Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires

le

Par : Lou SEMI
Titre : Étude des déviations de mortalité par partitionnement récursif et modèle linéaire généralisé.

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : SCOR
Signature :

J.T

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Directeur de Mémoire en entreprise :
Nom : Julien TOMAS
Signature :

J.T

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

Secrétariat :

Signature du responsable entreprise

J.T

Bibliothèque :

Signature du candidat

L.S

Résumé

La mortalité est au cœur du métier des compagnies de réassurane détenant du risque vie. Pour celles-ci, il est indispensable que les tables de mortalité soient calibrées au mieux afin de réduire la volatilité de leur résultat. Dans ce contexte, une analyse comparative de la mortalité observée et prédite est nécéssaire afin d'assurer un suivi de la mortalité du portefeuille.

Ce mémoire d'actuaire porte sur l'étude de la mortalité de la population assurée et de la population générale américaine à l'aide de deux algorithmes de partitionnement récursif. L'objectif sera d'identifier les segments de la population qui dévient significativement par rapport aux comportements capturés par la table de l'industrie et par rapport à la population générale américaine.

Dans un premier temps, nous allons étudier les algorithmes de partitionnement récursif que sont le GLM Tree et le CART. Le GLM tree repose en grande partie sur l'algorithme générique MOB développé par Zeileis et al. (2008) et utilisé pour répartir les données en plusieurs groupes qui diffèrent en termes des paramètres du modèle. Nous commencerons donc par décrire l'algorithme générique MOB et analyserons en détail ses différentes étapes. Ensuite, nous présenterons la méthode utilisée dans notre cas pour étudier la mortalité et capturer ses déviations. De même, l'algorithme non-paramétrique CART, développé par Breiman et al. (1984) fera l'objet d'une description de sa méthodologie. Nous présenterons finalement les différences entre ces deux algorithmes.

Dans un second temps, à la suite de cette première partie théorique, nous passerons aux applications. D'une part, les algorithmes GLM Tree et CART seront appliqués aux données de l'industrie américaine où la mortalité observée sera comparée à celle prédite par la table réglementaire 2015 VBT. Cette première application aura pour but de capturer des déviations significatives en fonction de divers paramètres que sont l'âge, le sexe, le montant assuré, le type de produit, la durée passée dans le contrat ainsi que la classe de risque (liée au statut fumeur ou non de l'individu). D'autre part, nous mènerons une analyse plus générale et démographique sur la population américaine. L'objectif reste toujours celui de capturer des déviations, mais cette fois par rapport à la mortalité moyenne et en utilisant d'autres variables telles que l'ethnie, le niveau d'éducation et les causes de décès.

Mots-clés : *Partitionnement récursif*, *Modèle Linéaire Généralisé*, *Arbre de décision*, *Déviations de mortalité*.

Abstract

Mortality is at the heart of the business of reinsurance companies holding life risks. For the latter, it is essential to calibrate as well as possible the mortality tables in order to reduce the volatility of their results. In this context, a comparative analysis of the observed and predicted mortality is necessary in order to ensure a follow-up of the portfolio mortality.

This actuarial thesis focuses on the study of the mortality of the insured population and the general american population using two recursive partitioning algorithms. The objective will be to identify segments of the population that differ significantly from the behaviors captured by the industry and general US population tables.

First, we will study the recursive partitioning algorithms GLM Tree and CART. The GLM tree is largely based on the MOB general algorithm developed by Zeileis et al. (2008) and used to partition the data into groups that differ in terms of the model parameters. We will therefore start by describing the generic MOB algorithm and analyzing in detail its different steps. Then, we will present the method used in our case to study the mortality and capture its deviations. Similarly, the non parametric algorithm CART, developed by Breiman et al. (1984) will be described in terms of its methodology. We will finally present the differences between these two algorithms.

Second, after this first theoretical part, we will move to the applications. On the one hand, the GLM Tree and CART algorithms will be applied to the US industry data where the observed mortality will be compared to the one predicted by the standard table 2015 VBT. The objective of this first application is to capture significant differences based on various parameters such as age, gender, amount insured, type of product, duration of the contract as well as risk class (related to the smoking status of the individual). On the other hand, we will carry out a more general analysis on the american population. The objective remains the same, that of capturing deviations but in relation to the average mortality this time and using other variables such as ethnicity, education level and causes of death.

Key words: *Recursive Partitioning*, *Generalized Linear Models*, *Decision Tree*, *Mortality Deviations*.

Note de Synthèse

Les compagnies de réassurance possédant du risque vie se doivent d'étudier la mortalité de leur portefeuille. Lorsque les taux de décès sont mal calibrés par les compagnies d'assurance, cela se répercute sur la compagnie de réassurance. Dans un tel contexte, il est indispensable pour le réassureur d'étudier la mortalité de ses portefeuilles afin de déterminer les segments de la population pour lesquels la mortalité dévie des hypothèses et prédictions faites. Outre le cas où les taux sont bien calibrés, il y aura soit un risque de mortalité ou de longévité pour les différents profils du portefeuille considéré. Que la mortalité soit plus ou moins élevée par rapport à la prédiction, le résultat du réassureur se verra impacter. Il paraît alors nécéssaire de régulièrement étudier la mortalité de son portefeuille et de mettre en place des actions de management en cas d'impact non négligeable sur la compagnie de réassurance. Ces actions de management sont définis dans le traité de réassurance signé entre l'assureur et le réassureur.

Le risque vie occupe donc une place centrale dans le résultat des compagnies de réassurance. Ici, nous allons nous concentrer sur la mortalité et étudier ses déviations au travers du ratio A/E (pour Actual/Expected) en nombre de décès. L'idée est d'identifier les segments de la population pour lesquels la mortalité observée ne concorde pas à celle prédite.

Une première application sera réalisée sur la population assurée américaine. Les données proviennnent de la SOA (Society Of Actuaries) et sont aggrégées sur la période allant de 2003 à 2013. Cette base contient les données de différentes compagnies américaines volontaires. Les résultats obtenus ne sauraient être représentatifs d'une compagnie en particulier. Ces observations seront comparées aux données de prédiction issues de la table règlementaire 2015 VBT (Valuation Basic Table). Cette table est construite de sorte à rajouter une marge sur les taux de mortalité afin de protéger les assureurs. De plus, elle repose sur des montants. Les compagnies d'assurance vie utilisent des tables de mortalité d'évaluation pour déterminer le montant des liquidités qu'elles sont tenues par la loi de mettre de côté pour les sinistres et les prestations.

Les premiers résultats obtenus sont issus du partitionnement avec l'algorithme GLM Tree. Celui-ci repose sur l'algorithme générique MOB, Zeileis et al. (2008) utilisé pour diviser les données en des groupes qui diffèrent en termes des paramètres du modèles. La particularité de cet algorithme repose sur son lien avec les modèles linéaires généralisés. Il est possible d'utiliser différentes lois de distribution pour modéliser nos données. Les tests pour le partitionnement basés sur les Mfluctuations Tests diffèrent selon que la variable de partitionnement utilisée est numérique ou catégorielle.

Dans notre cadre d'étude, le nombre de décès A est supposé être distribué de manière binomiale. La déviation est ainsi capturée par l'intercept β_0 definit par :

$$\beta_0 = \ln\left(\frac{\mathbb{E}(A)}{E}\right),$$

Où E correspond au nombre de décès espéré.

Ce choix de la loi binomiale nécéssite une approximation dans notre modèle. En effet, manipuler

des taux de décès q de l'ordre de 10^{-3} permet de faire l'approximation suivante :

$$1 - q \approx 1.$$

Cette approximation impacte donc la précision de la déviation calculée sans toutefois remettre en cause l'analyse. Il aurait été possible de se passer de cette approximation en utilisant la loi de poisson pour modéliser le nombre de décès. Cependant, la conception actuelle de l'algorithme ne permet pas d'intégrer les poids de manière convenable afin de définir la taille des groupes. En effet, les groupes avec cette loi de distribution sont basés sur le nombre de lignes de la base et non sur l'exposition.

Nous obtenons l'arbre de la figure 1 à l'issu du partitionnement de nos données avec le GLM Tree.

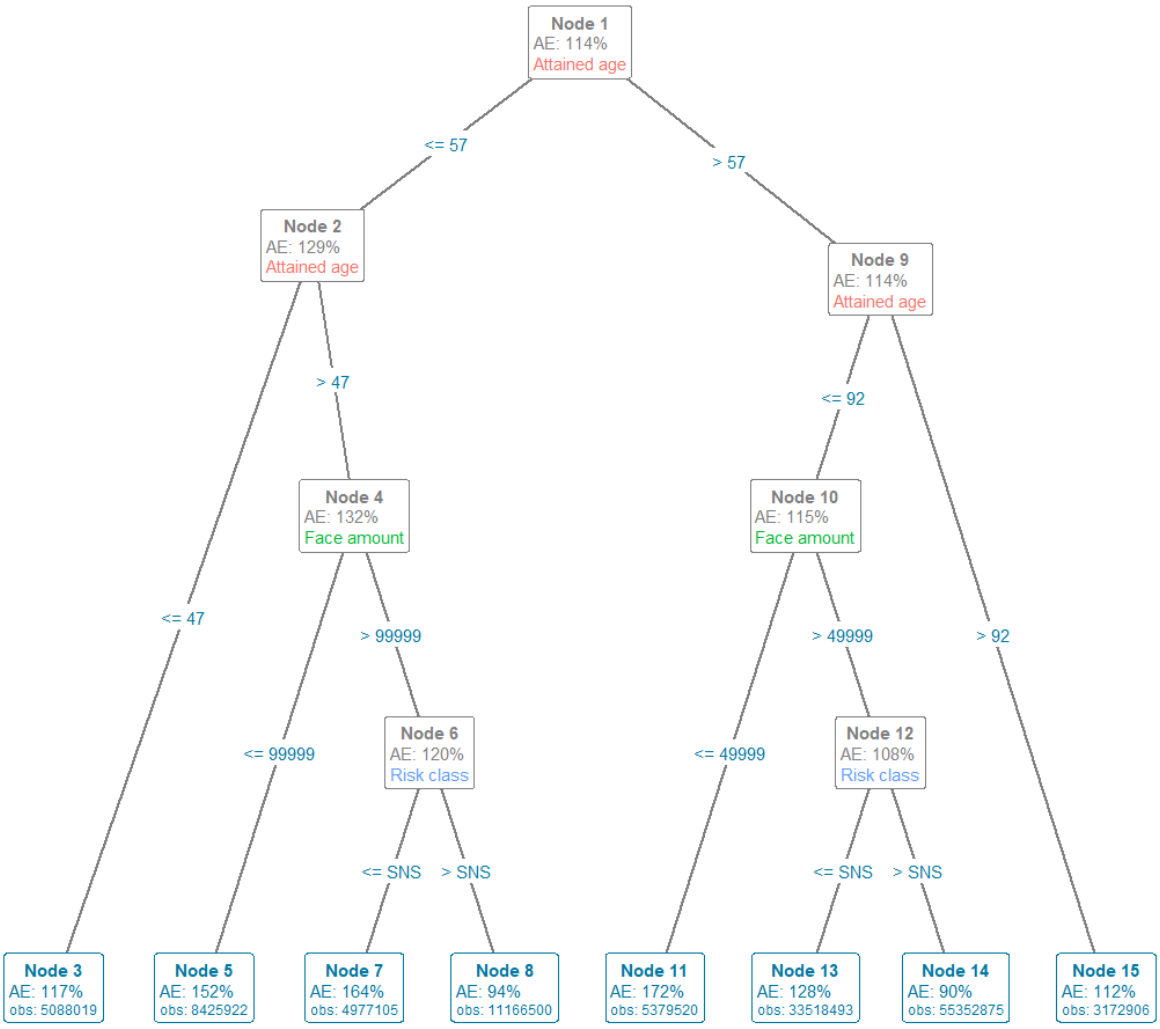


FIGURE 1: Arbre issu du partitionnement des données

Au niveau global, la première déviation obtenue s'élève à 114% : la table de référence sous estime et ne capture donc pas de façon adéquate le comportement de mortalité des assurés qui sont représentés dans cette base.

L'ordre des variables utilisées pour le découpage montre quel différentiel de mortalité est le plus important. Parmi les six variables de partitionnement utilisées dans la fonction (Attained age, Face amount band, Gender, Risk class, Duration et Insurance plan), seules trois sont utilisées pour le partitionnement. Il s'agit de :

- Attained age (x2),
- Face amount,
- Risk class.

Nous observons de fortes déviations pour tous les âges en regardant les segments suivants :

– Celui des faibles montants assurés (les nœuds 5 et 11).

– Celui des personnes en moins bonne santé (SS + PS + SNS) ayant de grands montants assurés (les nœuds 7 et 13).

Notons que les différents paramètres dans cet arbre ont été choisis de sorte à avoir une taille raisonnable pour les différents groupes. La stratégie de post-élagage en utilisant un critère d'information tel que le BIC ou le AIC n'est pas possible dans notre cas. En effet, la régression se fait uniquement selon l'intercept. Il n y a donc pas de variables explicatives sujettes aux critères d'information. L'alternative basée sur l'utilisation d'une faible valeur du niveau de confiance α et proposée par Zeileis et al. (2008) va se reveler inutile. En effet, l'arbre construit est stable selon les valeurs de ce paramètre α. Les tests de sensibilité des autres paramètres *minsize* et *maxdepth* pour respectivemet le minimum d'observations dans un nœud et la profondeur maximale de l'arbre vont donner des structures légèrement différentes de l'arbre initial. Toutefois, les résultats resteront inchangés.

Le graphique 2 apporte des informations supplémentaires à propos des feuilles de l'arbre.

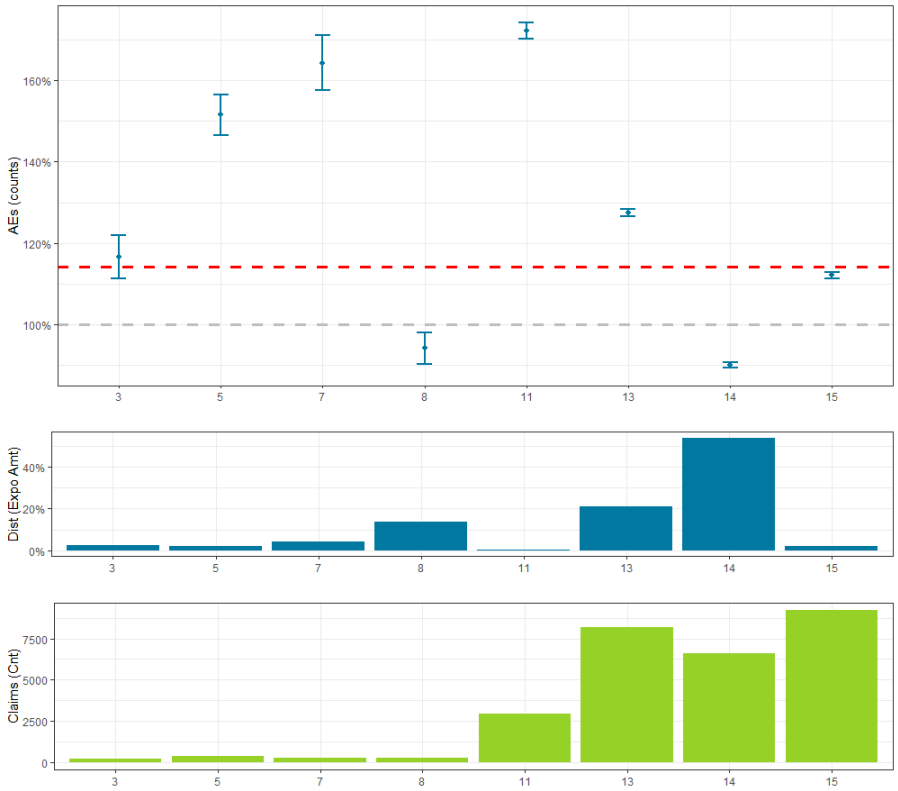


Figure 2: Ratio des décès observés et prédits par rapport à la table 2015 VBT

Outre les déviations observées et leurs intervalles de confiance, on y retrouve pour chacun des nœuds terminaux, la proportion des montants assurés et le nombre de décès. La ligne grise en pointillé

illustre le ratio A/E de 100% qui indique l'adéquation parfaite entre les tables de l'industrie et la 2015 VBT. La ligne rouge en pointillé est le ratio A/E à 114% qui illustre la déviation globale de la table par rapport à la table réglementaire. Dans ce cas, les intervalles de confiance proposés par Liddell (1984) sont construits non à partir d'approximation mais à l'aide d'une formule exacte. Ces derniers permettront d'évaluer le caractère significatif ou non des déviations obtenues.

