Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Madame / Monsieur Khoé-cuong Damien NGUYEN

Titre du mémoire :
Construction de tables françaises de mortalité en dépendance à partir de données américaines

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

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

Membres présents du jury de la filière : Signature :	Entreprise : Nom : KPMG Signature :
	Directeur de mémoire en entreprise Nom : Johan STEVENY Signature : Johan Steveny
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Valentin BARDOUX Signature : V Bardoux
	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 : Johan Steveny Signature du candidat : Nguyen

Résumé

Depuis les années 1950, l'espérance de vie des habitants des États développés a connu une croissance significative, poussée par la surreprésentation de la génération issue du baby-boom et par le vieillissement général des populations. Pour le peuple français, les chiffres et les projections de l'Institut National de la Statistique et des Études Économiques (INSEE) sont sans appel : la pyramide des âges connaît, depuis les années 70, un phénomène de rectangularisation qui devrait se poursuivre dans l'avenir.

Consécutivement à ces rebondissements démographiques, une portion conséquente des seniors se trouve en situation de dépendance. Alors que l'assurance joue un rôle majeur sur ce périmètre, les assureurs français qui s'y sont positionnés font face à plusieurs contraintes au sein de toutes leurs pratiques métiers. Des définitions hétéroclites de la perte d'autonomie dans la sphère française, conjuguées à des tailles de portefeuille dépendance insuffisantes aboutissent à une estimation difficile des lois de probabilité qui régissent le devenir vraisemblable des assurés. Face à de tels défis, le Comité Consultatif du Secteur Financier (CCSF) a réagi en début d'année 2024 et a publié ses recommandations sur le secteur en proposant notamment de rendre obligatoire l'assurance dépendance sur le périmètre de la perte d'autonomie dite «lourde».

En conséquence, ce mémoire se propose d'estimer une loi de mortalité française en dépendance lourde. Des taux bruts de létalité en perte d'autonomie sont estimés à partir de la base de données américaine LTC 2000-2011 de la Society Of Actuaries (SOA). Cette ressource est une mine d'informations qui s'avère utile face à la carence en données à laquelle font face les assureurs français. Plusieurs méthodes d'estimation sont employées et sont comparées entre elles : estimateur de Kaplan-Meier, arbre de survie et forêt aléatoire de survie. Les outils dédiés à l'analyse de l'intelligibilité globale des forêts aléatoires sont spécialement utilisés afin d'extraire l'information sur les variables explicatives. Ces probabilités sont ensuite lissées, positionnées et fermées pour correspondre à l'environnement de l'Hexagone. Pour ce faire, plusieurs modèles, qui sont davantage adressés à de la mortalité plus «standard», sont appliqués.

Les performances brutes de la forêt aléatoire de survie sont très faiblement supérieures à celles de l'arbre de survie et de l'estimateur de Kaplan-Meier d'après des indicateurs comme l'intégrale du score de Brier. Par ailleurs, l'estimateur de Kaplan-Meier traduit de manière beaucoup plus convaincante la mortalité des premiers mois passés en perte d'autonomie pour les âges d'incidence inférieurs à 75 ans en comparaison des estimateurs hérités du machine learning pour lesquels la létalité semble minorée de façon notable. Cette observation se répercute sur le positionnement à la population française et ce dernier ne semble pas suffisant pour capturer la létalité en dépendance dans l'Hexagone.

Mots-clés : *dépendance*, *actes de la vie quotidienne*, *incidence en dépendance*, *survie en dépendance*, *tables de mortalité en dépendance*, *estimateur de Kaplan-Meier*, *arbre de survie*, *forêt de survie aléatoire*, *lissage de Whittaker-Henderson*, *modèle de Brass*, *modèle de Denuit & Goderniaux*, *modèle multi-états*, *prime*, *provision pour risques croissants*, *ratio de mortalité standardisé*, *SOA*, *INSEE*, *CCSF*.

Abstract

Since the 1950s, the life expectancy of people living in developed countries has increased significantly, driven by the over-representation of the baby-boom generation and the general ageing of populations. For the French people, the figures and projections of the French national institute for statistical and economic studies (INSEE) are clear : the age pyramid has been experiencing a rectangularization phenomenon since the 1970s, which is expected to continue in the future.

As a result of these demographic developments, a significant portion of seniors are in a situation of dependency (or otherwise called loss of autonomy). While insurance plays a major role in this area, French insurers who have positioned themselves there face several constraints in all their business practices. Heterogeneous definitions of loss of autonomy in the French sphere, combined with insufficient dependency portfolio sizes, make it difficult to estimate the probability laws governing the likely future of policyholders. Faced with such challenges, the French advisory committee for financial sector (CCSF) reacted at the beginning of 2024 and published its recommendations on the sector, notably proposing to make dependency insurance mandatory in the scope of so-called 'heavy' dependency.

As a result, this report sets out to estimate a French mortality rate for heavy dependency. Crude mortality rates for loss of autonomy are estimated from the Society Of Actuaries (SOA) LTC 2000-2011 American database. This resource is a mine of information to be exploited, which is proving useful given the lack of data faced by French insurers. Several estimation methods are used and compared : Kaplan-Meier estimator, survival tree and random survival forest. Tools dedicated to analysing the global intelligibility of random forests are specifically used to extract information about the explanatory variables. These probabilities are then smoothed, positioned and closed to match the French environment. To do this, several models are applied, which are aimed more at more 'standard' lethality.

The raw performance of the random survival forest is very slightly better than that of the survival tree and the Kaplan-Meier estimator based on indicators such as the integral of the Brier score. In addition, the Kaplan-Meier estimator gives a much more convincing account of mortality in the first few months of loss of autonomy for incidence ages under 75, compared with estimators inherited from machine learning, for which lethality seems to be significantly reduced. This observation has repercussions on the positioning to the French population, which does not seem sufficient to capture the lethality of dependency in France.

Keywords : *long-term care*, *activities of daily living*, *long-term care incidence*, *long-term care survival*, *long-term care mortality table*, *Kaplan-Meier estimator*, *survival tree*, *random survival forest*, *Whittaker-Henderson smoothing*, *Brass model*, *Denuit & Goderniaux model*, *multi-state model*, *insurance premium*, *standardized mortality ratio*, *SOA*, *INSEE*, *CCSF*.

Note de Synthèse

Les travaux de ce mémoire se proposent de construire une loi de mortalité française sur le périmètre de la dépendance lourde à partir de la base américaine LTC 2000-2011 de la Society Of Actuaries (SOA) et de comparer la performance de plusieurs estimateurs.

Définitions de la perte d'autonomie et contexte de l'assurance dépendance

En France, plusieurs définitions de la perte d'autonomie existent. Les plus connues parmi ces dernières reposent sur la grille AGGIR ou sur le système des Actes de la Vie Quotidienne (AVQ). Des équivalences plus ou moins pertinentes peuvent être établies entre ces différentes caractérisations. Les diverses analyses effectuées dans ce mémoire sont confrontées à de multiples définitions de la dépendance qui seront considérées analogues systématiquement. Parmi les différentes caractérisations de la dépendance, celle qui figure dans le label Garantie Assurance Dépendance (GAD), étiquette accordée aux produits d'assurance dépendance respectant 9 critères particuliers, est l'une des plus populaires. En réaction à la pénétration modeste de la dépendance lourde au sein de l'écosystème assurantiel, le tampon GAD a été instauré. Sous ce dernier, la caractérisation de la dépendance totale se décline de 3 façons : perte d'autonomie physique, cognitive ou mixte.

Néanmoins, alors que ce label rencontre un certain succès de 2013 à 2020, le nombre de souscriptions de contrats portant cette étiquette a fortement diminué depuis le début de la décennie actuelle. Face aux difficultés rencontrées par les assureurs français sur les contrats de dépendance, la force publique s'est récemment manifestée. Début 2024, le Comité Consultatif du Secteur Financier (CCSF) a publié 3 recommandations dont la principale consiste à mettre en place un contrat dépendance solidaire obligatoire sur le périmètre de la perte d'autonomie lourde.

Données employées

Plusieurs jeux de données sont employés afin de mener à bien les travaux présentés dans ce mémoire et les principaux sont les suivants :

— la base de données SOA Long-Term Care (LTC) 2000-2011, reflet de 80% des polices en vigueur aux États-Unis sur la période d'observation 2000-2011. Ce jeu de données sert la construction des taux bruts américains de décès en perte d'autonomie.

— les tables d'incidence, de maintien en perte d'autonomie et de décès en perte d'autonomie estimées par le groupe de travail Qalydays *.

Le jeu de données SOA LTC 2000-2011 a été construit grâce à la contribution de 22 assureurs du marché américain. Par souci de comparabilité entre les données SOA et les lois Qalydays, seuls les contrats dits ≪Tax-Qualified≫ sont considérés dans l'étude. Une fois que

∗. https ://www.ressources-actuarielles.net/qalydays

la base de données soit filtrée et réarrangée, plusieurs remarques peuvent être formulées sur les 73 596 lignes du jeu de données mis à jour :

— la base de données est composé d'une majorité de femmes, à hauteur de deux tiers.

— contrairement à la France qui privilégie le soin des seniors en perte d'autonomie à domicile, la part de traitement à domicile et en établissement semble approximativement équirépartie sur le jeu de données américain. De plus, l'organisation des institutions de soins françaises et américaines diffère sur plusieurs points. 99% des seniors français demeurant en institution sont canalisés dans les EHPAD, les USLD et les résidences autonomies, tandis que du côté américain, 60% de la population en établissement réside dans des nursing homes.

— aux États-Unis, la prise en charge de la dépendance cognitive lourde semble plus préoccupante que celle de la dépendance physique. En effet, les pathologies relatives, comme la démence d'Alzheimer ou l'accident vasculaire cérébral, semblent responsables de la plus longue durée d'indemnisation en perte d'autonomie.

Une fois les taux américains de décès estimés, ces derniers sont positionnés et fermés à partir de la loi de décès en perte d'autonomie du groupe de travail Qalydays, cette dernière étant assimilée à des taux bruts observés sur le portefeuille d'un assureur dépendance reconnue par le label GAD. En 2018, l'équipe Qalydays publie plusieurs lois de probabilités sur la circonférence de la dépendance. La construction d'un modèle à état qui couvre la perte d'autonomie en a immanquablement besoin, dont notamment des lois de mortalité en dépendance selon le sexe et le type de dépendance lourde (cognitive, physique ou mixte). La définition de la dépendance qui a été adoptée pour la calibration de ces lois repose sur la définition du codage médical. La spécificité la plus remarquable de ces lois de décès estimées se situe au niveau du premier mois d'ancienneté en dépendance. D'un côté, la mortalité y est très importante et de l'autre, le modèle qui a servi à estimer ces taux de décès sous-estime la létalité réelle au cours de ces 30 premiers jours. À titre d'exemple pour les lois Qalydays, l'espérance de vie résiduelle à 80 ans chez les hommes est estimée à 3,8 ans dans la cohorte sans ancienneté, contre 4,4 ans pour la cohorte avec 1 mois d'ancienneté.

Diagnostic des estimateurs

Trois estimateurs de la survie brute sont employés pour caractériser la loi de décès en dépendance : l'estimateur de Kaplan-Meier, l'arbre de survie et la forêt aléatoire de survie. La loi de mortalité devra seulement dépendre du sexe, de l'âge d'incidence en perte d'autonomie et de l'ancienneté en dépendance.

L'arbre de survie construit est élagué par une validation croisée et finit par présenter deux niveaux. En outre, seules les variables *Diagnosis_Category*, *Gender* et *ClaimType* apparaissent dans les découpes de l'arbre. Par conséquent, l'arbre de survie se réduit simplement à un estimateur de Kaplan-Meier qui dépend uniquement du genre de la personne âgée en perte d'autonomie. La forêt aléatoire de survie est l'extension naturelle de l'arbre de survie. L'examen de la forêt aléatoire de survie s'effectue principalement à partir de la VIMP, de la profondeur minimale et des valeurs de Shapley. Pour la construction de la forêt, toutes les variables sont retenues et les trois plus importantes semblent être *Diagnosis_Category*, *Gender* et *ClaimType*. La variable *IncurredAgeBucket* n'apparaît pas parmi les plus capitales pour les prédictions.

Variables	VIMP	Profondeur minimale	Valeurs de Shapley
Diagnosis_Category	1	3	1
Gender	2	1	2
ClaimType	3	4	4
StateAbbr	4	7	7
Infl_Rider_Bucket	5	6	6
IncurredAgeBucket	6	5	5
GroupIndicator	7	8	9
Cov_Type_Bucket	8	10	10
Region	9	9	8
Incurred Year	10	2	3

TABLE 1 – Comparaison de la hiérarchisation de l'importance des variables déterminée par la VIMP, la profondeur minimale et les valeurs de Shapley

Les performances des différents estimateurs employés sont évaluées à partir du C-index, du score de Brier et de l'AUC cumulative/dynamique. Le pouvoir prédictif de la forêt aléatoire de survie semble très légèrement supérieur à celui de l'arbre de survie et de l'estimateur de Kaplan-Meier.

Estimateurs	C-index de Harrell	Intégrale du score de Brier	Moyenne des C/D AUC
Arbre de survie	0,640	0,103	0,652
Forêt aléatoire de survie	0,645	0,080	0,675

TABLE 2 – Performances prédictives de l'arbre de survie et de la forêt de survie

≪Traditionnellement≫, la fonction de survie et la fonction de hasard cumulative d'un même estimateur partagent la même loi. Cependant, ce n'est pas le cas pour la forêt aléatoire de survie qui délivre deux estimateurs différents, nommés RSF1 et RSF2, à partir des deux quantités énoncées respectivement à la phrase précédente. L'estimateur RSF1 se rapporte à une survie supérieure en comparaison de l'estimateur RSF2.

Lissage de Whittaker-Henderson

Le lissage de Whittaker-Henderson, qui consiste en la conjugaison d'un critère de fidélité et de régularité, est appliqué aux estimateurs de Kaplan-Meier et à ceux issus de la forêt aléatoire. La détermination des paramètres optimaux utilise le critère de validation croisée généralisée et le critère d'information d'Akaike corrigé. Malgré le réglage des paramètres, le lissage réalisé semble faire l'objet de sur-apprentissage. L'aplanissement des estimateurs de la forêt aléatoire de survie est globalement plus satisfaisant en se fiant au critère du R^2. Par ailleurs, tous les estimateurs minorent la mortalité des probabilités brutes de létalité.

Métriques de diagnostic	KM (F/H)	RSF1 (F/H)	RSF2 (F/H)
P-valeur du tests des signes	0,52/0,52	0,52/0,52	0,52/0,52
P-valeur du test de Wilcoxon	0,63/0,76	0,23/0,48	0,31/0,68
χ^2	13,5/11,2	30,07/17,02	31,67/19,05
MAPE	49/50	98/85	89/80
R^2	0,47/0,23	0,49/0,57	0,57/0,63
SMR	1,02/1,02	1,04/1,03	1,03/1,03

TABLE 3 – Diagnostic du lissage des estimateurs de Kaplan-Meier et de la forêt aléatoire de survie

Modèle de Brass pour le positionnement des tables de mortalité

Le modèle de Brass à deux paramètres est employé afin d'adapter les taux de décès américains estimés à la sphère française. Il est usuellement utilisé pour positionner des tables de mortalité plus standards qui dépendent de l'âge atteint et de l'année calendaire. Résultats des ajustements : le positionnement de Brass paraît loin d'être suffisant pour adapter les taux précédemment lissés à la sphère française. La valeur des SMR de RSF1 et de RSF2 suggère que la létalité des lois positionnées est très sous-estimée en comparaison des probabilités estimées par Qalydays. La mortalité autour de la première année passée en dépendance est systématiquement sous-évaluée.

Méthode de Denuit & Goderniaux pour l'extrapolation des tables de mortalité

L'extrapolation des lois de décès estimées intervient sur les anciennetés ainsi que sur les tranches d'âges d'entrée en dépendance. Le modèle de DENUIT et GODERNIAUX, 2005 est employé afin de fermer les anciennetés. Cette méthode permet notamment un rehaussement de la mortalité plus précoce en comparaison des tables Qalydays. L'extrapolation sur les tranches d'âges est, quant à elle, réalisée en appliquant un coefficient de passage. À l'issue de tous les ajustements, les estimées de Kaplan-Meier présentent une forme plus régulière que les probabilités des forêts aléatoires de survie. En cause, la sélection de certaines anciennetés de raccord «prématurées» et un creux de mortalité plus différé dans le temps pour les estimateurs des forêts aléatoires.

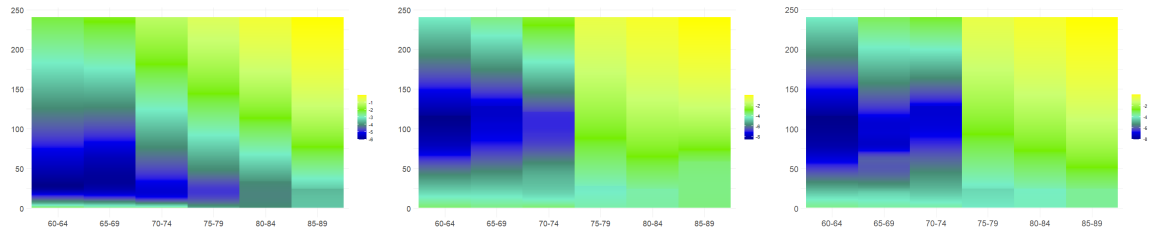


FIGURE 1 – À l'issue de tous les ajustements : carte de chaleur du logarithme des taux de décès féminins par tranche d'âge d'entrée en dépendance sur les 240 premiers mois d'ancienneté en perte d'autonomie. De gauche à droite : estimateur de Kaplan-Meier, RSF1 et RSF2.

Applications : calcul de primes et de provisions pour risques croissants (PRC)

Un modèle illness-death à trois états est considéré afin de calculer les primes et la PRC.

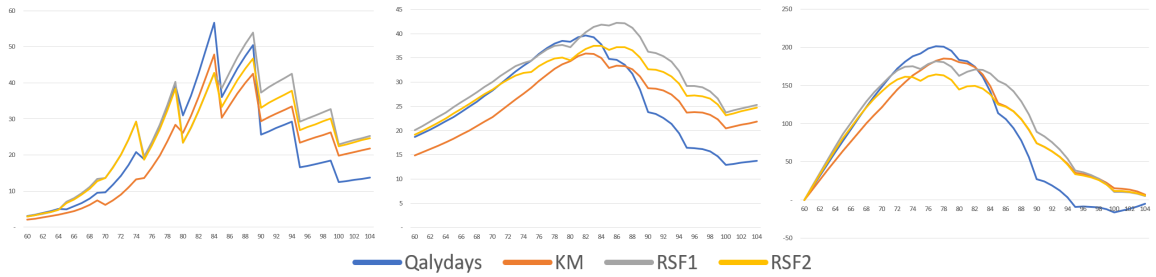


FIGURE 2 – De gauche à droite : prime de risque, prime pure nivelée et ${}_tPRC_{60}$ en fonction de l'âge

Synthesis note

The aim of this dissertation is to construct a French mortality law for heavy dependency based on the American LTC 2000-2011 database of the Society Of Actuaries (SOA) and compare the performance of several estimators.

Definitions of loss of autonomy and the context of long-term care insurance

In France, there are several definitions of loss of autonomy. The best known of these are based on the AGGIR grid or the activities of daily living (ADL) system. More or less relevant equivalences can be established between these different characterisations. The various analyses carried out in this dissertation are confronted with multiple definitions of dependency which will be considered to be systematically analogous. Among the various characterisations of LTC, the one contained in the Garantie Assurance Dépendance (GAD) label, awarded to LTC insurance products that meet 9 specific criteria, is one of the most popular. In response to the modest penetration of serious dependency within the insurance ecosystem, the GAD stamp was introduced. Under this stamp, total dependency is defined in 3 ways : loss of physical, cognitive or mixed autonomy.

However, while this label met with some success between 2013 and 2020, the number of policies bearing this label has fallen sharply since the start of the current decade. Faced with the difficulties encountered by French insurers in the field of long-term care contracts, the authorities have recently stepped in. At the beginning of 2024, the Comité Consultatif du Secteur Financier (CCSF) published 3 recommendations, the main one being the introduction of a compulsory long-term care contract covering severe loss of autonomy.

Data used

Several datasets are used to carry out the work presented in this thesis and the main ones are as follows :

— the SOA Long-Term Care (LTC) 2000-2011 database, reflecting 80% of policies in force in the United States over the observation period 2000-2011. This dataset is used to construct US crude rates of death due to loss of independence.

— the tables for incidence, maintenance in loss of autonomy and death in loss of autonomy estimated by the working group *.

The SOA LTC 2000-2011 dataset was built with contributions from 22 insurers in the US market. In order to ensure comparability between the SOA data and the Qalydays laws, only 'Tax-Qualified' contracts are considered in the study. Once the database has been filtered and rearranged, several remarks can be made about the 73,596 rows in the updated dataset :

— the database is two-thirds female.

∗. https ://www.ressources-actuarielles.net/qalydays

— unlike France, which gives priority to caring for frail older people at home, the proportion of treatment provided at home and in institutions appears to be roughly evenly distributed in the American dataset. In addition, the organisation of French and American care institutions differs on several points. 99% of French senior citizens living in institutions are channelled into EHPAD, USLD and independent living residences, whereas in the US, 60% of the institutionalized population resides in nursing homes.

— in the United States, the treatment of severe cognitive dependency seems to be a greater cause for concern than that of physical dependency. Indeed, relative pathologies, such as Alzheimer's dementia or stroke, seem to be responsible for the longest period of compensation for loss of autonomy.

Once the American death rates have been estimated, they are positioned and closed on the basis of the Qalydays working group's law of death due to loss of autonomy, the latter being assimilated to the gross rates observed on the portfolio of a long-term care insurer recognised by the GAD label. In 2018, this team published several probability laws on the circumference of dependency. The construction of a state model that covers loss of autonomy inevitably requires them, including in particular laws of death in dependency according to sex and type of heavy dependency (cognitive, physical or mixed). The definition of dependency adopted for the calibration of these laws is based on the medical coding definition. The most remarkable specificity of these estimated laws of death lies in the first month of dependency. On the one hand, mortality is very high in this period, and on the other, the model used to estimate these death rates underestimates actual mortality during the first 30 days. For example, for the Qalydays laws, residual life expectancy at age 80 for men is estimated at 3.8 years in the cohort with no seniority, compared with 4.4 years for the cohort with 1 month's seniority.

Diagnosis of estimators

Three crude survival estimators are used to characterise the law of death in dependency : the Kaplan-Meier estimator, the survival tree and the random survival forest. The mortality law will depend only on sex, age at onset of loss of autonomy and length of time in dependency.

The survival tree constructed is pruned by cross-validation and ends up presenting two levels. In addition, only the variables *Diagnosis_Category*, *Gender* and *ClaimType* appear in the tree cuts. As a result, the survival tree reduces simply to a Kaplan-Meier estimator that depends only on the gender of the frail elderly person. The random survival forest is the natural extension of the survival tree. The random survival forest is examined mainly on the basis of the VIMP, minimal depth and Shapley values. For the construction of the forest, all the variables are retained and the three most important seem to be *Diagnosis_Category*, *Gender* and *ClaimType*. The variable *IncurredAgeBucket* does not appear to be among the most important for predictions.

Variables	VIMP	Minimal depth	Shapley values
Diagnosis_Category	1	3	1
Gender	2	1	2
ClaimType	3	4	4
StateAbbr	4	7	7
Infl_Rider_Bucket	5	6	6
IncurredAgeBucket	6	5	5
GroupIndicator	7	8	9
Cov_Type_Bucket	8	10	10
Region	9	9	8
Incurred Year	10	2	3

TABLE 4 – Comparison of the hierarchy of variable importance determined by the VIMP, minimal depth and Shapley values

The performance of the various estimators used was assessed on the basis of the C-index, the Brier score and the cumulative/dynamic AUC. The predictive power of the random survival forest seems very slightly superior to that of the survival tree and the Kaplan-Meier estimator.

Estimators	Harrell C-index	Integrated Brier score	Mean of the AUC over the time
Survival tree	0,640	0,103	0,652
Random Survival Forest	0,645	0,080	0,675

TABLE 5 – Predictive performance of the survival tree and the survival forest

Traditionally, the survival function and the cumulative hazard function of the same estimator share the same law. However, this is not the case for the random survival forest, which delivers two different estimators, named RSF1 and RSF2, from the two quantities stated respectively in the previous sentence. The RSF1 estimator relates to superior survival compared with the RSF2 estimator.

Whittaker-Henderson smoothing

Whittaker-Henderson smoothing, which combines a fidelity and regularity criterion, is applied to Kaplan-Meier and random forest estimators. Optimal parameters are determined using the generalized cross-validation criterion and the corrected Akaike information criterion. Despite parameter tuning, smoothing appears to be overlearned. The smoothing of the estimators from random survival forest is generally more satisfactory if we rely on the R^2 criterion. Furthermore, all the estimators reduce the mortality of the crude lethality probabilities.

Diagnostic metrics	KM (F/H)	RSF1 (F/H)	RSF2 (F/H)
P-value of sign tests	0,52/0,52	0,52/0,52	0,52/0,52
P-value of the Wilcoxon test	0,63/0,76	0,23/0,48	0,31/0,68
χ^2	13,5/11,2	30,07/17,02	31,67/19,05
MAPE	49/50	98/85	89/80
R^2	0,47/0,23	0,49/0,57	0,57/0,63
SMR	1,02/1,02	1,04/1,03	1,03/1,03

TABLE 6 – Diagnosis of smoothing Kaplan-Meier estimators and random survival forest

Brass model for positioning mortality tables

The two-parameter Brass model is used to adapt estimated American death rates to the French sphere. It is usually used to position mortality tables that depend on attained age and calendar year. Adjustment results : Brass' positioning seems far from sufficient to adapt the previously smoothed rates to the French sphere. The SMR values of RSF1 and RSF2 suggest that the lethality of the positioned laws is very much underestimated compared with the probabilities estimated by Qalydays. Mortality around the first year of dependency is systematically underestimated.

Denuit Goderniaux method for extrapolating mortality tables

The extrapolation of the estimated laws of death is carried out on the length of service as well as on the age brackets of entry into dependency. The DENUIT et GODERNIAUX, 2005 model is used to close the seniorities. In particular, this method makes it possible to increase mortality earlier than with the Qalydays tables. Extrapolation across age groups is achieved by applying a pass-through coefficient. After all the adjustments, the Kaplan-Meier estimates have a more regular shape than the probabilities from the random survival forests. This is due to the selection of certain 'premature' connection ages and a more delayed mortality trough for the random forest estimators.

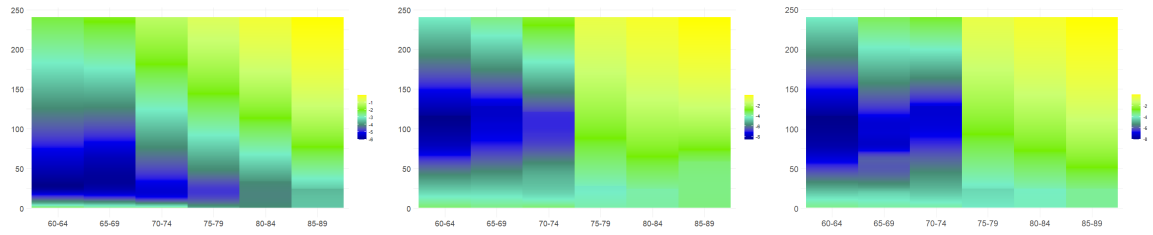


FIGURE 3 – After all adjustments : heat map of the logarithm of female death rates by age bracket at onset of dependency over the first 240 months of loss of autonomy. From left to right : Kaplan-Meier estimator, RSF1 and RSF2.

Applications : calculating premiums and PRC

A three-state illness-death model is used to calculate the premiums and the CRP.

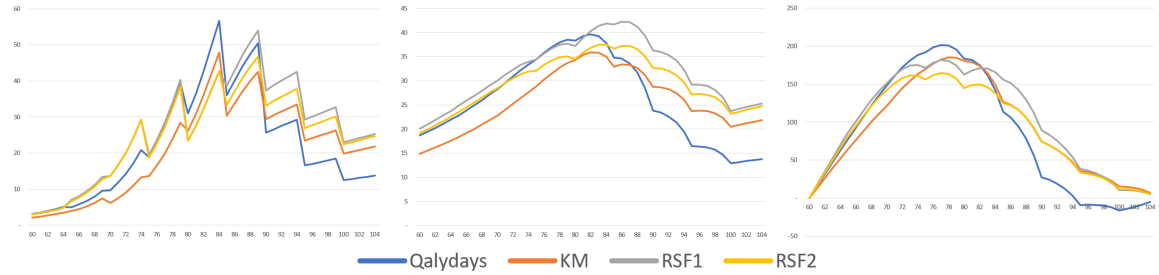


FIGURE 4 – From left to right : risk premium, pure premium and ${}_tPRC_{60}$ as a function of age

Remerciements

Je souhaite tout d'abord ardemment remercier mes maîtres d'apprentissage, Johan STEVENY et Valentin BARDOUX. Merci du fond du cœur pour votre accompagnement et votre disponibilité permanente. Vos recommandations et vos conseils consciencieux m'ont fait me challenger au quotidien et ont fait fleurir mon mémoire dans la bonne direction, j'en suis convaincu.

À l'ensemble de la Business Unit Actuariat de KPMG France, la bienveillance et la chaleur de chacun d'entre vous m'ont permis de m'épanouir quotidiennement dans mon travail et je vous adresse un gigantesque merci. Merci à ceux avec lesquels j'ai échangé sur mon mémoire et, bien sûr, un grand merci à la team des joueurs de baby-foot qui m'a permis de décompresser à chacune de nos pauses.

Des remerciements chaleureux vont également à mon tuteur académique et professeur de dépendance à l'ISUP, Guillaume BIESSY. Merci de m'avoir orienté sur le sujet de la perte d'autonomie pour mon mémoire. Malgré nos points de suivi en petit nombre, votre bienveillance et vos bons conseils m'ont toujours permis, j'en suis persuadé, de rendre mon mémoire meilleur.

Je suis excessivement reconnaissant envers Cindy CORNUAILLE, professeur de prévoyance à l'ISUP, qui m'a permis de trouver une base de données en assurance dépendance et sans qui ce mémoire n'aurait pu voir le jour.

Plus globalement, je tiens également à exprimer ma reconnaissance envers toute l'équipe pédagogique et administrative de l'ISUP, qui m'a guidé tout au long de mon Master et qui m'a fait découvrir le merveilleux macrocosme actuariel sous tous ses angles.

De chauds remerciements à ma chère camarade de l'ISUP, Coline KETTANEH. Tout au long de notre scolarité isupienne, tu as été continuellement de bons conseils et d'une excellente aide pendant chacun de nos cours. Je te respecte profondément comme personne.

Pour finir, mes remerciements les plus immenses et les plus spéciaux vont à ma famille. Merci à ma tante, Phuong NGUYEN, pour nos divers échanges sur le thème de la dépendance et pour tes encouragements sempiternels. À ma mère, Thi-Thanh-Lan NGUYEN, et à mon père, Khoe Dung NGUYEN, merci pour votre soutien indéfectible tout au long de mes études et pour l'éducation que vous m'avez prodiguée, qui ont fait que je suis devenu l'homme que je suis.

À tous, merci infiniment.

Table des matières

Introduction

D'ici 2050, la France comptera près de 4 millions de seniors en perte d'autonomie. L'assurance dépendance, confrontée à une prise en charge complexe et souvent insuffisante pour les assurés, est au cœur des préoccupations depuis plusieurs années. Depuis les années 2010, le marché de l'assurance perte d'autonomie français attend des encouragements du gouvernement. Le 24 janvier 2024, le Comité Consultatif du Secteur Financier secoue l'actualité réglementaire morose de ces dernières années en publiant trois recommandations. Former un contrat de dépendance solidaire sur le pourtour de la dépendance totale constitue sa proposition phare.

La démographie française ainsi que celle des pays développés n'a cessé de se transformer depuis l'après-guerre. Les envolées médicales de ces deux derniers siècles ainsi que l'amélioration des conditions de vie en bonne santé ont repoussé les limites de l'espérance de vie humaine sans discontinuer. L'une des conséquences : la compression de la mortalité et de la morbidité. Ces évolutions incarnent l'essence même du risque de longévité et du risque de morbidité qui pèsent sur la solvabilité des protagonistes du marché de l'assurance pour la perte d'autonomie. En outre, des similarités démographiques sont particulièrement visibles entre les États-Unis et la France.

Alors que la candeur du marché de l'assurance dépendance est plus prononcée du côté de l'Hexagone, l'idée d'employer la profondeur d'historique américaine afin de construire des lois de probabilités pour la perte d'autonomie germe au sein de la communauté actuarielle française. Ce mémoire se destine à complémenter les travaux déjà réalisés sur ce sujet et s'attelle à l'estimation d'une loi française de décès en dépendance à partir, entre autres, du jeu de données américain LTC 2000-2011 de la Society of Actuaries. Pour ce faire, une forêt aléatoire de survie est notamment employée pour l'évaluation des taux bruts de mortalité. Ses performances sont comparées à un estimateur plus rudimentaire : l'estimateur de Kaplan-Meier.

Le mémoire suit l'architecture suivante. Le premier chapitre expose un panorama de l'assurance dépendance tant sur la physionomie américaine que française. Le chapitre 2 consiste en l'introduction des données qui sont employées. Dans le chapitre suivant, l'estimateur de Kaplan-Meier, l'arbre de survie et la forêt de survie aléatoire servent l'évaluation des taux bruts de décès et ces derniers sont ensuite aplanis par le lissage de Whittaker-Henderson. La méthode de Brass, qui est employée pour transposer les lois estimées au périmètre français, et l'étape finale d'extrapolation des probabilités figurent dans l'avant-dernier chapitre. La mutation des lois américaines à la circonférence française est rendue possible par l'existence de loi française de mortalité en dépendance en open source. Un calcul de primes et de PRC conclut les travaux.

Chapitre 1

Le contexte de l'assurance dépendance

1.1 Panorama démographique : États-Unis et France

Depuis les années 1950, les populations occidentales et celles des pays développés voient leur espérance de vie bondir d'année en année. À titre d'exemple, la longévité française gagne un trimestre en moyenne chaque année depuis la période d'après-guerre. La raison principale : la sophistication du traitement des maladies cardio-vasculaires dont les bénéfices sont des gains majeurs d'espérance de vie aux grands âges. La montée de l'incidence en dépendance incarne l'un des revers de la médaille et ne se limite pas à la France. Le vieillissement de la population touche tous les pays du monde et se répercute sur le ratio de dépendance démographique.

