANCHER F., [2023]. Modèle de durée. Cours ISFA
BESSE P., [2023]. Science des données – Apprentissage statistique. Cours de l'INSA

Travaux universitaires

SAUTREUIL M. (thèse 2021) : *Contributions à la détection de marqueurs et à l'analyse de survie en oncologie*

DEVAUX A. (thèse 2022) : *Modélisation et prédiction dynamique individuelle d'événements de santé à partir de données longitudinales multivariées*

LE FAOU Y. (thèse 2019) : *Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé*

JUDD T. (mémoire 2018) : *Modélisation de la durée de maintien en arrêt de travail*

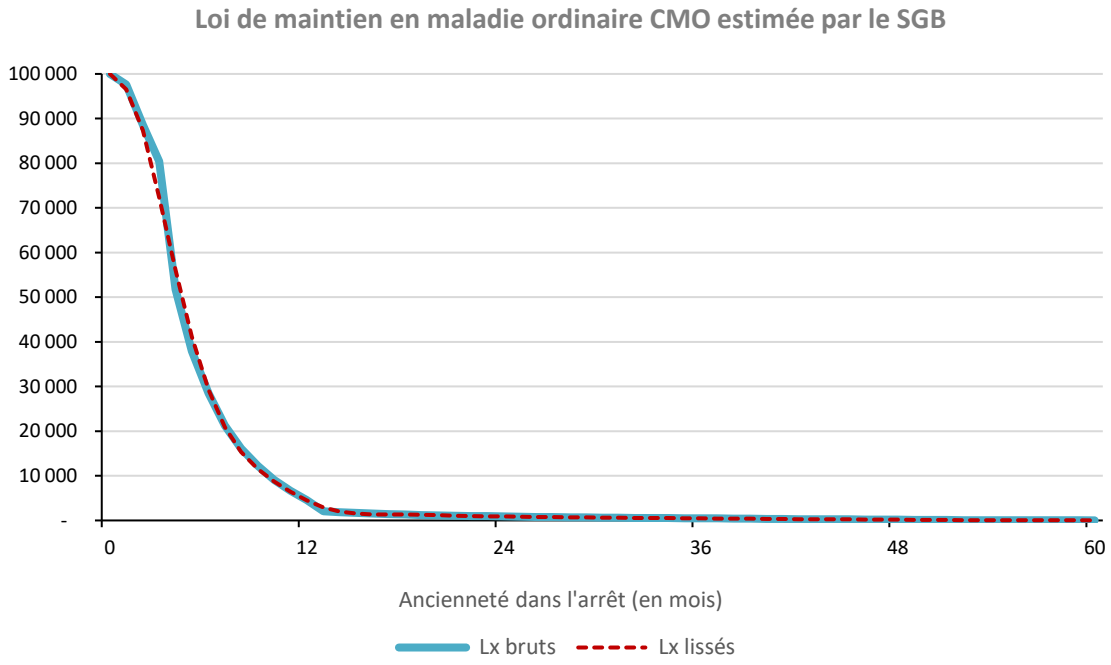
NGUYEN K. (mémoire 2018) : *Méthodes de provisionnement du maintien en incapacité des contrats dits franchises courtes*

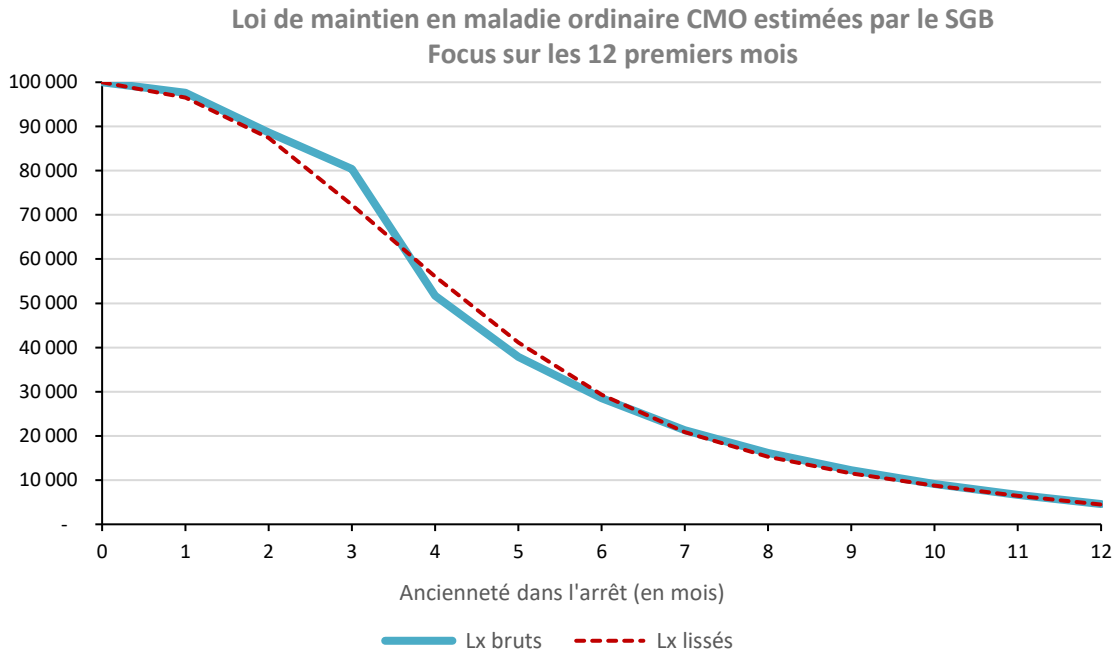
KOYE GK. (mémoire 2019) : *Comparaison des méthodes classiques et alternatives avec le machine learning pour la construction d'une table de mortalité d'expérience*