En observant le nœud 13 (avec les individus fumeurs aux montants élevés), il se trouve que la déviation affichée (128%) est supérieure au ratio global de 114% et encore plus au ratio à 100%. En outre, les montants pour ce nœud représentent plus de 20% des montants assurés et la sinistralité correspondante est également très forte (> 7500). Ce type de segment est préoccupant car assez coûteux pour une compagnie. Il doit donc être surveillé.

Nous allons donc l'étudier au plus près afin d'estimer l'impact de ce genre de déviation. Rappelons que ce nœud 13 concerne les individus âgés de 62 à 92 ans, qui font partie de la classe de risque la moins saine (les fumeurs donc) et qui ont des montants assurés élevés. Pour ces individus, la déviation obtenue est de 128%. En s'intéressant au taux de mortalité observé dans nos données par rapport au taux prédit par la table 2015 VBT, on s'aperçoit que le taux de mortalité observé est généralement au dessus de celui prédit par la table.

Ces différents taux de mortalité seront utilisés pour tarifer un contrat d'assurance en cas de décès. L'idée est de faire des hypothèses sur les montants et les taux d'actualisation afin de déterminer les flux potentiels que devraient débourser une compagnie sur une projection à 30 ans. Nous obtenons alors la figure 3.

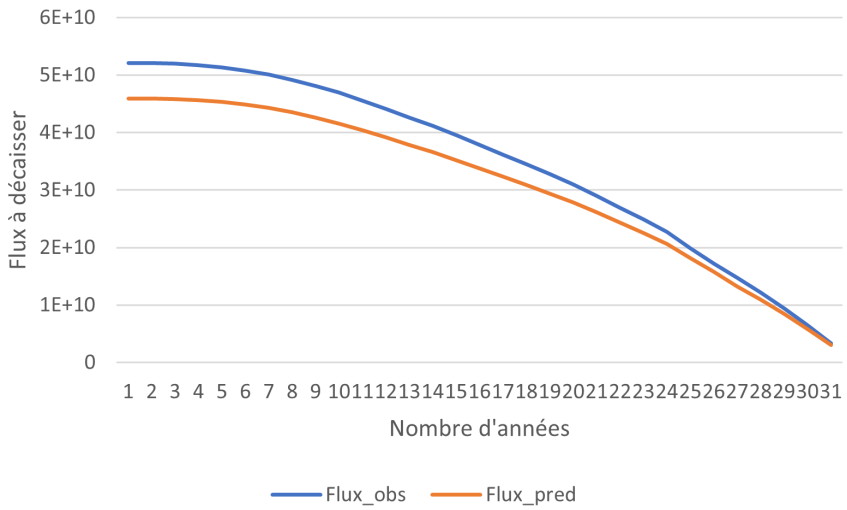


Figure 3: Comparaison des flux observés et prédits en fonction du nombre d'années projetées

On s'aperçoit qu'avec une déviation de mortalité à 128% dans ce nœud, les flux observés, c'est à dire, ceux à réellement décaisser pour couvrir les bénéficiaires des assurés sont nettement plus élevés que ceux qu'on aurait prédit avec la table standard. L'utilisation de taux différents conduit à un écart dans les flux s'élevant à 11%. Un écart de mortalité de presque 30% peut donc conduire à des pertes considérables pour une compagnie d'assurance. D'où l'importance de calibrer au mieux nos taux de mortalité. Après une telle étude, il est possible de mettre en place des actions de management afin de gérer ce segment de la population.

Afin d'évaluer la pertinence du GLM Tree, la méthode CART introduite par Breiman et al. (1984) a été appliquée aux mêmes données. Il s'agit d'un algorithme fondé sur les arbres qui cherche à diviser

localement les données en plus petits segments en fonction de différentes valeurs et combinaisons de prédicteurs. CART identifie les divisions les plus performantes, puis répète ce processus régulièrement jusqu'à obtenir le résultat idéal. Il en résulte un arbre de décision représenté par une série de divisions binaires débouchant sur des nœuds terminaux appelés feuilles.

Les résultats avec ce second algorithme ne sont pas aussi satisfaisants qu'espéré. En effet, les groupes formés à l'issu du partitionnement sont crées sur la base du nombre de lignes et non de l'exposition. Les résultats en termes de déviations ne concordent que partiellement à ceux obtenus précédemment. Toutefois, les variables utilisées pour le partitionnement sont exactement les mêmes dans les deux arbres. Ce dernier résultat prouve que la mortalité et les déviations qui en découlent sont plus sensibles à certaines variables telles que la classe de risque, l'âge ainsi que le montant assuré.

N'ayant pas pu confirmer la totalité des résultats issus du GLM Tree avec le CART, une seconde application a été menée. L'idée est d'appliquer le GLM Tree sur des données dont les résultats de mortalité sont connus d'avance. Il sera donc question d'étudier les déviations de mortalité de la population nationale américaine par rapport à une mortalité par âge et par sexe. Une seconde partie incluant les causes de décès a également été implémentée mais ne sera pas discuter danc cette synthèse.

L'idée étant toujours d'utiliser le modèle binomial à travers l'algortihme de partitionnement récursif GLM Tree. Les résultats de cette seconde application devraient rejoindre ceux mentionnés dans la littérature au sujet de la mortalité américaine et ainsi permettre d'affirmer de la cohérence et pertinence de cet algorithme. Pour ce faire, les données de la CDC et du Census Bureau sur la période allant de 2000 à 2015 ont été utilisées et aggrégées. L'arbre obtenu est visible à la figure 4.

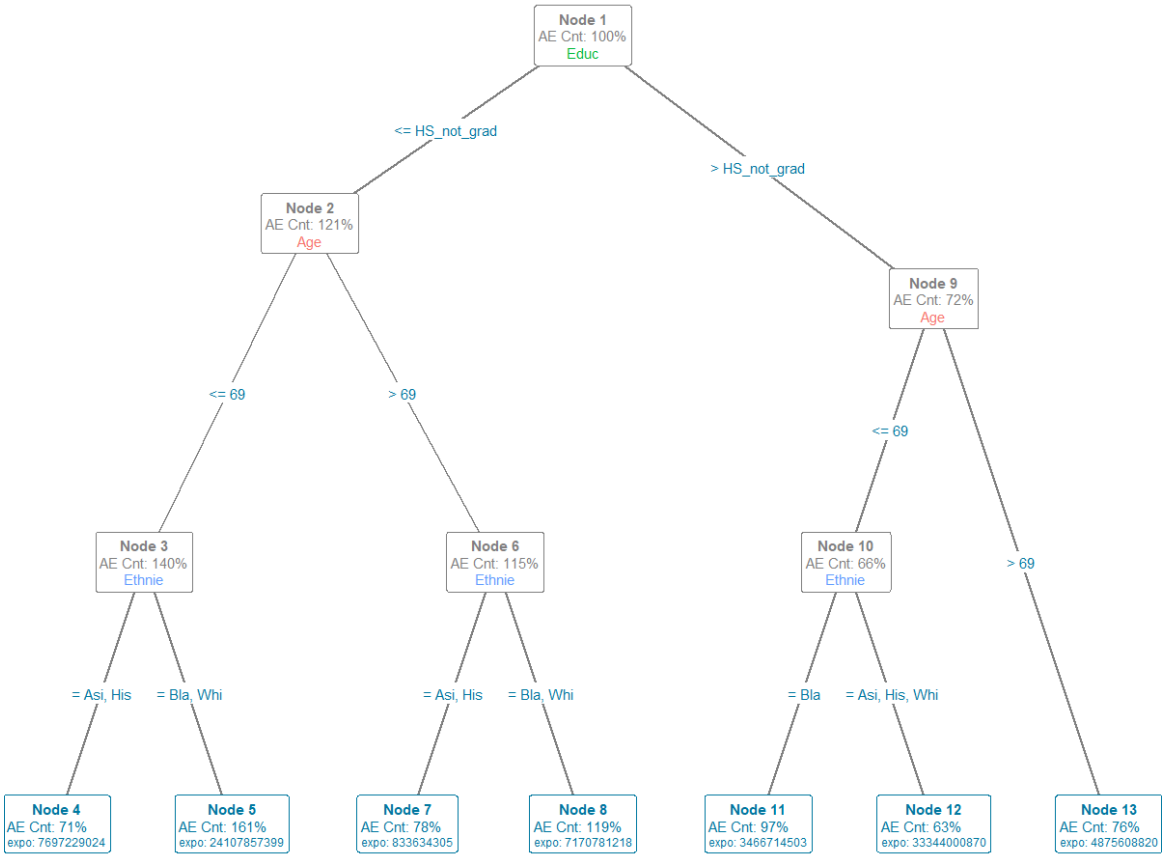


FIGURE 4: Arbre issu du partitionnement des données sans les causes de décès

La variable *ethnie* légalement renseignée dans cette base n'est utilisée qu'à des fins illustratives dans cette seconde application. En tarification, son utilisation poserait un réel problème d'éthique. Les résultats obtenus dans cette partie ne sauraient donc servir de base afin de proposer un quelconque contrat d'assurance.

La mortalité de la base sera donc comparée à la mortalité moyenne par âge et par sexe. Évidemment, la déviation globale est à 100%. 3 variables de partitionnement sur 4 (sexe, ethnie, âge et niveau d'éducation) ont été utilisées dans cet arbre. Des instabilités selon le niveau d'éducation, l'âge et l'ethnie ont été détectées. La mortalité la plus faible par rapport à la moyenne est obtenue pour les individus les plus éduqués ayant au au moins obtenu l'équivalent du Baccalauréat (nœud 9) ainsi que les américains d'origine asiatique et hispanique (nœuds 4 et 7). D'autre part, on remarque pour les autres ethnies, les américains noirs et blancs, que la mortalité est cette fois ci supérieure à la moyenne (nœud 5 et 8). Ce résultat s'accentue plus pour les américains noirs pour lesquels la mortalité bien qu'inférieure à la moyenne, reste supérieure à celle des autres ethnies comme c'est le cas dans le nœud 11. Ces résultats rejoignent directement ceux déjà obtenus dans la littérature qui indiquent la sous mortalité des hispaniques par rapport au reste de la population ainsi que la mortalité plus forte chez les américains noirs peu importe le niveau d'études.

En conclusion, nous sommes partis d'une approche de partitionnement récursif à l'aide des GLM et d'un modèle binomial afin de capturer les déviations de mortalité par rapport à une population de référence. Cette étude de la mortalité a prouvé qu'il était important de bien estimer ses taux de mortalité pour permettre une tarification au plus juste. Et que les écarts de déviation pouvaient conduire à des pertes énormes. La même application avec l'algorithme CART n'a pas été très concluante mais a tout de même révéler l'importance de certaines variables qui influencent la mortalité. La pertinence du GLM Tree a pu être démontrée par la seconde application sur la population nationale américaine. En effet, les résultats issus de cette application correspondent globalement à ceux mentionnés dans la littérature au sujet de la mortalité américaine.

Toutefois, il est important de relever les limites et faiblesses de ce travail. En effet, nous sommes partis sur l'utilisation du modèle binomial et avons fait une approximation afin de capturer la déviation qui nous intéressait. Cette approximation affecte la précision de la déviation calculée et n'aurait pas été nécéssaire en utilisant la loi de poisson. Il est également possible de complexifier un peu plus le modèle choisit en y intégrant un ou plusieurs régresseurs ou sinon de travailler cette fois avec les années d'observation afin d'analyser des tendances de mortalité des différents groupes.

Synthesis note

Reinsurance companies that own life risk need to study the mortality of their portfolio. When the mortality rates are poorly calibrated by the insurance companies, this has an impact on the reinsurance company. In such a context, it is essential for the reinsurer to study the mortality of its portfolios in order to determine the segments of the population for which the mortality deviates from the assumptions and predictions made. Aside from the case where rates are well-calibrated, there will be either a mortality risk or a longevity risk for the different profiles in the portfolio being considered. Whether the mortality is higher or lower than the prediction, the result of the reinsurer will be impacted. It therefore seems essential to regularly study the mortality of its portfolio and to implement management actions in the event of a non-negligible impact on the reinsurance company. These management actions are defined in the reinsurance treaty signed between the insurer and the reinsurer.

Life risk therefore occupies a central place in the results of reinsurance companies. Here, we will concentrate on mortality and study its variations through the A/E ratio (for Actual/Expected) in number of deaths. The idea is to identify the segments of the population for which the observed mortality does not match the predicted one.

A first application will be conducted on the U.S. insured population. The data comes from the SOA (Society Of Actuaries) and is aggregated over the period from 2003 to 2013. This database contains data from various voluntary american companies. The results obtained are not representative of any particular company. These observations will be compared to the predictions from the 2015 Valuation Basic Table (VBT). A valuation mortality table has a safety margin integrated into the mortality rates to protect the insurers. Moreover, it is based on amounts. Life insurance companies use valuation mortality tables to determine the amount of liquid assets they are required by law to set aside for claims and benefits.

The first results obtained are from the partitioning with the GLM Tree algorithm. This is based on the MOB algorithm, Zeileis et al., 2008 used to divide the data into groups that differ in terms of the model parameters. The particularity of this algorithm lies in its link with the generalized linear models. It is possible to use different distribution laws to model our data. The tests for the partitioning based on the Mfluctuations Tests will differ depending on whether the partitioning variable used is numerical or categorical.

In our study framework, we assumed that the number of deaths A, was binomial. The deviation is thus captured by the intercept β_0 defined by:

$$\beta_0 = \ln\left(\frac{\mathbb{E}(A)}{E}\right),$$

where E is the expected number of deaths.

This choice of the binomial distribution requires an approximation in our model. Indeed, manip-

ulating mortality rates q of the order of 10^{-3} allows us to make the following approximation:

$$1 - q \approx 1.$$

This approximation impacts the accuracy of the calculated deviation without compromising the analysis. We could have avoided this approximation by using the poisson law to model the number of deaths. However, the current design of the algorithm does not allow to integrate the weights in a suitable way to define the size of the groups. Indeed, the groups with this distribution law are based on the number of rows in the base and not on the exposure.