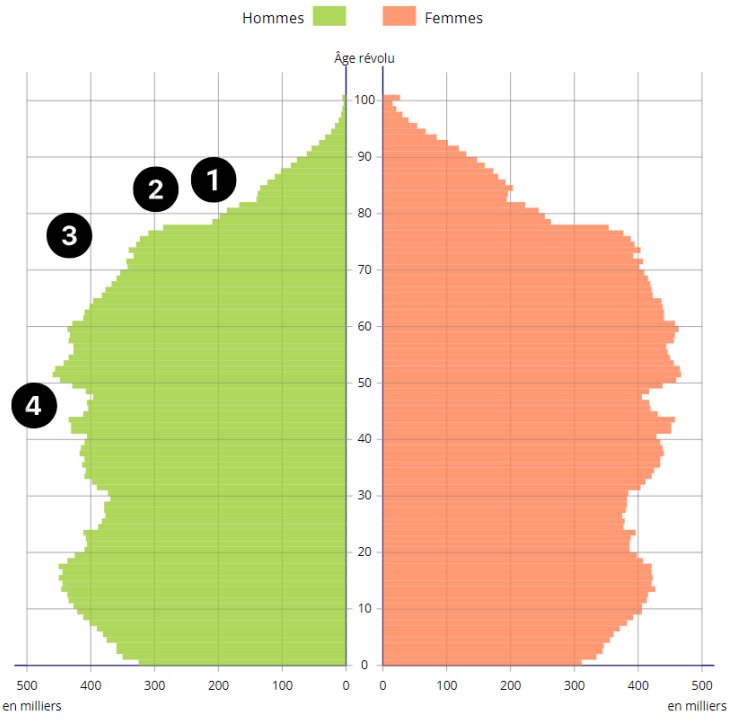


FIGURE 1.1 – Pyramide des âges de la France en 2024. Source : INSEE, 2024b

L'évolution de l'âge au sein de la population des aires développées fait face à deux phénomènes :

— un phénomène structurel : la population vieillit.
— un phénomène conjoncturel : la génération du baby-boom est surreprésentée. À l'heure d'aujourd'hui, les femmes et les hommes de l'après-guerre encore en vie ont dépassé les 70 ans.

Plusieurs moments démographiques se distinguent et sont mis en valeur sur la Figure 1.1 :

— période 1 : natalité portée par les classes creuses nées durant la Première Guerre mondiale
— période 2 : déficit de naissance consécutif à la Seconde Guerre mondiale
— période 3 : baby-boom
— période 4 : fin du baby-boom

L'Hexagone et les États-Unis sont exposés au phénomène du papy-boom. Ce dernier dénomme la montée du nombre de retraités dans les pays développés, événement démographique résultant du baby-boom de l'après-guerre ainsi que de l'accroissement continu de l'espérance de vie. Malgré une espérance de vie à la naissance plus importante d'environ 5 années en France par rapport aux États-Unis, la forme de la pyramide des âges est très similaire entre les deux pays. Ainsi et à l'instar de l'Hexagone, le pays qui détient le 1er PIB mondial n'est pas épargné par l'élargissement de sa population dépendante. Par conséquent, il convient également de se pencher sur les évolutions temporelles de l'intensité et de la durée passée dans un état pathologique incapacitant, état souvent associé aux âges vénérables. En 2019, 2,4 millions de Français présentaient un état de perte d'autonomie, mais ce chiffre pourrait s'élever respectivement à environ 3 et 4 millions d'ici 2030 et 2050 d'après l'Institut National de la Statistique et des Études Économiques (INSEE).

1.1.1 Phénomènes de compression de la mortalité et de compression de la morbidité

Au sein de l'univers assurantiel, le risque dépendance conjugue deux risques à minima : le risque de longévité et le risque de morbidité. Le risque de longévité correspond au risque financier consécutif à l'allongement de la durée de vie moyenne à l'échelle française et mondiale. Quant au risque de morbidité, il se définit comme le risque financier subordonné à la détérioration de l'état de santé des assurés. Ces deux risques sont d'ailleurs captés dans la formule standard de Solvabilité 2 pour la dépendance. Par ailleurs, la mortalité, comme adverse de la longévité, s'oppose par essence à la morbidité. Mourir plus jeune implique potentiellement de passer moins de temps dans un état de santé dégradé. À l'inverse : la morbidité influe-t-elle négativement ou positivement sur la mortalité ? Ainsi, il est intéressant de se pencher sur les interactions de ces deux forces antagonistes. Ces dernières connaissent chacun un phénomène de compression qui est détaillé par la suite.

Compression de la mortalité

La compression de la mortalité se définit comme l'amincissement de la variance de l'âge au décès. Ce phénomène prend place depuis plusieurs années et conduit à la rectangularisation de la fonction de survie de la population dans les pays développés. La rectangularisation se caractérise par la dilatation de la fonction de survie vers la droite. En outre, l'existence de limites théoriques à la diminution de cette variance a été démontrée.

Compression de la morbidité

La compression de la morbidité comprend, quant à elle, l'abaissement du temps passé en morbidité et en incapacité chronique. Cette tendance est le résultat d'un équilibre entre l'incidence en morbidité, les taux de décès et les taux de rétablissement. Les mécanismes sous-jacents à la compression de la morbidité sont multiples. Dans la littérature, les progrès médicaux et les modes de vie plus sains sont avancés comme les causes responsables de ce phénomène.

Pour se rendre compte du poids de l'évolution de la morbidité et de la mortalité au fil du temps, on introduit l'incrément de survie et le décrément de morbidité. L'incrément de survie correspond à l'évolution en années de l'espérance de vie, en supposant que les taux d'incidence en incapacité restent constants au cours du temps. De l'autre côté du spectre, le décrément de morbidité est défini comme la variation en années du temps passé en incapacité, en présumant de la constance temporelle des taux de survie. Le tableau ci-dessous compare l'espérance de vie et les longévités en incapacité pour des individus âgés de 65 ans aux États-Unis entre 1984 et 2004.

	1984	2004	Variation	Incrément de survie	Décrément de morbidité
			Hommes		
Espérance de vie	14,41	16,67	2,26	2,26	-
Durée ADL/CI	1,64	1,26	-0,39	0,44	0,83
Durée ADL	1,19	0,98	-0,21	0,32	0,53
Durée CI	1,09	0,79	-0,30	0,31	0,62
			Femmes		
Espérance de vie	18,66	19,50	0,84	0,84	-
Durée ADL/CI	3,26	2,29	-0,97	0,24	1,21
Durée ADL	2,32	1,88	-0,44	0,17	0,61
Durée CI	2,43	1,55	-0,88	0,18	1,06

TABLE 1.1 – Composantes des variations d'espérance de vie et de longévité en incapacité moyenne au sens HIPAA ADL/CI et HIPAA ADL pour les individus âgés de 65 ans aux États-Unis pour les années 1984 et 2004. Source : STALLARD et YASHIN, 2016

Précisons quelques notations du tableau ci-dessus. ADL est l'abréviation de Activities of Daily Living, se traduisant par activités de la vie quotidienne en français, tandis que CI signifie Cognitive Impairment ou déficience cognitive. Ces sigles synthétisent l'incapacité relative aux actes de la vie quotidienne et/ou à la déficience cognitive au sens de l'Health Insurance Portability and Accountability Act (HIPAA).

D'après le Tableau 1.1 et comme il était attendu, l'espérance de vie s'est allongée en 20 ans tandis que le temps passé en incapacité a diminué. Ces variations comprennent des différences d'échelles importantes entre les deux sexes et c'est également le cas pour les incréments de survie et les décréments de morbidité. Comme le nom de ces derniers pouvait l'indiquer, ils ont un effet antagoniste, du moins sur les durées passées en incapacité. Par conséquent, les deux types de compressions sont fonctions du sexe et pèsent sur les hypothèses des modèles actuariels dépendance. Notons tout de même que ces résultats sont vieux de plus d'une vingtaine d'années et qu'il est nécessaire de prendre du recul vis-à-vis de ces chiffres. En outre, il n'est pas certain que la compression de la morbidité se poursuive, l'amincissement de la morbidité étant compensé par l'augmentation naturelle de la morbidité à incidence constante.

Des changements de population d'étude ou de définition peuvent également amener à

des conclusions très différentes. À titre d'exemple, Crimmins et Beltrán-Sánchez, 2011 affichent un allongement de la morbidité chez les plus de 65 ans entre 1998 et 2008 pour les hommes et les femmes. Dans cette étude, la morbidité est définie comme la perte de mobilité fonctionnelle et l'échantillon américain considéré est issu du National Health Interview Survey.

1.1.2 Spécificités démographiques françaises et projection INSEE de la population française à horizon 2070

D'après l'INSEE, la population française serait constituée de plus d'une personne âgée de 65 ans ou plus pour deux personnes âgées de 20 à 64 ans.

L'INSEE projette à fréquence quinquennale la population française. Cette section reprend la dernière prévision en date à horizon 2070 de l'institut (voir Algava et Blanpain, 2021). Pour projeter le peuple français sur 50 ans, l'INSEE emploie la méthode des composantes qui consiste à modéliser la fécondité, la mortalité et le solde migratoire de façon scindée. À chaque composante, est associée une hypothèse parmi trois possibles : hypothèse centrale, haute ou basse. Les hypothèses sont relatives à l'indice conjoncturel de fécondité (ICF), qui correspond à la somme des taux de fécondité par âge observés une année donnée, et l'âge moyen de maternité. Définir un scénario revient à fixer ces hypothèses. Par la suite, le scénario central sera l'un des seuls à être examiné, ce dernier correspondant synthétiquement au scénario « moyen » et qui considère uniquement les hypothèses centrales. On se rend compte que la faible ancienneté des chiffres présentés par la suite pèse sur leur pertinence. C'est particulièrement remarquable en ce qui concerne les hypothèses migratoires pour lesquelles le solde migratoire est fixé à 70 000 dans le cadre de l'hypothèse centrale. Depuis quelques années, ce chiffre semble déjà dévier avec des valeurs à 183 000 depuis 2021 et que l'on peut, par exemple, expliquer par le conflit russo-ukrainien. Le graphe suivant montre clairement des tendances hautes du solde migratoire depuis 2017, face à un solde naturel en recul depuis plusieurs années.

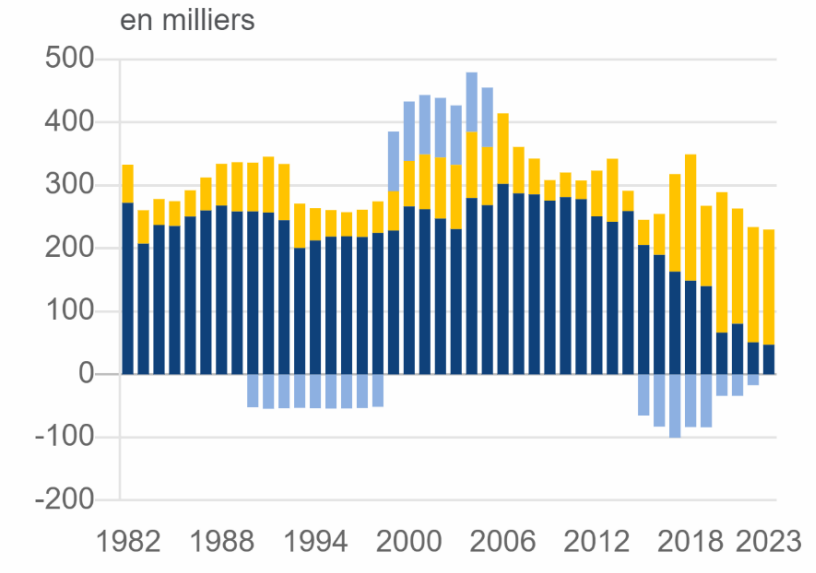


Figure 1.2 – Composantes de la croissance démographique entre 1982 et 2023. Le solde naturel est en bleu foncé, le solde migratoire en jaune et l'ajustement en bleu clair. Source : INSEE, 2024a

Par la suite, on introduit les résultats de la projection INSEE en commençant par la pyramide des âges française qui devrait connaître d'importants changements de forme sous le scénario central.

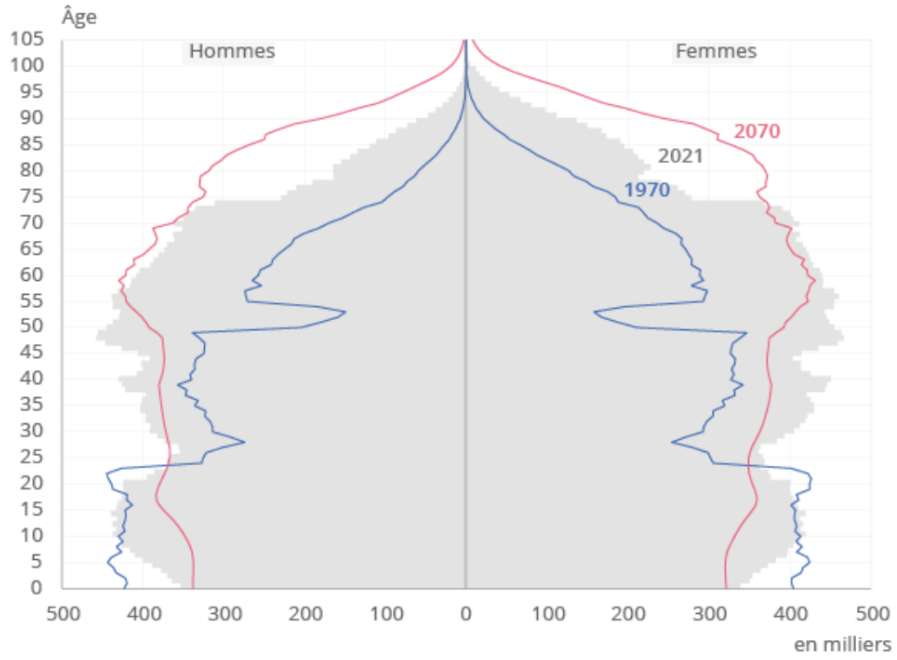


FIGURE 1.3 – Pyramide française des âges par sexe en 1970, en 2021 et projection à horizon 2070 de la pyramide sous le scénario central). Source : INSEE, 2021

En premier lieu, il apparaît que la pyramide prend de moins en moins une forme triangulaire au fil du temps. Le tétraèdre devrait connaître trois changements de formes majeurs entre 2021 et 2070 :

— un élargissement aux grands âges. Les raisons : la hausse actuelle et à venir de l'espérance de vie ainsi que la montée de la population des plus de 75 ans d'ici 2070.

— une compression de la base et du tronc. Les causes : la génération des années 2010 en nombre moins important comparativement aux dernières générations du baby-boom et un ICF en 2021 inférieur à ceux du passé.

— un rééquilibrage entre les taux de survie des femmes et des hommes. En particulier, ce rapprochement devrait se produire aux âges de forte mortalité.

De façon plus précise, la population des plus de 75 ans devrait s'élever d'un incrément de 5,7 millions d'individus, tandis que la population des moins de 60 ans devrait décroître de 5 millions. La cohorte des 60-74 ans devrait, quant à elle, rester stable.

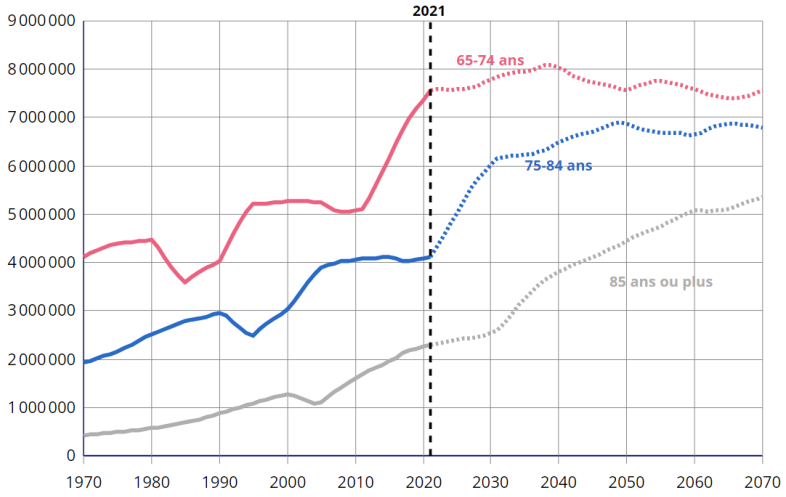


FIGURE 1.4 – Évolution de la population française des plus de 65 ans de 1970 à 2070 sous le scénario central. Source : INSEE, 2021

Il apparaît très clairement que les plus de 65 ans connaîtront une hausse durable de leur effectif. Cet accroissement, qui devrait avoir lieu à la même vitesse que sur les deux dernières décennies, est notamment poussé par les plus de 75 ans. Il est alors intéressant de se pencher sur le ratio de dépendance démographique afin de confronter les retraités à la population active. Le ratio de dépendance ainsi que l'espérance de vie en bonne santé sont des indicateurs, aujourd'hui clés, pour suivre les évolutions de la dépendance économique, qui englobe la dépendance au sens médical. Plusieurs définitions permettent de qualifier le ratio de dépendance. Dans sa caractérisation la plus usuelle, il correspond au rapport du nombre d'individus qui ne travaillent pas sur celui de la population active. Autre définition possible qui coïncide avec la caractérisation de l'INSEE : rapporter la population des plus de 65 ans à celle dont l'âge est compris entre 20 et 64 ans. En outre, cette définition peut diverger car les termes « population active » peuvent recouvrir des cohortes différentes d'un pays à un autre. Ainsi, le rapport s'élèverait à 1 sur 5 pour la population américaine et pourrait monter à 1 sur 4 en 2026. Du côté français, le ratio de dépendance français valait 37% en 2021 et devrait monter à 51% en 2040 selon le scénario central selon les estimations de INSEE, 2021. En outre, il devrait être proche de cette valeur peu importe le scénario considéré, et plus précisément appartenir à l'intervalle $[0,48; 0,53]$.

En 2025, le marché français de l'assurance souffle bientôt ses 40 bougies. Son âge se rapproche petit à petit de l'espérance de vie française et, malgré l'existence de quelques statistiques, dénote des manques encore persistants dans les données dépendance. Les résultats de l'INSEE sont des ressources précieuses afin d'établir une vision à l'échelle macroscopique de la démographie française. Des chiffres plus récents sont également disponibles à la maille régionale et départementale (voir CAZAUBIEL, 2022 pour les lecteurs les plus curieux).

1.2 Comment définir la dépendance ?

Aux États-Unis et en France, la dépendance est considérée comme un état dans lequel il est impossible d'effectuer seul et intégralement un ou plusieurs des actes de la vie courante. Plusieurs niveaux de dépendance existent et la dépendance totale (ou lourde) est généralement

distinguée de la dépendance partielle. Plusieurs référentiels permettent de caractériser l'état de perte d'autonomie. Le référentiel des Actes de la Vie Quotidienne (AVQ) est le plus commun à travers le monde. Dans ce cadre, pour qu'un assuré soit considéré comme dépendant, ce dernier doit être incapable de réaliser un nombre x d'AVQ sur une liste d'un nombre y d'AVQ. Dans la suite, cet état est synthétisé par la notation « xAVQy ». La liste y est différemment composée en fonction du pays examiné. Par ailleurs, cette liste ne diffère que d'un seul AVQ entre les États-Unis et la France. Les AVQs qui y apparaissent sont les suivants :

— faire sa toilette
— s'alimenter
— s'habiller
— être continent
— effectuer ses transferts (déplacements d'un lieu à un autre)
— se déplacer au sein de son domicile (France), aller et sortir des toilettes (États-Unis)

Le x et le y sont différents entre perte d'autonomie lourde et perte d'autonomie partielle. Ces deux états correspondent à des « sévérités » différentes de dépendance.

D'apparence simple, cette caractérisation peut être complexifiée par tout assureur grâce à l'ajout de certains tests. Par exemple, le test cognitif de Folstein, aussi appelé test MMS, permet d'évaluer le niveau de démence d'un individu, autrement dit, son niveau de perte d'autonomie cognitive. En outre, la définition de la dépendance change du tout au tout en modifiant la valeur du x et/ou du y mentionnés plus haut. C'est ainsi qu'à l'échelle du globe, chaque pays privilégie certaines définitions plutôt que d'autres. Ces divergences existent également entre différents assureurs d'un même pays. Par la suite, les spécificités françaises et américaines de la définition sont exposées.

1.2.1 Définition française de la dépendance : système AVQ, grille AGGIR et codage médical

En France, la perte d'autonomie est reconnue comme un état permanent et irréversible. Deux systèmes de formalisation permettent usuellement de définir cet état : le référentiel AVQ et le référentiel de la grille AGGIR. Le référentiel du codage médical, dont se sert le personnel médical en établissements hospitaliers et qui est employé dans une partie des données exploitées dans la suite du mémoire (voir Section 2.1 du Chapitre 2), sera également présenté.

Dans le système AVQ, les quatre premiers AVQs cités plus hauts sont les actes usuellement reconnus dans la définition française. En France, un dépendant total est généralement caractérisé par un état 3AVQ4 ou 4AVQ4, ou par la définition inscrite au label GAD. Petite parenthèse sur ce label (voir France Assureurs, s. d. pour plus de détails) : cette étiquette est accordée à des produits d'assurance en perte d'autonomie lourde qui respectent 9 critères. Parmi ces critères, l'un d'entre eux consiste à adopter une définition commune de la dépendance lourde basée sur le système AVQ. L'assuré en perte d'autonomie est alors dans l'impossibilité de réaliser certains actes de la vie quotidienne sur une liste de 5 actes qui comprend les transferts, les déplacements en intérieur, l'alimentation, la toilette et l'habillage. La garantie minimale doit couvrir les 3 déclinaisons suivantes de la dépendance lourde :

— La dépendance physique (dp) est caractérisée par l'état 4AVQ5. L'individu est incapable d'accomplir 4 actes de la vie quotidienne sur 5 sans l'assistance physique d'un tiers.
— La dépendance cognitive lourde (dem) correspond à l'état 2AVQ5 cumulé à un score

au test MMS de moins de 10. Les fonctions cognitives de l'individu sont altérées et il a constamment besoin d'être surveillé ou d'être incité pour la réalisation de ces 2 AVQs.

— La dépendance mixte (demdp) est associée à l'état 3AVQ5 cumulé à un score au test MMS de moins de 15. Le dépendant est incapable d'accomplir ces 3 AVQs sans l'assistance physique d'un tiers.

L'état d'un dépendant partiel est, quant à lui, estimé généralement à 2AVQ4.

En complément du système AVQ, la grille AGGIR consiste en un système de notation différent et se positionne comme la référence gouvernementale quant à la définition de la dépendance. L'État français emploie notamment ce référentiel afin de calculer le montant d'aide publique en cas d'entrée en dépendance. Par ailleurs, certaines compagnies utilisent également la grille afin de tarifier leur produit d'assurance. La classification AGGIR distingue 6 groupes iso-ressources. Un groupe iso-ressources rassemble des personnes qui nécessitent le même besoin d'aides. Plus le numéro de la classe GIR est élevé, plus le stade de dépendance est important. La dépendance totale circonscrit les niveaux GIR1 et GIR2 tandis que la dépendance partielle inclut le GIR3 et le GIR4. Pour déterminer le niveau de GIR d'un individu, des variables dites « discriminantes » sont employées et permettent de qualifier la perte d'autonomie physique et psychique ainsi que le contexte environnemental et social de l'individu. Une note entre A, B et C est attribuée à chacune de ces variables, puis un algorithme permet d'affecter le niveau de GIR en fonction de toutes les cotations. Ces variables sont au nombre de 10 et reprennent partiellement les AVQs du référentiel AVQs.

Il n'existe aucune équivalence parfaite entre les deux systèmes de définitions. Par ailleurs, SCOR Global Life, 2012 mutualise les deux systèmes de définition et en réalise un classement par degré de dépendance. De la dépendance la plus sévère à la plus légère, il en ressort :

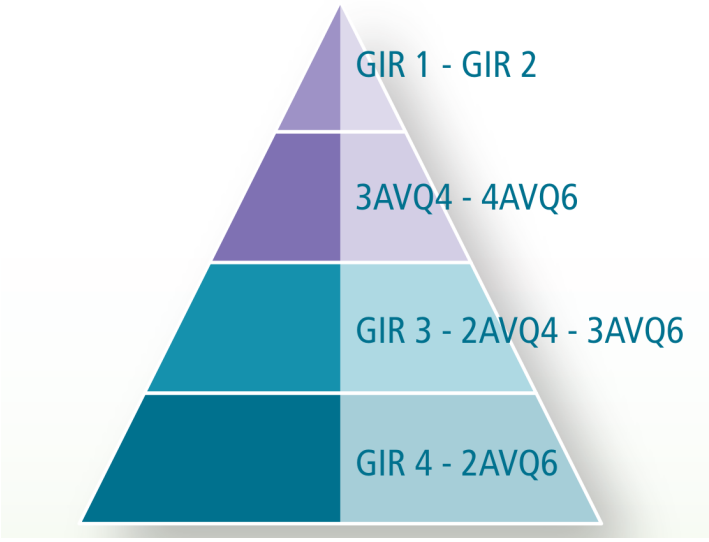


Figure 1.5 – Classement des définitions de la dépendance par sévérité. Source : SCOR Global Life, 2012

Pour finir, synthétisons les points clés du référentiel relatif au codage médical de la perte d'autonomie à l'hôpital, pour lequel il est encore plus difficile d'y trouver des équivalences avec les deux systèmes de notation précédemment exposés. Le codage en question est disponible sur le site internet *ICD-10 Version :2008* 2008. Cette fin de sous-section reprend les observations et les conclusions de Schwarzinger, 2019 afin de présenter cette classification et de motiver

sa comparaison au référentiel AVQ.

Le codage médical de la perte d'autonomie définit deux types de dépendance, on cite :

— Le codage médical d'une « démence » au sens large (maladies d'Alzheimer et apparentées) permet d'identifier toute perte d'autonomie dite « cognitive » (Schwarzinger et al., 2018). De plus, le codage médical permet de préciser le niveau de déficit cognitif, notamment l'existence d'un déficit cognitif sévère (CIM-10 : F00.xx2 ; F01.xx2 ; F02.xx2 ; F03.xx2), i.e., une perte d'autonomie cognitive « totale ». L'existence d'un déficit cognitif sévère est appréciée par la nécessité de la personne de recourir constamment à la surveillance ou l'incitation d'un tiers pour réaliser les actes élémentaires de la vie quotidienne.
— Le codage médical d'un état « grabataire » permet d'identifier une perte d'autonomie « physique » totale. En effet, un état grabataire (CIM-10 : R26.30) est défini par l'« état d'une personne confinée au lit ou au fauteuil par sa maladie, incapable de subvenir seule sans aide et en toute sécurité à ses besoins alimentaires, d'hygiène personnelle, d'élimination et d'exonération, de transfert et de déplacement ».

Dans son étude, Schwarzinger soulève les aspects de validation du caractère "total" de la dépendance identifiée à l'hôpital. L'auteur conclut que le codage médical de la perte d'autonomie au sens général, c'est-à-dire qu'elle soit de nature cognitive ou physique, s'apparente essentiellement à de la dépendance totale. En outre, on remarque que les concepts d'incitation et de surveillance sont des notions récurrentes qui reviennent dans la définition hospitalière et GAD de la démence cognitive (sévère). De plus, l'identification du niveau de dépendance semble fondamentalement fonction du questionnaire en lui-même.

Pour étayer plus finement ces courtes définitions, le personnel hospitalier mesure généralement 6 AVQs chez un patient. Parmi cette liste, 5 d'entre eux s'avèrent coïncider avec les 5 ou 6 actes de la vie quotidienne considérés dans le référentiel AVQ usuel et dans le label GAD. En effet, l'AVQ "Déplacement et locomotion" de la définition hospitalière semble capter les 2 AVQs relatifs aux transferts et aux déplacements intérieurs de la caractérisation usuelle et parmi tous les référentiels, l'AVQ "Relation et communication" apparaît uniquement en structure hospitalière. Dans une analyse plus subtile qui porte sur les définitions précises de chaque AVQ, la liste des AVQ hospitaliers s'accorde particulièrement avec les 4 AVQs du label GAD qui portent sur la perte d'autonomie à l'échelle physique. D'autre part, la continence est l'AVQ qui pèse le plus parmi la liste entière dans la détermination du niveau d'autonomie, tandis que les AVQs à propos de l'alimentation, du comportement ou de la communication informent surtout sur la perte d'autonomie. Des formes d'équivalence apparaissent donc entre les définitions hospitalières et GAD.

1.2.2 Spécificités de la dépendance au sens américain : système AVQ

Cette sous-section s'appuie principalement sur le mémoire d'actuariat de CORNUAILLE, s. d., vers lequel on redirige pour davantage de détails sur le marché de l'assurance américaine dépendance.

Aux États-Unis, la perte d'autonomie n'est pas soumise au critère de permanence. Ainsi, un dépendant peut tout à fait retourner dans l'état d'autonomie. Cette transition rentre ainsi dans les hypothèses des modèles à état dépendance au niveau américain, tandis que ce passage n'est pas du tout capturé dans les hypothèses françaises. Un 2nd point essentiel de divergence

franco-américaine est à noter. Depuis le 1er janvier 1997, deux types de contrats existent aux États-Unis : les contrats tax-qualified et les contrats non tax-qualified. Les contrats tax-qualified doivent notamment respecter certains critères et standards de couverture. Le critère relatif aux conditions de déclenchement de la prestation d'assurance fixe la définition de la dépendance considérée. La prestation est amorcée si l'assuré présente une déficience chronique au sens de l'HIPAA, c'est-à-dire si l'assuré se trouve dans l'un des 3 états suivants, on cite :

— incapable de réaliser, sans l'assistance d'une tierce personne, au moins 2 actes de la vie quotidienne pendant une période évaluée à au moins 90 jours, du fait de déficiences fonctionnelles.

— ayant un niveau de déficiences fonctionnelles équivalentes, approuvé par le Secretary of the Treasury en consultation avec le Secretary of Health and Human Services.

— nécessitant la surveillance d'une tierce personne du fait d'une déficience cognitive sévère.

D'autres différences entre États-Unis et France apparaissent relativement à la construction des produits d'assurance et seront détaillées en Section 1.5.

1.3 Le marché français de l'assurance dépendance

Le risque dépendance, comme conjugaison d'un risque biométrique et d'un risque financier, s'apprécie au long terme et peut présenter une volatilité importante. Depuis plusieurs années, les acteurs du microcosme de l'assurance dépendance sont confrontés à la problématique du manque de données. Par ailleurs, cette contrainte se résorbe progressivement grâce à la création de connaissances actuarielles fiables sur le sujet.

1.3.1 Chiffres de la DREES : répartition du sexe, de l'âge et du lieu de résidence en perte d'autonomie

Les femmes passent 50% de temps en plus en dépendance par rapport aux hommes et cette tendance se retranscrit également dans les souscriptions. Cette différence se recopie partiellement dans la Figure 1.6. De surcroît, les femmes sont également les premières à souscrire un contrat d'assurance dépendance. Que cela soit à domicile ou en établissement, globalement 3 dépendants sur 4 sont des femmes. Les maladies neurologiques et cardiovasculaires sont les principales causes menant l'homme à la perte d'autonomie. De l'autre côté du spectre, c'est la maladie d'Alzheimer qui conduit les femmes en dépendance. En outre et sans surprise, les individus qui présentent un âge très avancé sont les plus touchés par la perte d'autonomie.

En 2015, les seniors en perte d'autonomie résidaient majoritairement à domicile. Ces derniers étaient au nombre de 1 948 700 sur les 2 488 900 dépendants d'après l'enquête EHPA 2015 de la DREES, 2015.

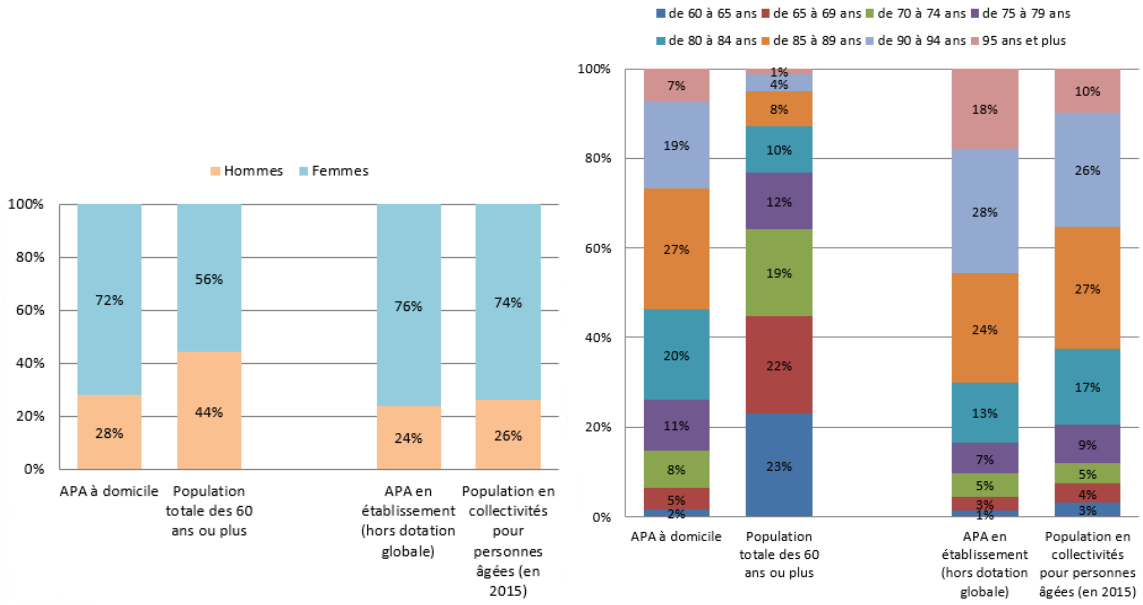


FIGURE 1.6 – Répartition par sexe et par âge des bénéficiaires de l'APA, en décembre 2019. Source : DREES, 2020

Les établissements français de résidence des seniors sont répartis usuellement en quatre catégories : EHPAD, EHPA, USLD et résidence autonomie. Il est intéressant de se pencher sur les divergences marquées entre le système français et l'organisation structurale des établissements de soins américains. Il est possible d'établir certaines équivalences entre les différents types d'institutions des deux nations en se référant simplement aux définitions qui leur sont attribuées. Aux États-Unis, les 3 classes de lieux de résidence pour personnes âgées, dans lesquels les individus de la base de données américaine employée dans la suite du mémoire vivent, et leurs pendants français sont les suivants :

— nursing home (NH) : les EHPADs, les maisons de retraite. De plus, les skilled nursing facilities (SNF), la réplique américaine des USLD, sont souvent intégrés à des nursing homes. Notons également que les seniors en perte d'autonomie qui vivent en résidence autonomie ont l'obligation de signer une convention avec un EHPAD.

— assisted living facilities (ALF) : les EHPA

— home health care (HHC) : les soins à domicile

L'enquête EHPA 2019 de la DREES, 2019, qui collecte des données sur l'activité des structures médico-sociales accueillant des personnes âgées, permet d'accéder à la répartition des seniors en perte d'autonomie par catégorie d'établissement.

La Table 1.2 suscite plusieurs commentaires. Les établissements concentrent une majorité de dépendants lourds, et pour cause, les seniors en dépendance peu sévère résident essentiellement à leur domicile. Il apparaît qu'un type d'établissement domine la rétention de seniors en perte d'autonomie : ce sont les EHPADs. Ces derniers capturent 91% des personnes âgées dépendantes qui résident en établissement. Le cumul des EHPADs, des USLDs et des résidences autonomie capte quant à lui 99% de cette même population. Comme ce sera vu dans le Chapitre 2, cette répartition dénote complètement avec le jeu de données américain qui sera employé dans les travaux du mémoire.