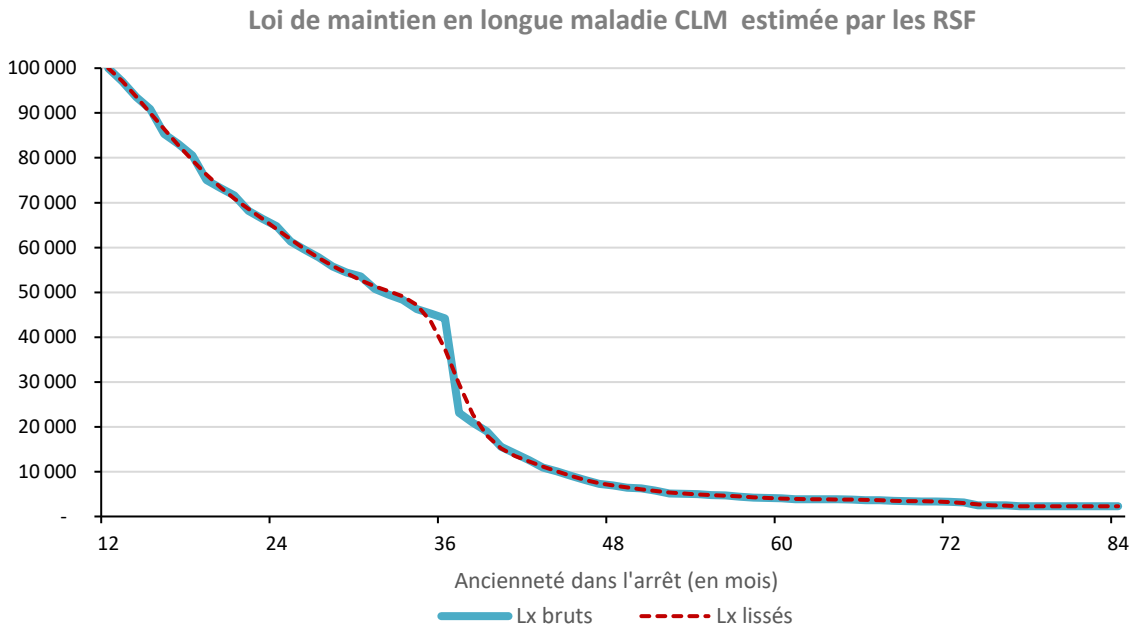
CARAYON R. (mémoire 2019) *: Construction de lois de maintien en arrêt de travail pour les Collectivités locales*

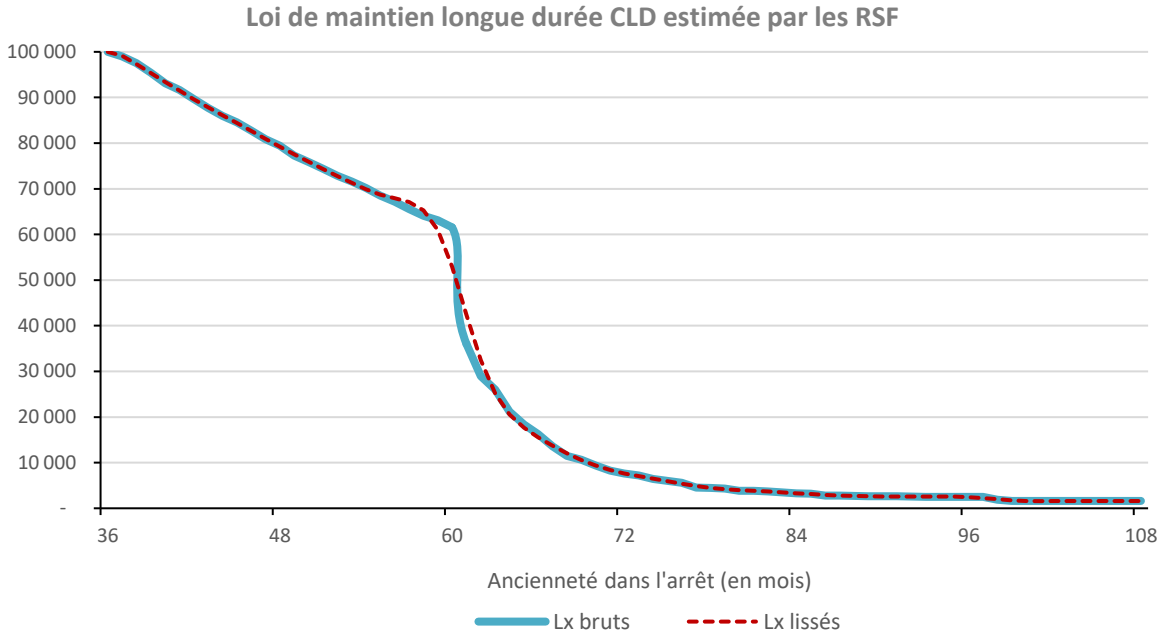
ANNEXE

Résultats du lissage des tables estimées par l'algorithme d'apprentissage supervisé






Loi de maintien en longue maladie CLM estimée par les RSF
100 000
90 000
80 000
70 000
60 000
50 000
40 000
30 000
20 000
10 000
-
12
24
36
48
60
72
84
Ancienneté dans l'arrêt (en mois)
Lx bruts
Lx lissés


Loi de maintien longue durée CLD estimée par les RSF
100 000
90 000
80 000
70 000
60 000
50 000
40 000
30 000
20 000
10 000
-
36
48
60
72
84
96
108
Ancienneté dans l'arrêt (en mois)
Lx bruts
Lx lissés