We obtain the tree of the figure 5 from the partitioning of the data with the GLM Tree.

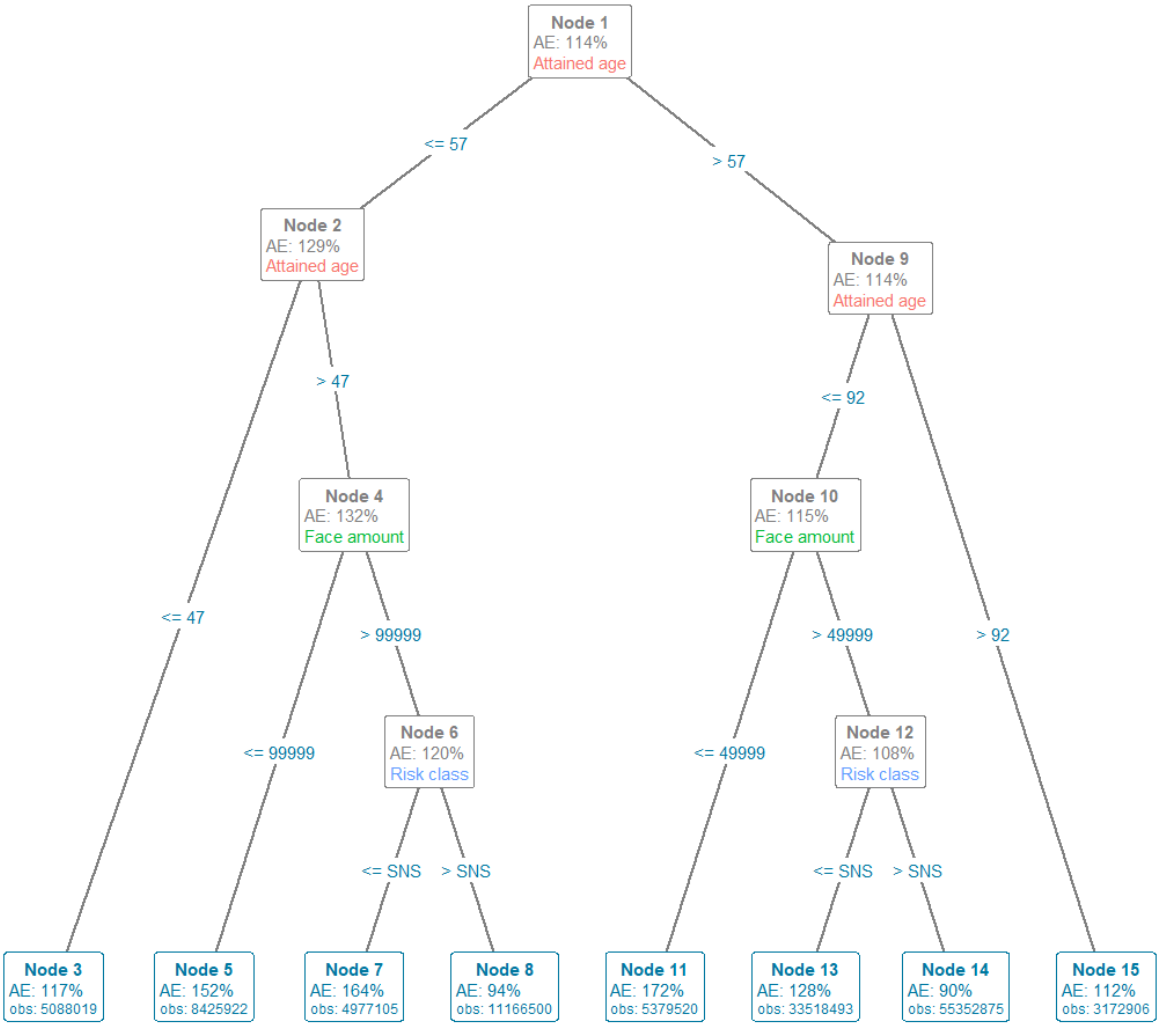


Figure 5: Tree from data partitioning

At the global level, the first deviation obtained is 114%: the reference table underestimates and therefore does not adequately capture the mortality behavior of the insurers represented in this database.

The order of the variables used for partitioning shows which mortality differential is most important. Among the six partitioning variables used in the function (Attained age, Face amount band, Gender, Risk class, Duration and Insurance plan), only three are used for partitioning. These are :

- Attained age (x2),
- Face amount,
- Risk class.

We observe strong variations for all ages when looking at the following segments:

– The low amount insured segments (nodes 5 and 11).

– Those in poorer health (SS + PS + SNS) with large amounts (nodes 7 and 13).

Note that the different parameters in this tree have been chosen to have a reasonable size for the different groups. The post pruning strategy using an information criterion such as the BIC or the AIC is not possible in our case. Indeed, the regression is only done according to the intercept. There are therefore no explanatory variables subject to information criteria. The alternative based on the use of a low value of the confidence level α and proposed by Zeileis et al. (2008) will turn out to be useless. Indeed, the constructed tree is stable according to the values of this parameter α. The sensitivity tests of the other parameters *minsize* and *maxdepth* for respectively the minimum of observations in a node and the maximum depth of the tree will give slightly different structures from the initial tree. However, the conclusions will remain unchanged.

The graph 6 provides additional information about the leaves of the tree. In addition to the observed deviations and their confidence intervals, it shows the proportion of insured amounts and the number of deaths for each of the terminal nodes.

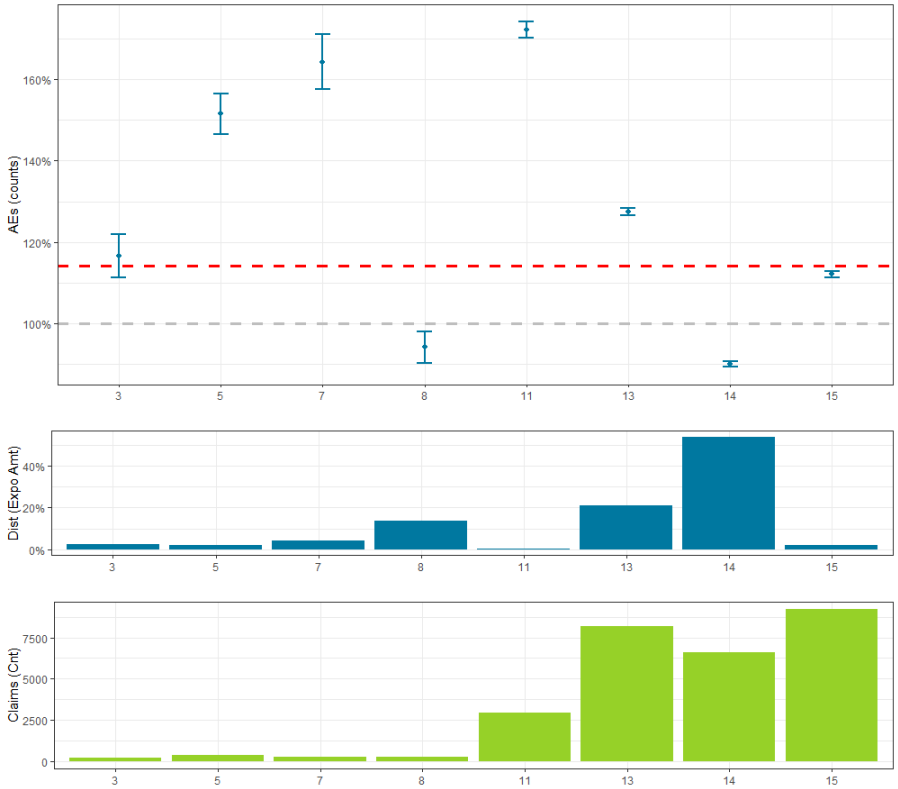


Figure 6: Ratio of observed and predicted deaths by 2015 VBT table

The gray dotted line illustrates the A/E ratio of 100% which indicates the perfect match between the industry tables and the 2015 VBT. The red dotted line is the 114% A/E ratio that illustrates the overall deviation of the table from the regulatory table. In this case, the confidence intervals proposed by Liddell (1984) are constructed not from approximations but from an exact formula. These will allow to evaluate the significance or not of the obtained deviations.

Looking at node 13 (with smokers with large amounts), it turns out that the displayed deviation (128%) is greater than the overall ratio of 114%, and even greater than the ratio at 100%. In addition, the amounts for this node represent more than 20% of the insured amounts and the corresponding claims experience is also very high (>7500). This type of segment is quite costly for a company. It must therefore be monitored.

We will therefore study it more closely in order to estimate the impact of this kind of deviation. Recall that this node 13 concerns individuals aged between 62 and 92, who are in the most vulnerable risk class (smokers) and who have high insured amounts. For these individuals, the deviation obtained is 128%. If we look at the mortality rate observed in our data compared to the rate predicted by the 2015 VBT table, we see that the mortality rate observed is generally higher than the rate predicted by the table.

These different mortality rates will be used to price an insurance death contract. The idea is to make assumptions about the amounts and discount rates in order to determine the potential flows that a company should pay out over a 30-year projection. We then obtain the figure 7.

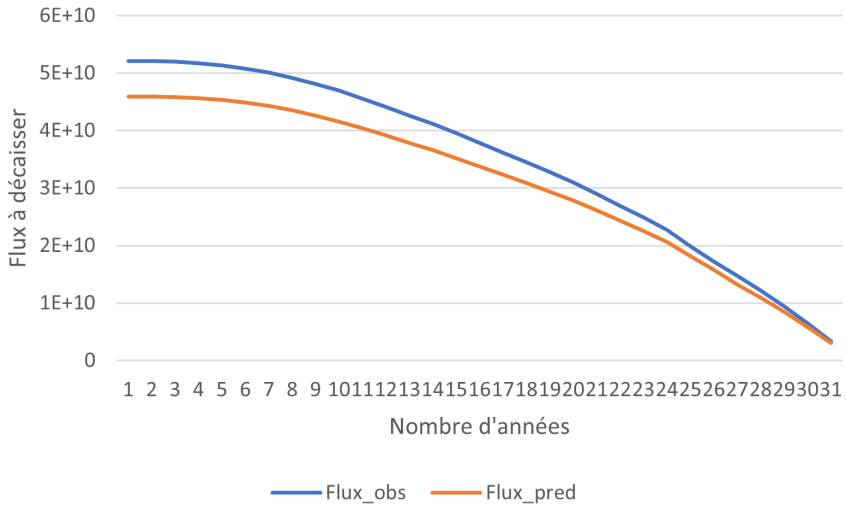


Figure 7: Comparison of observed and predicted flows according to the number of years projected

It can be seen that with a mortality deviation of 128% in this node, the observed flows, those that must actually be paid out to cover the beneficiaries of policyholders, are much higher than those that would have been assumed with the standard table. The use of different rates leads to a difference in flows amounting to 11%. A mortality gap of almost 30% can therefore lead to considerable losses for an insurance company. Hence the importance of calibrating our mortality rates as well as possible. After such a study, it is possible to set up management actions in order to manage this segment of the population.

In order to assess the relevance of the GLM Tree, the CART method introduced by Breiman et al. (1984) was applied to the same data. This is a tree-based algorithm that seeks to locally divide the data into smaller segments based on different values and combinations of predictors. CART identifies

the best performing splits and then repeats this process regularly until the ideal result is achieved. The result is a decision tree represented by a series of binary splits leading to terminal nodes called leaves.

The results with this second algorithm are not as satisfactory as expected. Indeed, the groups formed after partitioning are created on the basis of the number of lines and not on the basis of the exposure. The results in terms of deviations are only partially consistent with those obtained previously. However, the variables used for partitioning are exactly the same in both trees. This last result proves that mortality and the resulting deviations are more sensitive to certain variables such as risk class, age and the amount insured.

Since we were unable to confirm all of our results from GLM Tree with CART, a second application was conducted. The idea is to apply the GLM Tree to data whose mortality results are known in advance. It will thus be a question of studying the differences in mortality of the American national population in relation to a mortality by age and sex. A second part including the causes of deaths was also carried out but will not be discussed in this synthesis.

The idea is still to use the binomial model through the recursive partitioning algorithm GLM Tree. The results of this second application should join those mentioned in the literature about the American mortality and thus make it possible to affirm the consistency and relevance of this algorithm. For this purpose, CDC and Census Bureau data from 2000 to 2015 were used and aggregated. The resulting tree can be seen in Figure 8.

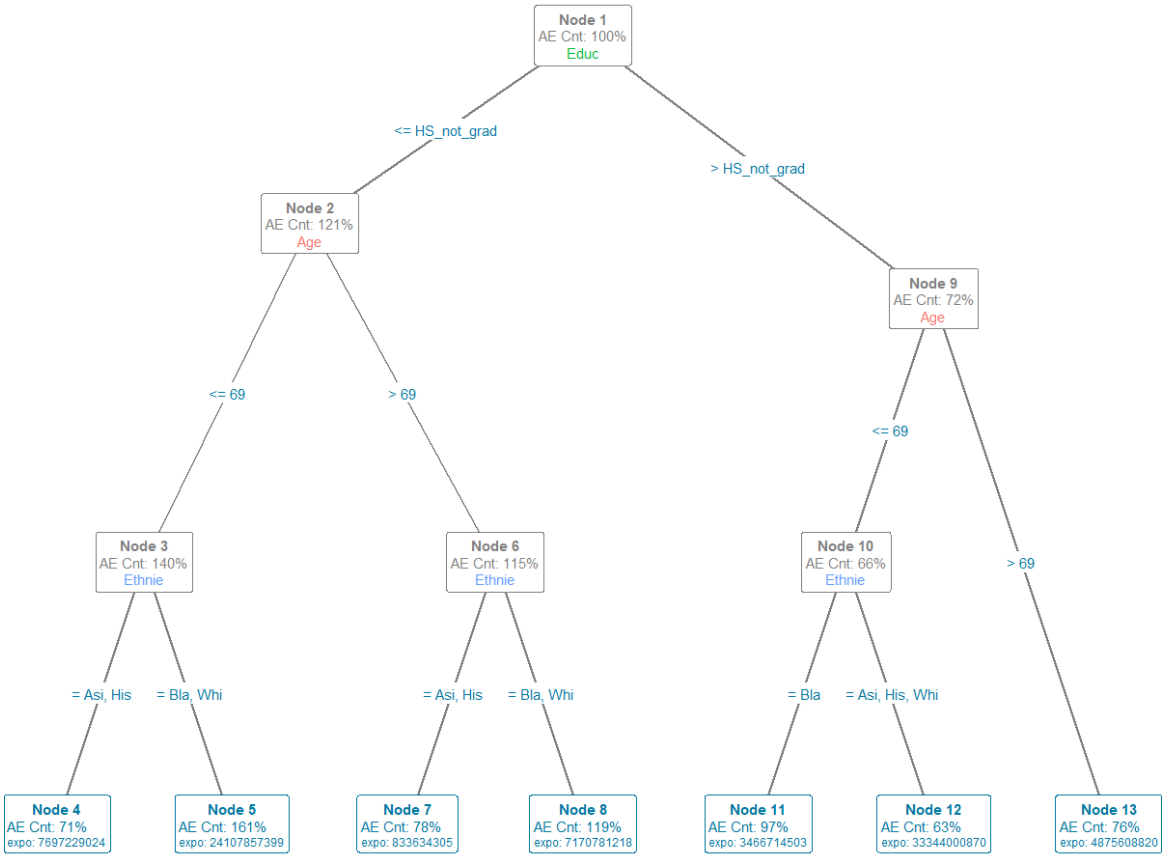


Figure 8: Tree resulting from the partitioning of the data without the causes of the deaths

The variable *ethnie*, also mentioned in this database, is used only for illustrative purposes in this second application. In pricing, its use would raise a real ethical problem. The results obtained in this part cannot be used as a basis to propose any insurance contracts.