	GIR1 ou 2	GIR3 ou 4	Seniors	% de seniors en GIR1 ou 2	% de seniors en GIR3 ou 4
EHPAD	324 112	230 744	594 700	55%	39%
EHPA	749	2 596	5 900	13%	44%
USLD	23 959	5 364	29 800	80%	18%
Résidence autonomie	1 394	22 908	99 600	1%	23%
Total	350 214	261 612	730 000	48%	36%

TABLE 1.2 – Répartition française des résidents seniors au 31/12/2019 selon leur niveau de dépendance et par catégorie d'établissement. Source : DREES, 2019

1.3.2 Évolution de la population dépendante

En 2019, la population française était formée de 2,4 millions de dépendants, dont 1,3 million qui étaient bénéficiaires de l'Allocation Autonomie Personnalisée (APA). D'après le *Rapport de la concertation grand âge et autonomie* (voir LIBAULT, 2019), le nombre de bénéficiaires de l'APA pourrait s'élever à 1,6 millions en 2030 et atteindre 2,2 millions en 2050. D'ici 2030, la hausse annuelle des effectifs de cette population serait de 20 000 et monterait à 40 000 entre les années 2030 et 2040, fort de l'arrivée aux grands âges des premières naissances du baby-boom.

Alors que la moitié des Français se disent préoccupés par le risque de leur propre perte d'autonomie liée à l'âge, seulement 10% ont souscrit un contrat d'assurance dépendance. De plus, sur un échantillon de 1013 personnes interrogées, 13% des 22-44 ans déclarent avoir souscrit à un contrat d'assurance tandis que ce pourcentage tombe à 8% chez les 45-75 ans. Ces chiffres sont issus de l'étude OpinionWay qui a été réalisée à la demande de France Assureurs. La crainte de l'entrée en dépendance pèse sur les Français alors que le reste à charge moyen d'un dépendant en établissement s'élève à 1875 euros mensuels (source : FRANCE ASSUREURS, 2021) et que les Français estiment ne pas être suffisamment indemnisés. La défiance face à cette assurance est invoquée pour expliquer le faible taux de souscription, défiance qui conjugue plusieurs raisons : niveaux de remboursement et prise en charge faible, mauvaise image de l'assurance, contrats complexes, niveaux de cotisations trop élevés et manque d'informations.

1.3.3 Enjeux financiers et APA

L'APA est une aide financière publique qui dépend de la grille AGGIR. Le montant de l'allocation dépend du revenu du bénéficiaire. Fin décembre 2020, les prestations ont été versées à 1,3 million de bénéficiaires âgés de plus de 60 ans. L'enjeu financier est majeur. Ce dernier s'élève à 30 milliards d'euros d'après les chiffres de la Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES), soit l'équivalent de 1,4 point de PIB. Ce montant pourrait grandir jusqu'à atteindre 2,8% de PIB, dont 2,1% correspondant à de la dépense publique. En 2014, ces dépenses couvrent les champs de la santé, de la dépendance et de l'hébergement qui pèsent respectivement pour 12,2 milliards, 10,7 milliards et 7,1 milliards d'euros. En comparaison, le taux de reste à charge vaut respectivement 1%, 22% et 54% et se veut très sporadique entre les 3 champs. Même si la sécurité sociale et les collectivités locales supportent pour beaucoup le montant des dépenses, la prestation moyenne versée monte à seulement 533 euros. Malgré le fait que la France emploie massivement son PIB au financement

des retraites, elle concède un sérieux retard quant à la prise en charge de ses dépendants.

1.3.4 Dépendance totale et label GAD

Depuis plusieurs années, la couverture de la dépendance totale connaît un phénomène de convergence, alors que le comité consultatif du secteur financier a publié ses recommandations sur le sujet en janvier 2024. Cette convergence double s'installe pour la dépendance lourde : les taux de passages d'un état à un autre convergent et l'offre dépendance se concentre sur cette couverture. Également quelques chiffres sur cette gravité de dépendance : la durée d'indemnisation monte à 4 ans en moyenne et les dépendants de moins de 2 ans d'ancienneté connaissent de la surmortalité.

L'instauration du label Garantie Assurance Dépendance (GAD) va dans le sens de la pénétration prononcée de la dépendance lourde au sein de l'écosystème assurantiel dépendance. Le label GAD n'a aucune valeur légale, mais a pour objectif de donner de la visibilité aux produits. L'assuré qui souscrit à ce produit d'assurance se voit promis un certain niveau de garantie. Par ailleurs, le label n'a pas le succès escompté sur le marché.

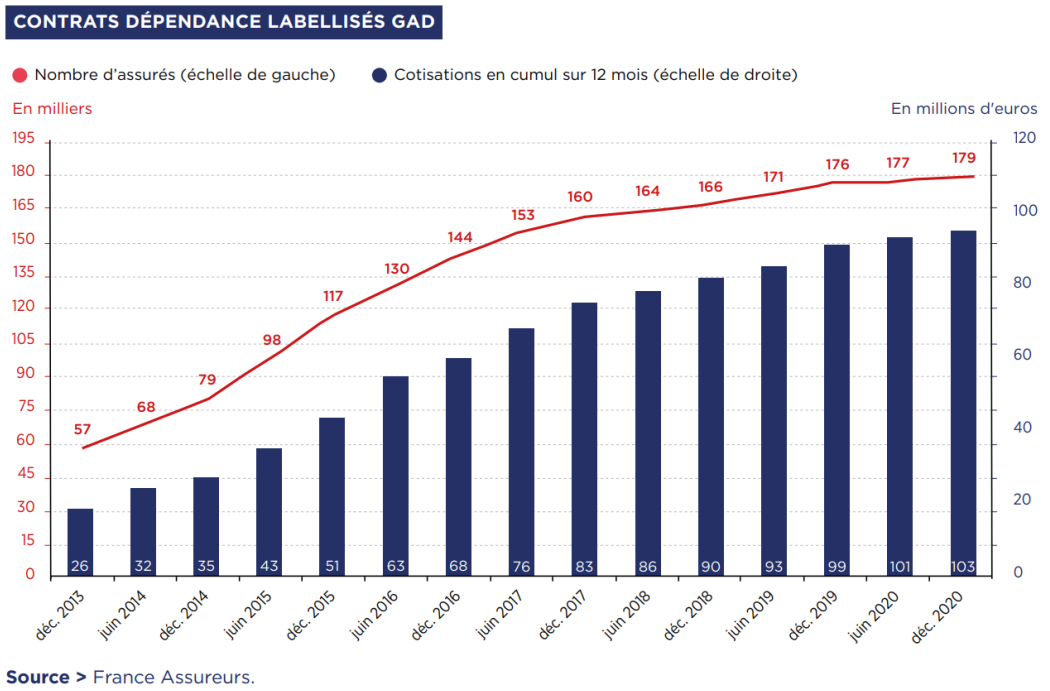


FIGURE 1.7 – Évolution du nombre d'assurés ayant souscrit à des contrats dépendance labellisés GAD et des cotisations relatives entre 2013 et 2020. Source : FRANCE ASSUREURS, 2021

Depuis la mise en place du label en 2013, le nombre de contrats dépendance reconnus par cette marque a augmenté à vitesse globalement constante. Par ailleurs, l'année 2020 a marqué un affaiblissement de la hauteur des incréments annuels du nombre de contrats. Les chiffres de France Assureurs arborent une montée de seulement 2 % du nombre de souscriptions de contrats labellisés GAD en 2020, alors que ce chiffre se portait à 6% pour l'année 2019. Par ailleurs, il est nécessaire de relativiser sur les chiffres suivants étant donné la période sanitaire mouvementée qu'ont vécue le monde et la France entre 2020 et 2022. Ce point apparaît notamment comme la cause de la réduction de 30% des souscriptions nouvelles.

Toujours d'après France Assureurs, ces contrats représentent 51% des nouvelles souscriptions de contrats prévoyant une garantie unique dépendance.

Par ailleurs, la dépendance totale ne se cantonne pas à la caractérisation inscrite au label GAD et le spectre des définitions apparaît très étendu au sein du macrocosme assurantiel. Le tableau suivant expose la répartition du nombre de contrats et du pourcentage de personnes couvertes en fonction de la définition de la dépendance totale au sein d'un échantillon de 25 contrats à adhésion individuelle :

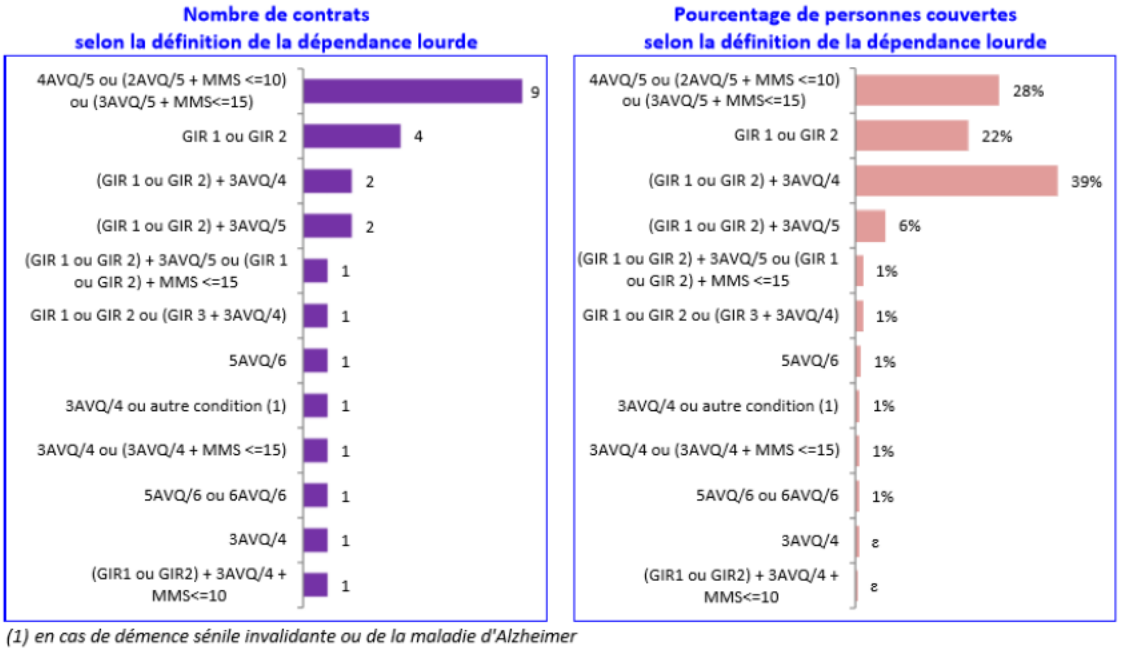


FIGURE 1.8 – Répartition du nombre de contrats et du pourcentage de personnes couvertes en fonction de la définition de la dépendance totale. Source : France Assureurs (2016)

Les contrats qui opèrent avec la définition de la dépendance lourde inscrite au label GAD sont les plus représentés. En outre, le constat n'est pas le même pour la proportion de personnes couvertes. Cette cohorte n'est qu'à la seconde place en représentant 28% du total. Présenter un état 3AVQ4 et GIR1/GIR 2 est la caractérisation la plus prisée à ce niveau-là. Prendre du recul par rapport à ces chiffres de 2016 s'impose de nouveau vis-à-vis de la dissonance temporelle : nous sommes en 2024 et il aurait été intéressant d'avoir ces chiffres pour l'année 2020 afin de les comparer aux chiffres de la Figure 1.7.

1.3.5 Dépendance partielle

À l'opposé de la convergence du marché sur le volet de la dépendance lourde, la dépendance partielle connaît un contexte de divergence. Cette couverture correspond à un risque nouveau. Les taux relatifs n'y sont pas encore bien définis malgré des acteurs du marché qui introduisent des prestations mensuelles sur ce pan dans la majorité des contrats depuis 2020. Les connaissances sont particulièrement concentrées aux grands âges. On estime que 30 ans d'historique de données seraient nécessaires afin d'assurer la convergence de la calibration des taux de passage ainsi que l'évaluation des garanties. La prudence est alors de rigueur dans ce marché encore naissant, d'autant plus que les progrès de la médecine sur le traitement des maladies

neurodégénératives amplifient la divergence. En outre, le champ des définitions de cette échelle de dépendance est encore plus épars que celui relatif à la perte d'autonomie totale. Ci-dessous le pendant de la Figure 1.8 pour la dépendance partielle.

Mesure de la dépendance partielle	Nombre de contrats	Nombre de personnes couvertes
2AVQ/4	1	1%
2AVQ/5	2	4%
3AVQ/5	1	1%
3AVQ/6 ou 4AVQ/6	1	ε
4AVQ/6	2	1%
GIR 3	2	ε
GIR 3 ou GIR 4	1	14%
GIR 3 + 2AVQ/5	1	4%
GIR 3 + 2AVQ/4	3	32%
(GIR 3 ou GIR 4) + 2AVQ/4	1	3%
3AVQ/5 ou (2AVQ/5+MMS <=15)	1	1%
2AVQ/4 + MMS = [16,23]	1	ε
(GIR1 ou GIR2 ou GIR3) + 2AVQ/4 + MMS<=15	1	ε
(GIR 1 ou GIR 2 ou GIR 3) + (3AVQ/5 ou 4AVQ/5 ou MMS <=10)	1	15%
GIR 3 + 2AVQ/4 ou GIR 3 + MMS <=15	1	1%
Non renseigné	10	17%
Total	**30**	**95 %**

FIGURE 1.9 – Répartition du nombre de contrats et du pourcentage de personnes couvertes en fonction de la définition de la dépendance partielle. Source : France Assureurs (2016)

L'état GIR3 et 2AVQ4 ressort comme le plus représenté et se présente comme la caractérisation usuelle de cette intensité de dépendance.

1.3.6 Action de la force publique et recommandations de la CCSF du 24 janvier 2024

Alors que la France paraît encore insuffisamment préparée à relever le défi du financement de la dépendance, les pouvoirs publics commencent à se mobiliser. Depuis plusieurs dizaines d'années, le monde assurantiel pour la prise en charge de la perte d'autonomie attend des signes d'aide du gouvernement. Ce dernier se manifeste depuis peu. En effet, face aux défis sociaux et financiers consécutifs aux évolutions de la démographie française, le Parlement a convergé sur une réponse : l'instauration d'un système assurantiel obligatoire pour la dépendance. Cette idée est consécutive à 3 raisons :

— l'assurance n'est pas en mesure de proposer une couverture large et efficace qui repose sur un équilibre entre mutualisation et individualisation.

— la couverture dépendance ne capte pas la classe moyenne à cause de l'antisélection. Dans l'univers assurantiel, l'antisélection ou sélection adverse correspond à l'asymétrie de l'information détenue entre l'assureur et l'assuré.

— le reste à charge pour les bénéficiaires de l'APA semble trop important.

Par ailleurs, la fondation Concorde, laboratoire d'idées françaises, soulève la nécessité de trouver un équilibre entre pouvoirs publics et assureurs. Le 24 janvier 2024, la CCSF a publié un communiqué de presse qui dévoile 3 recommandations phares au sujet de l'assurance dépendance (voir CCSF, 2024 pour davantage de détails) :

1. mettre en place un contrat dépendance solidaire obligatoire sur le périmètre de la perte d'autonomie lourde.
2. créer une gouvernance collégiale qui mélange les partenaires sociaux, les représentants d'associations, les représentants des assureurs et les pouvoirs publics. Cette structure aurait en charge la mise en œuvre et la supervision du contrat ci-dessus.
3. créer une grille tarifaire unique.

En raison de l'incapacité des pouvoirs publics à assumer la prise en charge, la CCSF a exploré des solutions de financement alternatives pour la dépendance, impliquant le secteur privé. Les recommandations, qui en découlent, sont portées par le constat de contrats d'assurance dépendance coûteux et générateurs de litiges, ainsi que par le reste à charge élevé pour les ménages, comme évoqué plus haut.

Des propositions similaires avaient déjà été avancées par France Assureurs dans son livre blanc *Construire une nouvelle solution solidaire et transparente face à la dépendance liée à l'âge* (voir France Assureurs, 2021) publié à l'occasion des élections présidentielles de 2022. France Assureurs proposait d'instaurer un nouveau contrat de complémentaire santé responsable comme solution aux défis de la dépendance. Les recommandations de la CCSF rejoignent sur plusieurs points les dires de France Assureurs.

1.4 Un bref contexte américain dépendance

Tout comme pour l'Hexagone, l'assurance dépendance américaine ne rencontre pas un franc succès. Alors que le nombre de polices d'assurance dépendance montait à 750 000 en 2002, ce chiffre s'est graduellement éteint pour atteindre les 100 000 en 2016. La raison : la montée en puissance des produits dits « combo », qui, comme leur nom l'indique, fusionnent une couverture dépendance à d'autres contrats comme des contrats d'assurance-vie. L'avantage pour l'assureur : atténuer son risque.

La complexité des produits perte d'autonomie réside dans la multiplicité des risques qu'elle supporte, ces derniers n'étant pas tous intrinsèques à la dépendance en elle-même. Économies régionales, recherches médicales et évolutions réglementaires sont autant de facteurs qui modulent cette discipline actuarielle.

Par ailleurs, l'assurance n'est pas régulée sur le périmètre de la nation, mais plutôt par les lois et les régulations des juridictions constitutionnelles. D'après l'expérience américaine, les réserves de police des 10 ou 20 dernières années ne suffisent pas à projeter les sinistres. Les rachats en deçà des prévisions n'arrivent pas à financer les sinistres attendus des polices restantes en raison de réserves plus faibles. Beaucoup de contrats proposent une option « Cost of Living Adjustment (COLA) », cette dernière apparaît systématiquement dans les contrats Tax-Qualified.

1.5 Comparaison de l'assurance dépendance française et américaine

	France	**États-Unis**
Système public	APA	Medicare-Medicaid
Type de prestation	Forfaitaire	Remboursement
Durée de la couverture	Viagère	Viagère
Durée de la prestation	Viagère	De 1 an à viagère
Durée de paiement de la prime	Viagère	Viagère
Structure de la prime	Nivelée, révisable	Nivelée, révisabilité si acceptation des régulateurs
Sélection médicale	Simplifiée	Complète
Conditions de souscription	Questionnaire médical parfois complétée par des demandes d'informations complémentaires	Examen médical, tests neurologiques, demande d'historique médical
Modalités de définition du niveau de garantie	Dépendance totale VS dépendance partielle	Le niveau de garantie est défini en fonction du lieu de résidence
Franchise	90 jours	De 90 à 360 jours
Période de carence	Accident : 0, maladie : 1 an, démence : 3 ans	Limitée
Définition de la dépendance	AGGIR ou 3 AVQ 4 ou définition GAD ; la perte d'autonomie est permanente et irréversible	2 AVQ 5 et démence ; pas de critère de permanence requis
Âge limite de souscription	75 ans le plus souvent	84 ans le plus souvent

TABLE 1.3 – Comparaison des produits d'assurance français et américain. Source : SCOR et CORNUAILLE, s. d.

Chapitre 2

Construction et analyse des données

Ce chapitre s'applique à présenter, filtrer, afficher et expliquer les différents jeux de données qui seront employés ultérieurement. Afin de mener à bien nos travaux, 2 sources de données sont employées :

— les tables d'incidence, de maintien et de mortalité estimées par le groupe de travail Qalydays (disponible en open source [†]).

— la base de données américaine SOA LTC 2000-2011.

2.1 Lois de probabilité estimées par le groupe Qalydays

2.1.1 Présentation générale des tables de probabilité et de Qalydays

Des tables françaises d'incidence et de maintien en dépendance sont disponibles en open source et ont été calibrées par le groupe de travail Qalydays. Ce dernier est formé de chercheurs qui couvrent les champs de la médecine, de l'ingénierie, de l'actuariat et de l'épidémiologie. De par leurs travaux, l'équipe de recherche souhaite impacter positivement les champs de la prévention et de la prévoyance du risque de perte d'autonomie. En 2018, Qalydays a publié plusieurs études sur la construction de lois qui permettent de probabiliser les différentes transitions dans un modèle à état qui couvre l'assurance dépendance. Ces tables se déclinent selon le sexe et la catégorie de perte d'autonomie parmi la dépendance lourde cognitive, physique ou toutes causes (c'est-à-dire cognitive et physique). Les auteurs caractérisent la dépendance lourde par la définition issue du codage médical (voir sous-section 1.2.1 du Chapitre 1). Dans la suite de cette section, la dépendance lourde/totale se référera systématiquement à la dépendance lourde/totale au sens du codage médical. Les tables mentionnées sont les suivantes :

— une table d'incidence en dépendance par âge atteint.

— une table de mortalité en dépendance par âge d'entrée en dépendance ainsi que par ancienneté en perte d'autonomie.

— une table de décès prospective pour une population qui n'a jamais connu l'état de dépendance, par âge et par année.

†. https ://www.ressources-actuarielles.net/qalydays

Dans la suite du chapitre, seule la loi de mortalité en dépendance sera examinée. Les lois de transitions et la méthodologie de construction de ces dernières sont disponibles sur le site internet *Lois biométriques pour le risque de perte d'autonomie en France* 2020.

2.1.2 Loi de décès en dépendance lourde toutes causes (ou mixte)

Cette sous-section présente les spécificités, jugées intéressantes, relatives aux tables de décès en dépendance lourde toutes causes ainsi qu'à leur construction (voir Planchet et al., 2018a). Les tables de décès en dépendance sont des tables prospectives qui dépendent de l'âge d'entrée en perte d'autonomie (compris entre 60 et 110 ans) ainsi que de l'ancienneté en dépendance (compris entre 0 et 611 mois). Elles ont été construites à partir des bases nationales d'hospitalisation PMSI 2008-2013, avec 2010-2012 comme période de référence. Notons que cette période d'observation est relativement courte pour estimer des lois dépendance et qu'il est usuellement conseillé de calibrer de telles probabilités sur une durée de 10 ans afin de restreindre l'impact de l'évolution du risque dans le temps. Cette base de données capture les informations médicales et administratives des hospitalisations en court séjour (MCO), en soins de suite et de réadaptation (SSR), à domicile (HAD) et en psychiatrie (voir *Descriptif du contenu des bases de données PMSI* 2023 pour la documentation relative). Notons que la fiabilité des tables Qalydays tient avant tout pour les âges d'entrées en dépendance inférieurs à 95 ans. Au-delà, les taux de décès ont été extrapolés. En outre, une révision de ces tables datant de 2020 a introduit une correction des taux de mortalité pour les âges postérieurs à 95 ans.

Afin de construire ces tables, le groupe de travail est passé par plusieurs étapes d'estimation et a notamment embelli leurs travaux sur 3 points :

— réalisation d'un ajustement spécifique des taux de première année en raison de la mortalité très importante qui intervient à l'entrée en dépendance.
— positionnement des taux au-delà de la première année par l'ajout d'une surmortalité aux taux issus de la table TD 88/90.
— extrapolation et mensualisation des probabilités conditionnelles de décès ajustées ainsi obtenues.

Il est intéressant de se soucier des espérances de vie résiduelles à 80 ans de ces lois présentées dans le Tableau 2.1. Ces dernières sont particulièrement faibles à cause de l'estimation de taux de décès très importants au cours du premier mois de perte d'autonomie. À cela s'ajoute des tendances de leur modèle à sous-estimer la mortalité au cours du premier mois. Ces différences apparaissent distinctement entre les espérances sans ancienneté et avec un mois écoulé en dépendance. D'autre part, l'écart d'espérance de vie entre les femmes et les hommes s'amenuise en considérant ce mois d'ancienneté. En outre, les espérances de vie sont globalement croissantes sur la première année écoulée en dépendance, tout âge confondu à partir de 60 ans.

Les différences de survie entre les deux sexes y sont les plus visibles aux âges proches de 60 ans et aux faibles anciennetés d'après les taux de décès et les espérances de vie résiduelles. La fonction de survie décroît très rapidement avec l'ancienneté, étant donné que nous sommes à des âges élevés. La Figure 2.1, quant à elle, résume l'information portée par ces lois par le prisme de l'espérance de vie résiduelle, de la médiane et du 3ème quartile pour la mortalité au moment de l'entrée en dépendance. Toutes les quantités connaissent plus ou moins une décroissance systématique, ce qui épouse l'intuition usuelle. L'espérance résiduelle nous intéressera particulièrement afin de comparer les deux sexes qui présentent des différences

Lois de probabilité et paramètres employés	Hommes	Femmes
Qalydays : aucune ancienneté en dépendance	3,8	4,7
Qalydays : 1 mois d'ancienneté en dépendance	4,4	5,1
TH002/TF002	7	9

TABLE 2.1 – Espérance de vie résiduelle à 80 ans selon différents paramétrages. Source : PLANCHET et al., 2018a

significatives en dépendance toutes causes d'après PLANCHET et al., 2018a.

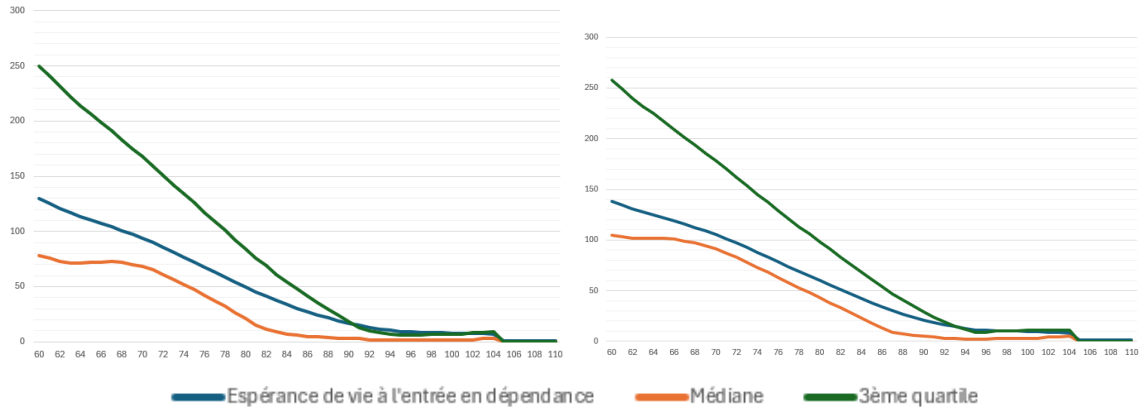


FIGURE 2.1 – À l'entrée en dépendance : espérance de vie résiduelle, médiane de la mortalité et 3ème quartile de la mortalité exprimée en mois et en fonction de l'âge d'entrée en perte d'autonomie. À gauche pour les hommes et à droite pour les femmes. Source des données : *Lois biométriques pour le risque de perte d'autonomie en France* 2020

2.2 Base de données SOA LTC 2000-2011

2.2.1 Présentation

Le jeu de données SOA LTC 2000-2011 a été construit à l'initiative de la Society of Actuaries (SOA) dans le cadre de l'établissement de tables d'expérience relatives aux soins de longue durée aux États-Unis. La description qui suit de cette base de données est tirée des rapports suivants de la SOA :

— *Long-Term Care Intercompany Experience Study – Aggregate Database 2000-2011 Report*, SOCIETY OF ACTUARIES, 2015b
— *Long-Term Care Experience Basic Table Development*, SOCIETY OF ACTUARIES, 2015a
— *Caveats for Use of Long-Term Care Experience Basic Tables*, SOCIETY OF ACTUARIES, 2017

Les données ont été recueillies auprès de 22 assureurs américains et la base complète est le reflet de 80% des polices en vigueur. Avant d'être utilisées, les données sont passées par une phase d'évaluation de leur qualité. Les variables qui y figurent sont également agrégées par

souci de confidentialité. La période d'étude s'étend de 2000 à 2011. Le jeu entier est formé de 3 bases et seule celle relative aux fins de sinistres nous intéressera par la suite. Les variables qui y apparaissent sont les suivantes :

- *GroupIndicator* : est-ce que le contrat est individuel ou collectif ?
- *Gender* : genre
- *IncurredAgeBucket* : intervalle de longueur 5 ans qui contient l'âge à la survenance du sinistre
- *IncurredYear* : année de survenance du sinistre
- *ClaimType* : lieu de vie en dépendance
- *Region* : région géographique des États-Unis où la police d'assurance a été souscrite
- *StateAbbr* : état fédéral dans lequel la police d'assurance a été souscrite
- *Diagnosis_Category* : pathologie initiale au début du sinistre
- *TQ Status* : statut fiscal du contrat
- *Cov_Type_Bucket* : est-ce que l'assurance est "comprehensive" ou non ? Une assurance dite "comprehensive" donne lieu à des prestations en établissement et à domicile, tandis qu'une assurance dite "non-comprehensive" inclut exclusivement des prestations en établissement
- *Infl_Rider_Bucket* : options d'augmentation des prestations
- *EP_Bucket* : durée de la franchise en jours
- *Max_Ben_Bucket* : durée maximale de versement des prestations prévues au contrat selon le lieu de vie
- *ClaimDuration* : durée du sinistre en mois. Cette durée comprend la période de franchise
- *Terminations* : nombre de décès et de rétablissement
- *Deaths* : nombre de décès

À noter : les données qui sont parvenues à la SOA n'ont pas été altérées afin d'adhérer à l'application des règles Safe Harbor.

Plusieurs termes issus du lexique assurantiel sont clarifiés ci-dessous :

- elimination period ou franchise en français : la franchise est un délai contractuel qui sépare la survenance du sinistre et le début de l'indemnisation. En France, plus de 50% des garanties dépendance présentent une franchise, qui s'élève usuellement à 90 jours.
- service date ou date de service en français : la date de service minimale (respectivement maximale) correspond à la date à partir de laquelle l'assuré commence (respectivement finis) à percevoir ses prestations.
- l'exposition en sinistre correspond à la durée passée en sinistre exposée à l'éventualité d'une fin de sinistre. Plus simplement, l'exposition constitue la durée qui sépare la date de service maximale de la date de service minimale. La franchise n'est pas comptée dans l'exposition.

Pour continuer sur cette dernière notion, voici un tableau des valeurs d'exposition entre hommes et femmes qui mettent en valeur les dissemblances entre les deux genres :

Plusieurs détails sont à mettre en valeur vis-à-vis de la base de fins de sinistres :

- lors de la collecte des données, les sinistres séparés de six mois ou moins d'une même police ont été combinés en un seul sinistre par souci de cohérence avec la base sur les incidences.

Genre	Mois passés en dépendance	Mois passés en dépendance (en % du total)	Nombre de fins de sinistres	Nombre de fins de sinistres (en % du total)
Femmes	3 119 379	69,8	86 660	60,0
Hommes	1 347 469	30,2	57 782	40,0
Total	4 466 848	100	144 442	100

TABLE 2.2 – Exposition en dépendance et nombre de fins de sinistre chez les femmes et chez les hommes. Source : SOCIETY OF ACTUARIES, 2015a

— lors de la récolte de données chez certains participants, certains éléments n'étaient pas disponibles ou bien présentaient une qualité insuffisante. Par exemple, la méthode de distribution des ventes est l'une des variables qui ne figure pas dans la base de fin de sinistres pour cette raison.
— la SOA définit la date d'incidence d'un sinistre individuel comme la première date de service diminuée de la période d'élimination. Cette caractérisation est motivée par la multiplicité des définitions existantes entre les différents assureurs fournisseurs de données.
— parmi les données, seule une poignée de contrats est âgée de plus de 25 ans et présente des durées passées en sinistres supérieures à 4 ans.
— la répartition entre les décès et les retours en autonomie dans le jeu de données paraît raisonnable d'après la SOA.

La base de données est composée de 5 millions d'observations. À titre de comparaison, 2,7 millions de Français sont couverts contre la dépendance auprès de sociétés d'assurance fin 2017, dont 1,6 million pour un contrat avec pour seule garantie la dépendance (source : France Assureurs).

2.2.2 Hypothèses pour la construction des lois américaines et retraitement de la base de données SOA LTC 2000-2011

Plusieurs réflexions entourent les hypothèses qui seront considérées pour la construction des tables de décès.

1er point : segmentation de la table par genre

Il paraît intéressant de séparer l'étude des femmes et des hommes pour deux raisons. La première : les hommes semblent beaucoup plus exposés au phénomène de censure par décès que les femmes. En particulier, cette observation est valable pour la base SOA via la comparaison des deux graphes de la Figure 2.3, en particulier pour les tranches d'âge 60-64, 65-69, 70-74 et 75-79. D'ailleurs, il serait intéressant d'obtenir de telles statistiques dans le cadre français afin de motiver notre démarche. La 2nde raison repose sur les différences de mortalité qui existent entre les hommes et les femmes dans la population générale et qui subsistent dans la population dépendante.

2ème point : différences entre la définition américaine de la dépendance et celle au sens du codage médical

L'absence d'une variable "niveau de perte d'autonomie" est l'un des inconvénients principaux de la base SOA et il est alors impossible d'y distinguer les populations par sévérité en dépendance. Ainsi, seuls les contrats Tax-qualified sont considérés dans notre étude afin de capter uniquement les dépendants incapables de réaliser 2 actes de la vie quotidienne, en déficience fonctionnelle ou en déficience cognitive sévère. Dans une certaine mesure, cette définition se rapproche de la caractérisation de la perte d'autonomie au sens GAD, et donc par extension de celle du codage médical. On fera donc l'hypothèse que nous arrivons à capter approximativement les dépendants américains qui rentrent dans la définition du label GAD.

3ème point : différences entre les lieux de résidence considérés dans la base de données SOA LTC 2000-2011 et les lieux d'hospitalisations compris dans le jeu de données PMSI 2008-2013

En effet, même si les deux sources de données couvrent le périmètre des soins à domicile et en établissement, il n'est pas possible de pointer avec précision le type des établissements dans lesquels se sont déroulés les MCO, les SSR et les PSY. De l'autre côté du spectre, la base de données SOA indique clairement à chaque ligne le lieu de résidence : NH, HHC, ou ALF. Par la suite, aucun retraitement ne sera réalisé sur la variable *ClaimType* de la base SOA.

Création de nouvelles variables et contrôle de la qualité de la base de données SOA LTC 2000-2011

En dehors des retraitements relatifs à la comparabilité des deux bases de données, d'autres changements doivent être effectués :

— Les lignes où le nombre de sinistres est strictement supérieur à 1 doivent être dupliquées et une variable "indicatrice de censure" (voir Section 3.2 du Chapitre 3) doit être créée afin d'adhérer au format des arguments des fonctions de la librairie *survival* de R.

— La suppression des lignes de valeur "Unknown" ne se limite pas à la seule variable *EP_Bucket*. En effet, les lignes qui présentent des valeurs inconnues dans la colonne *TQ Status* seront également éliminées.

Pour le contrôle de la qualité des données, les points suivants sont vérifiés :

— la date de fin de sinistre est comprise dans la période 2000-2011.

— le nombre de morts est inférieur au nombre de sinistres.

— le nombre de fins de sinistres est inférieur au nombre de sinistres.

À l'issue de l'ensemble des ajustements réalisés sur la base américaine, la nouvelle base contient 73 596 lignes.

2.2.3 Base de données SOA LTC 2000-2016

Le jeu de données employé existe dans une version plus neuve et permet de couvrir les années 2000 à 2016 *. Alors qu'il est traditionnellement conseillé d'employer une base de données dans sa version la plus récente, cette dernière n'est pas exploitée pour deux raisons principales. La première réside dans l'un des objectifs du mémoire qui consiste à recourir à une forêt aléatoire de survie. Cependant, il n'est pas possible de construire une indicatrice de censure/décès à partir des données de 2016. Toutefois, le calcul de l'estimateur de Hoem est tout de même réalisable et permettrait de bâtir une loi de survie en dépendance. Second argument : la mouture de 2011 présente davantage de covariables que la version 2016. Parmi les plus importantes, *TQ Status* ne figure pas dans la dernière mise à jour. Par conséquent, il

∗. https ://www.soa.org/resources/experience-studies/2020/2000-2016-ltc-aggregate-database/

en devient impossible de remonter à une quelconque sévérité en dépendance des individus qui figurent dans la base de 2016. Les variables *StateAbbr*, *Cov_Type_Bucket*, *Infl_Rider_Bucket* ne se trouvent également pas dans cette version. Par ailleurs, ces dernières présenteront une influence sur les prédictions bien inférieure à d'autres variables explicatives comme il sera montré dans le Chapitre 3. Ultimement, le nombre de modalités considérées pour *Diagnosis_Category* est réduit de moitié, alors qu'il s'agit vraisemblablement de la variable explicative qui pèse le plus dans les prédictions du Chapitre 3.

2.2.4 Statistiques descriptives sur la base retraitée

Les statistiques sur la perte d'autonomie en France sont assez rares. Il est essentiel de confronter, dans la mesure du possible, les chiffres de la base retraitée à ceux dont nous disposons dans l'Hexagone afin d'identifier les différences qui interviennent entre les deux nations et entre plusieurs sous-populations. Ces points de confrontation sont en rapport avec le genre, la durée passée en perte d'autonomie, l'âge d'incidence en dépendance, le lieu de résidence et la pathologie/le diagnostic. On renvoie vers Dupourqué et al., 2019 pour davantage de statistiques sur la base SOA intégrale.

Genre

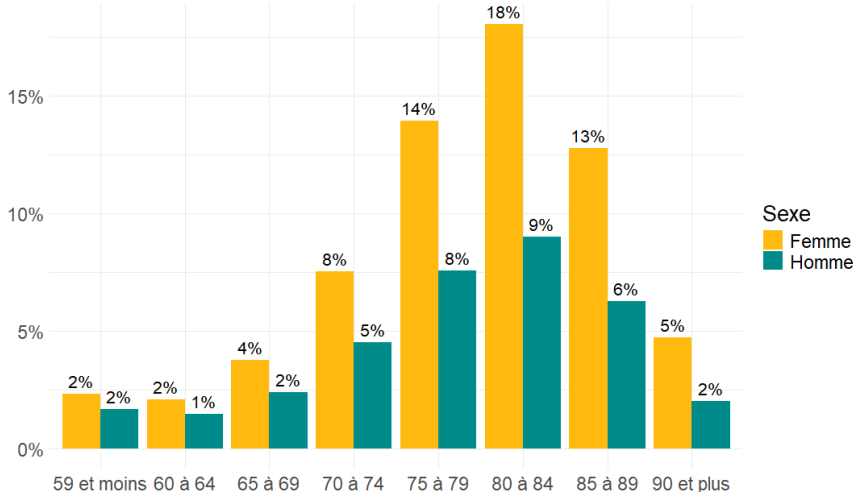


Figure 2.2 – Répartition des genres dans la base de données retraitée SOA LTC 2000-2011 par tranche d'âge d'entrée en dépendance

Les femmes américaines sont davantage touchées par la perte d'autonomie à l'instar des femmes françaises et sont approximativement deux fois plus nombreuses que les hommes aux États-Unis.

Temps passé en autonomie

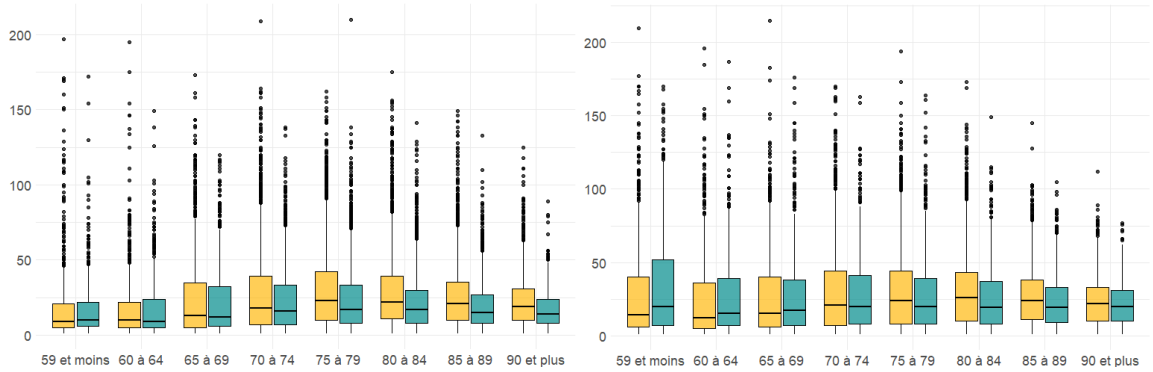


FIGURE 2.3 – Durées passées (en mois) en sinistre par tranche d'âge d'entrée en dépendance dans la base de données retraitée SOA LTC 2000-2011 ; à gauche pour les données uniquement censurées et à droite pour les données non censurées ; en jaune pour les femmes et en bleu pour les hommes

Les hommes sont beaucoup plus censurés comparativement aux femmes. En effet, on peut remarquer que les durées passées en sinistre dépendance chez les hommes connaissent une augmentation plus marquée en passant des observations censurées aux observations non-censurées en comparaison des femmes. Au-delà de la confrontation hommes/femmes, on remarque que pour les deux genres, l'écart entre la médiane et le 1er quartile a tendance à s'allonger en passant des données censurées aux données non-censurées. La médiane ne dépassant pas les deux ans passés en dépendance, cette observation serait associée à une surmortalité aux faibles anciennetés en dépendance ne dépassant pas deux ans.

Lieu de résidence

La répartition des résidents seniors par nature d'établissement de résidence diverge de façon marquée avec celle observée sur toute la France. En effet, 19% des Américains seniors en perte d'autonomie de la base demeurent en ALF contre 29% qui habitent en NH. Ces fractions rompent avec les chiffres français puisque 99% des seniors de France demeurant en institution de soins sont concentrés au sein des EHPAD, des USLD et des résidences autonomies. La discordance entre ces chiffres peut s'expliquer par plusieurs raisons :

— comme décrit dans le Chapitre 1, l'organisation des institutions de soins et la caractérisation de la perte d'autonomie diffère entre les États-Unis et la France.
— comparé à l'Hexagone, le système de soins américain est davantage tourné vers le privé. Le coût consécutif à un séjour dans une nursing home est également bien plus important par rapport à celui d'un ALF ou bien d'un HHC.
— ultimement, certaines différences culturelles et sociales pourraient également intervenir.

D'autre part, d'après l'enquête EHPA DREES, 2015 et les projections démographiques Omphale de l'INSEE, les seniors français en perte d'autonomie résident pour 80% d'entre eux à domicile et cette proportion devrait se maintenir jusqu'à horizon 2070. Ce chiffre dénote avec les 46-47% en home health care dans la base américaine.

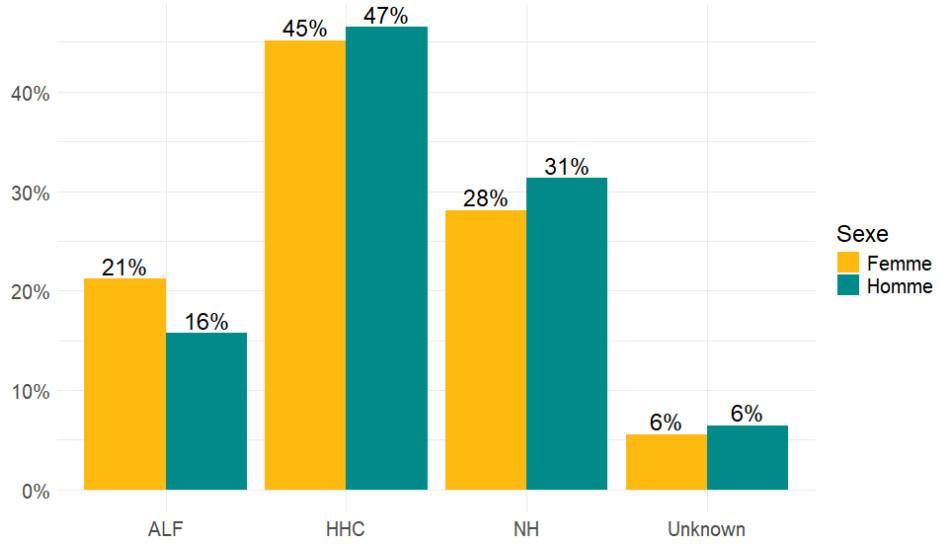


FIGURE 2.4 – Répartition des genres dans la base de données retraitée SOA LTC 2000-2011 par lieu de résidence

Par conséquent, la construction de nos tables de décès se heurte à l'équivalent d'un effet d'antisélection puisque les lois estimées dans le Chapitre 3 sont appliquées à un portefeuille qui se veut français. L'adaptation des tables de mortalité au cadre français se révèle ainsi primordiale afin de pallier cette singularité.

En outre, la modalité *Unknown* est notamment associée à une durée passée en perte d'autonomie beaucoup plus faible comparativement aux individus dont le lieu de résidence est connu comme montré en Annexe A.1 Cette sous-population pourrait donc présenter des caractéristiques bien particulières. Le choix sera fait de conserver ces valeurs manquantes comme une modalité à part entière.

Diagnostic médical

En France, la répartition des pathologies/diagnostics responsables de la perte d'autonomie est assez éparse, mais est tout de même marquée par la prépondérance des syndromes de démence comme la maladie d'Alzheimer suivie généralement des cancers, des maladies cardiovasculaires et des autres troubles neuro-psychiatriques. Ces trois dernières pathologies sont usuellement la conséquence d'au moins un dixième chacune des incidences totales en dépendance. Néanmoins, il est assez difficile de dénicher des chiffres précis sur la dispersion des pathologies au sein de la population dépendante française. Par exemple, la DREES a supprimé le volet relatif aux pathologies et à la morbidité de l'enquête EHPA car la récupération de ces informations était trop demandeuse pour les établissements de soins. Afin d'identifier des ordres de grandeur, on peut tout de même se pencher sur l'enquête CARE de la DREES, 2016 qui présente, entre autres, la part des seniors de plus de 65 ans résidant en institution (maisons de retraite, EHPAD ou USLD) et déclarant des maladies ou des problèmes de santé chroniques. Comparativement aux trois pathologies les plus représentées dans la base SOA (voir Figure 2.5), la répartition de ces mêmes pathologies dans l'enquête CARE 2016 est la suivante :

— maladie d'Alzheimer ou autre démence : 29%
— maladie cardiovasculaire : 37% (dont 8% d'accidents vasculaires cérébraux parmi ces 37%)

— Cancer : 5%

Ces trois pathologies seront mises en évidence sur les deux figures suivantes. Par ailleurs, les chiffres sont à remettre en question : les individus considérés pouvant se trouver dans un état polypathologique et, comme énoncé plus haut, les seniors en perte d'autonomie ne sont que 20% à vivre chez eux. En outre, l'apparition de ces trois troubles n'est pas décorrélée et présente certains facteurs de risque communs.

Dans le jeu de données SOA, notons que les lignes qui présentent le diagnostic de maladie cardiovasculaire rassemblent toutes les maladies cardiovasculaires en dehors de l'accident vasculaire cérébral et de l'hypertension puisque ces derniers forment des modalités distinctes. De plus, la liste des troubles responsables de la perte d'autonomie des sinistrés du jeu de données de la SOA comporte 18 modalités différentes, listées sur les Figures 2.5 et 2.6. La dépendance d'ordre cognitive couvre le trouble d'Alzheimer, les troubles psychiques, l'accident vasculaire cérébral et les pathologies du système nerveux et des organes sensoriels, tandis que les autres pathologies sont initiatrices d'une perte d'autonomie physique. Cette distinction est mise en évidence sur les Figures 2.5 et 2.6. D'après Dupourqué et al., 2019, la durée moyenne de l'état de dépendance serait de 15 mois si la perte d'autonomie est d'ordre physique et passe à 33 mois si elle est d'ordre cognitive.

Comme montré en Figure 2.5 et à l'instar de la circonférence dépendance de France, la pathologie cognitive d'Alzheimer est le trouble responsable du plus d'incidences en perte d'autonomie, à hauteur de 30%, dans le jeu de données. Les cancers et les accidents vasculaires cérébraux sont les causes qui arrivent juste derrière. L'équivalent de la Figure 2.5 distinctement pour les résidents à domicile et pour les résidents en établissement est disponible en Annexe A.2. En établissement, la maladie d'Alzheimer y est encore davantage représentée à hauteur de 35% et les patients sont également plus touchés par les accidents vasculaires cérébraux.

Les proportions de seniors atteints de la démence d'Alzheimer et de cancer sont très proches entre la base SOA et l'enquête DREES, 2016. Par ailleurs, les personnes âgées souffrant plus généralement de maladies cardiovasculaires, accidents cardiovasculaires et hypertensions compris, ne sont pas identiquement représentées des deux côtés et il en est de même pour les conclusions médicales d'un cancer dans une moindre mesure. En dehors des trois pathologies mises en valeur plus haut, il est plus difficile de comparer la dispersion des autres diagnostics qui apparaissent dans le jeu de données SOA car la typologie des maladies diffère de celle qui est employée dans l'enquête DREES, 2016. En outre, ces pathologies sont nettement moins représentées dans la base américaine. Les boxplots de la Figure 2.5 indiquent clairement que la durée passée en sinistre dépendance diffère en fonction du diagnostic médical de l'assuré. La pathologie d'Alzheimer est associée à la durée passée en dépendance la plus importante en médiane, suivie de l'accident vasculaire cérébral. Le cancer arrive en dernière position. Les troubles associés à la perte d'autonomie d'ordre cognitive sont bien ceux responsables des plus longs temps passés en dépendance.

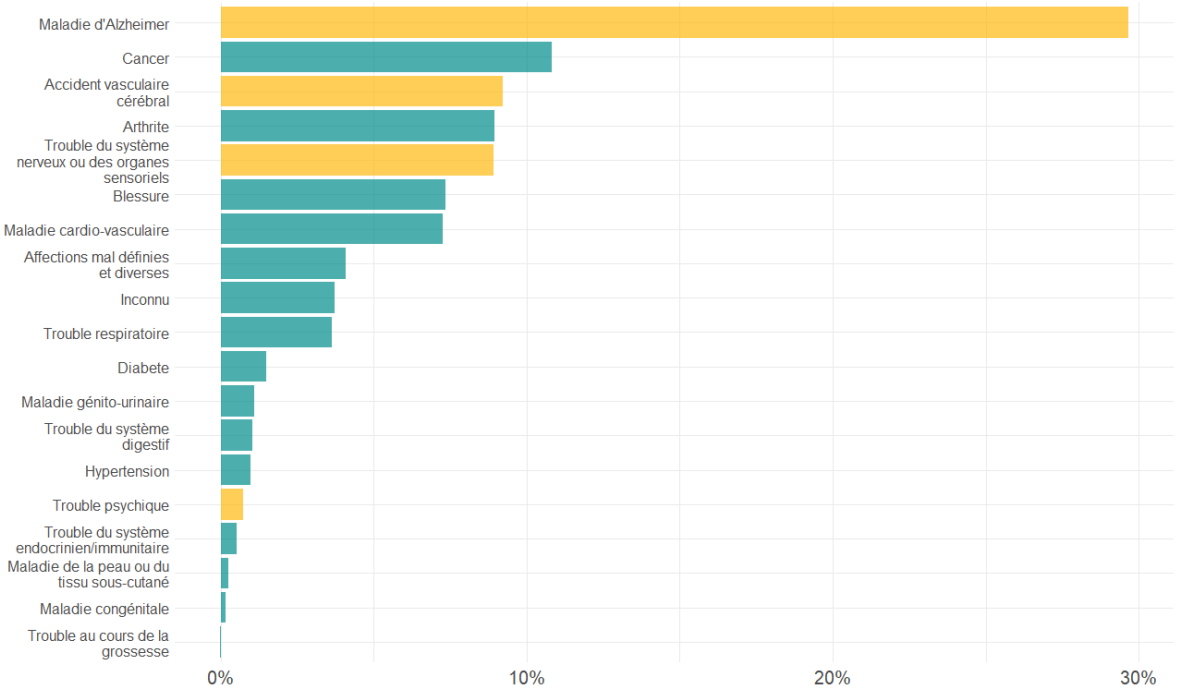


FIGURE 2.5 – Répartition des pathologies dans la base de données retraitée SOA LTC 2000-2011. Les diagnostics relatifs à la perte d'autonomie d'ordre cognitive sont mis en évidence en jaune.

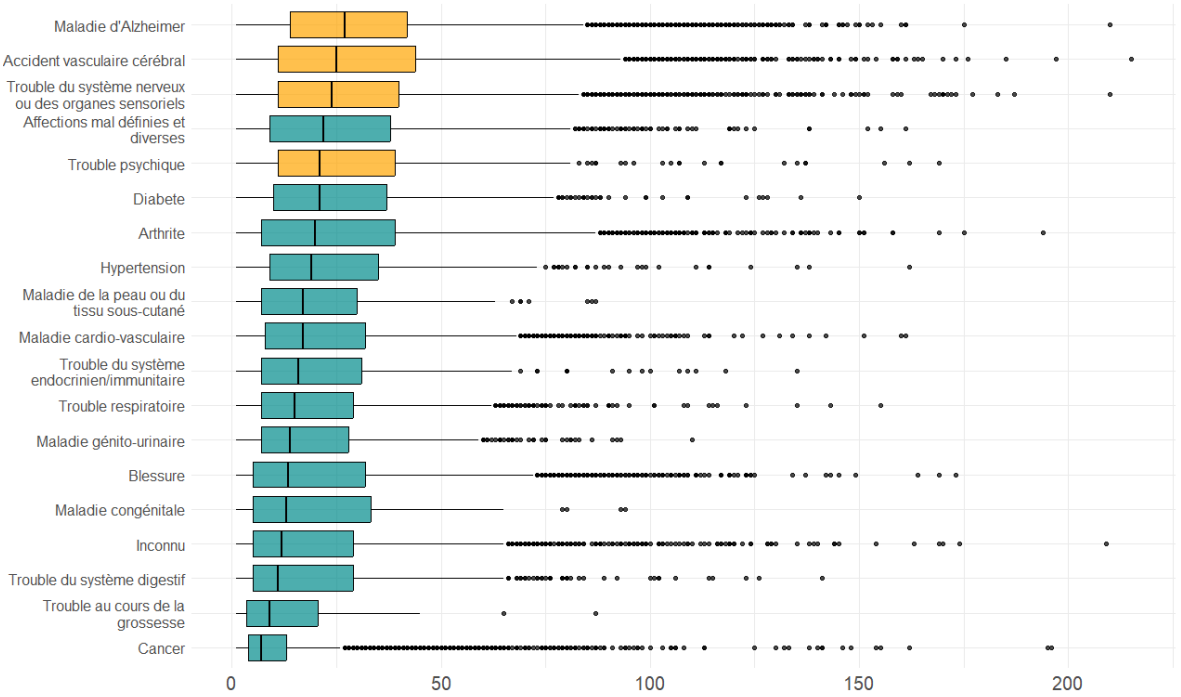


FIGURE 2.6 – Durées en mois passées en sinistre en fonction de la pathologie dans la base de données retraitée SOA LTC 2000-2011. Les boxplots sont rangés par durée médiane croissante de bas en haut. Les diagnostics relatifs à la perte d'autonomie d'ordre cognitive sont mis en évidence en jaune.

Diagnostic	Montant payé (en millions de $)	Pourcentage du total (%)
Maladie d'Alzheimer	1 156	25
Trouble psychique	635	14
Cancer	543	12
Maladie cardiovasculaire	361	8

TABLE 2.3 – Répartition du montant de sinistres dépendance payés en fonction du diagnostic de la maladie responsable de l'entrée en perte d'autonomie d'après les données SOA 2011. Source : DUPOURQUÉ et al., 2019

D'après la Table 2.3, les sinistrés entrés en dépendance en raison de la maladie d'Alzheimer, d'un trouble psychique ou d'un accident vasculaire cérébral capturent 51% du montant total de sinistres payés. En comparaison des Figures 2.5 et 2.6, il n'est pas étonnant de retrouver la maladie d'Alzheimer et l'AVC en haut du classement. Par ailleurs, les troubles psychiques semblent être l'un des diagnostics qui revient le moins dans les données tandis que les dépendants atteints de cancer sortent rapidement de l'état de dépendance, ce qui supposerait un haut montant de sinistre moyen pour la prise en charge de ces deux maladies. C'est effectivement le cas pour les cancers. D'après LE FIGARO, 2022, les dépenses américaines pour le traitement de ces derniers s'élèveraient à 200 milliards par an et un tel coût pourrait être porté par l'explosion du coût des médicaments.

Chapitre 3

Estimation de la loi américaine de mortalité en dépendance

Cette étude rentre dans le cadre de l'analyse de survie et nécessite la mise en place de modèles de durée. Dans ce cadre, T, une variable aléatoire, représente la durée qui sépare l'entrée en dépendance du décès et constituera notre objet d'étude. Nombre d'outils mathématiques gravitent autour de cette grandeur et permettent de la caractériser, dont notamment la fonction survie et la fonction de hasard. Afin d'estimer ces quantités, une forêt aléatoire de survie sera employée et ses performances seront comparées à celles d'un estimateur de Kaplan-Meier et d'un arbre de survie. Utiliser cette multiplicité d'estimateurs permet de couvrir à la fois des objectifs de prédictions et d'interprétation. Il sera ainsi intéressant d'examiner la pertinence des outils de machine learning pour cette tâche.

3.1 Quantités d'intérêt : fonction de survie et fonction de hasard

La fonction de survie et la fonction de hasard sont des outils incontournables de l'analyse de survie et sont notées respectivement S et h dans la suite. Généralement, les estimateurs de ces différentes quantités sont signifiés par l'ajout de l'adjectif « empirique ». Dans la suite et par abus de langage, la qualification « empirique » sera omise étant donné que seuls des estimateurs seront manipulés. Au sein de la discipline actuarielle, la fonction de hasard peut également prendre le nom de hasard, de taux de mortalité instantanée ou bien de force de mortalité. La fonction de survie et la fonction de hasard sont définies comme suit :

$$S(t) = \mathbb{P}(T > t) \quad ; \quad h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = -\log'(S(t))$$

La fonction de hasard cumulative H sera également utilisée dans la section 3.5. Elle se définit comme :

$$H(t) = \begin{cases} \sum_{t_i \leq t} h(t_i) & \text{dans le cas discret} \\ \int_0^t h(s)\,\mathrm{d}s = -\log(S)) & \text{dans le cas continu} \end{cases}$$

où $(t_i)_{i=1,\dots,n}$ représente les temps de décès réellement observés dans les données.

3.2 Phénomènes de censure

Les phénomènes de censure sont usuellement rencontrés dans les modèles de durée en assurance, et plus particulièrement ceux de censure à droite.

Pour définir la censure (définition reprise de Biessy, 2024), prenons l'exemple d'une étude de mortalité. Soient T_γ et C_γ les durées de vie associées respectivement aux processus de vie et de mort Z^T et Z^C dépendant de covariables notés γ. T_γ est dite censurée à gauche par C_γ lorsque la variable $Y_\gamma = max(T_\gamma, C_\gamma)$ est observée à la place de T_γ. Dans le cas de la censure à droite, c'est $Y_\gamma = min(T_\gamma, C_\gamma)$. C_γ est également appelée variable censurante. Afin d'indiquer la présence ou non de censure, on définit l'indicatrice de censure comme $\delta_\gamma = \mathbb{1}_{T_\gamma < C_\gamma}$ et $\delta_\gamma = \mathbb{1}_{T_\gamma \geq C_\gamma}$ pour respectivement la censure à gauche et la censure à droite.

3.3 Estimateur de Kaplan-Meier (KM)

L'estimateur de Kaplan-Meier $\hat{S}^{KM}$, introduit par Kaplan et Meier, 1958, est un estimateur de la fonction de survie de T. La suite aléatoire des observations des événements de décès sera notée $(T_i)_{i=1,\dots,n}, n \in \mathbb{N}$. Par ailleurs, les bases de données sujettes aux études d'analyse de survie présentent des contraintes de censure. Les épisodes de censure à droite sont en particulier pris en compte par l'estimateur de Kaplan-Meier. Par conséquent, les T_i ne sont pas obligatoirement observés et, à la place, la suite des observations censurées $(Y_i)_{i=1,\dots,n}$ et des indicatrices de censure $(\delta_i)_{i=1,\dots,n}$ sont présentes dans le jeu de données. Plus explicitement pour δ_i :

$$\delta_i = \begin{cases} 1 & \text{si le décès est observé} \\ 0 & \text{si le décès est censuré} \end{cases}$$

En notant $(Y(i))_{i=1,\dots,n}$ la suite des statistiques d'ordres de $(Y_i)_{i=1,\dots,n}$, l'estimateur naturel de la probabilité de décès $\hat{q_i}$ entre $[Y(i), Y(i+1)]$ est défini par :

$$\hat{q_i} = \frac{d_i}{E_i}$$

avec

— d_i le nombre de décès observés au temps $Y(i)$

— E_i le nombre d'individus exposés au risque de décès juste avant $Y(i)$

Dans ce cadre, l'estimateur de Kaplan-Meier, évalué en $t \in [Y(i); Y(i+1)]$, admet la forme suivante :

$$\hat{S}^{KM}(t) = \prod_{j=1}^{i}(1 - \hat{q_j}) = \prod_{j=1}^{i}(1 - \frac{d_j}{E_j})$$

3.4 Arbre CART (Classification And Regression Tree) et arbre de survie

Pour commencer, les définitions d'arbre CART et d'arbre de survie (ce dernier sera abrégé ST) sont présentées avant d'enchaîner sur les forêts aléatoires ainsi que leur adaptation à l'analyse de survie dans la Section 3.5.

3.4.1 Arbre CART

Breiman, 1984 initie les modèles d'arbre de décision. L'arbre CART est un arbre de décision défini comme un ensemble de règles de coupure (ou découpe) qui s'apparente à des questions à choix binaire sur les covariables. L'algorithme aboutit à un partitionnement du jeu de données inséré en entrée. Un arbre de décision est composé des éléments suivants :

- nœud de décision : nœud qui consiste en une question à choix multiple (2 choix dans le cadre des arbres CART) et qui admet plusieurs embranchements relatifs aux réponses. Les données correspondant au nœud sont partitionnées en fonction de la réponse et chaque embranchement devient ensuite un nouveau nœud, exception faite pour les nœuds feuilles. Exemple : soient Y la variable aléatoire réponse et $\mathbf{X}$ le vecteur de variables explicatives qui composent un jeu de donnée. La règle de décision est de la forme $\{X_i < c \in \mathbb{R}\}$ (resp. $\{X_i \in C\}$ avec C un ensemble de modalités) si X_i est de nature quantitative (resp. qualitative).
- nœud racine : premier nœud de l'arbre de décision.
- feuille ou nœud feuille : nœud terminal qui ne donne pas lieu à de nouveaux embranchements.
- branche : sous-arbre qui débute depuis l'un des nœuds de décision de l'arbre complet.
- arbre maximal : arbre de décision dont l'ensemble des feuilles correspond à la partition des covariables la plus fine possible.

La création de groupes homogènes, représentés par les nœuds feuilles, constitue l'objectif de l'arbre CART, et de l'arbre de décision plus généralement. Pour ce faire, on définit un indicateur quantitatif de précision, nommé pureté, qui permet d'optimiser la discrimination engendrée par chaque nœud. Une pureté est affectée à chaque nœud. Plus ce dernier est pur, mieux il sépare les embranchements qui prennent la suite de ce nœud. Les mesures d'impureté classiques sont l'indice de Gini et l'entropie pour la classification tandis que l'erreur quadratique moyenne (EQM) et l'erreur absolue (EA) font référence pour de la prédiction.

Une fois l'arborescence construite grâce aux mesures d'impureté, l'arbre doit passer par une phase d'élagage qui consiste à supprimer certaines de ses branches. En effet, l'arbre optimisé repose sur un compromis biais-variance : l'arbre maximal admet un faible biais et une variance importante et, à contrario, l'arbre qui se résume au nœud racine est marqué par un biais important et une variance faible. L'arbre final sera composé idéalement d'un nombre de nœuds approprié qui échappe au sur-apprentissage.

Les arbres de décision présentent plusieurs avantages en termes d'interprétabilité et de stabilité (voir Genuer et Poggi, 2017 pour plus de détails) :

- découpes compétitives : il s'agit de la suite ordonnée des variables explicatives par réduction décroissante de l'hétérogénéité de toutes les découpes d'un nœud. Les découpes compétitives ne comprennent pas la « meilleure » coupure. Ces coupures sont moins

optimales par rapport à la « meilleure ». Les découpes compétitives peuvent être privilégiées à cette dernière et permettent de classer les règles de partitionnement par importance.

— valeur manquante : dans le cas d'une prédiction pour laquelle une covariable manque et apparaît dans les coupures de l'arbre, la 2nd découpe compétitive consécutive ou les découpes de substitution (voir Breiman, 1984 pour davantage de détails sur les découpes de substitution) pourront être employées à la place de la « meilleure ».

— interprétabilité : un indice d'importance des variables peut être défini (voir sous-section 3.6.2) comme la réduction d'hétérogénéité à chaque découpe.

— valeur aberrante : les valeurs aberrantes influencent essentiellement la feuille dans laquelle elle finit. L'instabilité de l'arbre est également une mesure du poids des données aberrantes et/ou influentes.

— complexité algorithmique : elle est quadratique dans le pire des cas.

3.4.2 Arbre de survie

L'arbre de survie est un modèle de survie. Il s'agit d'un arbre CART dont la règle de découpe repose sur le critère log-rank. Par la suite, on détaille cette mesure d'impureté ainsi que le calcul des différentes quantités de survie, repris de Lu, 2023. Les notations T, $(T_i)_{i=1,...,n}$, $(Y_i)_{i=1,...,n}$ et $(\delta_i)_{i=1,...,n}$, introduites dans la sous-section 3.3, sont toujours considérées afin d'établir le cadre de l'arbre de la forêt de survie aléatoire. Soit $\{t_1, ..., t_m\}$ la subdivision des temps de décès observés. Sans perte de généralité, on considérera uniquement des covariables $\mathbf{X}$ quantitatives pour définir les quantités qui vont suivre. Les nœuds filles sont notés par la suite $L = \{X_i \leq c\}$ et $R = \{X_i > c\}$ avec c un réel. Ces notations permettent d'introduire les quantités $d_{j,k}$ et $W_{j,k}$ qui correspondent respectivement au nombre de décès et au nombre d'individus à risque au temps t_j pour la branche $k = L, R$. On note également $W_j = W_{j,L} + W_{j,R}$ et $d_j = d_{j,L} + d_{j,R}$. À partir de ces quantités, la statistique de coupure log-rank s'écrit :

$$L(X, c) = \frac{\sum_{j=1}^{m}(d_{j,L} - W_{j,L}\frac{d_j}{W_j})}{\sqrt{\sum_{j=1}^{m}\frac{W_{j,L}}{W_j}(1 - \frac{W_{j,L}}{W_j})(\frac{W_j - d_j}{W_j - 1})d_j}}$$

Plus $|L(X, c)|$ est grand, plus la différence de survie est importante entre la branche gauche et la branche droite. On cherche donc à maximiser $|L(X, c)|$.

3.4.3 Quantités de survie résultantes de l'arbre de survie

Soient $\{t_{1,h}, ..., t_{m(h),h}\}$ la subdivision des temps de décès observés pour un nœud fille h. La fonction de hasard cumulative et la fonction de survie sont définies à partir de l'estimateur de Nelson-Aalen et de l'estimateur de Kaplan-Meier :

$$H_h(t) = \sum_{t_{j,h} \leq t} \frac{d_{j,h}}{W_{j,h}} \quad ; \quad S_h(t) = \prod_{t_{j,h} \leq t} (1 - \frac{d_{j,h}}{Y_{j,h}})$$

3.5 Forêt de survie aléatoire

3.5.1 Forêt aléatoire « classique » et forêt aléatoire de survie (RSF)

Forêt aléatoire

Breiman, 2001 est l'article genèse de la forêt aléatoire. Une forêt aléatoire est une méthode d'apprentissage non-paramétrique. Elle consiste à agréger les résultats de plusieurs arbres de décision non élagués et calibrés sur des échantillons bootstrap du jeu de données initial. La forêt aléatoire de survie, introduite par Ishwaran et al., 2008, reprend les grandes lignes de la version classique d'une forêt aléatoire en employant des arbres de survie. Comme décrit par Huchon, 2021, la construction se décompose en 3 grandes étapes :

1. création de $B \in \mathbb{N}$ échantillons bootstrap. Les échantillons bootstrap non inclus sont qualifiés de out-of-bag (OOB).
2. application d'un type d'arbres décisionnels correspondant au type de données considérées pour chaque échantillon bootstrap $b = 1, ..., B$.
3. agrégation de l'information des nœuds terminaux pour obtenir le prédicteur final.

La forêt aléatoire présente des avantages et des inconvénients par rapport à l'arbre de décision.

— inconvénients : la forêt aléatoire perd les propriétés relatives aux valeurs aberrantes.
— avantages : stabilité, interprétabilité (la notion d'importance des variables par permutation remplace l'importance des variables « simple »).

Forêt aléatoire de survie et quantités de survie résultantes

Dans le cadre des forêts aléatoires de survie, les estimateurs $H_h(t)$ et $S_h(t)$ relatifs à chaque arbre existent sous deux versions différentes : in-bag (IB) et out-of-bag (OOB) qui correspondent aux estimateurs évalués sur les données respectivement inclus et exclus dans les échantillons bootstrappés. À variables explicatives fixées, les estimateurs de la fonction hasard cumulative $\bar{H}$ et la fonction de survie $\bar{S}$ pour la forêt correspondent simplement à la moyenne sur tous les arbres de ces quantités, qu'ils soient en version IB ou OOB.

Remarque : il est important de prédire la fonction de survie et la forêt aléatoire distinctement car ces dernières ne caractérisent pas la même loi de probabilité. Dans le cadre usuel, $H(t) = -\log(S(t))$. Par ailleurs, la concavité du log et l'inégalité de Jensen indiquent que $-\log(\bar{S}) \geq \bar{H}$. Autrement dit, caractériser l'estimation de la loi par la fonction de survie revient à sous-estimer la loi de décès comparativement à l'emploi de la fonction de hasard cumulative. Dans la suite, la « première » version (abrégée v1 ou RSF1) de la loi de décès en dépendance sera caractérisée par l'estimation initiale de la fonction de survie, tandis que la « seconde » version (abrégée v2 ou RSF2) sera particularisée par l'estimation initiale de la fonction de hasard. À titre d'exemple, la fonction de survie qui est issue de la fonction de hasard initialement estimée correspond à la seconde version de la loi de probabilité.

3.6 Construction et diagnostic de l'arbre de survie et de la forêt aléatoire de survie

3.6.1 Diagnostic de l'arbre de survie par validation croisée

La construction de l'arbre de survie a été réalisée sur Python, contrairement à l'estimateur de Kaplan-Meier et de la forêt aléatoire de survie qui sont issus de codes R. La règle de découpe log-rank est implémentée dans le package *sksurv.tree*. À la place d'employer une procédure d'élagage, la construction de l'arbre sera optimisée par une validation croisée K-fold (voir Berrar et al., 2019 pour plus de détails). Les hyperparamètres, qui seront testés entre les valeurs 2 et 12, sont :

— la profondeur maximale, définit comme le nombre maximal de couches de l'arbre
— le nombre minimal d'échantillons requis afin de séparer un nœud
— le nombre minimal d'échantillons dans une feuille

Par la suite, les métriques employées pour la validation croisée sont détaillées brièvement et on renvoie vers la bibliographie relative pour davantage de détails à leur propos :

— le C-index pour les données censurées à droite basé sur la probabilité inverse des poids de censure (et sera noté par la suite C-index$_{IPCW}$). Cette métrique est adaptée aux données censurées à droite et repose, comme son nom l'indique, sur les probabilités inverses des poids de censure (voir Uno et al., 2011).
— l'estimateur de l'AUC cumulative/dynamique (C/D AUC) pour des données censurées à droite (et sera noté par la suite $\widehat{AUC}(t)$). On renvoie vers Uno et al., 2007, Hung et Chiang, 2010 et Lambert et Chevret, 2016.