The mortality of the base will thus be compared to the average mortality by age and gender. Obviously, the global deviation is 100%. Three out of four partitioning variables (gender, ethnicity, age and education level) are used in this tree. Instabilities related to education level, age and ethnicity have been detected. The lowest mortality relative to the mean is obtained for the most educated individuals with at least the equivalent of a Bachelor's degree (node 9) as well as for asian Americans and hispanics (nodes 4 and 7). On the other hand, we note that for the other ethnic groups, black and white Americans, mortality is higher than average (nodes 5 and 8). This result is even more pronounced for black Americans, for whom mortality, although lower than average, is still higher than the mortality of other ethnic groups, as shown in node 11. These results are in direct agreement with those already obtained in the literature, which indicate the under-mortality of hispanics compared to the other groups, as well as the higher mortality among black Americans, regardless of their level of education.

In conclusion, we started with a recursive partitioning approach using glm and a binomial model to capture mortality deviations from a reference population. This mortality study showed that it was important to estimate mortality rates correctly to ensure fair and appropriate pricing. And that the deviations could lead to huge losses. The same application with the CART algorithm was not very conclusive but did reveal the importance of certain variables that influence mortality. The relevance of the GLM Tree has been demonstrated by the second application on the American national population. Indeed, the results from this application are broadly consistent with those reported in the literature on US mortality.

However, it is important to note the limitations and weaknesses of this work. Indeed, we started with the use of the binomial model and made an approximation in order to capture the deviation. This approximation affects the precision of the calculated deviations and would not have been necessary using the poisson law. It is also possible to make the model a little more complex by integrating one or more regressors or to work with the years of observation in order to analyze the mortality trends of the different groups.

Remerciements

Je tiens à remercier la compagnie de réassurance SCOR pour l'opportunité qui m'a été donnée de travailler dans un cadre agréable et motivant.

Plus particulièrement, je remercie Julien Tomas, actuaire et mon tuteur pédagogique à la SCOR, pour son encadrement tout le long de ce mémoire. J'ai pu apprécier sa disponibilité, ses précieux conseils et son sens aigu de la pédagogie.

Je suis également reconnaissante envers Christophe Dutang pour ses relectures attentives et ses conseils avisés depuis l'université Paris Dauphine.

Je remercie l'ensemble des membres de l'équipe Knowledge et BRM pour le temps qu'ils m'ont consacré et pour toutes les connaissances que j'ai acquises.

Enfin, je remercie ma famille pour son soutien au quotidien.

Table des matières

Introduction

Les compagnies d'assurance vie fournissent des contrats en rapport aux risques liés à la personne : mortalité, longévité, dépendance. Une bonne connaissance des risques encourus par les personnes assurées est donc indispensable afin de fixer le montant de la prime. Cette dernière doit être suffisante afin de permettre à la compagnie de couvrir les sinistres futurs et d'assurer sa prospérité. Cependant, selon le principe de prudence et dans le but de stabiliser ses fonds propres dans le temps, la compagnie d'assurance va transférer une partie de son risque à un réassureur via un traité de réassurance. Dès lors que'un traité est signé, les compagnies de réassurance se doivent à leur tour d'étudier les risques biométriques de leurs différents portefeuilles. Cette étude des risques biométriques, en particulier de la mortalité, va permettre d'identifier les segments du portefeuille dont les estimations s'écartent de la mortalité observée et impactent négativement le résultat du réassureur.

C'est dans ce contexte que s'inscrit ce mémoire d'actuaire qui se veut d'étudier deux algorithmes de partitionnement afin de détecter les déviations de mortalité.
Nous utiliserons un premier algorithme dit de partitionnement récursif implémenté par Zeileis et al. (2008) à travers la fonction `glmtree()` du package **partykit** sous le langage R, R Core Team (2021). Cette approche combine les Modèles Linéaires Généralisés (GLM) aux arbres de décision et mène à la création d'un arbre facile à interpréter et construit à l'aide de variables de partitionnement à spécifier. Le second algorithme utilisé est le CART implémenté par Breiman et al. (1984) à travers la fonction `rpart()` du package R du même nom, Therneau et Atkinson (2019). Celui-ci permet de construire un arbre de décision binaire dans le cadre de la régression ou de la classification. Les résultats issus de cette approche classique seront comparés à ceux obtenus avec le GLM Tree afin de tester la pertinence de ce dernier.

En vue d'identifier les segments problématiques dont les prédictions s'écartent de la réalité, deux applications sont proposées. La première porte sur les déviations de mortalité sur les données de l'industrie américaine par rapport à la table réglementaire 2015 VBT (Valuation Basic Table) et la seconde sur les données de la population nationale américaine. Ces déviations seront capturées à l'aide du ratio A/E (Actual/Expected) en nombre de décès et en montant assuré.

Cette étude permet d'identifier les segments qui présentent des déviations par rapport aux hypothèses best estimate. Elle permet également d'étudier les variables importantes dans le partitionnement. Ces résultats sont importants pour une compagnie de réassurance. En effet, ils permettront de surveiller ces segments de la population et de mettre en place des actions de management pour les gérer.

L'étude se décompose ainsi en quatre parties qui constituent les différents chapitres. Ainsi, le chapitre 1 précisera les notions importantes liées à l'étude de la mortalité. La description des algorithmes et leurs différentes étapes seront présentées dans le chapitre 2. Puis au chapitre 3, viendra la première application sur les données de l'industrie américaine. Le dernier chapitre sera consacré à l'application sur les données nationales américaines. Enfin, un rappel des résultats et des pistes d'amélioration

seront présentés en conclusion.

Chapitre 1

Etude de la mortalité - Définitions et notations

1.1 Définitions et notations des termes de mortalité

Dans cette partie, les bases de l'étude de la mortalité seront posées. Il s'agira de définir les termes utilisés pour étudier et modéliser la mortalité. Les définitions et notations qui vont suivre seront utiles pour comprendre les concepts clés régissant la mortalité.

Le phénomène de durée

Dans les problématiques actuarielles et plus précisement en assurance vie, la modélisation d'un phénomène de durée est un outil central. On appelle durée de vie, une variable aléatoire T, positive ou nulle, représentant la durée s'écoulant entre deux évènements, Planchet et Théron (2011). Ces évènements peuvent être un décès, une entrée en chômage, en incapacité, en invalidité.
Les modèles de durée constituent un sous ensemble de méthodes statistiques dont le champ d'application est très large. On distingue entre autres :

- La démographie, la médecine : durée de la vie humaine, durée entre le début d'une maladie et la guérison.
- L'assurance, l'économie : durée en incapacité, durée entre deux sinistres, durée d'un épisode de chômage, durée de vie d'une entreprise.
- La fiabilité : durée de fonctionnement d'un composant réparable ou non.

Soulignons la spécificité des modèles de durée par rapport aux méthodes classiques. En effet, les distributions de ces modèles dépendent de variables explicatives pouvant dépendre du temps (âge, niveau d'éducation, niveau de revenu).
De plus, les données étudiées sont des échantillons de courte durée parfois dépendantes et elles sont la plupart du temps incomplètes : on parle de censure et de troncature.
La censure : elle se manifeste lorsque l'assuré n'a pas subi l'évènement à la fin de la période d'observation. La durée de vie observée n'est pas complète.
La troncature : elle diffère de la censure dans le sens où elle concerne l'échantillonnage lui-même. On perd complètement l'information sur les observations en dehors de la plage.

Il est commode de considérer les variables T_x représentant la durée de vie résiduelle d'un individu conditionnellement au fait qu'il soit vivant à l'âge x, $T_x = [T - x|T > x]$.
L'interprétation conduit à définir des représentations de la loi non plus au travers de la fonction de répartition, mais au travers de la fonction de survie et de la fonction de hasard.

Présentation d'une table de mortalité

Les tables sont établies selon les observations statistiques d'une population donnée à un moment donné. Une table de mortalité contient généralement 3 colonnes :
- L'âge, noté x.
- Le nombre de vivants à l'âge x, noté l_x.
- Le nombre de décès entre les âges x et $x+1$, noté d_x.

On en déduit les relations suivantes :

$$d_x = l_x - l_{x+1},$$
$$l_{w+1} = 0,$$

avec w correspondant à l'âge limite de la table. De la même façon, le nombre de décès entre x et $x+t$ est noté ${}_td_x = l_x - l_{x+t}$.

Durée de vie

Soit un individu d'âge x et T_x sa durée de vie future.
La loi de la variable aléatoire T_x est caractérisée par sa fonction de répartition :

$$F_x(T) = \mathbb{P}\left(T_x \leq t\right) = \mathbb{P}(T \leq t + x|T > x), t \geq 0.$$

Supposons que la loi de T_x admet une densité f_x. Notons par S_x sa fonction de survie telle que :

$$S_x(t) = 1 - F_x(t) = \mathbb{P}[T_x > t], t \geq 0.$$

La fonction de survie $S_x(t)$ est, pour t fixé, la probabilité de survivre jusqu'à l'instant t pour un individu d'âge x. Pour l'étude de la durée de vie, le terme fonction de survie paraît donc plus adapté et plus facilement interprétable que la fonction de répartition.

Les probabilités de décès et de survie q_x et p_x

À partir des données de la table de mortalité, on peut déduire les probabilités q_x et p_x définies par :
- q_x : La probabilité pour qu'un individu d'âge x décède avant d'atteindre l'âge $x+1$.
- p_x : La probabilité pour qu'un individu d'âge x soit encore en vie à l'âge $x+1$.

q_x et p_x s'écrivent de la manière suivante :

$$p_x = \frac{l_{x+1}}{l_x}, q_x = \frac{l_x - l_{x+1}}{l_x}.$$

La valeur de la fonction de répartition $F_x(t)$ représente la probabilité de décès d'un individu d'âge x avant le temps t, notée :

$${}_tq_x = F_x(t).$$

La probabilité de survie est la probabilité qu'un individu d'âge x vive encore plus de t années :

$$ {}_tp_x = 1 - F_x(t). $$

Par convention, l'indice t sera omis lorsque celui-ci vaut 1.

Exposition au risque

L'exposition au risque mesure le temps au cours duquel les individus de la population sont exposés au risque de mortalité. Elle représente le temps moyen vécu par les individus d'âge x au cours d'une période t.

$$ E_{x,t} = \int_0^t l_{x+\alpha} d\alpha. $$

Taux de mortalité instantané ou force de mortalité

Le taux instantané de mortalité, intensité de mortalité ou encore force de mortalité d'un individu d'âge x à un instant t est définie comme suit :

$$ {}_t\mu_x = \lim_{\Delta \to 0^+} \frac{F_x(t+\Delta) - F_x(t)}{\Delta(1 - F_x(t))} = \lim_{\Delta \to 0^+} \frac{{}_{t+\Delta}q_x - {}_t q_x}{\Delta(1 - {}_t q_x)}. $$

Cela signifie que la probabilité conditionnelle de décès d'un individu d'âge x sur une courte période $[t; t+\Delta]$, à condition d'avoir survécu jusqu'à t, est proportionnelle à la longueur de cette période avec un cœfficient $\mu(x,t)$. On note :

$$ {}_t\mu_x = \frac{\frac{d({}_tp_x)}{dt}}{{}_tp_x} = -\frac{d(log({}_tp_x))}{dt}. $$

La relation fondamentale suivante peut être déduite :

$$ {}_tq_x = 1 - \exp(-\int_0^t {}_s\mu_x ds). $$

Le quotient ${}_tq_x$ représente une probabilité tandis que les taux de mortalité sont exprimés en inverse de l'unité de temps. Ils décomptent des décès par personne exposée au risque et par unité de temps. Cette distinction conduit à certaines relations avec la fonction de hasard, appelée dans ce contexte ≪taux instantané de mortalité≫, Planchet (2020) :

$$ h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} lnS(t). $$

La fonction de hasard détermine ainsi entièrement la loi de T. On peut donc déduire une nouvelle expression de la fonction de survie :

$$ S_x(t) = \exp(-\int_x^{x+t} h(u)du) = \exp(-\int_0^t h(u+x)du). $$

Dans la suite, l'hypothèse suivante est adoptée :

$$ \forall \alpha, \forall t, \tau \in [0,1[, {}_{t+\tau}\mu_{x+\alpha} = {}_t\mu_x. $$

Cela suppose donc que la force de mortalité est constante. Ainsi, il est par exemple supposé que la force de mortalité à l'âge exact $x + \frac{1}{2}$ et au temps exact $t + \frac{1}{2}$ est simplement ${}_t\mu_x$.

Par ailleurs, cette hypothèse n'est pas sans conséquence. En effet, si ${}_td_x$ est traitée comme une variable aléatoire $D_{x,t}$, et l'exposition centrale $E_{x,t}$ comme fixée, alors selon Scott (1981), $D_{x,t}$ a une distribution de poisson :

$$D_{x,t} \sim P(E_{x,t} \times {}_t\mu_x).$$

Taux central de mortalité

Le taux central de mortalité pour $x, t \in \mathbb{N}$, est obtenu en rapportant le nombre de décès à l'âge x et au temps t à l'effectif moyen d'âge x durant l'année t. Avec l'exposition définie précédemment, ce taux se note :

$${}_tm_x = \frac{{}_td_x}{E_{x,t}}.$$

Les quotients de mortalité sont des probabilités, et représentent des décès par personne sous risque et par unité de temps. Le taux instantané de mortalité est en lien direct avec le taux central de mortalité :

$$\mu_x = \lim_{h \to 0} {}_hm_x.$$

On a aussi, sous l'hypothèse précédente de constance par morceaux des taux instantanés de mortalité que le taux d'évolution annuel de la mortalité correspond à la variation relative du taux :

$${}_tm_x = {}_t\mu_x = \frac{{}_td_x}{E_{x,t}},$$

$${}_tq_x = 1 - \exp(-{}_t\mu_x).$$

Le diagramme de Lexis

Aux alentours de 1870, des démographes, particulièrement en Allemagne, ont éprouvé le besoin de représenter, sur un graphique, la dynamique des populations. Le diagrammme de Lexis est un outil essentiellement utilisé en actuariat et en démographie permettant une représentation des trajectoires de vie des individus, leur naissance et leur décès lorsqu'il survient.

Bien que comportant deux axes, il permet l'utilisation de trois coordonnées distinctes : l'âge, l'année civile et l'année de naissance. Ces trois données sont évidemment liées et la connaissance de deux d'entre elles permet d'en déduire la troisième. L'axe des abscisses représente le temps (année calendaire) et l'axe des ordonnées représente l'âge de l'individu. La vie des individus d'une même génération au cours du temps est représentée par une ligne diagonale qui coupe l'axe des abscisses à la date de naissance de l'individu. Dans ce système de coordonnées, l'existence d'un individu est représentée par une ligne de vie qui s'interrompt lors du décès de celui-ci en un point appelé : point mortuaire.

La figure 1.1 représente la vie d'un individu né en t_0 et qui décède en t_1 à l'âge x dans le diagramme de Lexis.