Les résultats de la cross-validation sont les suivants. Le C-index et et l'AUC s'accordent tous les deux et assurent que la valeur 2 semble adéquate pour tous les hyperparamètres. L'arbre résultant est affiché sur la Figure 3.1.

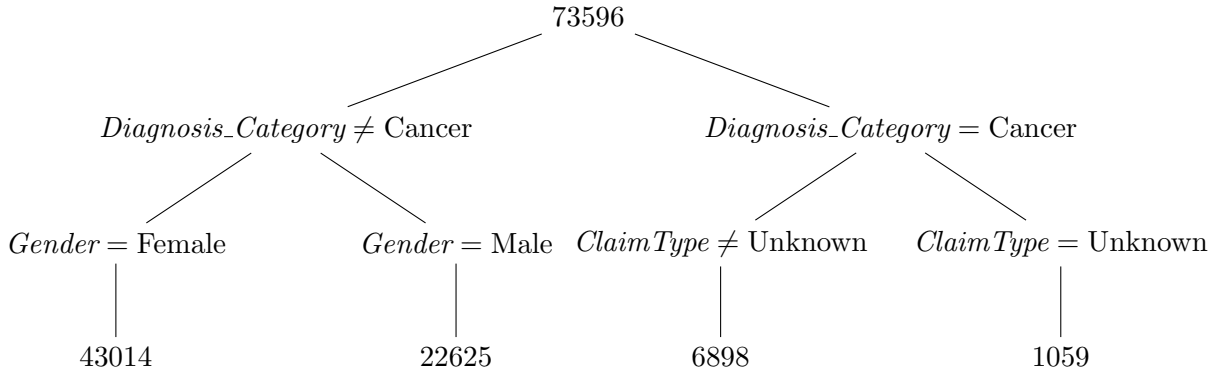


Figure 3.1 – Arbre de survie construit. Les quantités qui figurent au niveau du nœud racine et des nœuds feuilles correspondent au nombre d'individus compris dans chacun de ces nœuds.

L'arbre construit est à deux étages et scinde la base de données selon les variables *Diagnosis_Category*, *Gender* et *ClaimType*. Le cancer serait associé à un temps passé en perte d'autonomie notablement inférieur à l'ensemble des autres pathologies, en toute cohérence avec les Figures 2.5 et 2.6. Les échantillons pour lesquels *ClaimType* est inconnu sont également associés à la durée en dépendance la plus faible par rapport aux autres lieux de résidence. L'arbre est donc organisé par mortalité croissante de gauche à droite. Notons que l'analyse

de l'intelligibilité globale d'un arbre de décision à partir des découpes de l'arbre et de leur ordre peut se révéler biaisée. Il est nécessaire de se tourner vers des métriques de diagnostics plus élaborées comme ce sera le cas dans la sous-section 3.6.2 dédiée aux forêts aléatoires de survie.

L'âge d'incidence en dépendance n'apparaît à aucun moment comme règle de décision dans l'arbre ce qui signifie que, à caractéristiques identiques, des seniors entrées en dépendance à des âges différents, présenteraient la même loi de mortalité vis-à-vis de l'arbre de survie. Plus globalement, l'arbre paraît assez réducteur dans sa discrimination. Par exemple, le groupe des femmes non diagnostiquées d'un cancer représente plus de la moitié de toute la base de données. Il pourrait être envisageable d'augmenter la profondeur de l'arbre afin de scinder davantage le jeu de données, malgré les résultats de la validation croisée. Par la suite, l'arbre de survie ne sera pas ajusté à la population française à l'inverse des estimateurs de Kaplan-Meier et de la forêt aléatoire de survie.

3.6.2 Diagnostic de la forêt aléatoire de survie

Les estimations de nos lois de survie doivent uniquement dépendre des variables de genre et d'âge d'entrée en perte d'autonomie. Étant donné que la forêt aléatoire de survie est construite sur un ensemble de variables explicatives qui ne se réduit pas au genre et à l'âge d'entrée en perte d'autonomie, il est nécessaire d'adapter nos estimations. Pour ce faire, toutes les covariables différentes des deux citées précédemment seront considérées comme manquantes lors de la phase d'estimation. À la place, des valeurs seront imputées selon la méthode de Ishwaran et al., 2008. La procédure n'est pas détaillée par la suite, mais deux de ses particularités sont à mettre en avant :

— d'une part, les valeurs sont imputées de façon aléatoire, sans pour autant employer une loi uniforme. Ainsi, une couche d'aléa se rajoute aux estimations.
— d'autre part et point positif, cette méthode est une alternative intéressante à l'utilisation des découpes de substitution et permet de réduire drastiquement le temps d'exécution de la gestion des valeurs manquantes.

La forêt aléatoire est construite avec 500 arbres. Le diagnostic des ajustements sera réalisé à partir de différentes métriques parmi lesquelles l'importance des variables, les valeurs de Shapley ou encore le score de Brier (voir sous-section 3.7.1 pour la définition du score de Brier).

Erreur Out-Of-Bag (EOOB)

Les définitions de l'EOOB sont reprises de l'article Genuer et Poggi, 2017 vers lequel on renvoie pour davantage de détails. L'erreur Out-Of-Bag, à comprendre comme l'erreur sur les données en dehors des échantillons bootstrap, est une métrique usuellement regardée pour les forêts aléatoires « classiques » ainsi que pour les forêts aléatoires de survie. Son calcul suit les étapes suivantes :

1. pour chaque observation de la base de données d'apprentissage, une prédiction est calculée à partir des arbres de décision qui ont été construits sur un échantillon bootstrap qui ne contient pas cette observation.
2. l'erreur OOB est ensuite calculée à partir de ces prédictions en employant des métriques d'erreur usuelles, l'erreur quadratique moyenne par exemple

D'après Ishwaran et Lu, 2018, l'expression de l'erreur OOB de prédiction, ou encore

appelé erreur de prédiction (PE) ou taux d'erreur, s'obtient simplement à partir de la formule suivante dans le cas des forêts de survie aléatoires :

$$\text{PE} = 1 - \text{C-index}$$

On renvoie vers la sous-section 3.7.1 à propos de la définition du C-index. Une erreur de prédiction égale à $0,5$ signifie que la prédiction n'est pas meilleure qu'une prédiction purement aléatoire. D'un autre côté, un taux d'erreur nul permet de qualifier la prédiction de parfaite.

L'erreur de prédiction, présentée en Figure 3.2, est inférieure à $0,5$, décroît avec le nombre d'arbres et semble admettre une limite pour un grand nombre d'arbres. La décroissance de l'erreur de prédiction, à hauteur de 2%, est assez modeste.

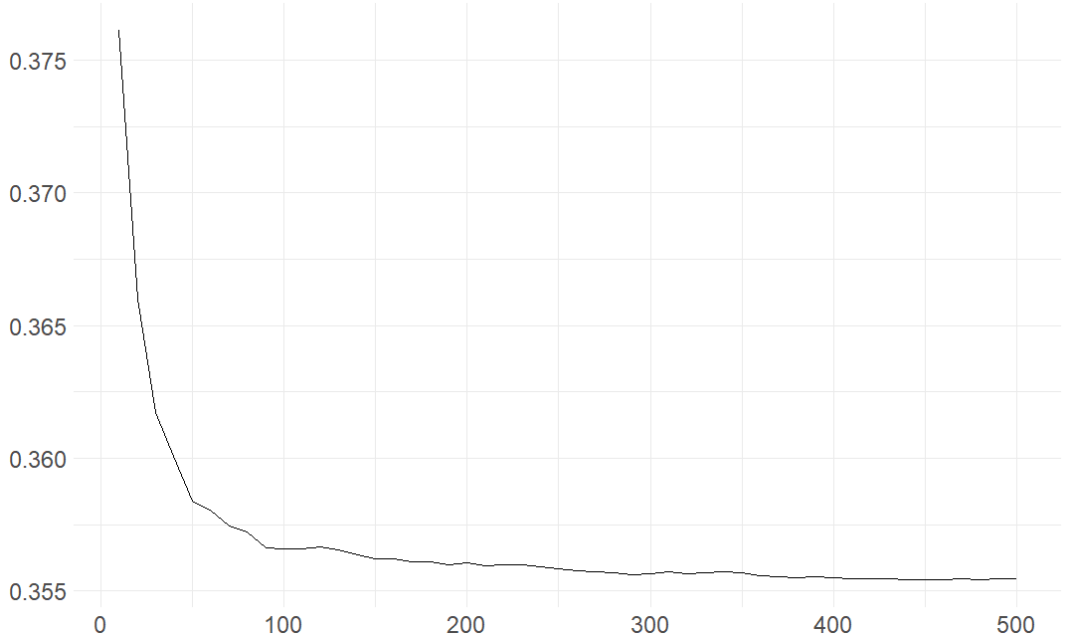


FIGURE 3.2 – Erreur de prédiction OOB en fonction du nombre d'arbres compris dans la forêt aléatoire de survie

Importance des variables

L'importance d'une variable explicative se comprend comme l'effet de cette variable sur la prédiction (voir GENUER et POGGI, 2017) et sa mesure permet de sélectionner les covariables à intégrer dans la construction de la forêt. Plusieurs métriques d'importance des variables explicatives existent, dont l'importance par permutation (VIMP) de BREIMAN, 2002, la profondeur minimale (ISHWARAN et al., 2010 et ISHWARAN et al., 2011) et les valeurs de Shapley. Cette dernière sera introduite plus en détails par la suite. En général, ces métries sont accompagnées du calcul d'un seuil au-dessus duquel les variables sont sélectionnées.

L'importance par permutation repose sur l'erreur de prédiction et se positionne comme une alternative computationnellement moins coûteuse par rapport à la validation croisée. Une importance positive et élevée traduit un pouvoir prédictif élevé pour la variable en question. À contrario, une covariable qui présente une importance par permutation négative est considérée comme une variable qui « bruite » les prédictions. Pour ce qui est de la profondeur minimale,

celle-ci fait l'hypothèse que les variables à impact fort sur la prédiction sont celles qui divisent les nœuds près de la racine le plus fréquemment.

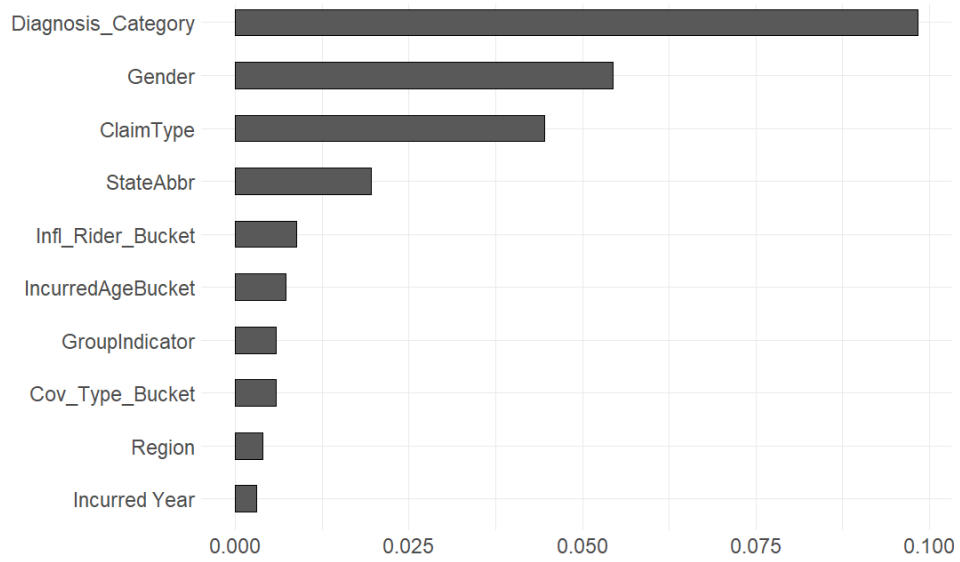


FIGURE 3.3 – VIMP de toutes les variables explicatives

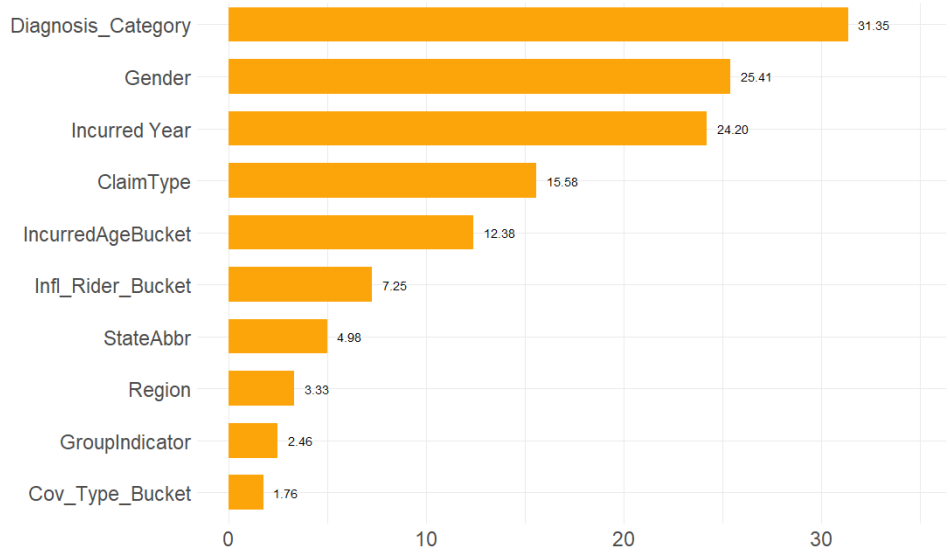


FIGURE 3.4 – Moyenne des valeurs de Shapley en valeur absolue pour toutes les variables explicatives

D'après les Figures 3.3 et 3.4, *Diagnosis_Category* arrive en première place suivie de *Gender*. Toutes les covariables présentent également une VIMP positive. Le tableau 3.1 oppose les différents classements d'importance des variables en considérant la VIMP comme référence arbitraire pour ordonner la disposition des covariables. Le triptyque ne s'accorde pas parfaitement et en particulier sur les covariables *StateAbbr*, *GroupIndicator* et *IncurredYear*. Les trois variables qui semblent peser le plus dans la construction de la forêt aléatoire sont *Diagnosis_Category*, *Gender* et *ClaimType* ce qui paraît correspondre avec les différentes statistiques présentées dans les Chapitre 1 et 2. Quant à la variable *IncurredAgeBucket*, cette dernière ne figure pas parmi les plus capitale alors qu'elle est généralement déterminante sur les tables

Variables	VIMP	Profondeur minimale	Valeurs de Shapley
Diagnosis_Category	1	3	1
Gender	2	1	2
ClaimType	3	4	4
StateAbbr	4	7	7
Infl_Rider_Bucket	5	6	6
IncurredAgeBucket	6	5	5
GroupIndicator	7	8	9
Cov_Type_Bucket	8	10	10
Region	9	9	8
Incurred Year	10	2	3

TABLE 3.1 – Comparaison de la hiérarchisation de l'importance des variables déterminée par la VIMP, la profondeur minimale et les valeurs de Shapley

françaises de mortalité en dépendance.

De plus, le calcul du seuil relatif à la profondeur minimale informe que toutes les variables explicatives ont un impact sur la prédiction au sens de la profondeur minimale (voir Annexe A.3). Par conséquent, toutes les variables explicatives sont conservées pour la construction de la forêt. Une analyse plus fine pourra être réalisée en utilisant la statistique Z de FLEMING et HARRINGTON, 2013.

Une ultime version étoffée de l'importance des variables consiste à se pencher sur une mouture qui dépend du temps. Pour ce faire, la VIMP avec comme fonction de perte le score de Brier dépendant du temps (voir BREIMAN, 2001 et FISHER et al., 2019) est une métrique qui s'y prête bien et est affichée en Figure 3.5. L'allure de la VIMP *Diagnosis_Category* s'y démarque particulièrement. Au-delà d'être la variable la plus importante pour la construction de la forêt de survie, *Diagnosis_Category* connaît une décroissance particulièrement pentue à partir de 30 mois d'ancienneté. Il est intéressant de noter que l'explosion de la VIMP de *Diagnosis_Category* à faible ancienneté semble portée par les dépendants cognitifs d'une perte d'autonomie physique. En effet, l'Annexe A.4 présente l'équivalent de la Figure 3.5 mais sur les dépendants cognitifs et physiques séparément et le met en évidence. Cette constatation rejoint totalement la Figure 2.6 : relativement à la perte d'autonomie cognitive, peu importe la pathologie considérée, les seniors atteints sont ceux qui restent le plus longtemps en perte d'autonomie tandis que les différences sont davantage prononcées parmi les diagnostics relatifs à la perte d'autonomie physique. La distinction entre dépendance d'ordre cognitive et d'ordre physique paraît justifiée pour les tables Qalydays. Plus généralement, toutes les

courbes explosent à faible ancienneté pour finir par doucement décroître. La variable *IncurredYear* semble être une exception parmi ses compères et son importance semble relativement stationnaire à partir de 60 mois d'ancienneté et croît légèrement à partir de 170 mois.

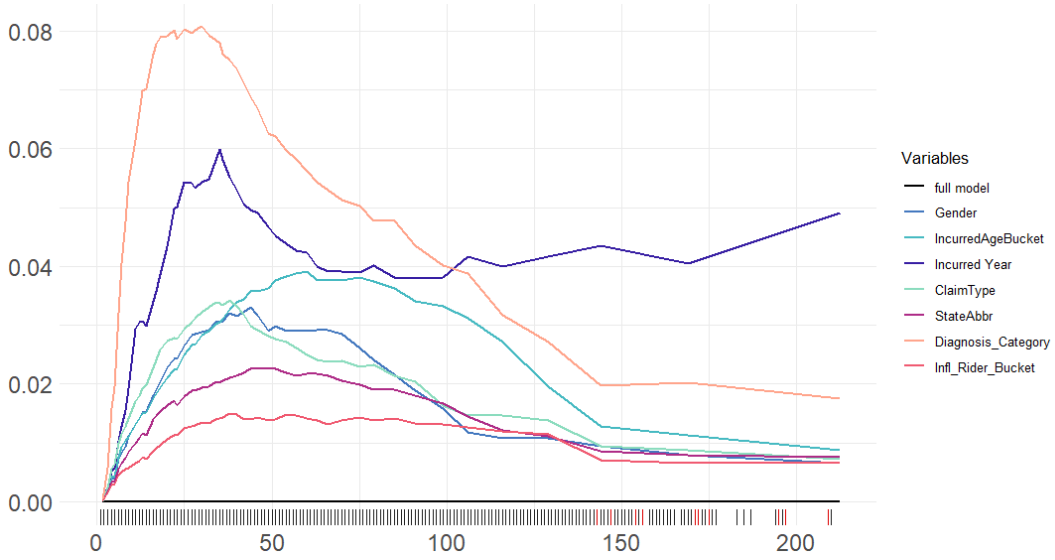


FIGURE 3.5 – Importance par permutation des variables en fonction de l'ancienneté en dépendance en mois avec le score de Brier comme fonction de perte

Valeurs de Shapley

Une exploration plus fine de l'importance des variables peut consister à regarder les valeurs de Shapley. Ces dernières connaissent une popularité grandissante pour l'examen de l'intelligibilité locale des modèles de régression. SHAPLEY, 1953 introduit pour la première fois son concept dans le cadre de la théorie des jeux coopératifs et c'est LUNDBERG, 2017 qui généralise son emploi pour l'interprétation locale des modèles prédictifs. L'idée est la suivante. Soit M le nombre de variables et S l'ensemble des variables. Les valeurs de Shapley $(\phi_i)_{i=0,\dots,M}$ correspondent aux contributions de chaque variable explicative à la prédiction et leur somme constitue la prédiction f_x de l'individu x. La première des contributions consiste tout simplement en la prévision moyenne $\phi_0 = \bar{f}_x$ du modèle tandis que toutes les autres dévient f_x de $\bar{f}_x$ et correspondent chacune à l'une des modalités de l'individu x :

$$f_x = \phi_0 + \sum_{i=1,\dots,M} \phi_i z_i$$

où $z_i \in \{0,1\}^M$ vaut respectivement 1 et 0 lorsque la variable est observée et lorsqu'elle est inconnue.

L'expression des contributions de Shapley est la suivante :

$$\phi_i(f_x) = \sum_{S \subseteq 1:M \setminus \{i\}} \frac{|S|!\,(M-|S|-1)!}{M!} \left(f_x(S \cup \{i\}) - f_x(S)\right)$$

Bien que les valeurs de Shapley soient initialement un outil d'analyse pour l'intelligibilité locale, une mesure de l'importance des variables s'en déduit. Cette dernière est définie comme

la moyenne de la valeur absolue de $(\phi_i)_{i=1,\dots,M}$. En outre, la détermination des contributions de Shapley est une tâche qui peut s'avérer numériquement lourde. Pour réduire le coût computationnel, la librairie *fastshap* de R a été employée et permet le calcul d'une approximation des valeurs de Shapley (voir Štrumbelj et Kononenko, 2014). Les résultats sont ceux de la Figure 3.4.

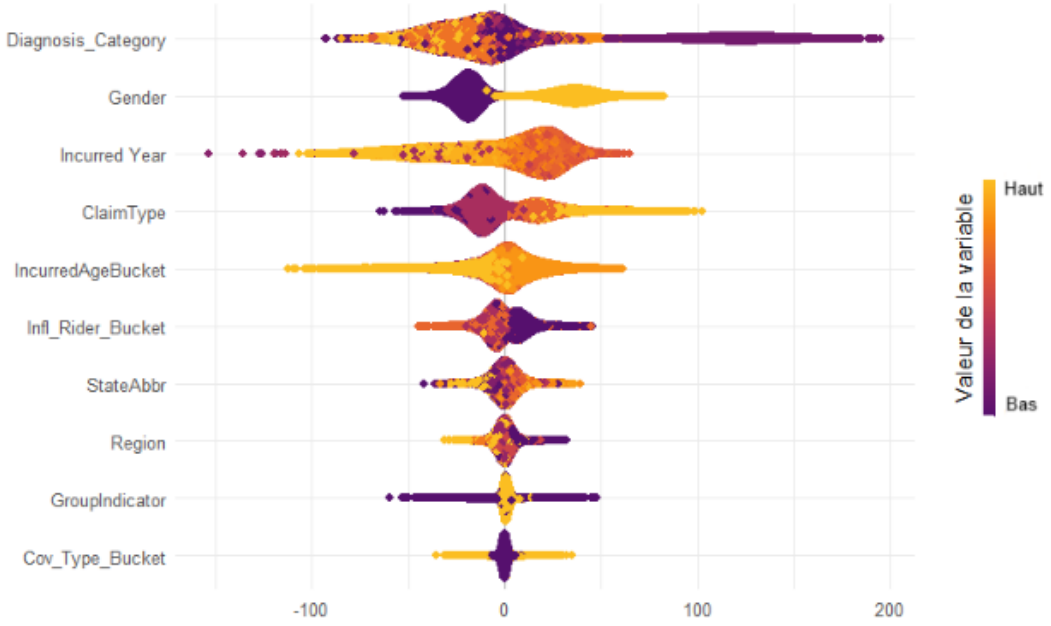


Figure 3.6 – Beeswarm plot des valeurs de Shapley. (NB : les variables catégorielles ont été encodées en chiffres afin de disposer d'une légende commune. Les correspondances sont disponibles en Annexe A.1)

La Figure 3.6 représente un « beeswarm plot » des valeurs de Shapley. Ce type de graphe consiste à afficher l'intégralité des contributions de Shapley pour toutes les modalités de chaque covariable. Un tel type de visualisation permet de distinguer dans quel sens chaque modalité fait dévier la prédiction moyenne $\bar{f}_x$. Une modalité qui présente des contributions de Shapley globalement positives (resp. négatives) est associée à une mortalité plus importante (resp. faible) par rapport à l'estimation moyenne. L'inconvénient indéniable du beeswarm plot : sa lecture est difficile voire impossible pour des variables explicatives qui présentent un nombre trop important de modalités. Dans notre cas, les covariables *Diagnosis_Category*, *Infl_Rider_Bucket*, *StateAbbr* et *Region* y sont confrontées. Un autre type de visualisation sera employé pour commenter *Diagnosis_Category* en Figure 3.7. Sur la Figure 3.6, les constatations marquantes sont les suivantes :

— pour *Incurred Year* : il semblerait que plus l'année d'incidence soit élevée, plus le risque de décès s'affaiblisse. Ce résultat s'aligne sur l'amélioration de l'état de santé de la population mondiale au fil du temps et paraît assez cohérent.

— pour *ClaimType* : contrairement à ce qui figure dans l'Annexe A.1, les résultats correspondent à l'intuition première qui consiste à penser que les dépendants les plus sévères et qui sont les plus susceptibles de perdre la vie sont davantage localisés en nursing homes. En outre, les résidents en ALF semblent toujours être ceux qui persistent le plus dans l'état de perte d'autonomie.

Score de risque pour les modalités de *Diagnosis_Category*

Afin de se pencher sur l'effet des modalités de *Diagnosis_Category*, un score de risque sera employé à la place du beeswarm plot. Le score de risque pour une modalité x_i est calculé comme l'espérance de la fonction de hasard cumulative pour laquelle la valeur de *Diagnosis_Category* est fixée à x_i (voir Spytek et al., 2023).

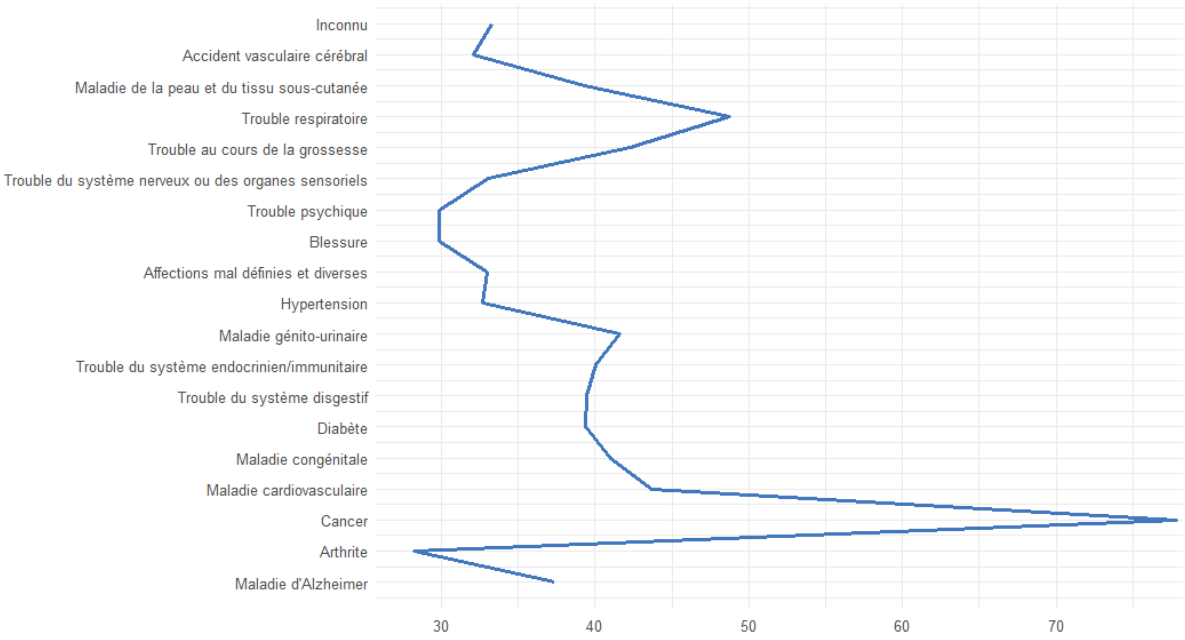


Figure 3.7 – Scores de risque associés à chaque modalité de la variable explicative *Diagnosis_Category*

D'après la Figure 3.7, les 3 scores de risques les moins élevés sont détenus par l'arthrite, le trouble psychique et la blessure. Ces dernières seraient donc associées au risque de mortalité le plus faible. En outre, les résultats dénotent quelque peu avec la Figure 2.6. Les personnes âgées atteintes de la maladie d'Alzheimer, d'un accident vasculaire cérébral ou de troubles du système nerveux ou des organes sensoriels ne sont pas les cohortes avec le risque de mortalité le plus faible.

Résidus de Cox-Snell

Cox et Snell, 1968 introduisent les résidus de Cox-Snell. Ces derniers permettent de juger de la qualité de l'ajustement du modèle. Le résidu de Cox-Snell pour une observation x_i est défini comme la fonction de hasard cumulative au temps t_i. Si le modèle de survie ajuste correctement les données d'entrée, alors les résidus de Cox-Snell doivent suivre une loi exponentielle de paramètre 1.

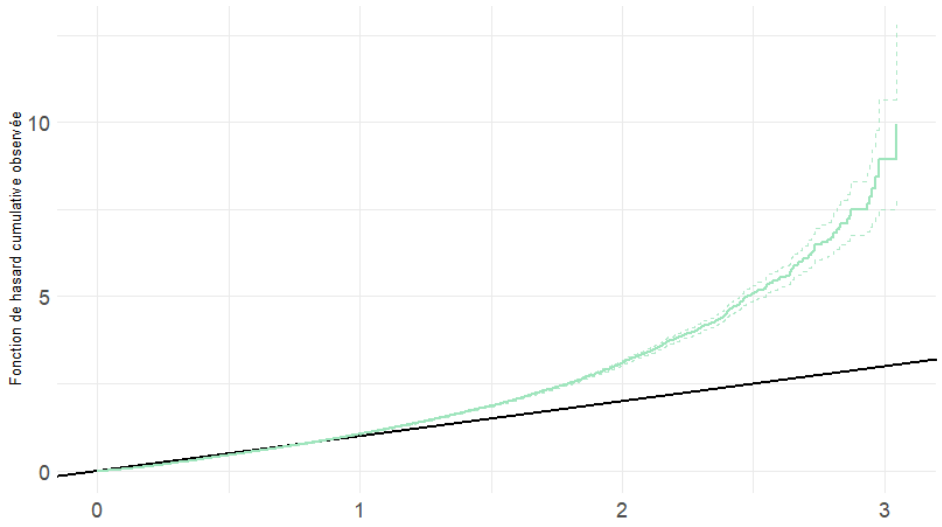


FIGURE 3.8 – Diagramme quantile-quantile exponentiel de paramètre 1 pour les résidus de Cox-Snell

Il apparaît très clairement que les résidus ne suivent pas une loi exponentielle. Il aurait été possible d'écarter certaines variables de la construction de la forêt aléatoire. Par ailleurs, on décide de toutes les conserver puisqu'aucune ne semble bruiter nos prédictions d'après les résultats de la VIMP et de la profondeur minimale.

3.7 Comparaison des performances du modèle de Kaplan-Meier, de l'arbre de survie et de la forêt aléatoire de survie

3.7.1 Définitions des métriques de performance employées pour l'arbre et la forêt

C-index de Harrell

L'indice de concordance de Harrell, ou C-index, est un indicateur très populaire de la qualité prédictive des forêts aléatoires « classiques » mais aussi des forêts aléatoires de survie. HARRELL et al., 1982 l'introduit dans sa toute première version, mais c'est bien l'extension de LU, 2023, vers laquelle on renvoie pour le détails du calcul du C-index de « survie », qui sera employée par la suite.

Le C-index est une métrique intéressante afin de classer les modèles entre eux. L'idée derrière la construction du C-index consiste à mesurer la capacité du modèle à ranger les individus entre eux. Pour ce faire, parmi toute paire d'observations qu'il est possible de créer à partir de la base de données, on cherche à savoir quel individu admet la mortalité la plus importante, cette mortalité étant synthétisée par la somme de la fonction cumulative de hasard sur l'ensemble des temps qui apparaissent dans la base de données (voir ISHWARAN et LU, 2018). Une fois ce classement réalisé, on le compare aux dates d'événements réellement observées.

Score de Brier et son intégrale

À l'instar du C-index de Harrell, le score de Brier est également une métrique appréciée pour l'évaluation de la performance prédictive (voir Brier, 1950 et Graf et al., 1999). Dans sa version adaptée pour l'analyse de survie de Lu, 2023, le score de Brier $\widehat{BS}$ est estimé de la façon suivante :

$$\widehat{\mathrm{BS}}(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{S}^2(t|\mathbf{X}_i)\mathbb{1}_{T_i\leq t}\delta_i}{\widehat{G}(T_i - |\mathbf{X}_i)} + \frac{\left(1-\widehat{S}(t|\mathbf{X}_i)\right)^2\mathbb{1}_{T_i>t}}{\widehat{G}(t|\mathbf{X}_i)}\right\}$$

avec $\widehat{S}(t|\mathbf{X})$ un estimateur de la fonction de survie et $\widehat{G}(t|\mathbf{X}_i)$ un estimateur de la fonction de survie des données censurées. La notation T_i- correspond à la limite à gauche du point T_i.

L'intégrale du score de Brier $IBS(\tau)$, ou encore appelé Continuous Rank Probability Score (CRPS), au temps τ est, quant à elle, définie par l'expression suivante :

$$IBS(\tau) = \frac{1}{\tau}\int_0^{\tau} BS(t)\,\mathrm{d}t$$

BS sera substituée par $\widehat{BS}$ dans le cadre de l'estimation de cette intégrale.

AUC cumulative/dynamique dépendant du temps

La courbe ROC est un outil graphique qui oppose le taux de vrais positifs au taux de faux positifs et permet de mesurer la performance d'un test binaire. Complémentairement à la courbe ROC, l'AUC est généralement calculé et correspond à l'aire sous la courbe ROC. Dans le cadre de l'analyse de survie, la sensibilité et la spécificité dépendent du temps. L'Annexe A.5 présente l'AUC de la forêt aléatoire en fonction de l'ancienneté en perte d'autonomie.

3.7.2 Synthèse des performances de chaque estimateur

Estimateurs	C-index de Harrell	Intégrale du score de Brier	Moyenne des C/D AUC
Arbre de survie	0,640	0,103	0,652
Forêt aléatoire de survie	0,645	0,080	0,675

Table 3.2 – Performances prédictives de l'arbre de survie et de la forêt de survie

D'après le C-index de Harrell et comme montré précédemment sur la Figure 3.2, la forêt aléatoire de survie semble seulement légèrement plus performante que l'arbre de survie. Par ailleurs, le score de Brier et son intégrale confèrent un avantage certain pour la forêt de survie. Cette constatation est également étayée par la Figure 3.9. Le CPRS de la forêt aléatoire est bien inférieur aux deux autres estimateurs.

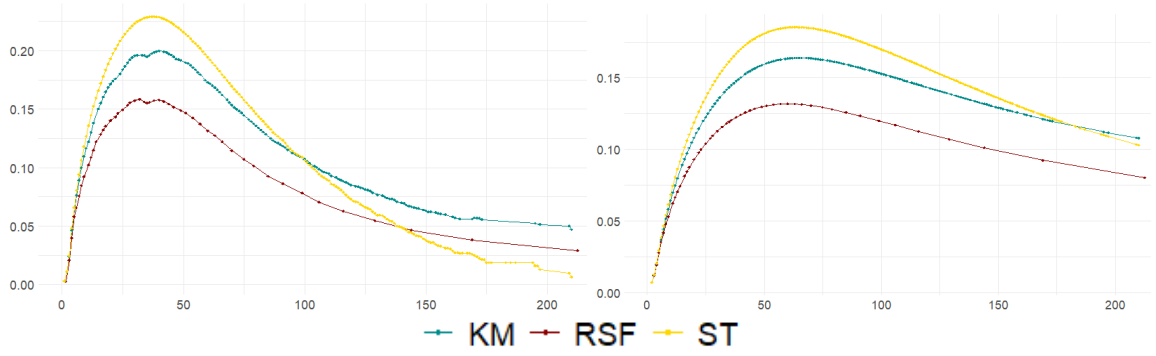


FIGURE 3.9 – Score de Brier à gauche et intégrale du score de Brier à droite en fonction de l'ancienneté en perte d'autonomie

3.7.3 Résultats : estimation des lois et statistiques

Sur la Figure 3.10, plusieurs courbes de survie de Kaplan-Meier se croisent avant les 50 mois d'ancienneté, que ce soit pour les hommes ou les femmes. À titre d'exemple, les courbes féminines des moins de 60 ans et des 60-64 ans présentent une mortalité accrue avant de passer les 25 mois en dépendance. Néanmoins, cette surmortalité est loin d'être caractéristique de toutes les tranches d'âges d'incidence en dépendance alors qu'il s'agit d'un phénomène usuellement observé chez les seniors en perte d'autonomie. Comparativement aux lois Qalydays, cette mortalité excessive intervient plus tardivement à faible ancienneté en dépendance ou n'est simplement pas constatée sur les estimations brutes de Kaplan-Meier. En effet et à titre d'exemple, l'espérance de vie résiduelle en considérant 1, 2 ou 3 mois d'ancienneté en perte d'autonomie n'est jamais supérieure à cette même quantité mais pour une durée nulle passée en perte d'autonomie. Il faut attendre 4 mois d'ancienneté chez les femmes entrées en dépendance à l'âge 60-64 ans pour que l'espérance de vie résiduelle devienne supérieure de seulement deux mois par rapport à la même cohorte mais sans ancienneté en perte d'autonomie. Cette tendance est semblable pour la forêt aléatoire v1 ou v2 sur les 4 premiers mois. Une surmortalité comparable à la Table 2.1 n'est donc aucunement observée. La documentation de la base de données SOA LTC 2000-2011 suppose que la franchise est déjà prise en compte dans la variable *Gender* mais ce point mériterait d'être davantage examiné.

Concernant la comparaison des deux estimateurs de survie de la forêt aléatoire, la version 1 de la fonction de survie sur-estime la survie comparativement à la version 2, comme attendu. Dans les deux cas, les probabilités de survie ont l'air de présenter une mortalité excessivement faible dans les grandes anciennetés. Ce point nécessite une attention particulière sur la fermeture de table du Chapitre 4. De plus, les anciennetés en dépendance ne dépassent pas communément les 200 mois tandis que certaines lois Qalydays peuvent encore être définies après 500 mois d'ancienneté. Par ailleurs, les courbes de survie font également apparaître un caractère beaucoup plus lisse que les estimations de Kaplan-Meier dans ces fortes anciennetés.

La fonction de survie de l'arbre de survie, quant à elle, coïncide avec un estimateur de Kaplan-Meier qui serait construit avec *Gender* comme seule variable explicative. Les deux courbes sont tout de même présentées en Annexe A.6

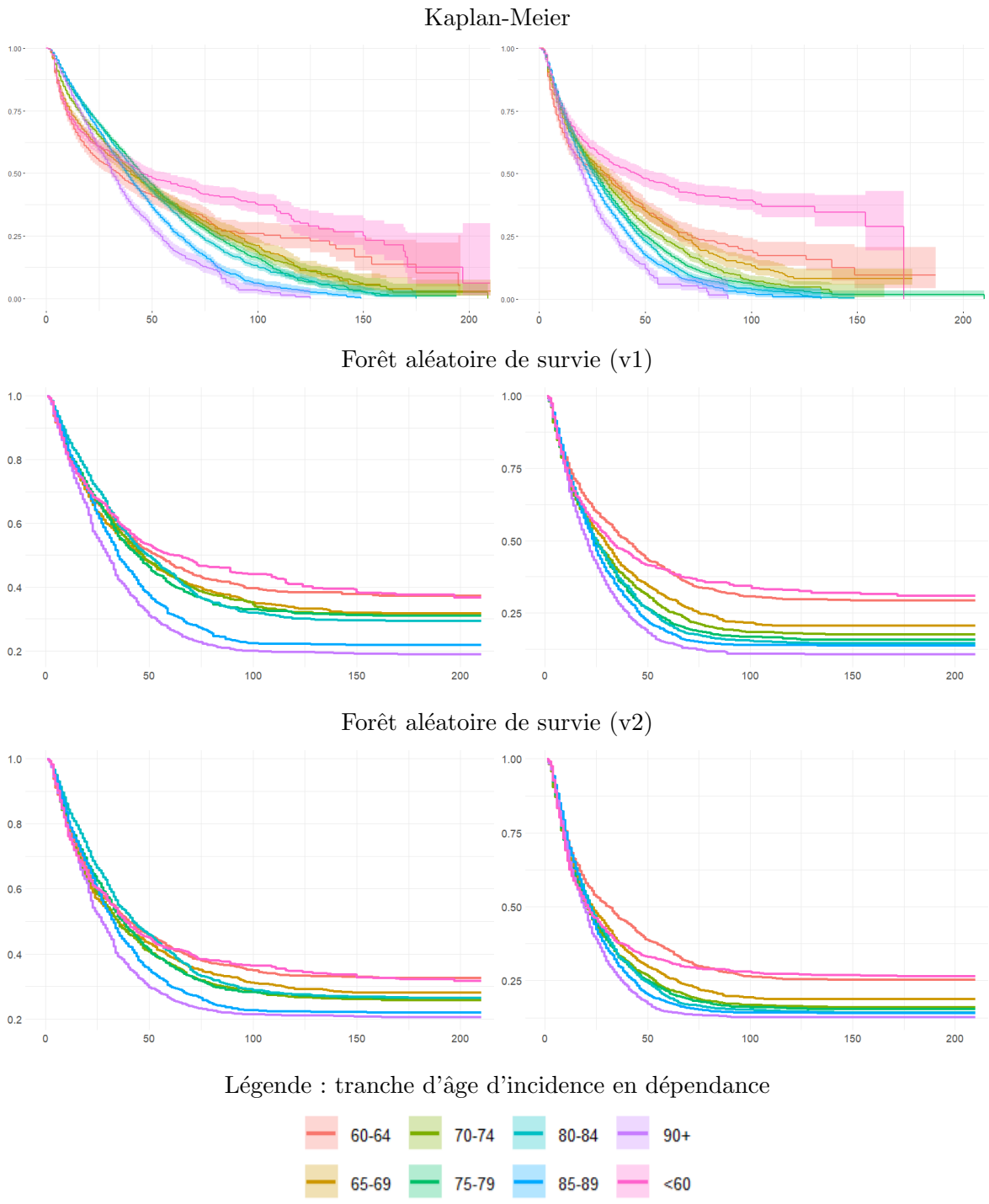


FIGURE 3.10 – Fonctions de survie par rapport à l'ancienneté en perte d'autonomie en mois, à gauche pour les femmes et à droite pour les hommes

Les chiffres du tableau 3.3 confirment plusieurs observations énoncées ci-dessus. Les moyennes, les médianes et les 3èmes quartiles masculins de la durée de survie résiduelle à l'entrée en perte d'autonomie sont systématiquement inférieurs à celles des femmes, exception faite pour l'estimateur de Kaplan-Meier pour les moins de 60 ans d'âge d'entrée. L'espérance de vie en perte d'autonomie et le 3ème quartile sont systématiquement décroissants avec l'âge d'incidence en dépendance. Les 3ème quartiles de relatifs à RSF1 sont également beaucoup plus

élevés à cause de la sous-mortalité aux anciennetés élevées. L'espérance de vie résiduelle est systématiquement plus élevée chez les estimations par forêt aléatoire. La cause : la forte survie en ancienneté élevée.

Âge d'incidence en dépendance	Espérance à l'entrée en dépendance (F/H)	Espérance à 4 mois d'ancienneté en dépendance (F/H)	Médiane à l'entrée en dépendance (F/H)	3ème quartile à l'entrée en dépendance (F/H)
Kaplan-Meier				
< 60	77/78	76/81	46/46	150/>172
60-64	62/52	62/53	34/29	111/72
65-69	55/46	54/47	40/31	91/67
70-74	55/39	54/37	42/28	84/58
75-79	53/36	52/34	45/27	80/51
80-84	51/33	50/30	43/26	73/48
85-89	43/29	43/26	38/23	63/43
≥ 90	37/24	36/21	31/20	54/36
Forêt aléatoire de survie (v1)				
< 60	102/83	103/86	62/32	>210/>210
60-64	98/82	101/83	54/40	>210/>210
65-69	89/64	90/66	47/29	>210/75
70-74	89/58	90/59	47/26	>210/63
75-79	87/54	87/54	45/25	>210/54
80-84	87/52	87/52	49/25	>210/53
85-89	70/49	69/48	35/24	81/46
≥ 90	62/42	62/41	30/21	66/40

TABLE 3.3 – Survie résiduelle en mois pour les femmes et les hommes

3.8 Taux de décès à 1 mois et lissage de Whittaker-Henderson

Les probabilités de décès à 1 mois se déduisent de la fonction de survie. La variable de durée étudiée étant une variable discrète, le graphe des probabilités de décès n'est ni lisse ni continu. Il est désormais nécessaire de d'aplanir les taux bruts. Les taux lissés serviront de référence au modèle de Brass dans le Chapitre 4. Plusieurs types de lissage existent et sont envisageables : lissage par moyenne mobile, lissage par splines cubiques, lissage de Whittaker-Henderson... C'est justement ce dernier qui sera employé. Le lissage de Whittaker-Henderson présente l'agréable avantage de se reposer sur des principes assez intuitifs. Cette méthode trouve un grand public au sein de la communauté actuarielle.

3.8.1 Présentation du lissage de Whittaker-Henderson

Le lissage de Whittaker-Henderson (voir Whittaker, 1922 et Henderson, 1924) est fondé sur le compromis de deux critères :

— un critère de fidélité : la courbe lissée épouse la courbe des taux bruts.

— un critère de régularité : la courbe est lisse.

Afin de formaliser ce cadre en dépendance, les notations suivantes sont introduites :

— t : le temps passé en perte d'autonomie

— t_{max} : l'ancienneté maximale en dépendance

— x : l'âge d'incidence en dépendance. Dans notre cas, x sera une tranche d'âge parmi 60-64, ..., 85-89 ans

— $(w_{x,t})_{t=1,...,t_{max}}$: une suite de poids positifs

— $(\tilde{q}_{x,t})_{t=1,...,t_{max}}$: les taux lissés

— $(\hat{q}_{x,t})_{t=1,...,t_{max}}$: les taux bruts

Le critère de fidélité

Le critère de fidélité F se définit comme la distance quadratique ci-dessous :

$$F(\tilde{q_x}) = \sum_{t=1}^{t_{max}} w_{x,t}(\tilde{q}_{x,t} - \hat{q_{x,t}})^2 = (\tilde{q_x} - \hat{q_x})^t W_x(\tilde{q_x} - \hat{q_x})$$

avec $W_x = diag((w_{x,t})_{t=1,...,t_{max}})$. La suite $(w_{x,t})_{t=1,...,t_{max}}$ est classiquement choisie comme l'exposition au risque.

Le critère de régularité

Le critère de régularité S s'exprime comme :

$$S(\tilde{q_x}) = \sum_{t=1}^{t_{max}-n} ((\Delta^n \tilde{q_x})_t)^2 = \tilde{q_x}^t K_z^t K_z \tilde{q_x}$$

avec K_n une matrice de taille $t_{max} - n \times t_{max}$. En effet, l'opération Δ^n étant linéaire, cela nous assure l'existence de K_n.

— Δ dénote l'opérateur différence telle que $(\Delta\tilde{q_x})_t = q_{x,\tilde{t+1}} - q_{\tilde{x},t}$
— Δ^n dénote l'opérateur qui applique n fois l'opérateur de différenciation tel que $(\Delta^n\tilde{q_x}) = \Delta(\Delta^{n-1}\tilde{q_x})$

Le critère de Whittaker-Henderson

Le critère de Whittaker-Henderson WH_h de paramètre h est défini comme :

$$WH_h(\tilde{q_x}) = F(\tilde{q_x}) + hS(\tilde{q_x})$$

Le lissage consiste en la minimisation du critère de Whittaker-Henderson sur $\tilde{q_x}$. Le paramètre h correspond au paramètre de lissage. Plus h est grand, plus la courbe sera régulière tandis qu'un h petit contraint la courbe lissée à coller le plus à la courbe des taux bruts.

3.8.2 Application et résultats

Le positionnement des lois de mortalité fait face à une complication limitante en raison de la petite taille de la base de données SOA après retraitement. Le jeu de données qui sert à l'estimation de chacune de nos cohortes ne présente pas toujours des anciennetés successives séparées par un seul mois d'ancienneté. Étant donné que le calcul des taux de décès nécessite la connaissance des valeurs successives de la fonction de survie ou de la fonction de hasard séparées d'un seul mois d'ancienneté, des pertes d'informations ont lieu. Afin d'illustrer cette contrainte, rappelons la définition des taux de décès par âge d'incidence en dépendance et prenons un exemple :

$$q_{x,t} = 1 - \frac{S(x,t+1)}{S(x,t)}$$

À titre d'exemple, si la fonction de survie aux anciennetés 22 mois et 24 mois est disponible mais n'est pas connue à 23 mois alors il ne sera pas possible de calculer un taux brut de décès pour les anciennetés 23 mois et 24 mois. Ces probabilités manquantes se trouvent bien évidemment en grosse proportion aux anciennetés élevées. À l'issue du Chapitre 4, les taux manquants sont imputés par extrapolation linéaire.

Les paramètres n et h sont optimisés à partir du critère de validation croisée généralisée (GCV) et du critère d'information d'Akaike corrigé (AICc). Hurvich et Tsai, 1995 préconise l'utilisation de l'AICc lorsque le rapport du nombre de paramètres sur le nombre d'observations est strictement inférieur à 40. Cette condition est effectivement vérifiée. D'autres critères pour la sélection des paramètres existent comme le BIC ou la vraisemblance marginale (voir Biessy, 2023). La GCV et l'AICc adoptent les expressions suivantes :

$$\text{GCV} = \frac{t_{max}}{(t_{max}-p)^2} \cdot \sum_{t=1}^{t_{max}} (\tilde{q}_{x,t} - \hat{q_{x,t}})^2$$

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{t_{max}-p-1} = t_{max}.\ln\left(\sum_{t=1}^{t_{max}} (\tilde{q}_{x,t} - \hat{q_{x,t}})^2\right) + 2p + \frac{2p(p+1)}{t_{max}-p-1}$$

où p représente le nombre de degrés de liberté du lissage non-paramétrique.

n et h sont respectivement testés parmi les valeurs $\{2;3\}$ et $\{10^{0,1k}\ ;\ k = 0, ..., 40\}$. Comparativement à la GCV, l'AICc apparaît comme un meilleur outil de décision au regard du sous et du sur-apprentissage. Les deux critères s'accordent unanimement sur $n = 3$.

Âge d'incidence	KM : h (F/H)	KM : p (F/H)	RSF1 : h (F/H)	RSF1 : p (F/H)	RSF2 : h (F/H)	RSF2 : p (F/H)
60-64	$10^{1,5}/10^{1,2}$	20/20	$10^{4}/10^{1,3}$	10/25	$10^{4}/10^{1,3}$	10/25
65-69	$10^{4}/10^{3,7}$	12/11	$10^{4}/10^{4}$	12/10	$10^{1,3}/10^{2,9}$	34/15
70-74	$10^{1,6}/10^{1,6}$	37/28	$10^{3,9}/10^{4}$	14/12	$10^{3,3}/10^{3,8}$	17/12
75-79	$10^{4}/10^{4}$	16/12	$10^{4}/10^{4}$	15/12	$10^{4}/10^{3,5}$	15/15
80-84	$10^{4}/10^{4}$	23/12	$10^{3,9}/10^{3,1}$	16/17	$10^{3,6}/10^{3,2}$	18/16
85-89	$10^{4}/10^{2,7}$	13/16	$10^{4}/10^{3,8}$	13/11	$10^{4}/10^{3,8}$	13/11

TABLE 3.4 – Paramètres de lissage h sélectionnés et nombre de degrés de liberté p (arrondi à l'unité) associé par âge d'incidence en perte d'autonomie

D'après les degrés de liberté affichés, les lissages semblent généralement sujets au sur-apprentissage. La létalité des 3 premiers mois de tous les estimateurs est systématiquement inférieure à celles des lois Qalydays tandis que celle en forte ancienneté semble sous-estimée. Par ailleurs, ce phénomène disparaît pour l'estimateur de Kaplan-Meier lorsqu'on dépasse la tranche d'âge des 70-74 ans, si bien que l'on finit par constater une force de mortalité très légèrement pentue voire plate. Les tendances à la surmortalité sont beaucoup plus éparses dans le temps que celles des taux Qalydays. Dans la suite, seuls les courbes relatives aux âges d'incidence 60-64 et 65-69 ans sont affichées, par souci de concision et de lisibilité.

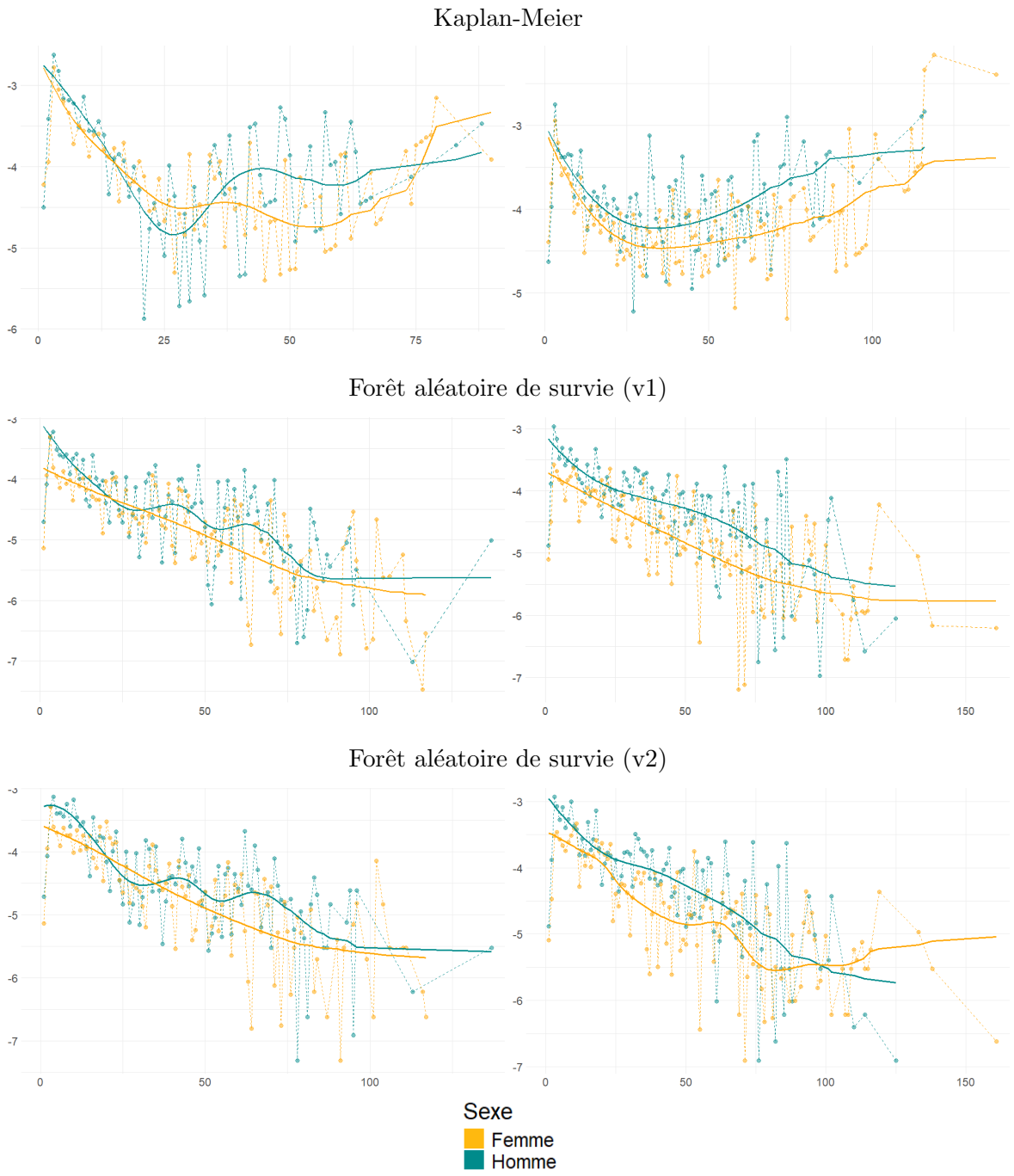


FIGURE 3.11 – Logarithme des taux de décès à 1 mois en dépendance en fonction de l'ancienneté en mois. Cohortes d'âge d'incidence en dépendance 60-64 ans et 65-69 ans respectivement à gauche et à droite. Les taux lissés sont en traits pleins tandis que les taux bruts forment un nuage de points.

3.8.3 Validation du lissage

Plusieurs critères sont employés afin de valider les tables lissées. La méthodologie employée est tirée de Tomas et Planchet, 2013 qui s'articule autour de la validation des ajustements de la mortalité et de ses tendances d'évolution. Les indicateurs de performance utilisés seront également employés dans le Chapitre 4. Le test des signes, le test des runs (voir Forfar et al., 1988 et Debón et al., 2006) ainsi que le test de Wilcoxon sont également réalisés afin de témoigner de la qualité du lissage et des écarts entre les ajustements et les taux bruts. On renvoie une nouvelle fois vers Tomas et Planchet, 2013 afin de connaître le détail de ces tests.

Métriques de diagnostic : χ^2, MAPE, R^2 et SMR

Le χ^2, le MAPE, le R^2 et le Standardized Mortality Ratio (SMR) sont calculés afin de vérifier la qualité du lissage des taux bruts. Les définitions de ces différentes quantités sont adaptées de Tomas et Planchet, 2013 à notre périmètre. Seule la définition du SMR est détaillée ci-dessous. Ce dernier se caractérise communément comme le ratio du nombre de décès observé sur le nombre de décès attendu. Dans notre cas, il s'agira plutôt du rapport entre le nombre de décès relatif aux taux bruts $\hat{q}_{x,t}$ et le nombre de décès relatifs aux taux ajustés $\tilde{q}_{x,t}$. Plus formellement, le SMR s'écrit dans notre cas :

$$SMR = \frac{\sum_{(x,t)} E_{x,t} . \hat{q}_{x,t}}{\sum_{(x,t)} E_{x,t} . \tilde{q}_{x,t}}$$

où $E_{x,t}$ représente l'exposition au risque.

Si le SMR est supérieur à 1, alors les décès ajustés sont sous-estimés. Dans le cas contraire, les décès sont sous-évalués. Au-delà d'être un outil de diagnostic, le SMR échafaude la méthode de positionnement de lois la plus simple qui existe et qui consiste à appliquer le SMR comme taux d'abattement à une table déjà construite.

Résidus de la réponse et résidus de Pearson

Les résidus de la réponse $r_{x,t}$ et les résidus de Pearson $r^P_{x,t}$ sont définis de la façon ci-contre :

$$r_{x,t} = \hat{q}_{x,t} - \tilde{q}_{x,t} \quad ; \quad r^P_{x,t} = \frac{E_{x,t}(\hat{q}_{x,t} - \tilde{q}_{x,t})}{\sqrt{\widehat{\mathbf{Var}}(E_{x,t}\tilde{q}_{x,t})}}$$

Ces résidus représentent la déviation des estimations des taux de mortalité face aux observations et apportent de l'information locale sur la qualité de l'ajustement. Des résidus successifs de même signe figurent un phénomène de surlissage local. Les résidus de Pearson affichés en Figure 3.12 et en Annexe A.7 ne présentent pas de tendances fortes, ce qui semble rassurant.

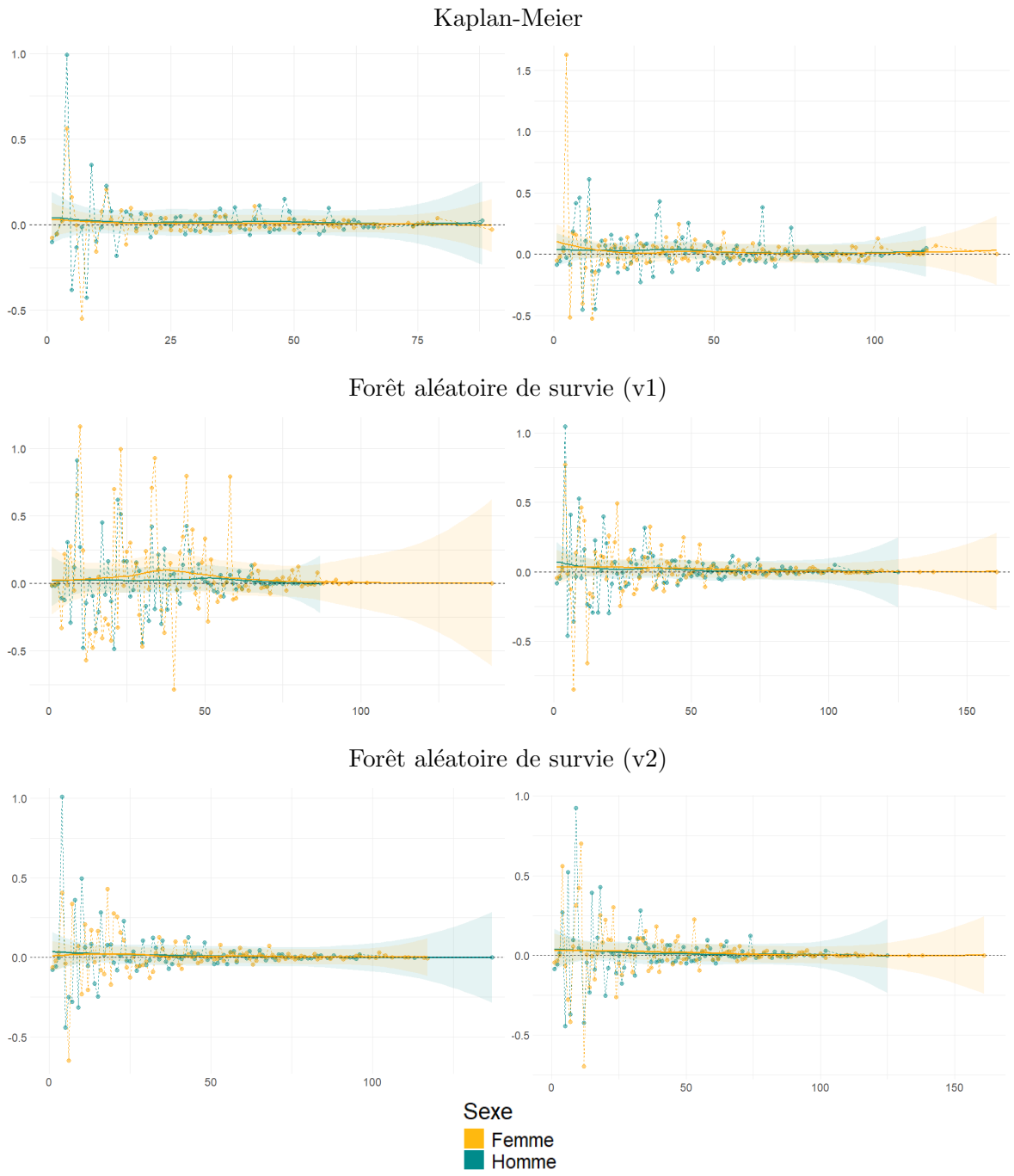


FIGURE 3.12 – Résidus de Pearson consécutifs au lissage des taux de décès à 1 mois en dépendance. Cohortes d'âge d'incidence en dépendance 60-64 ans et 65-69 ans respectivement à gauche et à droite.

Synthèse des résultats

Les chiffres du χ^2 donnent l'ascendant au lissage des taux estimés par KM. Le R^2, quant à lui, assure que le lissage des estimées de la forêt aléatoire explique davantage la variance de l'ajustement. Pour finir, tous les SMR dénotent une faible sous-estimation de la mortalité de la part des taux lissés. Pour l'ensemble des cohortes d'âge d'incidence, la p-valeur du test des signes et du test de Wilcoxon ne permet pas de rejeter l'hypothèse nulle au seuil 5%. Le test

des runs a également été pratiqué individuellement pour chaque tranche d'âge d'incidence en dépendance et les résultats sont unanimement rassurants.

Métriques de diagnostic	KM (F/H)	RSF1 (F/H)	RSF2 (F/H)
P-valeur du tests des signes	0,52/0,52	0,52/0,52	0,52/0,52
P-valeur du test de Wilcoxon	0,63/0,76	0,23/0,48	0,31/0,68
χ^2	13,5/11,2	30,07/17,02	31,67/19,05
MAPE	49/50	98/85	89/80
R^2	0,47/0,23	0,49/0,57	0,57/0,63
SMR	1,02/1,02	1,04/1,03	1,03/1,03

TABLE 3.5 – Diagnostic du lissage des estimateurs de Kaplan-Meier et de la forêt aléatoire de survie

Chapitre 4

Adaptation au cadre français : positionnement et fermeture de tables

Dans ce chapitre, les lois américaines seront adaptées grâce à un modèle relationnel de Brass à deux paramètres. Les méthodes d'adaptation et de validation des tables de mortalité sont reprises de Planchet et Tomas, 2014. Dans la suite, les taux de décès adaptés par le modèle de Brass sont notés $q_{x,t}$. Les probabilités de décès estimées dans le Chapitre 3 reprennent la même notation à laquelle on ajoute l'exposant ref pour les identifier. La 2nde section de ce chapitre s'attache à extrapoler les tables de mortalité selon la méthode de Denuit et Goderniaux, 2005. Le chapitre se conclut par un bilan de la construction des lois de mortalité qui s'étend depuis le Chapitre 3.

4.1 Positionnement : modèle de Brass

Le modèle de Brass, 1971 est un modèle généralement employé afin de positionner des taux de mortalité déjà existants sur une population particulière. Le modèle de Brass n'admet pas de garanties théoriques quant à la justesse des résultats. Par ailleurs, il a déjà montré à de nombreuses reprises des résultats empiriques satisfaisants. En outre, l'idée d'utiliser un modèle de Brass pour de la mortalité en dépendance a été suscitée par la lecture de RAPIOR, 2021 dont le sujet est l'estimation de taux de mortalité en dépendance sur le marché japonais à partir de données dépendance françaises.

4.1.1 Définition

Le modèle de Brass à deux paramètres est un modèle de régression qui s'écrit de la façon suivante :

$$\text{logit}(q_{x,t}) = \alpha + \beta.\text{logit}(q_{x,t}^{ref}) + \epsilon_{x,t}$$

où α, β sont les paramètres à optimiser et $\epsilon_{x,t}$ représente l'erreur commise par l'ajustement.

À la différence du modèle de Brass usuel, t représente ici le temps passé en dépendance au lieu de l'année calendaire, et x correspond à l'âge d'entrée en dépendance contre l'âge atteint pour le modèle habituel. Les paramètres sont obtenus par minimisation d'une distance entre les taux bruts observés et les taux de référence. Dans ce contexte, plusieurs distances peuvent être considérées :

$$\sum_{(\alpha,\beta)} |E_{x,t}(q^{ref}_{x,t} - q_{x,t})| \quad ; \quad \sum_{(\alpha,\beta)} E_{x,t}(q^{ref}_{x,t} - q_{x,t})^2$$

où $E_{x,t}$ représente l'exposition au risque. Ces distances représentent tout simplement des distances en norme 1 et 2 que l'on pondère par l'exposition au risque. Par la suite, c'est le critère en norme 2 qui sera privilégié.

4.1.2 Mise en oeuvre des ajustements

N'ayant pas à disposition des taux bruts français de décès en dépendance pour positionner les tables précédemment estimées, les probabilités lisses de mortalité du groupe Qalydays sont employées directement en remplacement. Il aurait été tout de même possible de simuler des décès grâce aux tables Qalydays à partir d'une loi de Poisson qui prend pour paramètre le produit de l'exposition avec les taux Qalydays, et de générer des taux bruts par extension. Par ailleurs, les probabilités brutes simulées présentent un risque non nul d'être égales à zéro et de ne pas permettre le calcul des logits. On fait donc l'hypothèse suivante pour le positionnement de Brass : les tables Qalydays sont appréciées comme des taux bruts observés d'un portefeuille d'assurance qui reconnaît la perte d'autonomie au sens du label GAD. Ne disposant pas de l'exposition sur laquelle les tables Qalydays ont été calibrées, on emploie l'exposition observée dans la base SOA LTC 2000-2016 comme poids du critère des moindres carrées. À l'instar du Chapitre 3, seuls les courbes relatives aux âges d'incidence 60-64 et 65-69 sont affichées sur la Figure 4.1 par souci de concision et de lisibilité.

Concernant le diagnostic graphique des courbes suivantes en comparaison des lois Qalydays : ces dernières connaissent toutes une sous-estimation de la létalité autour de la première année passée en dépendance, puis sur-évalue la mortalité pour ensuite de nouveau la minorer jusqu'à l'ancienneté maximale considérée. La mortalité aux environs de la première année est particulièrement sous-estimée pour les deux estimateurs de la forêt aléatoire.