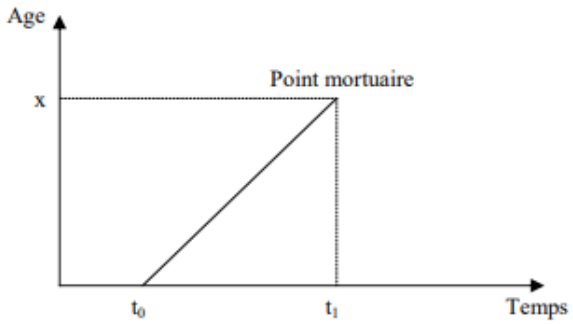


FIGURE 1.1: Parcours d'un individu sur le diagramme de Lexis

Le diagramme de Lexis peut également représenter les lignes de vie des personnes nées au cours d'une même année, ces personnes font partie d'une même génération (cohorte). En traçant une bande verticale entre $g+x$ et $g+x+1$, on retient les décès des individus de la génération $g+x$ et cela permet de repérer facilement les informations suivantes :

– L'effectif d'une génération à une date donnée (intersection avec les axes verticaux) ou un âge donné (intersection avec les axes horizontaux),

– Le nombre de décès ayant lieu au cours d'une année pour une génération d'individus.

La mesure naturelle de la mortalité consiste à comptabiliser les décès survenus au sein d'une génération à un âge x, ou au cours d'une période donnée (une année t par exemple).

La figure 1.2 permet d'estimer le nombre de décès à l'âge x dans la génération g. Ces décès sont localisés dans un parallèlogramme, intersection entre un couloir oblique (correspondant à la génération g) et un couloir horizontal (correspondant à l'âge x). On peut noter que ces décès se sont produits au cours des deux années $g+x$ et $g+x+1$.

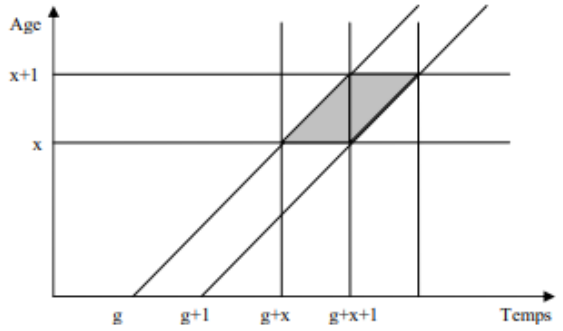


FIGURE 1.2: Nombre de décès à l'âge x parmi les individus de la génération g

À partir de la figure 1.3, il est possible d'évaluer le nombre de décès d'une année au sein d'une génération. Ces décès sont localisés dans un parallèlogramme, intersection entre un couloir oblique (correspondant à la génération g) et un couloir vertical (correspondant à l'année t). On peut noter que ces décès se sont produits aux âges $t-g$ et $t-g-1$.

Le diagramme de Lexis sert donc à déterminer correctement les taux bruts de mortalité dans un contexte où il est rare de disposer d'une information exacte sur les âges et dates de décès (âge entier

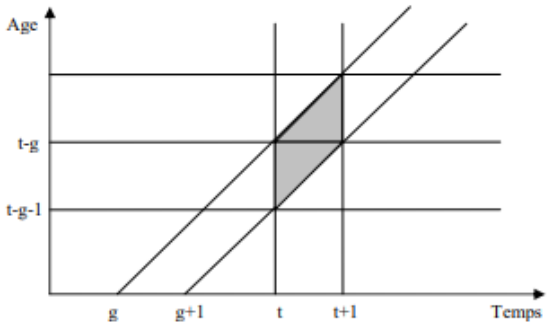


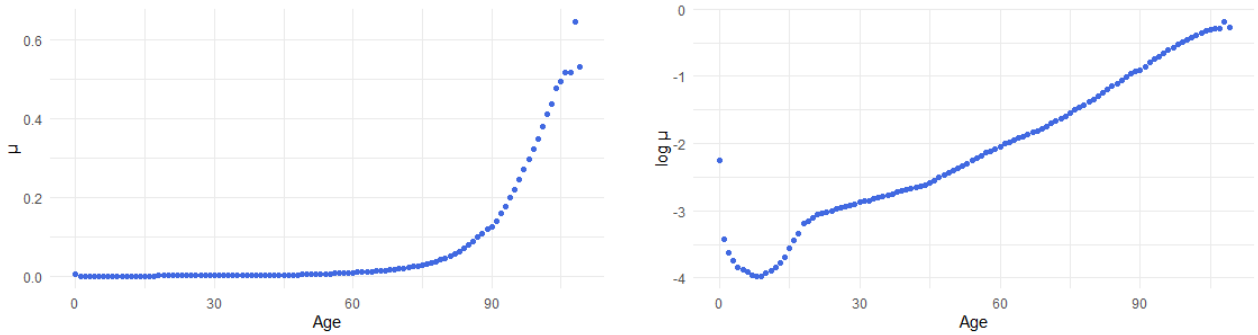
FIGURE 1.3: Nombre de décès parmi les individus de la génération g au cours de l'année t

et année entière).

1.2 La structure en temps et en âge de la mortalité : l'exemple des USA

Il sera maintenant question d'analyser la dynamique de mortalité en fonction de l'âge et du temps. Les applications prochaines se focaliseront sur les USA, raison pour laquelle nous étudierons dès à présent la mortalité de la population américaine.

La figure 1.4 représente la force de mortalité μ de la population américaine et son logarithme $\log(\mu)$ en fonction de l'âge pour l'année 2019. On observe dans ce cas, d'une part, une croissance exponentielle de la force de mortalité en fonction de l'âge, et d'autre part, qu'à partir d'un certain âge le logarithme de la force de mortalité semble bien être une fonction linéaire de l'âge.



(a) Force de mortalité

(b) Logarithme de la force de mortalité

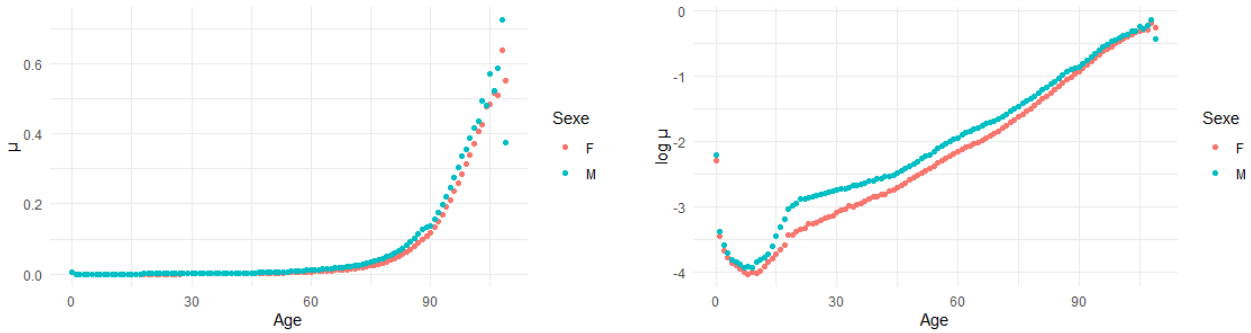
FIGURE 1.4: Force de mortalité de la population totale américaine (a) et son logarithme (b) en fonction de l'âge en 2019 — données provenant de la Human Mortality Database

Néanmoins, le logarithme de la force de mortalité présente une structure en fonction de l'âge assez particulière. On distingue à cet effet :

- La mortalité infantile : la force de mortalité des enfants de moins d'un an est largement supérieure à celle des enfants âgés entre deux et quinze ans. Cela fait écho aux multiples maladies qu'il est possible d'avoir à la naissance ;

- Un creux de mortalité pour les âges inférieurs à quinze ans ;
- La célèbre "bosse des accidents", attribuée aux accidents de la route, à la drogue, à l'alcool et aux suicides, et qui concerne surtout les individus âgés de 15 à 25 ans ;
- À partir de 30 ans, une croissance approximativement linéaire de la force de mortalité à l'échelle logarithmique.

La figure 1.5 représente la force de mortalité μ de la population américaine et son logarithme $\log(\mu)$ en fonction de l'âge et du sexe pour l'année 2019. Elle permet d'observer que les femmes présentent un niveau de force de mortalité inférieur à celui des hommes. Cet écart est notable pour les âges supérieurs à 15 ans. On remarque notamment que la "bosse des accidents", très marquée chez les hommes, l'est moins chez les femmes. Les hommes ont beaucoup plus de comportements à risque que les femmes. En ce qui concerne la mortalité infantile, elle semble être du même ordre de grandeur entre les deux sexes. En allant vers les âges les plus élevés, on s'aperçoit que la structure de la mortalité des américains est moins linéaire que celle des américaines. Cependant, dès 90 ans, une convergence de la mortalité pour les deux sexes est observée.



(a) Force de mortalité

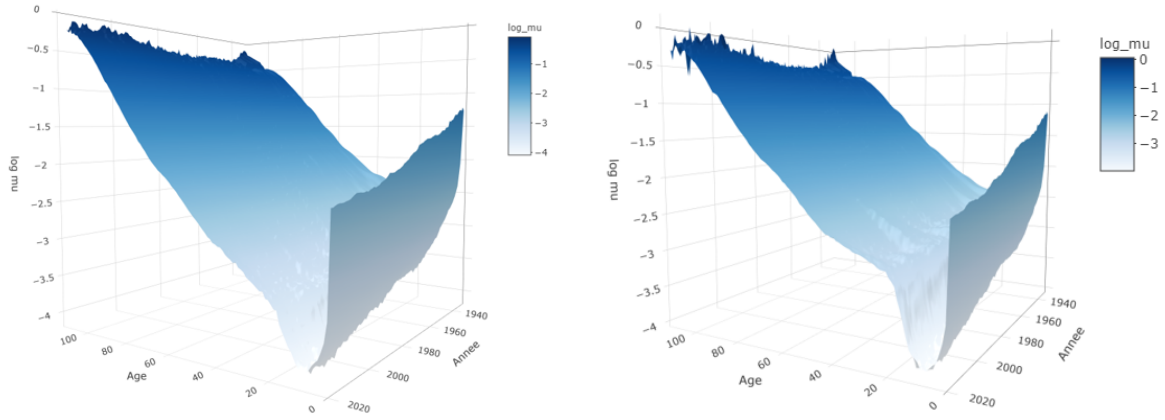
(b) Logarithme de la force de mortalité

FIGURE 1.5: Force de mortalité (a) et son logarithme (b) en fonction de l'âge et du sexe en 2019 pour la population américaine— données provenant de la Human Mortality Database

En plus d'une structure en âge, la force de mortalité possède également une structure en temps. Sur les dernières années, une baisse assez marquée de la mortalité humaine au cours du temps a été observée. Cela a été principalement dû au progrès de la médecine, à un large accès aux soins par les populations et à l'accroissement des revenus.
Dans la plupart des pays industrialisés, la mortalité chez les adultes et les personnes âgées révèle en effet une probabilité de décès annuelle décroissante BENJAMIN et SOLIMAN (1993) ; MACDONALD et al. (1998).

La figure 1.6 représente le logarithme de la force de mortalité des hommes et des femmes aux États-Unis en fonction de l'âge sur la période temporelle 1933-2019. Elle permet d'observer une diminution de la mortalité au cours du temps, en particulier de la mortalité infantile et de celle des personnes âgées. De plus, la "bosse des accidents" sur toute la période étudiée est beaucoup plus marquée chez les hommes.
Cette figure met par ailleurs en évidence l'impact de certains événements historiques catastrophiques comme l'effet conjoint de la Première Guerre mondiale et de la pandémie de grippe espagnole de 1918, ainsi que celui de la Seconde Guerre mondiale. Les niveaux de force de mortalité constatés sont différents entre les hommes et les femmes. Les deux guerres mondiales par exemple ont eu un impact différent en fonction du sexe considéré : les hommes étant plus acteurs de ces événements que les

femmes.
Depuis la fin du 19^e siècle, les études portant sur la mortalité ont permis de remarquer que l'espérance de vie des femmes dépasse désormais celle des hommes dans le monde entier, Beltrán-Sánchez et al. (2015). Les démographes évoquèrent dès lors une sur-mortalité masculine plutôt qu'une sous-mortalité féminine. Selon eux, la nature et les conditions particulièrement difficiles du travail des hommes avaient un impact non négligeable sur leur survie. D'autre part, les travaux domestiques, traditionnellement attribués aux femmes leur offraient un cadre moins rude. Cependant, depuis la fin de la guerre, ces arguments ne suffisent plus pour expliquer ce phénomène persistant. Par exemple, en analysant le comportement des femmes concernant la consommation d'alcool et de tabac, il se trouve qu'elles ont un retard d'une vingtaine d'années par rapport aux hommes. Elles seraient alors moins exposées à cette mortalité précoce.



(a) Données des femmes (b) Données des hommes

Figure 1.6: Logarithme de la force de mortalité des femmes (a) et des hommes (b) en fonction de l'âge sur la période 1933-2019 pour la population américaine — données provenant de la Human Mortality Database

Il est important de remarquer la baisse de la mortalité infantile pour les deux sexes depuis 1940. Les principales raisons de cette baisse sont les progrès réalisés lors des accouchements et par rapport aux premiers soins prodigués aux nouveau-nés.
La poursuite de la baisse au 20^e siècle s'explique par le succès de la lutte contre les maladies infectieuses. La mortalité infantile due à la pneumonie/grippe a fortement diminué pour les nourrissons entre 1970 et 2016, Singh et Yu (2019). Entre 1970 et 2016, la mortalité due à des blessures non intentionnelles a été réduite de plus de moitié. Aujourd'hui aux USA, comme dans la plupart des pays développés, la mort des enfants est surtout devenue accidentelle.

1.3 La mortalité américaine par facteurs socio-économiques et démographiques

Le cadre décrit dans la suite permettra de comprendre les résultats des applications prochaines. La situation assez particulière aux USA conduit à l'étude de la mortalité selon plusieurs aspects.

Aux États-Unis, il existe une longue tradition qui consiste à étudier les différences de mortalité selon les mesures du statut socio-économique (SSE). Un avantage de longévité pour les groupes de statut socio-économique supérieur est bien établi dans la littérature. Plus récemment, la recherche

a montré que les différences de mortalité persistent non seulement aujourd'hui mais se sont aussi considérablement accentuées. Cette recherche a d'abord examiné les différences de mortalité de la population américaine par rapport aux mesures économiques au niveau géographique. Ensuite, il s'agissait de regarder la mortalité selon l'angle du niveau d'instruction. Enfin, la troisième approche a examiné la mortalité en fonction des revenus de carrière. Les trois approches s'accordent pour constaster que les différences de mortalité s'élargissent. Les disparités en matière de mortalité ont augmenté, que l'on utilise le revenu ou le niveau d'éducation comme indicateur pour rendre compte du SSE, Preston et Elo (1995), Manchester et Topoleski (2008).

Dans notre cas, il s'agira d'étudier deux des trois approches analysées dans la littérature, celles selon le niveau d'éducation et selon le revenu. Nous allons également tenter de cerner les différences de mortalité selon l'origine ethnique des individus en la combinant à l'étude par rapport au niveau d'instruction.

1.3.1 Les tendances par niveau d'éducation et par ethnie

Les récentes études s'accordent sur le fait que l'espérance de vie a augmenté le plus rapidement pour les personnes ayant un niveau d'instruction ou un revenu supérieur et que l'écart de longévité selon le statut socio-économique s'est élargi.

L'éducation a été utilisée comme le principal marqueur du statut socio-économique. Cependant, la principale difficulté avec ce facteur est qu'au fil du temps le niveau d'éducation atteint par les générations successives a augmenté, Dowd et Hamoudi (2014). Les générations avec des niveaux d'éducation inférieurs, deviennent une partie plus petite et plus fortement sélectionnée de leurs générations. Par conséquent, si la mortalité des moins instruits diminue plus lentement de génération en génération ou augmente réellement, alors il est difficile de séparer la part due à la plus grande sélectivité de ce groupe de la part dûe à d'autres causes.

De nombreuses études ont montré que les écarts de mortalité selon le niveau de scolarité se sont accrus au cours des dernières décennies. Bien qu'il ne soit pas difficile de comprendre les raisons pour lesquelles des individus plus instruits pourraient bénéficier de manière disproportionnée de la médecine, des baisses importantes de l'espérance de vie sont surprenantes et largement sans précédent dans les pays développés en dehors des périodes de guerres, Bound et al. (2014). De nombreux auteurs ont fait valoir que la prévalence accrue de maladies chroniques telles que le diabète entraîne une baisse de l'espérance de vie aux États-Unis et dans d'autres pays développés. En outre, l'augmentation de la prévalence du diabète s'est concentrée parmi les moins instruits.

Dans un article très médiatisé (article paru à la une du New York Times), Olshansky et al. (2012) ont utilisé des statistiques de l'état civil et des données de recensement américaines pour estimer les changements d'espérance de vie selon l'ethnie et le niveau d'éducation. Leur conclusion était qu'entre 1990 et 2008, l'espérance de vie à la naissance chez les hommes blancs ayant moins de 12 ans d'études diminuait de plus de 4 ans, tandis que celle des femmes blanches ayant un niveau d'instruction comparable diminuait de plus de 5 ans comme on peut le voir sur la figure 1.7.

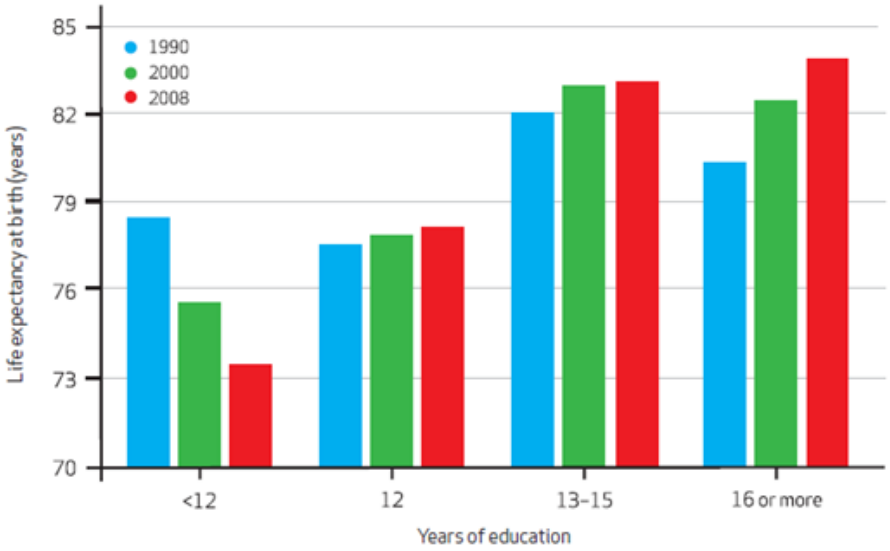


FIGURE 1.7: Espérance de vie à la naissance, selon le nombre d'années d'études à 25 ans pour les femmes blanches, 1990-2008. SOURCE : OLSHANSKY et al. (2012)

Cette étude a également révélé que la différence d'espérance de vie à la naissance entre les hommes ayant moins de 12 ans d'études et ceux ayant plus de 16 ans est passée de 13,4 ans en 1990 à 14,2 ans en 2008, tandis que pour les femmes, l'augmentation comparable était de 7,7 à 10,3 années.

La figure 1.8, indique l'évolution de l'espérance de vie à la naissance selon le nombre d'années d'études en différenciant par sexe et par ethnie.

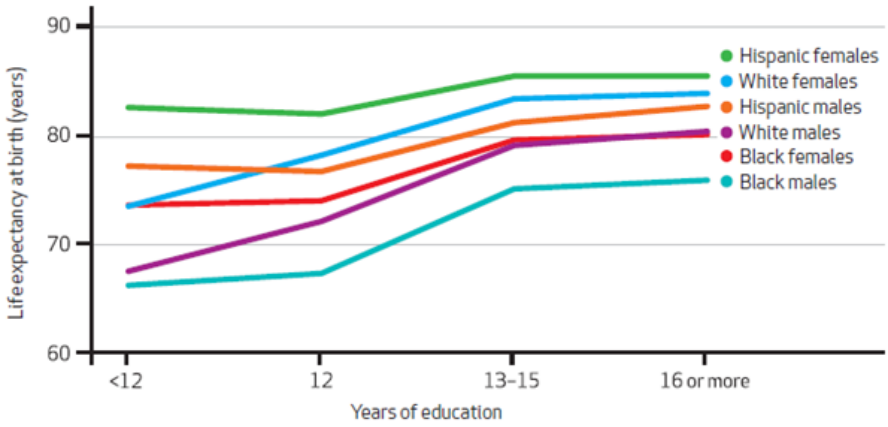


FIGURE 1.8: Espérance de vie à la naissance, selon le nombre d'années d'études à 25 ans, par ethnie et par sexe, 2008. SOURCE : OLSHANSKY et al. (2012)

Les auteurs ont constaté qu'en 2008, la différence d'espérance de vie entre les hommes noirs à haut et à faible niveau d'éducation (pour 12 ans d'études ou moins contre au moins 13 ans d'études) était de 8,4 ans, et entre les hommes blancs à haut et à faible niveau d'instruction, la différence était de 7,8 ans. Pour les femmes noires et blanches, l'espérance de vie est la même pour un niveau d'études de moins de 12 ans. Néanmoins, l'écart se creuse et semble constant en allant vers un nombre d'années d'études plus grand. En ce qui concerne les hommes noirs et blancs, en analysant les espérances de

vie, l'écart observé est encore plus large que celui des femmes. En outre, il apparaît que les femmes hispaniques se démarquent des autres catégories en ayant une espérance de vie supérieure et ce pour tous les niveaux d'éducation.

En remontant jusque dans les années 1970 aux États-Unis, des différences substantielles de mortalité entre les noirs américains et les blancs américains, toujours à l'avantage de ces derniers ont été observées. La figure 1.9 présente l'évolution de l'écart absolu d'espérance de vie à la naissance par sexe et par ethnie.

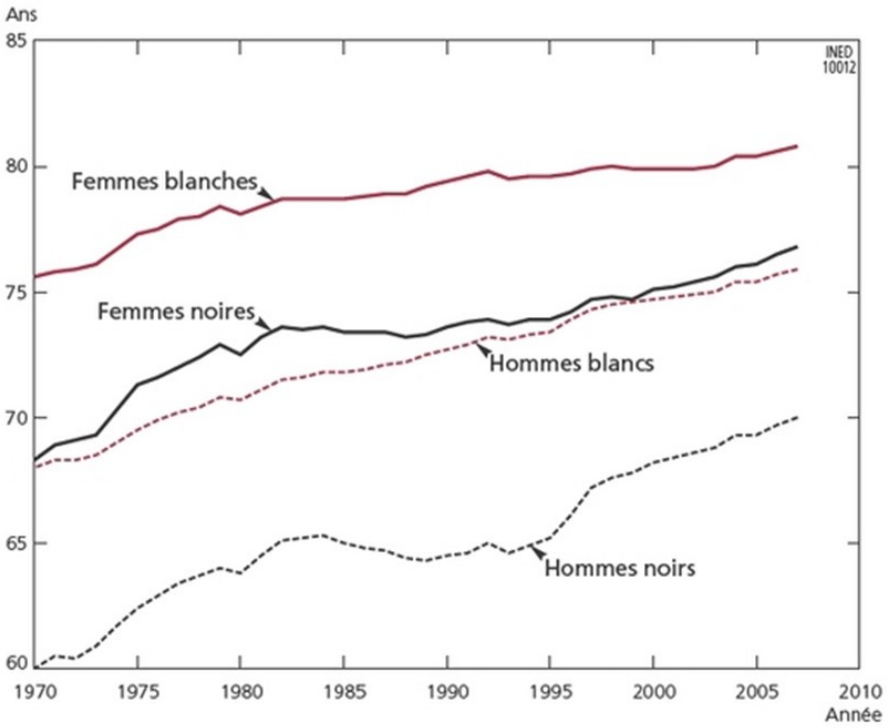


FIGURE 1.9: Espérance de vie à la naissance selon l'ethnie et le sexe aux États-Unis, de 1970 à 2007. SOURCE : ARIAS (2011)

Entre 1900 et 1982, l'écart s'est d'abord nettement réduit, passant de 14,6 à 5,7 ans tous sexes réunis. En revanche, les dix années suivantes ont été marquées par une recrudescence des disparités, due en grande partie à une augmentation de la mortalité liée au VIH/sida et aux homicides chez les hommes noirs (HARPER et al. (2007)). En 1993, l'écart atteignait ainsi 7,1 ans pour les deux sexes (8,5 ans pour les hommes et 5,8 ans pour les femmes). Depuis cette date, la situation s'est améliorée, de sorte qu'en 2007, la durée de vie moyenne des hommes et des femmes blancs étaient de 5,9 et de 4,0 ans supérieure à celle des hommes et des femmes noirs. Même s'il s'agit des écarts les plus faibles enregistrés depuis 1900 pour chaque sexe, ils demeurent considérables et ne sont que peu inférieurs (moins d'un an) à ceux qui prévalaient au début des années 1980.

À l'exception donc des personnes les plus âgées, les noirs américains ont le taux de mortalité le plus élevé de tous les groupes ethniques d'Amérique. Cette situation peut être attribuée en grande partie aux inégalités de statut économique, d'éducation et de profession, qui sont toutes liées au risque de mortalité. Mais les différences ethniques en matière de mortalité persistent même dans les études qui comparent des individus ayant des niveaux de revenus et d'éducation similaires. Le désavantage des noirs apparaît pour toutes les principales causes de décès : maladies cardiaques, cancers et accidents

vasculaires cérébraux, comme on peut le voir sur le tableau 1.1. Le désavantage est le plus important pour les décès dus à l'infection par le VIH, pour laquelle le risque pour les noirs est plus de sept fois supérieur au risque pour les américains blancs.

Cause of Death	White	Black	American Indian	Asian	Hispanic
All causes	852	1130	697	507	586
Heart disease	254	327	165	145	165
Cancers	198	250	127	125	121
Stroke	59	82	40	53	39
Injuries	36	38	60	18	31
Suicide	12	6	12	6	6
Cirrhosis of the liver	10	10	29	4	16
Homicide	4	21	8	3	8
HIV infection	3	24	3	1	7

TABLE 1.1: Taux de mortalité ajustés selon l'âge pour certaines causes de décès aux États-Unis, par race et origine hispanique, 2000. (Décès pour 100 000 personnes). SOURCE : CDC (2002)

Les taux de mortalité des amérindiens sont similaires à ceux des blancs pour la plupart des causes, mais sont sensiblement plus élevés pour la cirrhose du foie et pour les blessures, le suicide et l'homicide. En revanche, les asiatiques et les hispaniques ont des taux de mortalité beaucoup plus faibles que ce que l'on pourrait attendre de leur statut social et économique aux États-Unis. Leur avantage est particulièrement évident en ce qui concerne les principales causes de décès - les maladies cardiaques et les cancers. L'un des facteurs contribuant à la bonne santé des asiatiques et des hispaniques aux États-Unis est "l'avantage de l'immigrant". Plusieurs études ont montré que les migrants internationaux ont tendance à être des personnes particulièrement saines et optimistes, avec un meilleur régime alimentaire et des comportements plus positifs en matière de santé que les non-immigrants. En outre, les populations asiatiques et hispaniques comprennent une proportion plus élevée d'immigrants récents que les populations blanches non hispaniques ou noires non hispaniques. Mais l'avantage de l'immigrant en soi n'explique pas entièrement le paradoxe de la bonne santé des hispaniques et des asiatiques-américains.

1.3.2 Les différences de revenu

L'inégalité des revenus s'est notablement accrue aux États-Unis au cours des trois dernières décennies. L'opinion dominante parmi les économistes est que le changement technique axé sur les compétences et l'évolution des niveaux d'éducation se sont combinés pour jouer un rôle dominant dans l'inégalité des revenus et des gains, National Academies of SCIENCES et MEDICINE (2015).

Dans les années 1990, la technologie et la mondialisation ont provoqué une polarisation de l'emploi où la croissance de l'emploi était concentrée dans les emplois hautement qualifiés et à haut salaire et dans les emplois peu qualifiés et à bas salaire. En conséquence, les emplois moyennement qualifiés ont souffert. Bien que la demande de main-d'œuvre qualifiée ait continué de croître au fil du temps, l'augmentation du niveau d'éducation a ralenti. Il en a résulté une augmentation des revenus pour les personnes les plus éduquées, ce qui a provoqué une augmentation des inégalités de revenus.

Les résultats de la National Academies of SCIENCES et MEDICINE (2015) confirment les nombreuses autres études montrant que le gradient de l'espérance de vie selon le revenu a augmenté. Même parmi les personnes nées en 1930, les personnes qui se sont retrouvées dans le quintile supérieur des revenus (elles étaient dans le cinquième des revenus les plus élevés, d'après le montant qu'elles ont gagné entre 41 et 50 ans) ont une espérance de vie plus longue à 50 ans en moyenne que ceux du quintile inférieur. On estime que l'écart s'est considérablement creusé depuis lors parce que l'espérance de vie

des générations au bas de l'échelle des revenus est relativement stable ou même en baisse, alors que l'espérance de vie augmente rapidement au sommet. Par exemple, pour la cohorte d'américains nés en 1960 (qui ont atteint l'âge de 50 ans en 2010), les auteurs se sont demandés : « Comment les revenus à vie changeraient-ils si ces personnes de 50 ans étaient confrontées aux risques de mortalité de celles nées 30 ans plus tôt ? » Dans le cadre de l'étude des personnes nées en 1930 et survivant jusqu'à 50 ans, les travailleurs masculins dans la tranche inférieure des gains auraient une espérance de vie supplémentaire de 26,6 ans, de sorte qu'ils pourraient s'attendre à vivre jusqu'à 77 ans, en moyenne, comme on le constate sur la figure 1.10.