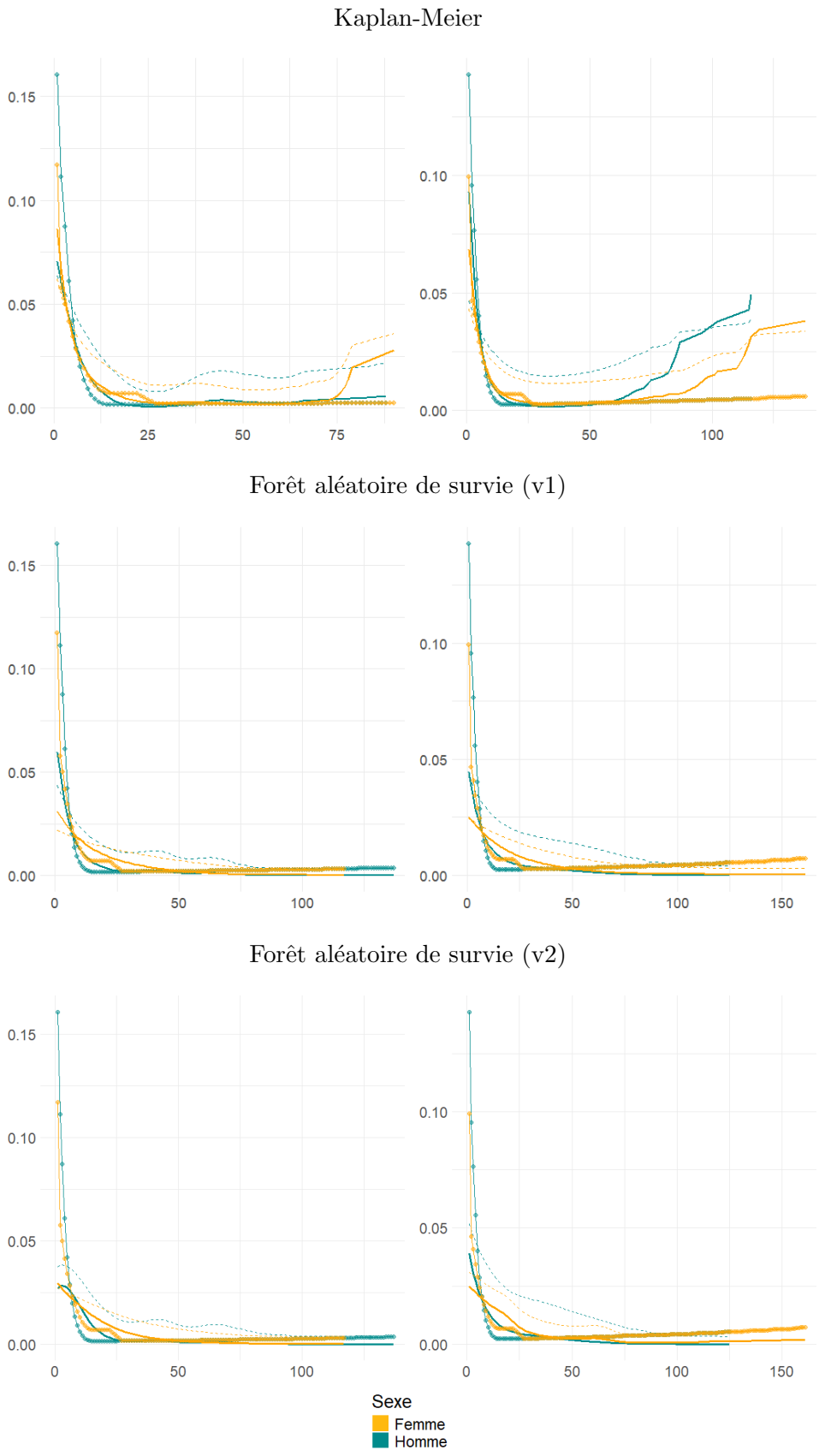


FIGURE 4.1 – Taux de décès à 1 mois en dépendance. Cohortes d'âge d'incidence en dépendance 60-64 ans et 65-69 ans respectivement à gauche et à droite. En trait plein : les ajustements de Brass. En pointillés : les taux estimés lissés. Les points reliés : les taux Qalydays.

4.1.3 Validation de l'ajustement

Afin de se convaincre du bien-fondé ou non des ajustements de Brass, les outils employés dans la sous-section 3.8.3 du Chapitre 3 sont repris. Leurs utilisation nécessitent seulement les changements suivants : les taux bruts deviennent les probabilités Qalydays et les taux ajustés grâce au modèle de Brass remplacent les taux lissés.

Les performances du positionnement de Brass délivrées dans la Table 4.1 sont très insuffisantes. Le χ^2 et le R^2 révèlent un ajustement assurément lacunaire. Le positionnement de Brass conduit à un SMR qui s'éloigne de l'unité et retranscrit une sous-estimation de la mortalité. Par conséquent, le modèle de Brass est loin d'être suffisant afin de positionner les taux américains à la mortalité en dépendance de l'Hexagone. Une piste de solution possible : modéliser différemment les taux de décès en première année et les probabilités pour des ancienetés plus élevées en amont du positionnement de Brass. À l'image d'une fermeture de table, les probabilités de décès en jeune ancienneté pourraient être adaptées. Afin de représenter plus fidèlement la mortalité des 365 premiers jours en perte d'autonomie, Planchet et al., 2018a lisse les taux de décès bruts de première année sur tous les âges d'entrée en perte d'autonomie avant de les interpoler mensuellement en considérant la linéarité des logits des taux mensuels de première année. Il aurait été intéressant d'employer une telle méthode, si la variable *IncurredAgeBucket* du jeu de données américain fournissait l'âge en année de l'entrée en dépendance au lieu d'un intervalle d'amplitude 5 ans. C'est pourquoi, il paraît peu raisonnable de réaliser un lissage sur seulement 6 valeurs dans notre cas de figure.

En outre, les SMR de RSF1 et de RSF2 sont très proches l'un de l'autre, ce qui laisse penser que la différence de mortalité intrinsèque aux estimations brutes a été atténuée, voire a disparu par la conjugaison du lissage de Whittaker-Henderson et du positionnement français. Contrairement aux forêts aléatoires, l'estimateur de Kaplan-Meier anticipe l'augmentation des décès de longue date. L'explosion erratique et précoce des taux en forte ancienneté pèse de façon modéré dans le SMR de Kaplan-Meier, étant donné la décroissance de l'exposition avec l'ancienneté en dépendance. Ce constat impose tout de même d'extrapoler correctement les taux. À titre de comparaison, Qalydays étend la loi de létalité des 60-64 ans d'âge d'incidence jusqu'à 556 mois d'ancienneté.

Malgré ces maigres performances, les estimations de Brass sont utilisées dans cet état pour la suite. D'autres modèles de positionnement existent, comme ceux détaillés dans Planchet et Tomas, 2014 et mériteraient d'être employés.

Métriques de diagnostic	KM (F/H)	RSF1 (F/H)	RSF2 (F/H)
χ^2 (en milliers)	150/112	175/144	178/145
MAPE	0,54/0,73	0,24/0,42	0,22/0,43
R^2	$< 0/< 0$	0,32/0,35	0,33/0,30
SMR	1,26/1,38	1,30/1,53	1,31/1,52

Table 4.1 – Diagnostic sur l'ensemble des ajustements de Brass

4.2 Extrapolation des tables de mortalité

L'extrapolation figure usuellement comme la dernière étape de construction d'une loi de décès. À cet égard et dans notre cas, la fermeture est double, en intervenant à la fois sur les tranches d'âge d'incidence en perte d'autonomie et sur la durée passée dans l'état de dépendance.

4.2.1 Extrapolation aux anciennetés élevées : modèle de Denuit & Goderniaux et convergence aux grands âges sur les tables Qalydays

Plusieurs méthodes de fermeture sont envisageables. Premièrement, il est possible d'employer les procédures d'extrapolation usuelles pour les tables de mortalité "classiques" qui dépendent de l'âge atteint. Quashie et Denuit, 2005 référence pléthore de procédures de fermeture dont parmi elles les modèles de Denuit et Goderniaux, 2005, Coale et Guo, 1989, Coale et Kisker, 1990, Lindbergson, 2001 ou encore celui de Thatcher, 1998. Comme le préconise Quashie et Denuit, 2005, le modèle de Denuit & Goderniaux (2005) sera retenu par la suite parmi les différentes méthodes à tester. Les atouts indéniables de ce modèle : l'introduction d'une contrainte de fermeture qui assure la concavité des quotients de mortalité aux grands âges, l'existence d'une tangente horizontale au point de la contrainte de fermeture et une non-décroissance des probabilités de décès. La sélection du point de raccordement s'effectue par optimisation du R^2. Plus explicitement, le modèle prend l'expression suivante dans notre cas et consiste à ajuster un modèle log-quadratique par la méthode des moindres carrés en introduisant une contrainte de fermeture :

$$\ln q_{x,t} = a_x + b_x t + c_x t^2 + \epsilon_{x,t} \quad \text{avec } \epsilon_{x,t} \sim \mathcal{N}(0, \sigma_x^2) \text{ i.i.d}$$

Le modèle impose une contrainte de fermeture sur les âges atteints. Dans notre cas, la contrainte de fermeture est naturellement modifiée pour porter sur les anciennetés à la place. Elle est reprise et adaptée à partir des lois Qalydays :

- $q_{60-64,555} = 1$; $q'_{60-64,555} = 0$
- $q_{65-69,495} = 1$; $q'_{65-69,495} = 0$
- $q_{70-74,435} = 1$; $q'_{70-74,435} = 0$
- $q_{75-79,375} = 1$; $q'_{75-79,375} = 0$
- $q_{80-84,315} = 1$; $q'_{80-84,315} = 0$
- $q_{85-89,255} = 1$; $q'_{85-89,255} = 0$

Les points de raccord sont testés à partir de 24 mois d'ancienneté jusqu'à l'ancienneté maximale pour laquelle nous avons encore des estimations. D'après l'enquête EHPA 2019 de la DREES, 2019, la durée de séjour en EHPAD, USLD et en EHPA admet 28 mois comme moyenne. Une borne inférieure de 24 mois à partir de laquelle sont testées les anciennetés de raccord paraît raisonnable.

Faire converger les taux estimés sur les probabilités Qalydays aurait également été envisageable. Par ailleurs, l'explosion des taux Qalydays à forte ancienneté étant assez tardive, il paraît préférable d'employer le modèle de Denuit & Goderniaux afin d'introduire un redressement des taux de décès plus anticipé. En outre, les estimées brutes de Kaplan-Meier

révélaient également un redressement des probabilités de décès plus précoce par rapport aux tables Qalydays.

Le R^2 et l'ancienneté de raccord sont présentés sur la Figure 4.2. Le R^2 est satisfaisant.

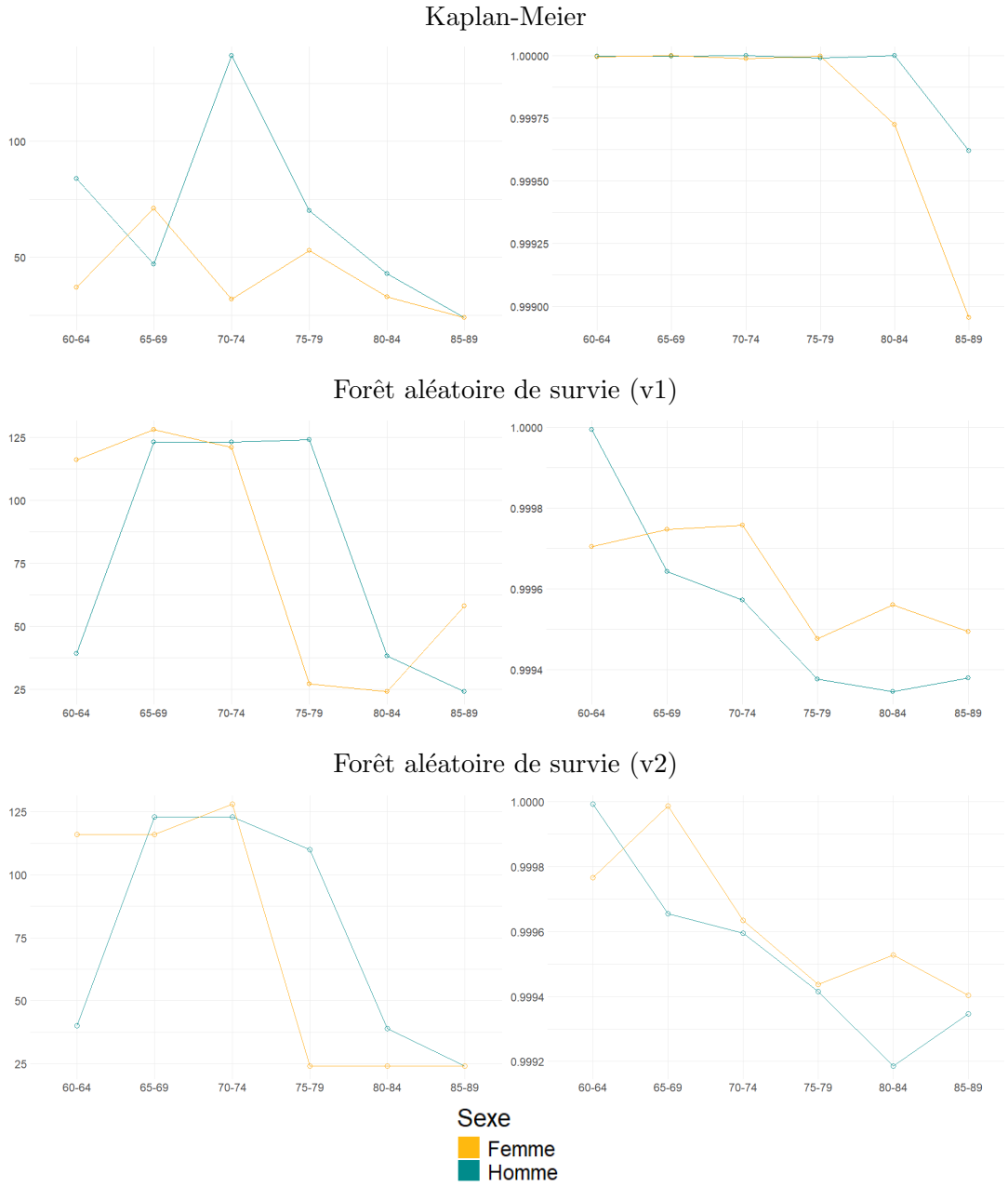


FIGURE 4.2 – Points de raccord et R^2 respectivement à gauche et à droite en fonction de l'âge d'incidence en perte d'autonomie

À l'issue de tous les précédents ajustements de ce chapitre, les probabilités de décès sont présentées en Figure 4.3. Plusieurs remarques peuvent être avancées à partir de ces graphiques. Tout d'abord, une meilleure régularité se détache des probabilités calculées initialement par l'estimateur de Kaplan-Meier. Le redressement des taux à forte ancienneté intervient de plus en plus prématurément dans l'ordre croissant des âges de survenue de la dépendance. À contrario, il est plus difficile de dégager une tendance nette en ce qui concerne les taux issus de

la forêt aléatoire. La dispersion plus «instable» des anciennetés de raccord pour les estimateurs de la forêt est sans doute responsable de cette représentation plus hétérogène. Les raccords qui peuvent sembler prématurés sont tout de même conservés. À titre de comparaison, les probabilités de décès Qalydays, pour les hommes entrés en dépendance entre 65 et 69 ans, commencent à croître à partir de la seconde année restée en perte d'autonomie. De plus, le mois qui connaît la mortalité la plus faible apparaît assez tardivement dans les estimations par forêt aléatoire et il n'est donc pas étonnant d'introduire tôt un point de raccord. Pour ce qui est de Kaplan-Meier, les anciennetés de raccords ne semblent pas effacer les creux de létalité observés. D'autre part, la surmortalité accrue lors de l'entrée en dépendance est concentrée sur une plage temporelle bien moins large pour l'estimateur de Kaplan-Meier par rapport à RSF1 et RSF2. Plus encore, l'amplitude de cette portion de temps semble s'allonger avec l'âge de survenue de la perte d'autonomie pour RSF1 du côté masculin. Plus généralement, la mortalité aggravée, pour une faible durée passée en dépendance, semble foncièrement disparaître en se décalant vers la droite des cartes de chaleurs.

Afin de comparer plus finement les différentes lois, il est également agréable de se pencher sur la Figure 4.4 qui présente les quantiles de survie résiduelle de chaque loi. Plusieurs commentaires peuvent être formulés :

— pour le 1er quartile, la survie en perte d'autonomie est systématiquement surestimée par l'ensemble des estimateurs, en particulier par la forêt aléatoire de survie dans le camp des seniors masculins. Le constat est plus hétérogène du côté de Kaplan-Meier qui apparaît plus satisfaisant avant 75 ans d'âge d'incidence notamment pour les hommes mais plus erratique, passé cette séniorité. En effet, les taux bruts qui avaient été estimés présentent une surmortalité très faible voire nulle, cause d'un positionnement perturbé.

— pour les médianes, celles de Kaplan-Meier se rapprochent davantage de celles de Qalydays. Les constatations sont plus partagées du côté de RSF1 et RSF2 qui subissent le choix des anciennetés de raccords pour la médiane. La sélection des raccords affecte de manière similaire le quantile d'ordre 0,75.

— pour le 3ème quartile, le résultat des ajustements de Kaplan-Meier surestime la mortalité. En cause : le modèle de Denuit et Goderniaux, 2005 qui incorpore un rebond plus précoce de la létalité.

— de façon plus globale, l'espérance de vie à l'entrée en dépendance montre des tendances très similaires au 3ème quartile, signe que la méthode d'extrapolation de Denuit et Goderniaux, 2005 n'est pas la plus adéquate.

— concernant les probabilités évaluées par Qalydays : la médiane, le 3ème quartile et l'espérance de vie résiduelle sont décroissants avec l'âge de survenue de la dépendance. Cette tendance n'est pas complètement reproduite par les estimateurs employés.

— tous les quantiles présentés sont relativement proches entre RSF1 et RSF2. Comme précédemment déduit de la Table 4.1, la combinaison de la méthode de Brass et de l'extrapolation de tables a réduit, pour la majorité des âges de survenue de la perte d'autonomie, la déviation initiale de létalité.

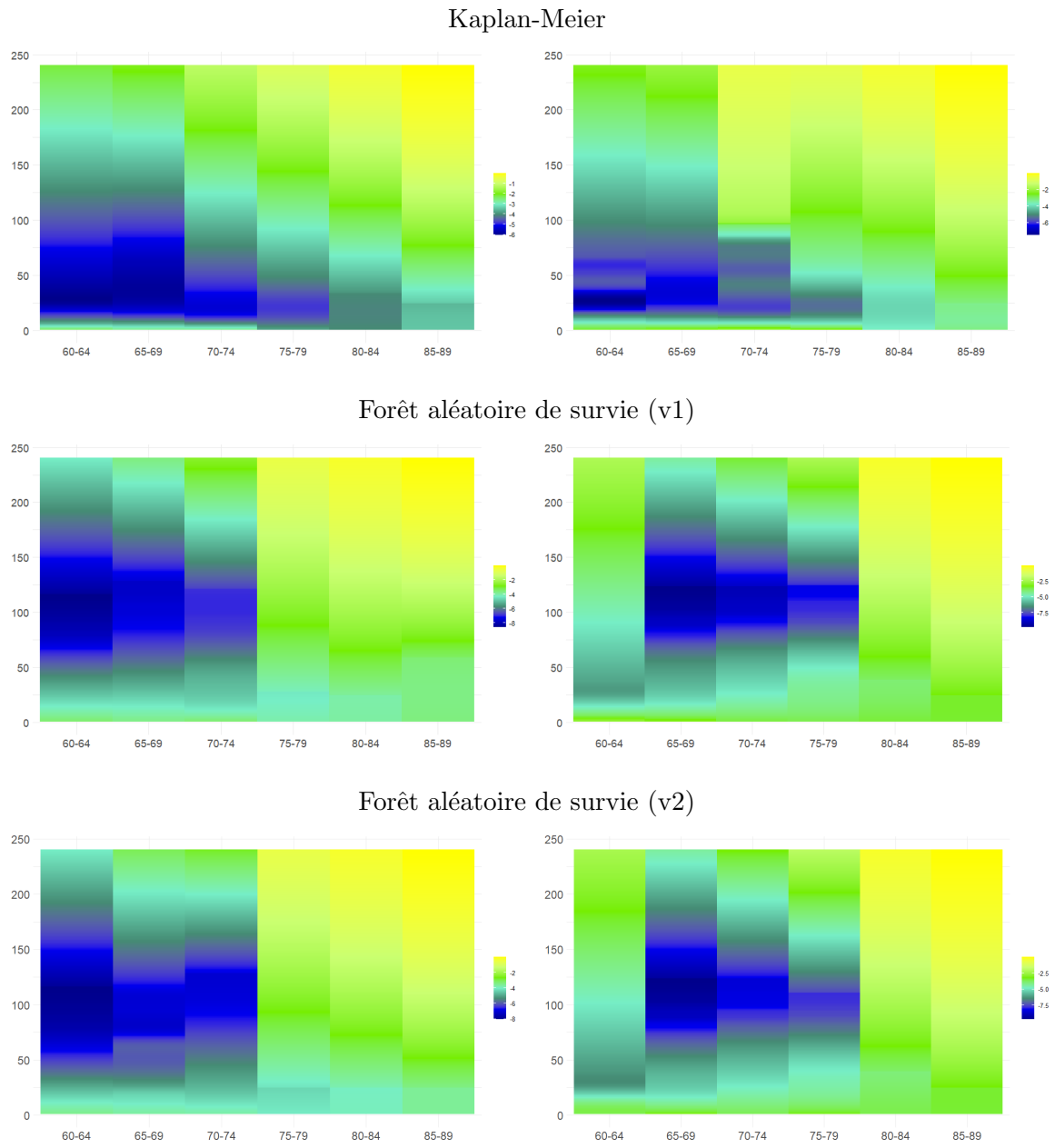


FIGURE 4.3 – Carte de chaleur du logarithme des taux de décès par tranche d'âge d'entrée en dépendance sur les 240 premiers mois d'ancienneté en perte d'autonomie. À gauche pour les femmes et à droite pour les hommes.

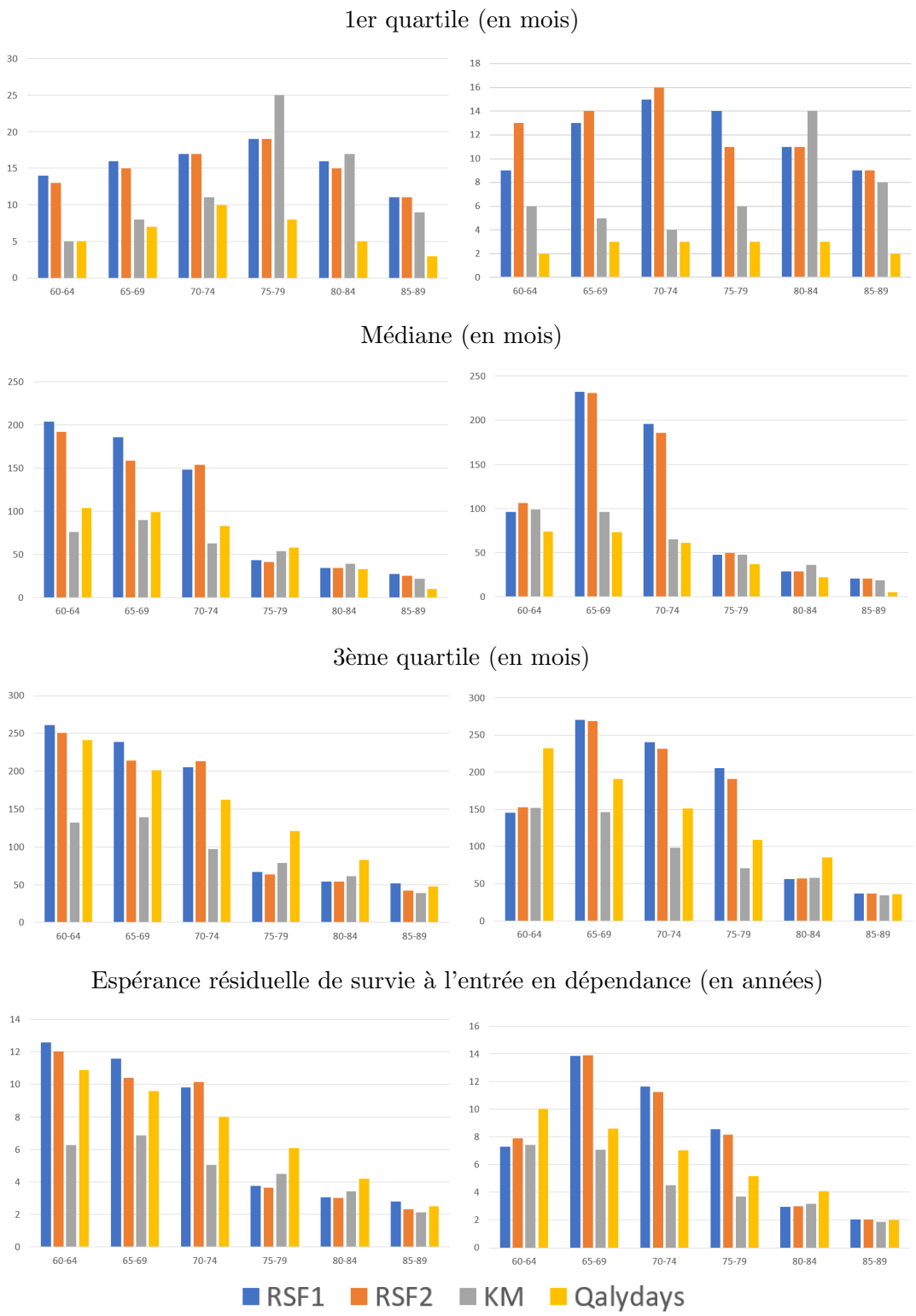


FIGURE 4.4 – Quantiles de la survie résiduelle par tranche d'âge d'entrée en perte d'autonomie, pour les femmes et les hommes de gauche à droite.

4.2.2 Extrapolation sur les tranches d'âges d'incidence en dépendance

Sur le périmètre de la perte d'autonomie, il convient généralement d'extrapoler jusqu'à la tranche d'âge d'entrée en dépendance $[100; 104[$. Dans cette optique, une méthode simple sera employée qui consiste à appliquer aux tables estimées les coefficients de passage des tables Qalydays entre les différentes tranches d'âges successives entre $[85; 90[$ et $[100; 104[$ à ancienneté en perte d'autonomie fixée.

4.3 Mise en perspective des estimations et des ajustements

En considérant les tables Qalydays comme point de comparaison français adéquat pour la modélisation de la survie en dépendance, l'ajustement des taux américains semble insatisfaisant. En outre, le contexte suivant est à prendre en compte :

- la majorité de la population dépendante française présente un âge compris entre 85 et 95 ans d'après la Table 1.6. En parallèle, la Figure 4.4 montre des résultats généralement satisfaisants pour la plage d'âge 85-89.
- la base SOA LTC 2000-2011 est un jeu de données composé pour environ deux tiers de femmes. Les résultats relatifs à cette population semblent donc plus robustes, comme pourrait le suggérer la Table 4.1.
- les tables Qalydays ont été estimées à partir des bases PMSI. Ces dernières présentent une prédominance de déficits cognitifs sévères face à la proportion d'état grabataire d'après Schwarzinger, 2019. Cette prépondérance est de l'ordre de 2 pour 1 tandis que du côté de la base américaine, les dépendants d'ordre cognitif et les dépendants d'ordre physique sont sensiblement équirépartis comme le montre la Figure 2.5. De plus et d'après la Figure 2.6, rappelons que les seniors en perte d'autonomie cognitive restent généralement plus longtemps dans l'état de dépendance. Il est tout de même difficile d'appréhender les conséquences résultantes sur les ajustements.
- la définition américaine de la dépendance diverge de celle tirée du codage médical ou de celle inscrite dans le label GAD. Il est difficile de quantifier les biais introduits.

Finalement, l'utilisation des estimateurs de la forêt aléatoire de survie n'est pas particulièrement justifiée. La surmortalité des premiers mois en perte d'autonomie est retranscrite de manière plus fidèle par l'estimateur de Kaplan-Meier ajusté sous 74 ans d'âge d'incidence. Pour RSF1 et RSF2 et malgré que leur létalité soit plus étalée dans le temps, la mortalité reste sous-évaluée comme peut le montrer la médiane toujours haute pour les seniors hommes entrés en dépendance entre 70 et 74 ans. Pour finir et dans l'ensemble, les estimateurs semblent assez prudents en jeune ancienneté mais le sont beaucoup moins à forte ancienneté.

Chapitre 5

Primes et PRC

L'objectif de ce chapitre consiste au calcul des primes et des provisions pour risques croissants en employant les différentes lois de mortalité en perte d'autonomie estimées.

5.1 Modèle dépendance à état

Un modèle illness-death est généralement considéré par les compagnies d'assurance pour modéliser l'état de santé évolutif de leurs assurés qui ont souscrit un produit dépendance. Seuls 3 états sont permis à l'assuré : autonome (A), dépendant (I) et décédé (D). Comme le montre la Figure 5.1 et à moins de rester dans le même état, seules 3 transitions y sont autorisées. Fixer la probabilité de retour en autonomie à 0 rend très simple ce modèle et évite toute contrariété en rapport avec la gestion d'une boucle entre l'état valide et l'état de perte d'autonomie. En outre, cette hypothèse est usuellement retenue en France, le manque de données ne permettant pas d'estimer une loi de retour en autonomie robuste.

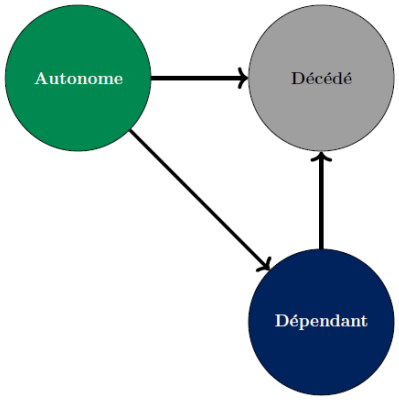


FIGURE 5.1 – Modèle illness-death pour la dépendance

Ultérieurement, les notations suivantes sont introduites, pour lesquelles x et t représentent respectivement le temps passé en autonomie et l'ancienneté en dépendance :

- $i(x)$ correspond à la probabilité d'entrée dans l'année en perte d'autonomie à l'âge x
- $q_{aa}(x)$ est la probabilité de décès dans l'année à l'âge x en état d'autonomie
- $q_i(x,t)$ est la probabilité pour un dépendant d'âge d'incidence en dépendance x et d'ancienneté t de décéder dans l'année

Le modèle illness-death sera employé dans la suite afin de probabiliser les différentes transitions possibles. Toutes les lois de probabilités sont arrêtées une fois l'âge de 105 ans atteint. Les différentes lois employées sont les suivantes :

— l'incidence et la mortalité des autonomes seront retranscrites par les lois Qalydays associées sur le périmètre de la perte d'autonomie mixte.
— la mortalité en dépendance sera modélisée par la loi Qalydays associée ainsi que par les lois estimées au Chapitre 4.
— les lois sont distinctes par sexe

Contraintes consécutives au format des lois Qalydays

Les lois Qalydays de mortalité et d'incidence se présentent sous un format spécifique :

— les tables de mortalité en autonomie sont des tables prospectives. Par souci de simplification, ces dernières ne seront pas employées dans leur intégralité. Seules les probabilités relatives à l'année 2017 serviront.
— la table d'incidence s'arrêtant à 100 ans d'âge, on fait le choix d'extrapoler linéairement les probabilités jusqu'à 105 ans d'âge en considérant très simplement la pente des taux d'entrée en dépendance sur la plage d'âge $]90; 100]$ ans. Cette décision est motivée ci-dessous, l'extrapolation aux grands âges des taux d'incidence en dépendance soulevant des enjeux majeurs pour les assureurs de la place.

D'après la littérature actuarielle, l'hypothèse la plus plébiscitée consiste à prolonger de façon exponentielle les taux d'incidence lissés aux grands âges, tout particulièrement pour les cas de démence ou pour des cohortes atteintes de la maladie d'Alzheimer (voir Corrada et al., 2010, Grasset et al., 2016 et Rocca et al., 2011). Par ailleurs, l'extrapolation des taux se confronte communément à la carence en données au-delà de 90 ans. Planchet et al., 2018b testent plusieurs méthodes d'extrapolation sur la plage d'âge 95-100 ans considérant que leurs taux d'incidence estimés sont robustes entre 50 et 95 ans. Il en ressort que la fermeture linéaire offre un meilleur compromis par rapport à une extrapolation exponentielle dans l'optique de surmonter le manque d'informations en plus de présenter plusieurs avantages, dont sa simplicité et l'introduction d'un biais limité. Utiliser un allongement exponentiel de ces probabilités revient également à rajouter une nouvelle couche de prudence par rapport à nos taux précédemment estimés.

5.2 Calculs de primes et de provisions

5.2.1 Définitions

On introduit les notations suivantes :

— i le taux actuariel
— $\nu = \frac{1}{1+i}$
— R le montant de rente annuel

Plusieurs quantités doivent être considérées. Premièrement, les probabilités de survie en dépendance et la probabilité de rester autonome pendant k années s'écrivent respectivement :

$$ {}_kp^i_{x,t} = \prod_{j=0}^{k-1} \left[1 - q_i(x, t+j)\right] \quad ; \quad {}_kp^{aa}_x = \prod_{j=0}^{k-1} \left[1 - q_{aa}(x+j) - i(x+j)\right] $$

Ces montants permettent de définir la valeur actuelle probable d'une rente viagère de 1 euro à la fin de chaque année passée en dépendance et au début de chaque année passée en autonomie :

$$a^i_{x,t} = \sum_{k=1}^{\infty} \nu^k \, {}_kp^i_{x,t} \quad ; \quad \ddot{a}^{aa}_x = \sum_{k=0}^{\infty} \nu^k \, {}_kp^{aa}_x$$

Dans le cas des produits collectifs, la prime s'écrit :

$$P_x = i(x) R \; a^i_{x,0}$$

L'engagement de l'assuré et de l'assureur s'écrivent alors respectivement :

$$\Pi_x = P_x \; \ddot{a}^{aa}_x \quad ; \quad \Pi'_x = \sum_{k=0}^{\infty} \nu^k \, i(x+k) \, {}_kp^{aa}_x \, R \, a^i_{x+k,0}$$

Par égalisation des engagements de l'assureur et de l'assuré, la formule de la prime pure est la suivante :

$$P_x = R \frac{\sum_{k=0}^{\infty} \nu^k \, i(x+k) \, {}_kp^{aa}_x \, a^i_{x+k,0}}{\sum_{k=0}^{\infty} \nu^k \, {}_kp^{aa}_x}$$

De par son expression, la prime pure est sensible à plusieurs facteurs de risque. L'âge et l'incidence en dépendance font évoluer à la hausse cette quantité tandis que la mortalité en autonomie, la létalité en perte d'autonomie et le taux actuariel minorent la prime pure.

L'expression de la prime unisexe est la suivante :

$$\Pi^u_x = \alpha_m(x)\Pi^m(x) + \alpha_f(x)\Pi^f(x)$$

où $\alpha_m(x)$ (respectivement $\alpha_f(x)$) représente la proportion d'hommes (respectivement de femmes) parmi les nouveaux assurés x et $\Pi^m(x)$ (respectivement $\Pi^f(x)$) la prime calculée pour les hommes (respectivement les femmes).

La gradation du risque dépendance avec l'âge et le nécessaire maintien sur toute la plage d'assurance du tarif oblige l'assureur à constituer une provision pour risques croissants (PRC). Cette dernière prend, quant à elle, l'expression suivante :

$${}_tPRC_x = \Pi'_x - \Pi_x = \sum_{k=0}^{\infty} \nu^k i(x+k) \; {}_kp^{aa}_{x+t}(1+r) R a^i_{x+t+k,0} - P_x \ddot{a}^{aa}_{x+t}$$

5.2.2 Résultats

Pour les calculs, le taux actuariel annuel est fixé à 1% et le montant de la rente est fixé à 500 €. Puisque les tables de mortalité ont été construites par tranche d'âge d'incidence, des discontinuités apparaissent tous les 5 ans sur les courbes suivantes. Les taux de décès Qalydays sont moyennés sur chaque tranche d'âge d'incidence pour convenir au format des tables estimées.

Les trois figures suivantes retranscrivent les différences de mortalité qui subsistent entre les différents estimateurs. L'agencement entre elles des courbes de primes de risque est le reflet de l'espérance de vie résiduelle de chaque estimateur. Les fermetures de tables semblent contraindre drastiquement la survie en dépendance, si bien que les primes calculées avec les probabilités de Kaplan-Meier sont parfois bien inférieures à celles calculées avec les lois Qalydays. La PRC calculée avec les tables Qalydays finit par devenir négative au bout de 95 ans d'âge suggérant que l'engagement assuré surplombe l'engagement assureur. En parallèle, la PRC calculée sur les 3 autres estimateurs semble converger tout en restant positive.