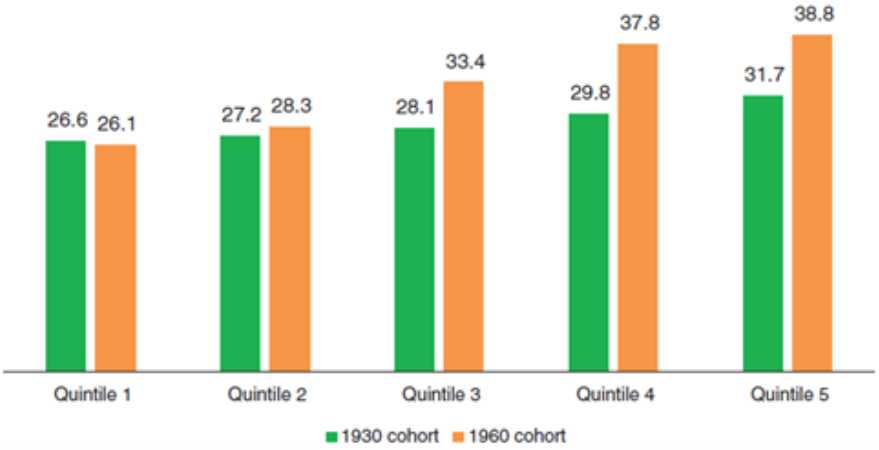


FIGURE 1.10: Espérance de vie estimée et projetée à 50 ans pour les hommes nés en 1930 et 1960, par quintile de revenu. LA SOURCE :National Academies of SCIENCES et MEDICINE (2015)

Pour les personnes nées en 1960, en supposant que les tendances de la moyenne et de la dispersion se poursuivent, l'espérance de vie à 50 ans est légèrement inférieure, à 26,1 ans. Cela signifie que pour une période de plus de 30 ans, il n'y aura pas eu de gains nets d'espérance de vie à 50 ans pour les hommes au bas de l'échelle des revenus, si ces projections se vérifient. Le résultat est différent au sommet de la distribution. Pour les hommes du quintile de revenu supérieur, l'espérance de vie à 50 ans pour la cohorte de 1930 est de 31,7 ans. Pour les personnes nées en 1960, l'espérance de vie à 50 ans devrait passer à 38,8 ans. En d'autres termes, entre la génération 1930 et la génération 1960, l'espérance de vie est à peu près inchangée pour les hommes au bas de l'échelle des revenus, mais augmente de plus de 7 ans pour ceux qui se trouvent au sommet. L'écart d'espérance de vie se creuse rapidement. Pour les hommes nés dans la cohorte 1930, l'espérance de vie du quintile le plus élevé à 50 ans est de 5,1 ans de plus que celle du quintile le plus bas. Pour les hommes nés dans la cohorte 1960, l'écart projeté s'élargit à 12,7 ans.

Pour les femmes, en regardant la figure 1.11 les résultats semblent encore plus prononcés. Elle suggère que l'espérance de vie à 50 ans pour les femmes au bas de la distribution des revenus diminue nettement entre celles nées en 1930 et les projections pour celles nées en 1960, passant de 32,3 ans à 28,3 ans. Au sommet de la distribution des revenus des femmes, cependant, l'espérance de vie devrait augmenter de plus de 5 ans. Il en résulte que l'écart d'espérance de vie entre les femmes à revenu élevé et les femmes à faible revenu devrait passer de 4 ans à 13,6 ans.

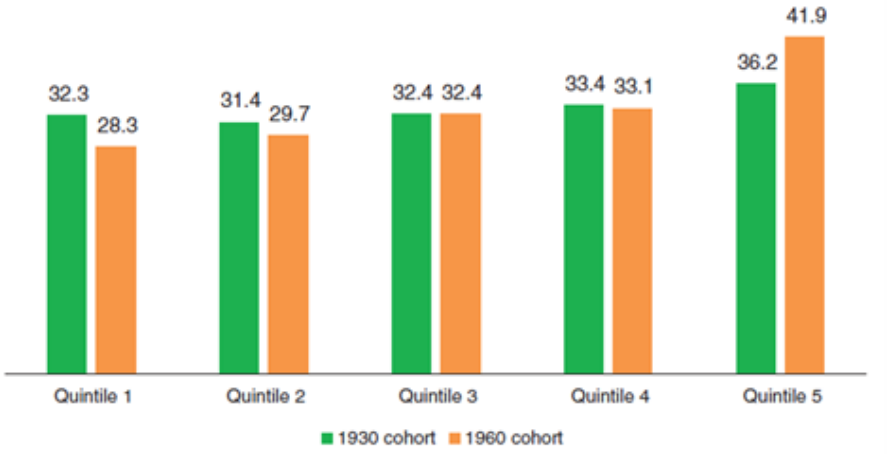


FIGURE 1.11: Espérance de vie estimée et projetée à 50 ans pour les femmes nées en 1930 et 1960, par quintile de revenu. SOURCE : National Academies of SCIENCES et MEDICINE (2015)

En somme, cette étude de la mortalité américaine a montré que l'espérance de vie a nettement augmenté aux Etats-Unis au cours du siècle dernier. Selon le dernier rapport publié par la National Academies of SCIENCES et MEDICINE (2015), il y a cinq décennies, l'espérance de vie moyenne aux États-Unis était de 67 ans pour les hommes et de 73 ans pour les femmes. Ces moyennes sont aujourd'hui de 76 et 81 ans, respectivement. Au cours des dernières décennies, plusieurs études dont celles énoncées dans ce mémoire, s'accordent à dire que le gradient de l'espérance de vie en fonction de l'éducation et du revenu s'est accentué. Ces écarts entre les personnes à revenu élevé et celles qui se situent plus bas dans la distribution socio-économique se sont accrus. L'écart se creuse également de façon notable en examinant les distributions selon l'ethnie. Il apparaît dès lors que les noirs américains ont un taux de mortalité plus élevé que les autres ethnies. Concernant les asiatiques et les hispaniques, les taux de mortalité sont beaucoup plus faibles, du fait de pratiques de vie plus saines.

Chapitre 2

Description des méthodes de partitionnement récursif

Dans cette partie assez théorique, il s'agira d'étudier les algorithmes qui seront appliqués à nos données.

Il s'agit tout d'abord du MOB (Model-Based recursive partitioning) qui combine les Modèles Linéaires Généralisés à la méthode des arbres de décision. Le but de cette approche est de partitionner les données en des groupes qui diffèrent en termes des paramètres du modèle. L'étude portera sur l'algorithme générique MOB provenant du package **parykit** duquel découle le principe des GLM Tree.

Ensuite, une alternative aux modèles paramétriques sera présenté au travers de l'algorithme CART. Ce dernier est sans doute le plus utilisé pour les arbres de classification et de régression à cause de sa facilité d'implémentation et d'interprétation. Cette approache sera implentée à l'aide du package **rpart** disponible sous R.

Le GLM Tree sera donc comparé à une approche concurrente, l'algorithme CART.

2.1 Généralités sur le MOB

Le MOB est un algorithme générique développé par Zeileis et al. (2008) et qui peut être utilisé pour différents modèles. Pour fixer la notation, considérons un modèle paramétrique $M(Y, \theta)$, où Y correspond aux observations et θ est le vecteur de paramètres à k dimensions. Ce modèle peut être un modèle de régression lorsque $Y = (y, x)$, c'est à dire que Y peut être divisé en une variable dépendante y et des régresseurs x. Un exemple de ce dernier cas pourrait être un modèle de régression linéaire $y = x^T\theta$ ou un modèle linéaire généralisé (GLM).

Dans cet algorithme, plutôt que d'ajuster un modèle global à un ensemble de données, des modèles locaux définis par partitionnement récursif seront estimés sur des sous-ensembles de données.

L'algorithme procède de la manière suivante :
(1) **Ajustement d'un modèle paramétrique** à un ensemble de données,
(2) **Test d'instabilité des paramètres** sur un ensemble de variables de partitionnement,
(3) S'il y a une certaine instabilité générale des paramètres, **partitionner le modèle** par rapport à la variable associée à la plus grande instabilité.

Enfin, on répète la procédure dans chacun des sous-échantillons résultants.

L'idée dans cet algorithme est de construire un arbre dans lequel chaque nœud est associé à un modèle de type M.

Dans la suite, les différentes étapes de l'algorithme MOB seront détaillées avant d'étudier le modèle appliqué à nos données à l'aide de la fonction `glmtree()`.

2.1.1 Estimation du modèle et des paramètres (1)

Dans cette première étape de l'algorithme, il s'agit d'ajuster le modèle une fois, à toutes les observations dans le nœud actuel en estimant $\widehat{\theta}$ via la minimisation de la fonction objectif :

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta).$$

Y_i est la $i^{ème}$ observation parmi n de la variable réponse Y.
Cette fonction objectif peut être soit la somme résiduelle des carrés, soit le négatif de la log-vraisemblance menant respectivement à une estimation des moindres carrés ordinaires (OLS) ou du maximum de vraisemblance (ML).

Sous de légères conditions de régularité, White (1994), l'estimation $\widehat{\theta}$ peut aussi être calculée en résolvant les conditions du premier ordre :

$$\sum_{i=1}^{n} \psi(Y_i, \theta) = 0,$$

où

$$\psi(Y, \theta) = \frac{\partial \Psi}{\partial \theta}.$$

ψ est la fonction de score ou la fonction d'estimation correspondant à $\Psi(Y, \theta)$. Des solutions analytiques pour $\widehat{\theta}$ ne sont disponibles que dans certains cas particuliers. Mais, pour de nombreux modèles d'intérêt, des algorithmes d'ajustement bien établis pour calculer $\widehat{\theta}$ sont disponibles (par exemple, l'estimation des moindres carrés ordinaires via la décomposition QR pour la régression linéaire ou le maximum de vraisemblance par moindres carrés pondérés itératifs pour les GLM).

La fonction de score évaluée au niveau des paramètres estimés $\widehat{\psi_i} = \psi(Y_i, \widehat{\theta})$ est ensuite contrôlée pour détecter les déviations systématiques par rapport à sa moyenne dans l'étape suivante.

2.1.2 Les tests d'instabilité des paramètres (2)

L'objectif de cette étape de l'algorithme consiste à déterminer si les paramètres du modèle ajusté sont stables pour chaque ordre particulier de la $j^{ème}$ variable de partitionnement Z_j. Cela revient à évaluer si la division de l'échantillon par rapport à l'une des variables Z_j peut capturer des instabilités dans les paramètres et ainsi améliorer l'ajustement.

Pour évaluer l'instabilité des paramètres, une idée naturelle est de vérifier si les scores $\widehat{\Psi_i}$ fluctuent de manière aléatoire ou présentent des déviations systématiques par rapport à leur moyenne. Ces

déviations peuvent être capturées par le processus de fluctuation empirique :

$$W_j(t) = \widehat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \widehat{\Psi}_\sigma(Z_{ij}), \quad (0 \leq t \leq 1). \tag{2.1}$$

où $\sigma(Z_{ij})$ est la permutation d'ordonnancement qui donne l'anti-rang de l'observation Z_{ij} dans le vecteur $Z_j = (Z_{1j}, ..., Z_{nj})^T$. Ainsi, $W_j(t)$ est simplement le processus de la somme partielle des scores ordonné par la variable Z_j, pondéré par le nombre d'observations n et une estimation appropriée $\widehat{J}$ de la matrice de covariance $cov(\psi(Y, \widehat{\theta}))$.

Ce cadre très général des tests de la stabilité des paramètres est appelé generalized Mfluctuation test. Il a été démontré qu'il englobe un grand nombre de tests de changement structurel proposé à la fois dans la littérature économétrique et statistique, notamment les tests CUSUM et MOSUM basés sur les MCO (Ploberger et Krämer (1992)) et les tests basés sur le score (Nyblom (1989)).

En principe, n'importe lequel de ces tests pourrait être utilisé dans l'algorithme de partitionnement récursif. Cependant, deux statistiques de test semblent être particulièrement intéressantes pour évaluer les variables de partitionnement Z_j qu'elles soient numériques ou catégorielles.

Évaluation des variables numériques

Pour capturer les instabilités sur une variable numérique Z_j , la fonction suivante est privilégiée :

$$\lambda_{sup_{LM}}(W_j) = \max_{i=\underline{i}..\bar{l}} \left(\frac{i}{n} . \frac{n-i}{n} \right) \left\| W_j(\frac{i}{n}) \right\|_2^2. \tag{2.2}$$

Elle représente le maximum de la norme L2 au carré du processus de fluctuation empirique pondéré par sa fonction de variance. Il s'agit de la statistique supLM d'Andrews (1993) qui peut être interprétée comme le supremum des statistiques LM par rapport à une alternative à un seul point de changement. Le point de changement potentiel est déplacé sur l'intervalle $[\underline{i}, \bar{l}]$ qui est typiquement défini en exigeant une certaine taille minimale de segment $\underline{i}$ et ensuite $\bar{l} = n - \underline{i}$.

Cette statistique de test est asymptotiquement équivalente au supremum de la statistique du rapport de vraisemblance mais a l'avantage que le modèle ne doit être ajusté qu'une seule fois sous l'hypothèse nulle de stabilité des paramètres.

Cette approche de l'évaluation des variables numériques de partitionnement combine les idées des algorithmes d'arbres de modèles (linéaires) tels que GUIDE (Loh (2002)) ou les arbres RD et RA de Potts et Sammut (2005), avec une méthodologie pour tester l'instabilité des paramètres (Zeileis (2005)). Les arbres GUIDE et RD/RA évaluent tous deux l'instabilité des paramètres selon les variables numériques de partitionnement. En outre, les arbres GUIDE et RA évaluent uniquement le signe des résidus, et non les scores du modèle complet. Enfin, les arbres RD utilisent une approximation de Bonferroni pour calculer les p-valeurs.

Ainsi, nous partons des mêmes idées, mais nous les intégrons dans un cadre pour tester la stabilité structurelle des modèles paramétriques généraux. Comme mentionné précédemment, n'importe quelle fonctionnelle λ pourrait être utilisée. Cependant, le $\lambda_{sup_{LM}}$ est particulièrement attrayant pour l'ajustement des modèles d'arbres. En effet, il réagit très bien contre les changements abrupts (Andrews et Ploberger (1994)) tels que capturés par le partitionnement, il évalue tous les points de changement concevables et est sensible aux changements de tous les éléments du vecteur de paramètres

θ. Le $\lambda_{sup_{LM}}$ sera donc utilisé dans ce cas pour évaluer l'instabilité des variables de partitionnement numériques.

Évaluation des variables catégorielles

Pour capturer l'instabilité par rapport à une variable catégorielle Z_j avec C niveaux ou catégories différents, une statistique différente est nécessaire. La statistique la plus naturelle, qui est insensible à l'ordre des niveaux C et à l'ordre des observations à l'intérieur de chaque niveau, est donnée par l'équation suivante :

$$\lambda_{\chi^2}(W_j) = \sum_{c=1}^{C} \frac{|I_c|}{n}^{-1} \left\| \Delta_{I_c} W_j(\frac{i}{n}) \right\|_2^2. \tag{2.3}$$

$\Delta_{I_c} W_j$ est l'incrément du processus de fluctuation empirique sur les observations dans la catégorie $c = 1, ..., C$ (avec les indices I_c associés), c'est-à-dire essentiellement la somme des scores de la catégorie c. La statistique de test est alors la somme pondérée de la norme L2 au carré des incréments, qui a une distribution asymptotique χ^2 avec $k.(C-1)$ degrés de liberté à partir de laquelle la p-valeur correspondante p_j peut être calculée.

L'avantage d'utiliser cette approche, basée sur les processus de fluctuation empirique de l'équation (2.1) avec les fonctionnelles des équations (2.2) et (2.3) est que les estimations des paramètres et les fonctions de score correspondantes ne doivent être calculées qu'une seule fois dans un nœud. Pour effectuer les tests d'instabilité des paramètres, les scores doivent simplement être réordonnés et agrégés en un test scalaire à chaque fois.