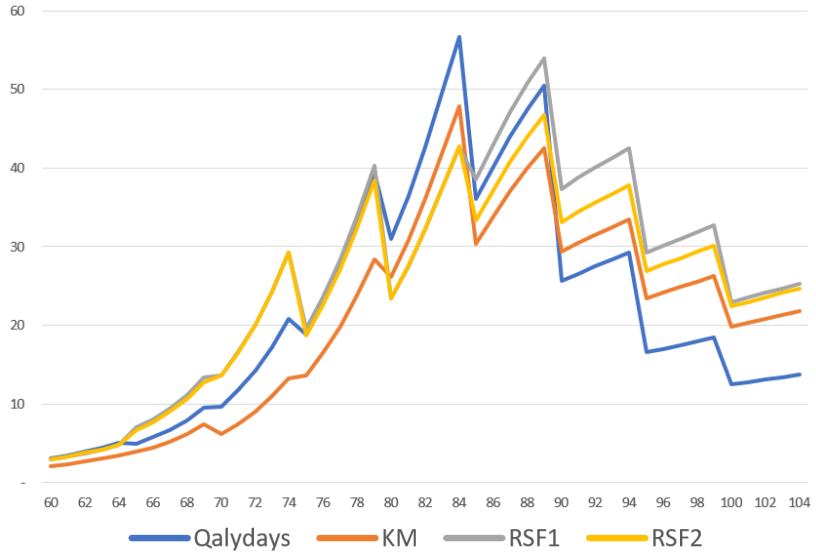


FIGURE 5.2 – Prime de risque en fonction de l'âge

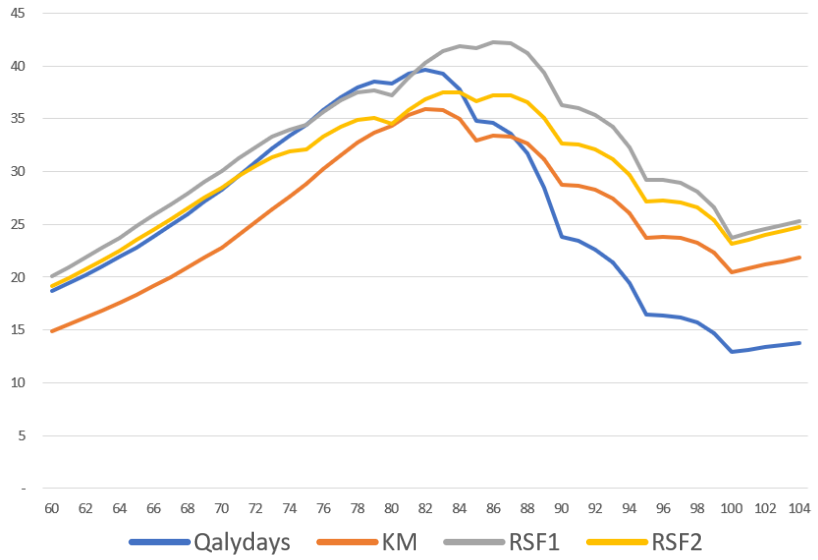


FIGURE 5.3 – Prime pure nivelée en fonction de l'âge

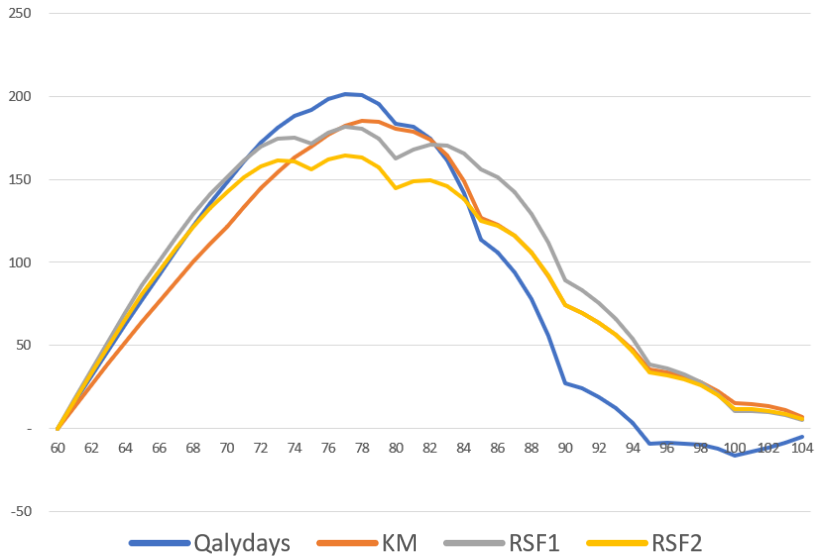


FIGURE 5.4 – PRC en fonction de l'âge pour un assuré qui a souscrit son assurance à 60 ans

Conclusion

Depuis plusieurs années, la sphère assurantielle française montre davantage d'attention aux contrats qui couvrent la perte d'autonomie lourde. La création du label GAD et les récentes recommandations de la CCSF en la matière ont insufflé de nouveaux élans à ce périmètre. La construction de lois de mortalité en dépendance totale est une préoccupation partagée par bon nombre d'acteurs de la place, si bien que le groupe de travail Dépendance de l'Institut des Actuaires s'est lui-même lancé sur de tels travaux. Ce mémoire poursuit également cet objectif par l'évaluation de taux américains de mortalité en dépendance qui sont ensuite positionnés à la population française. Face à la rareté des données dépendance sur l'aire de l'Hexagone, exploiter l'historique américain est la solution qui est explorée.

La construction de probabilité de décès brutes est l'opportunité d'y consacrer plusieurs méthodes d'estimation et de les comparer entre elles. Parmi ces dernières, les forêts aléatoires de survie présentent des performances très légèrement supérieures à l'arbre de survie et à l'estimateur de Kaplan-Meier d'après les mesures de C-index, de score de Brier ainsi que de l'AUC cumulative/dynamique. L'analyse de l'intelligibilité globale de la forêt aléatoire de survie désigne le diagnostic médical comme le facteur le plus significatif dans les prédictions du modèle. Cette constatation dénote avec notre calibration des lois qui dépend uniquement du sexe et de l'âge d'entrée en perte d'autonomie, alors que cette dernière variable n'apparaît que 5ème ou 6ème parmi les 10 variables explicatives qui figurent dans la base de données SOA LTC 2000-2011. La distinction entre la dépendance d'ordre physique et la dépendance d'ordre cognitive semble être un niveau de segmentation pertinent à employer.

Plusieurs méthodes d'ajustement de la mortalité, qui sont généralement dédiées à des cohortes plus standards sont employées comme le modèle de Brass, 1971 ou encore l'extrapolation de Denuit et Goderniaux, 2005. La méthodologie de positionnement mise en œuvre apparaît insuffisante et résulte systématiquement à un SMR supérieur à 1,26 pour tous les estimateurs. La sous-évaluation de la mortalité se cristallise premièrement autour de la première année en perte d'autonomie et secondement à forte ancienneté en perte d'autonomie. La fermeture des tables de mortalité remédie en partie à ces dissonances avec les lois françaises Qalydays.

Ultimement, les travaux de ce mémoire se consacrent exclusivement à la survie en perte d'autonomie totale. La dépendance partielle est un risque nouveau et tout aussi préoccupant, pour lequel la carence en données se manifeste d'autant plus. L'estimation des lois qui y sont relatives se présente également comme un sujet majeur. L'ajustement des probabilités de transitions entre l'état de dépendance partielle vers le stade lourd, et inversement, montre un grand intérêt pour étayer le modèle illness-death dépendance. Ce dernier pourrait se révéler de plus en plus réducteur face au vieillissement de la population et à l'envolée de la prévalence en perte d'autonomie dans les années à venir.

Table des acronymes

— ADL : Activities of Daily Living
— AGGIR : Autonomie Gérontologique Groupes Iso-Ressources
— AIC : Akaike Information Criterion
— ALF : Assisted Living Facilities
— APA : Allocation Personnalisée Autonomie
— AUC : Area Under the Curve
— AVC : Accident Vasculaire Cérébral
— AVQ : Acte de la Vie Quotidienne
— BIC : Bayesian Information Criterion
— BS : Brier Score
— CART : Classification And Regression Tree
— CCSF : Comité Consultatif du Secteur Financier
— CI : Cognitive Impairment
— COLA : Cost Of Living Adjustment
— CRPS : Continuous Rank Probability Score
— DREES : Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
— EA : Erreur Absolue
— EHPA : Établissement d'Hébergement pour Personnes Âgées
— EHPAD : Établissement d'Hébergement pour Personnes Âgées Dépendantes
— EOOB : Erreur Out-Of-Bag
— EP : Elimination Period
— EQM : Erreur Quadratique Moyenne
— GAD : Garantie Assurance Dépendance
— GCV : Generalized Cross Validation
— GIR : Groupes Iso-Ressources
— HAD : Hospitalisation À Domicile
— HHC : Home Health Care
— HIPAA : Health Insurance Portability and Accountability Act
— IB : In-Bag
— IBS : Intégrale du Score de Brier
— i.i.d : Indépendant et Identiquement Distribué
— INSEE : Institut National de la Statistique et des Études Économiques
— KM : Kaplan-Meier
— LIMRA : Life Insurance Marketing and Research Association
— LTC : Long-Term Care

— MAPE : Mean Absolute Percentage Error
— MCO : Médecine, Chirurgie, Obstétrique
— MMS : Mini Mental Score
— MP : Minimal Depth
— NH : Nursing Home
— OOB : Out-Of-Bag
— PIB : Produit Intérieur Brut
— PMSI : Programme de Médicalisation des Systèmes d'Information
— PRC : Provision pour Risques Croissants
— RSF : Random Survival Forest
— SMR : Standardized Mortality Ratio
— SNF : Skilled Nursing Facilities
— SSR : Soins de Suite et de Réadaptation
— ST : Survival Tree
— TQ : Tax-Qualified
— USLD : Unité de Soins Longue Durée
— VI : Variable Importance
— VIMP : Variable Importance Minimal Depth
— WH : Whittaker-Henderson

Bibliographie

Algava, É. et Blanpain, N. (2021). 68,1 millions d'habitants en 2070 : une population un peu plus nombreuse qu'en 2021, mais plus âgée. url : https://www.insee.fr/fr/statistiques/5893969.

Berrar, D. et al. (2019). Cross-validation.

Biessy, G. (2023). Revisiting whittaker-henderson smoothing. *arXiv preprint arXiv :2306.06932*.

Biessy, G. (2024). Assurance dépendance. Cours dispensé dans le cadre du M2 Actuariat ISUP.

Brass, W. (1971). On the scale of mortality. *Biological aspects of demography*.

Breiman, L. (1984). Classification and regression trees.

Breiman, L. (2001). Random forests. *Machine learning* 45, p. 5-32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA* 1.58, p. 3-42.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review* 78.1, p. 1-3.

Cazaubiel Arthur et El Guendouz, A. (2022). D'ici 2070, un tiers des régions perdraient des habitants. url : https://www.insee.fr/fr/statistiques/6658362.

CCSF (2024). Pour une meilleure protection des personnes dépendantes et de leur famille : Le Contrat Dépendance Solidaire. Communiqué de presse.

Coale, A. et Guo, G. (1989). Revised regional model life tables at very low levels of mortality. *Population index*, p. 613-643.

Coale, A. et Kisker, E. E. (1990). Defects in data on old age mortality in the United States : New procedures for calculating approximately accurate mortality schedules and life tables at the highest ages. *Asian and Pacific Population Forum*. T. 4. 1, p. 1-31.

CORNUAILLE, C. (s. d.). Le risque dépendance : l'expérience du marché américain peut-elle permettre d'anticiper les évolutions du marché français ? Mémoire d'Actuariat.

Corrada, M. M., Brookmeyer, R., Paganini-Hill, A., Berlau, D. et Kawas, C. H. (2010). Dementia incidence continues to increase with age in the oldest old : the 90+ study. *Annals of neurology* 67.1, p. 114-121.

Cox, D. R. et Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society : Series B (Methodological)* 30.2, p. 248-265.

Crimmins, E. M. et Beltrán-Sánchez, H. (2011). Mortality and morbidity trends : is there compression of morbidity ? *Journals of Gerontology Series B : Psychological Sciences and Social Sciences* 66.1, p. 75-86.

Debón, A., Montes, F. et Sala, R. (2006). A comparison of nonparametric methods in the graduation of mortality : Application to data from the Valencia region (Spain). *International Statistical Review* 74.2, p. 215-233.

Denuit, M. et Goderniaux, A.-C. (2005). Closing and projecting lifetables using log-linear models. *Bulletin of the Swiss Association of Actuaries*, p. 29.

Descriptif du contenu des bases de données PMSI (2023). URL : https://www.atih.sante.fr/bases-de-donnees/descriptif-du-contenu-des-bases-de-donnees-pmsi.

DREES (2015). Enquête EHPA.

DREES (2016). Enquêtes Capacités, Aides et REssources (CARE) des séniors-institutions, volet seniors.

DREES (2019). Enquête EHPA.

DREES (2020). Répartition par sexe et par âge des bénéficiaires de l'APA, d'aides ménagères et de l'ASH en établissement, en décembre 2019. URL : https://data.drees.solidarites-sante.gouv.fr/api/v2/catalog/datasets/les-caracteristiques-des-beneficiaires-de-l-aide-sociale-departementale-aux-pers/attachments/pa_beneficiaires_par_gir_sexe_et_age_apa_ash_aides_menageres_donnees_2020_xlsx.

Dupourqué, E., Planchet, F. et Sator, N. (2019). Actuarial aspects of long term care. Springer.

Fisher, A., Rudin, C. et Dominici, F. (2019). All models are wrong, but many are useful : Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20.177, p. 1-81.

Fleming, T. R. et Harrington, D. P. (2013). Counting processes and survival analysis. T. 625. John Wiley & Sons.

Forfar, D., McCutcheon, J. et Wilkie, A. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries* 115.1, p. 1-149.

France Assureurs (s. d.). Le label GAD assurance dépendance : l'accompagnement du bien vieillir et la prise en charge de la perte d'autonomie. URL : https://www.franceassureurs.fr/wp-content/uploads/2022/09/3-24_vf_plaquette-gad-2022.pdf.

France Assureurs (2021). Construire une nouvelle solution solidaire et transparente face à la dépendance liée à l'âge.

Genuer, R. et Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables.

Graf, E., Schmoor, C., Sauerbrei, W. et Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* 18.17-18, p. 2529-2545.

Grasset, L., Brayne, C., Joly, P., Jacqmin-Gadda, H., Peres, K., Foubert-Samier, A., Dartigues, J.-F. et Helmer, C. (2016). Trends in dementia incidence : evolution over a 10-year period in France. *Alzheimer's & Dementia* 12.3, p. 272-280.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. et Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama* 247.18, p. 2543-2546.

Henderson, R. (1924). A new method of graduation. *Transactions of the Actuarial Society of America* 25, p. 29-40.

Huchon, M. (2021). Sélection de variables à l'aide de forêts aléatoires pour données de survie de grande dimension.

Hung, H. et Chiang, C.-T. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* 38.1, p. 8-26.

Hurvich, C. M. et Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics*, p. 1077-1084.

ICD-10 Version :2008 (2008). URL : https://icd.who.int/browse10/2008/fr#/V.

INSEE (2021). Estimations de population et projections de population 2021-2070. URL : https://www.insee.fr/fr/statistiques/5893969?sommaire=5760764.

INSEE (2024a). Composantes de la croissance démographique. URL : https://www.insee.fr/fr/outil-interactif/5367857/details/20_DEM/21_POP/21D_Figure4.

INSEE (2024b). Pyramide des âges Données annuelles 2024. URL : https://www.insee.fr/fr/statistiques/2381472.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. et Lauer, M. S. (2008). Random survival forests.

Ishwaran, H., Kogalur, U. B., Chen, X. et Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining : The ASA Data Science Journal* 4.1, p. 115-132.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. et Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association* 105.489, p. 205-217.

Ishwaran, H. et Lu, M. (2018). Random survival forests.

Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* 53.282, p. 457-481.

Lambert, J. et Chevret, S. (2016). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical methods in medical research* 25.5, p. 2088-2102.

Le Figaro (2022). Le coût élevé des soins de cancer aux États-Unis n'améliore pas le taux de survie, selon une étude. *Le Figaro*.

Libault, D. (2019). Concertation Grand âge et autonomie. url : https://sante.gouv.fr/IMG/pdf/rapport_grand_age_autonomie.pdf.

Lindbergson, M. (2001). Mortality among the elderly in Sweden 1988–1997. *Scandinavian Actuarial Journal* 2001.1, p. 79-94.

Lois biométriques pour le risque de perte d'autonomie en France (2020). url : https://www.ressources-actuarielles.net/qalydays.

Lu, M. (2023). Random Survival Forests. url : https://www.randomforestsrc.org/articles/survival.html.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv :1705.07874*.

Planchet, F et Tomas, J (2014). Méthodes de positionnement : Aspects méthodologiques. *Note de travail de l'Institut des Actuaires II1291-15 v1* 7.

Planchet, F., Guibert, Q. et Schwarzinger, M. (2018a). Mesure De L'Espérance De Vie En Dépendance Totale En France. *Bulletin Français d'Actuariat*.

Planchet, F., Guibert, Q. et Schwarzinger, M. (2018b). Mesure du risque de perte d'autonomie totale en France métropolitaine. *Bulletin Français d'Actuariat*.

Quashie, A. et Denuit, M. (2005). Modèles d'extrapolation de la mortalité aux grands âges. *Institut des Sciences Actuarielles et Institut de Statistique, Université Catholique de Louvain, WP*.

RAPIOR, H. (2021). Création de tables de mortalité de dépendants dans le marché japonais, via la transposition de la dépendance en France. Mémoire d'Actuariat. ISFA.

Rocca, W. A., Petersen, R. C., Knopman, D. S., Hebert, L. E., Evans, D. A., Hall, K. S., Gao, S., Unverzagt, F. W., Langa, K. M., Larson, E. B. et al. (2011). Trends in the incidence and prevalence of Alzheimer's disease, dementia, and cognitive impairment in the United States. *Alzheimer's & Dementia* 7.1, p. 80-93.

Schwarzinger, M. (2019). Etude Qalydays : données source et retraitements pour l'étude du risque de perte d'autonomie. *Bulletin Français d'Actuariat*.

Schwarzinger, M., Pollock, B. G., Hasan, O. S., Dufouil, C., Rehm, J., Baillot, S, Guibert, Q., Planchet, F et Luchini, S (2018). Contribution of alcohol use disorders to the burden of dementia in France 2008–13 : a nationwide retrospective cohort study. *The Lancet Public Health* 3.3, e124-e132.

SCOR Global Life (2012). Assurance Dépendance.

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games* 2.

Society of Actuaries (2015a). Long Term Care Experience Basic Table Development.

Society of Actuaries (2015b). Long Term Care Intercompany Experience Study - Aggregate Database 2000-2011 Report.

Society of Actuaries (2017). Caveats for Use of Long Term Care Experience Basic Tables.

Spytek, M., Krzyziński, M., Langbein, S. H., Baniecki, H., Wright, M. N. et Biecek, P. (2023). survex : an R package for explaining machine learning survival models. *Bioinformatics* 39.12, btad723.

Stallard, P et Yashin, A (2016). LTC Morbidity Improvement Study : Estimates for the Non-Insured US Elderly Population Based on the National Long Term Care Survey 1984–2004. *Society of Actuaries, Schaumburg, IL*.

Štrumbelj, E. et Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41, p. 647-665.

Thatcher, A. (1998). The force of mortality at ages 80 to 120.

Tomas, J et Planchet, F (2013). Critères de validation : aspects méthodologiques. *Institut des Actuaires*.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. et Wei, L.-J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* 30.10, p. 1105-1117.

Uno, H., Cai, T., Tian, L. et Wei, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* 102.478, p. 527-537.

Whittaker, E. T. (1922). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society* 41, p. 63-75.

Annexe A

Annexes

A.1 Statistiques descriptives supplémentaires

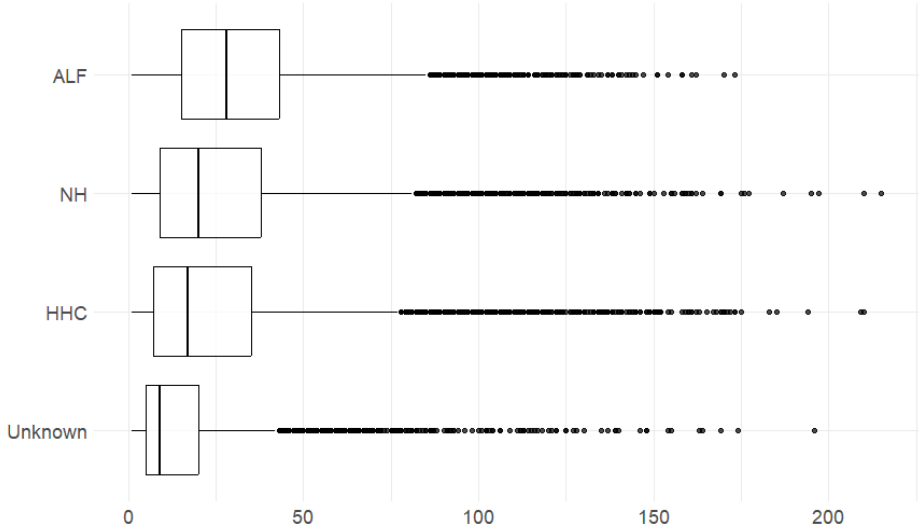
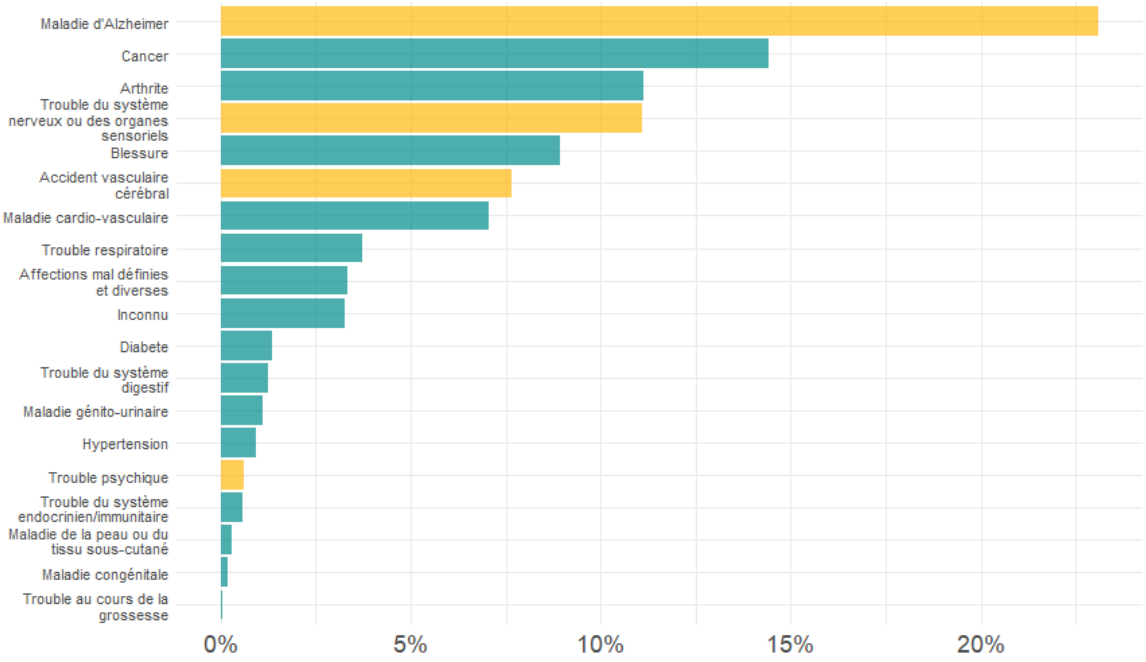


FIGURE A.1 – Durées en mois passées en sinistre en fonction du lieu de résidence dans la base de données retraitée SOA LTC 2000-2011

HHC



ALF, NH and Unknown

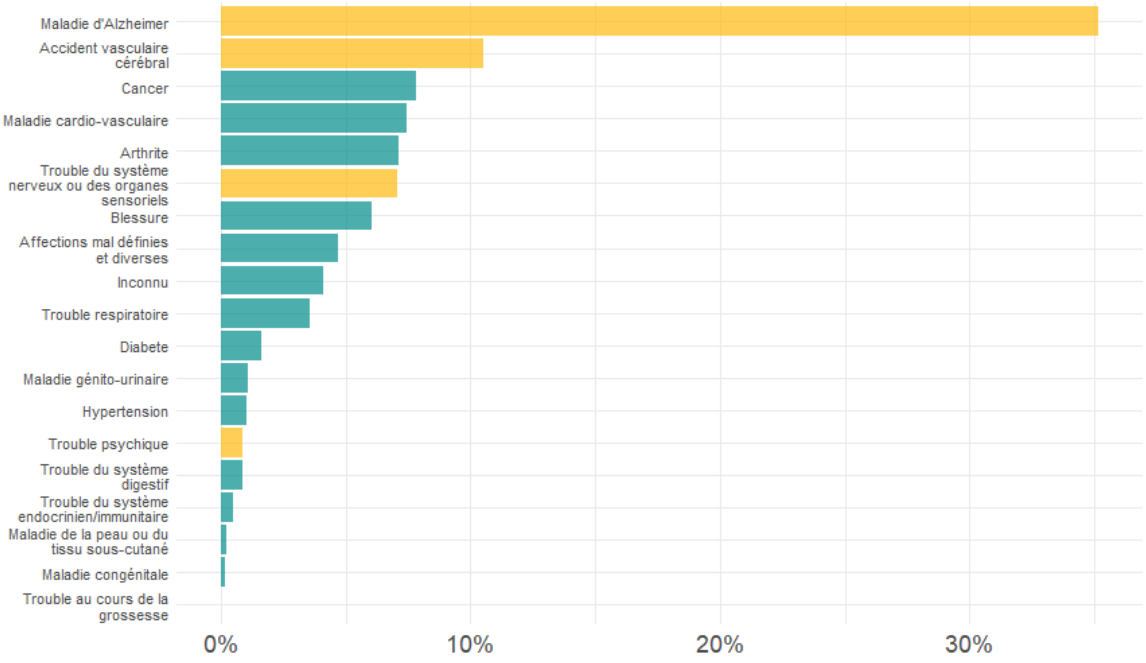


FIGURE A.2 – Répartition des pathologies dans la base de données retraitée SOA LTC 2000-2011. Les diagnostics relatifs à la perte d'autonomie d'ordre cognitive sont mis en évidence en jaune.

A.2 Examen et performances de l'arbre de survie et de la forêt aléatoire de survie

	Gender	*ClaimType*	*IncurredAgeBucket*	*Infl_Rider_Bucket*	*GroupIndicator*	*Cov_Type_Bucket*
1	Female	ALF	60 to 64	GPO	Group	Comprehensive
2	Male	HHC	65 to 69	Inflation	Individual	Other
3	...	NH	70 to 74	None	...	...
4	...	Unknown	75 to 79	Unknown	...	...
5	...	...	80 to 84	...	...	...
6	...	...	85 to 89	...	...	...

TABLE A.1 – Encodage numérique des variables qualitatives, qui comptent moins de 6 modalités, de la base de données SOA LTC 2000-2011

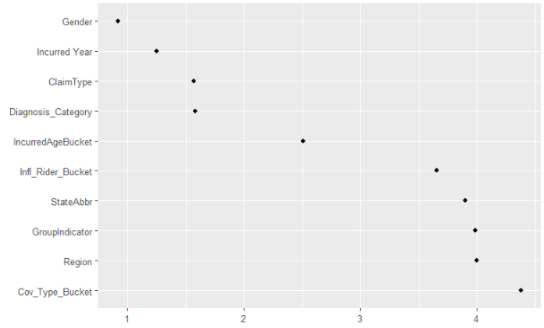


FIGURE A.3 – Profondeurs minimales de toutes les variables explicatives de la base de données SOA LTC 2000-2011

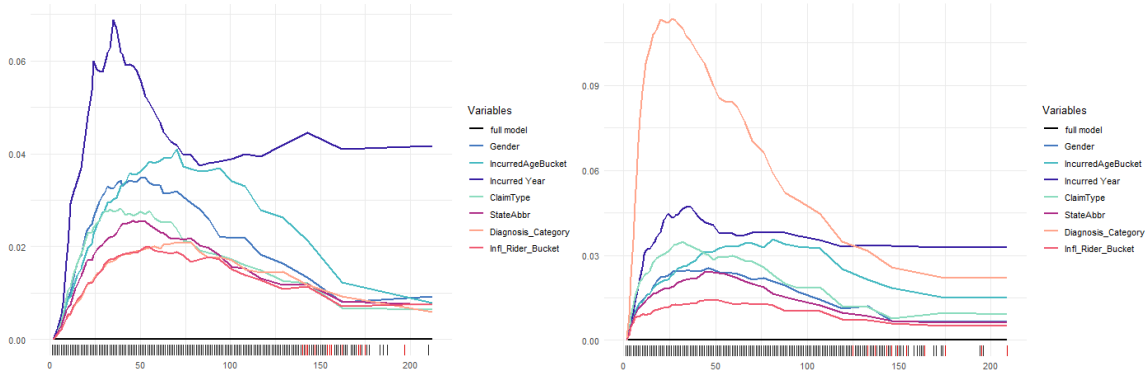


FIGURE A.4 – Importance par permutation des variables en fonction de l'ancienneté en dépendance en mois avec le score de Brier comme fonction de perte. De gauche à droite, pour les seniors atteints de dépendance cognitive et de dépendance physique.

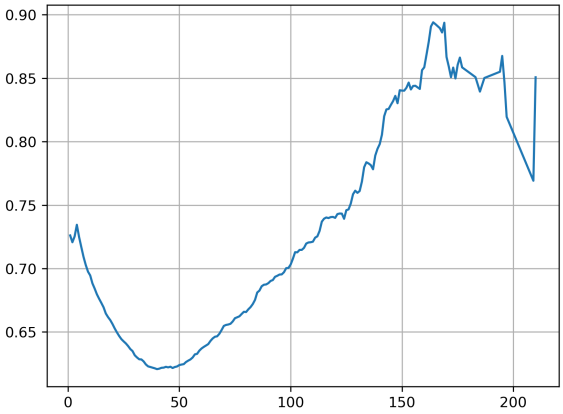


FIGURE A.5 – AUC cumulative/dynamique de la forêt aléatoire de survie en fonction du temps passé en perte d'autonomie

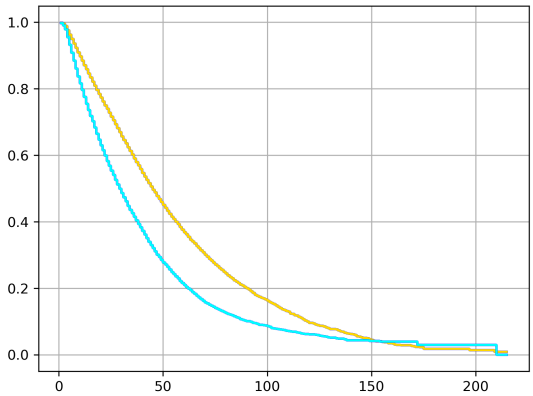


FIGURE A.6 – Fonctions de survie de l'arbre de survie par rapport à l'ancienneté en perte d'autonomie en mois, en orange pour les femmes et en turquoise pour les hommes

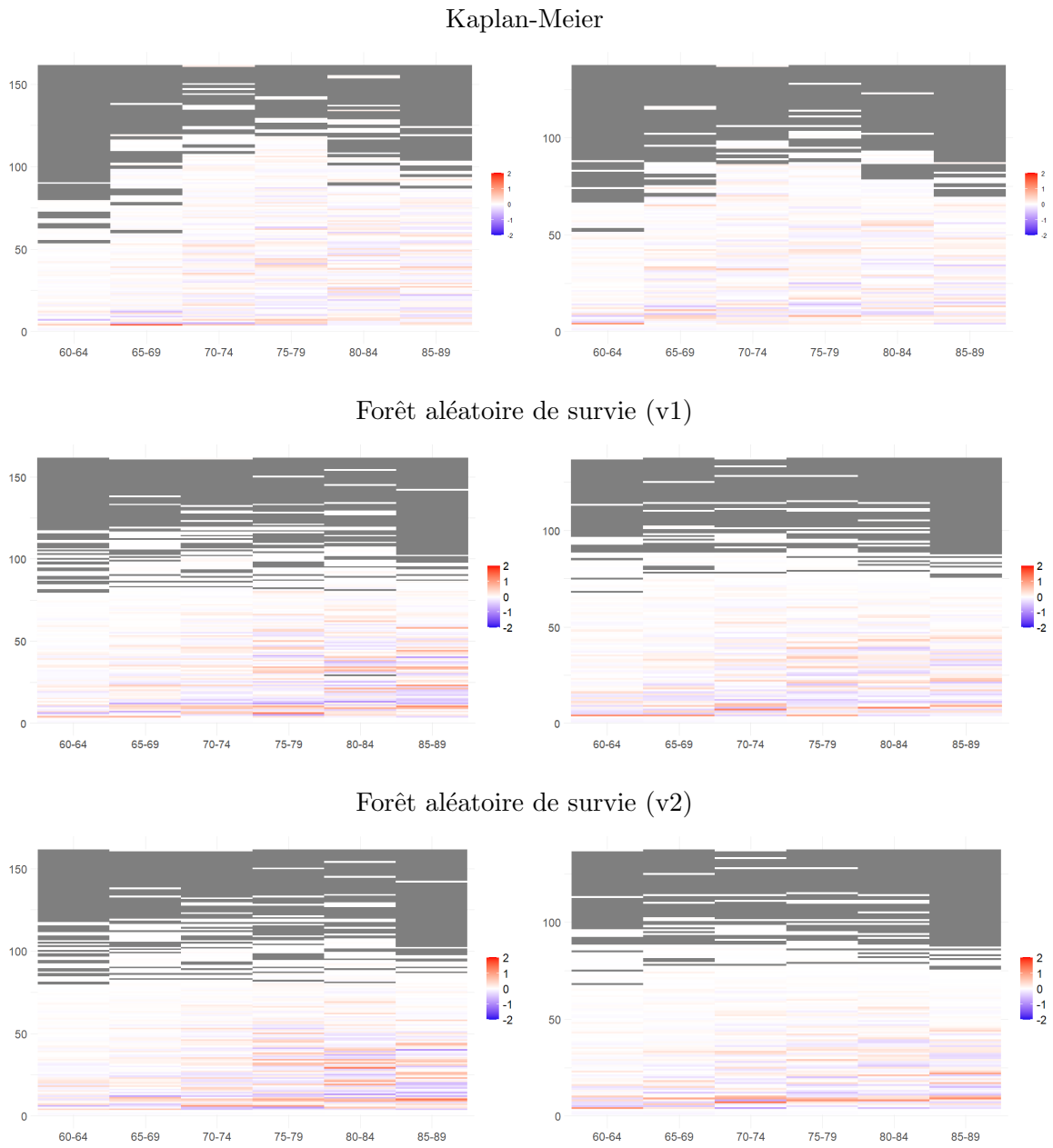


FIGURE A.7 – Résidus de Pearson consécutifs au lissage des taux de décès à 1 mois en dépendance. De gauche à droite respectivement pour les femmes et les hommes. Aucun taux n'est défini sur les plages grises.