Pour tester s'il existe une certaine instabilité globale dans le nœud actuel, il suffit de vérifier si la valeur minimale de la p-valeur $\min_{j=1..l} p_j$ tombe en dessous d'un niveau, seuil de signification α préspécifié. Si c'est le cas, la variable Z_j associée à la p-valeur minimale est choisie pour partitionner le modèle dans l'étape suivante de l'algorithme.

2.1.3 Partitionnement des données (3)

Dans cette étape de l'algorithme, le modèle ajusté doit être divisé par rapport à la variable Z_{j^*}, variable avec le plus d'instabilité à l'issue des tests précédents. Le modèle est réparti en B segments, où B peut être fixé ou choisit de manière adaptée. Pour un nombre fixe de segmentations, deux partitions rivales peuvent être comparées facilement en comparant la fonction objective segmentée $\sum_{b=1}^{B} \sum_{i \in I_b} \Psi(Y_i, \theta_b)$.

Une recherche exhaustive de toutes les partitions concevables avec B segments est requise pour trouver la partition optimale, mais peut s'avérer fastidieuse. C'est pourquoi plusieurs méthodes de recherche sont brièvement examinées pour les variables de partitionnement numériques et catégorielles, respectivement.

Partionnement des variables numériques

La recherche exhaustive d'un découpage en B = 2 segments est réalisable en $O(n)$ opérations. Pour $B > 2$, une recherche exhaustive serait d'ordre $O(n^{B-1})$.
Cependant, la partition optimale peut être trouvée en utilisant une approche de programmation dynamique d'ordre $O(n^2)$. Par ailleurs, on peut également utiliser des algorithmes itératifs.

Si B n'est pas fixe, mais doit être choisi de manière adaptée, diverses méthodes sont disponibles. En particulier, des critères d'information (AIC/BIC) peuvent être utilisés si les paramètres sont estimés par ML.

Partitionnement des variables catégorielles

Pour les variables catégorielles, le nombre de segments ne peut pas être plus grand que le nombre de catégories $B \leq C$. Deux approches simples consisteraient à soit, de toujours diviser en tous les niveaux B = C possibles, soit de toujours diviser en un nombre minimal de segments B = 2. Dans ce dernier cas, la recherche de la partition optimale est d'ordre $O(2^{C-1})$.

Pour les variables ordinales, il est également logique de se contenter de diviser dans l'ordre des niveaux, de sorte que la recherche d'une partition binaire est seulement d'ordre $O(C)$. Là encore, les critères d'information pourraient être une option pour déterminer de manière adaptée le nombre de divisions, bien que cela soit moins intuitif que pour les variables numériques.

En résumé, deux stratégies plausibles consisteraient soit à toujours utiliser des divisions binaires, c'est-à-dire à utiliser un B = 2 fixe, soit à déterminer B de manière adaptée pour les variables numériques tout en utilisant toujours B = C pour les variables catégorielles.

Ces 3 étapes sont répétées dans chacun des B nœuds fils issus du partitionnement d'un nœud parent. Si aucune autre instabilité significative ne peut être trouvée, la récursivité s'arrête. Dès lors, vient l'étape de l'élagage de l'arbre construit afin d'optimiser la profondeur de l'arbre ou le nombre minimal d'observations dans chaque nœud.

2.1.4 Élagage de l'arbre construit

Pour déterminer la taille optimale de l'arbre, on peut utiliser une stratégie de pré-élagage ou de post-élagage.
Dans le premier cas, l'algorithme s'arrête lorsqu'aucune instabilité significative des paramètres n'est détectée dans le nœud actuel (ou lorsque le nœud devient trop petit).
Dans le second cas, il faut d'abord construire un grand arbre (sous réserve uniquement d'une exigence minimale de la taille des nœuds), puis élaguer les scissions qui n'améliorent pas le modèle ; par exemple, à en juger par des critères d'information tels que l'AIC ou le BIC (Su et al. (2004)).

Par defaut, le pré-élagage est utilisé (via les p-valeurs corrigées de Bonferroni provenant des tests de fluctuation), mais un post-élagage basé sur l'AIC/BIC est également disponible.

2.2 Les principaux arguments de la fonction `mob()`

Pour représenter les arbres résultants, le package R **partykit** est utilisé et étendu avec une infrastructure générique pour les partitions récursives où les nœuds sont associés à des modèles statistiques. Par rapport à l'implémentation précédemment disponible dans le package **party**, la nouvelle implémentation est plus facile à étendre à de nouveaux modèles, et fournit plus de fonctionnalités pratiques. La fonction `mob()` est destinée à être la fonction de base qui peut être utilisée pour explorer rapidement de nouveaux modèles. Quant aux fonctions `lmtree()` et `glmtree()`, pour respectivement le

modèle linéaire et le modèle linéaire généralisé, elles seront les interfaces utilisateurs typiques facilitant les applications.

Le traitement de la formule dans la fonction `mob()` se fait de la "manière habituelle", c'est-à-dire avec des arguments tels que, **formula** pour la formule et **data** pour les données et éventuellement d'autres arguments tels que **weights** et **offset**.

Comme il peut y avoir trois groupes de variables (la réponse y, les régresseurs x et les variables de partitionnement z), le package **Formula** (Zeileis et Croissant (2010)) est utilisé pour traiter ces trois parties. Ainsi, la formule peut être du type $y \sim x_1 + \cdots + x_k | z_1 + \cdots + z_l$ où les variables à gauche de | spécifient les données Y et les variables à droite spécifient les variables de partitionnement Z_j. Comme indiqué ci-dessus, Y peut souvent être divisé en une réponse (y dans l'exemple ci-dessus) et des régresseurs ($x_1, \ldots, x_k$ dans l'exemple ci-dessus). S'il n'y a pas de régresseurs et que l'on utilise uniquement des ajustements constants, alors la formule peut être spécifiée comme suit $y \sim 1 \mid z_1 + \cdots + z_l$. Cette dernière formulation sera utilisée dans le cadre de cette étude.

Cette représentation de la formule n'est en fait qu'une spécification de groupes de variables et n'implique rien quant au type de modèle qui doit être ajusté aux données dans les nœuds de l'arbre. La fonction `mob()` ne sait rien du type de modèle et transmet juste les variables y et x à la fonction `fit`. Seules les variables de partitionnement Z sont utilisées en interne pour les tests d'instabilité des paramètres.

En plus de ces arguments assez classiques dans la fonction `mob()`, on retrouve dans le `mob_control()`, certains paramètres comme le niveau de confiance α, le minimum d'observations dans le nœud **minsize** ou encore la profondeur maximale de l'arbre **maxdepth**.

Plusieurs exemples d'application de la fonction `mob()` sont disponibles dans l'article publié par Zeileis et al. (2008).

2.3 Caractérisation du GLM

Au cours des trente dernières années, l'utilisation des Modèles Linéaires Généralisés (GLM) (Nelder et Wedderburn (1972)) a reçu beaucoup d'attention depuis les applications de McCullagh et Nelder (1989). Les GLM sont parfaitement adaptés à l'analyse de données que l'on rencontre généralement lorsqu'on s'intéresse à des sujets liés à l'assurance (vie et non-vie).

Le modèle linéaire généralisé est caractérisé par les trois quantités suivantes qui seront vues plus en détail par la suite :

- La variable réponse Y à laquelle est associée une loi de probabilité,
- Les variables explicatives $X_1 ... X_p$,
- la fonction de lien g qui décrit la relation entre la combinaison linéaire des variables explicatives et l'espérance de la variable réponse Y.

Le modèle s'écrit donc :

$$g(\mathbb{E}(Y|X)) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i. \tag{2.4}$$

2.3.1 La variable réponse Y

On considère que la variable Y est issue d'une structure de famille de loi exponentielle. C'est à dire que la loi de Y est dominée par une mesure de référence et que la vraisemblance de Y calculée en y par rapport à cette mesure s'écrit : $f_Y(y, w, \phi) = \exp\left(\frac{yw-bw}{a\phi} - c_\phi(y)\right)$. Cette formulation inclut les lois usuelles suivantes : gaussienne, poisson, binomiale, gamma, etc. Le paramètre w est appelé paramètre naturel de la famille exponentielle. En outre, la fonction $a(\phi) = \phi$ pour certaines lois ; ϕ est dans ce cas appelé paramètre de dispersion. Il s'agit d'un paramètre de nuisance qui intervient par exemple dans le cas où les variances des lois gaussiennes sont inconnues, mais vaut 1 pour des lois à un paramètre comme la poisson ou la bernouilli, etc.

Soit Y une variable aléatoire dont la loi appartient à la famille exponentielle, alors :

$$\begin{cases} \mathbb{E}(Y) = b'(w) \\ \mathbb{V}(Y) = b''(w) \times a(\phi) \end{cases}$$

Le tableau 2.1 ci-dessous fournit quelques exemples de lois de probabilité appartenant à la famille exponentielle.

Distribution/Quantité	w	$b(w)$	$a(\phi)$	$\mathbb{E}$	$\mathbb{V}$
Gaussienne	μ	$\frac{w^2}{2}$	$\phi = \sigma^2$	$\mu = w$	σ^2
Binomiale	$ln\frac{p}{1-p}$	$n\ ln(1+e^w)$	1	np	$np(1-p)$
Poisson	$\ln(\lambda)$	$\lambda = \exp(w)$	1	$\lambda = \exp(w)$	$\lambda = \exp(w)$

TABLE 2.1: Exemples de lois usuelles appartenant à la famille exponentielle

2.3.2 Les variables explicatives

La matrice X, appelée matrice de design, regroupe l'ensemble des observations des variables explicatives. Soit β, un vecteur de k paramètres. Le prédicteur linéaire, composante déterministe du modèle est le vecteur défini par :

$$\eta = X\beta.$$

Le β_0 que l'on observe dans la formule de la prédiction est appelé intercept, il représente la classe de référence. Le profil de référence sera celui qui regroupe toutes les variables explicatives de référence (c'est à dire les modalités les plus exposées dans la base de données).

2.3.3 Fonction de lien

Cette quantité exprime une relation fonctionnelle entre la variable à expliquer et les variables explicatives.
Soit $\mu_i = \mathbb{E}(Y_i); i = 1, ..., n$. On pose que $\forall i = 1..n,\ \eta_i = g(\mu_i)$ où g, la fonction de lien, est supposée monotone et différentiable. Ceci revient donc à écrire que $\forall i = 1..n,\ g(\mu_i) = x_i\beta$. La fonction de lien

qui associe la moyenne μ_i au paramètre naturel w_i est appelée fonction de lien canonique. Dans ce cas, $\forall i = 1, .., n,\ g(\mu_i) = w_i = x_i\beta$.

De plus, il est important de souligner que l'interprétation des résultats d'un modèle linéaire généralisé dépend de la fonction de lien choisie. Par exemple, si la fonction de lien choisie est la fonction identité, alors le modèle sera additif :

$$\mathbb{E}(Y|X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

En revanche, si la fonction de lien utilisée est le logarithme, le modèle sera multiplicatif :

$$\ln(\mathbb{E}(Y|X)) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \ \text{ et donc } \ \mathbb{E}(Y|X) = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right).$$

Le principe d'offset et d'exposition

Lors de la modèlisation des données d'expérience provenant de l'assurance vie, l'exposition est généralement prise en compte. L'exposition de chaque observation du portefeuille étudié impacte la probabilité de survenance d'un sinistre, et doit par conséquent être intégrée dans la modélisation.

L'offset θ, est un élément du prédicteur linéaire dont le coefficient est contraint à 1. Pour intégrer ce phénomène au Modèle Linéaire Généralisé, l'écriture du modèle est remplacée par :

$$g\left(\mathbb{E}[\frac{Y}{\theta}|X]\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

où θ correspond à l'offset.

Si notre variable Y est modélisée à l'aide d'une loi de poisson et donc que notre fonction de lien est la fonction de lien logarithmique, le modèle s'ecrit :

$$\ln(\mathbb{E}(Y|X) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \ln(\theta).$$

L'offset permet donc de rajouter une information supplémentaire dans le prédicteur linéaire. Dans le cas de la loi de poisson, en supposant que Y correspond au nombre de décès et θ à l'exposition, il permet de capturer la force de mortalité. En effet, l'utilisation de la fonction de lien logarithmique permet de faire rentrer l'exposition au dénominateur.

2.4 Ratio A/E

L'idée de cette étude étant de capturer des déviations via le ratio A/E en %, nous allons dans cette partie étudier ce dernier pour comprendre ce à quoi il correspond en pratique.

2.4.1 Ratio A/E ou SMR en médecine

Le ratio A/E (Actual/Expected) ou ratio standardisé de mortalité (SMR) dans les études de médecine indique si une population spécifique (par exemple, les patients d'un certain hôpital) a plus, moins ou autant de chances de mourir qu'une population de référence.

Les taux de mortalité bruts décrivent le nombre de personnes qui meurent dans une population donnée au cours d'une période spécifique, mais ne tiennent pas compte de la répartition par âge de la population par exemple. Il est clair que les populations ayant un nombre élevé de personnes âgées sont susceptibles d'avoir un taux de mortalité plus élevé. Par conséquent, il est souvent plus juste de comparer les hôpitaux ou les tendances du taux de mortalité dans le temps en tenant compte de la répartition par âge de la population.

Il est possible de comparer chaque catégorie individuelle d'âge et de sexe, mais avec des catégories supplémentaires, par exemple le statut fumeur/non-fumeur, les données deviennent ingérables. Au lieu de cela, la standardisation nous donne un seul chiffre pour comparer les taux de mortalité dans deux populations tout en tenant compte des différentes distributions de la population. Cela permet de transmettre l'information plus simplement, bien que la contrepartie soit la perte de certaines informations.

2.4.2 Le calcul du ratio A/E

Les données de mortalité peuvent être standardisées en utilisant la méthode directe ou indirecte. On utilise la méthode directe lorsqu'on connaît les taux spécifiques âge-sexe de la population étudiée et la structure âge-sexe de la population standard. De plus amples détails sur la standardisation directe et sur la façon de la calculer sont disponibles (Curtin et Klein (1995)).

La méthode indirecte pour calculer le ratio A/E est utilisée lorsque les taux spécifiques par âge pour la population étudiée sont inconnus ou non disponibles. Cette méthode utilise donc le nombre observé de décès dans la population étudiée et le compare au nombre de décès qui serait attendu si la distribution par âge était la même que celle de la population standard. C'est cette seconde approche qui sera utilisée pour calculer notre ratio dans les applications suivantes. Pour minimiser les biais, la distribution par âge de la population standard doit être aussi similaire que possible à celle de la population étudiée.

Un avantage de l'utilisation du A/E est que la variance des taux indirectement standardisés est plus faible que celle des taux directement standardisés, ce qui donne des estimations plus précises. Par conséquent, lorsque le nombre de décès est faible, ce qui est souvent le cas dans les hôpitaux, et dans des portefeuilles d'assurance vie, le A/E est privilégié.

Le ratio A/E est généralement calculé en utilisant des catégories spécifiques à l'âge et au sexe. Cependant, des variables supplémentaires peuvent être rajoutées.

Le ratio A/E se calcule de la manière suivante :

$$A/E = \frac{nombre\ de\ décès\ observé}{nombre\ de\ décès\ attendu}.$$

Parfois, le A/E est exprimé après avoir été multiplié par 100. Dans ce cas :

- $A/E < 100$ indique un nombre de décès inférieur à celui attendu.

- $A/E = 100$ indique que les décès observés sont égaux aux décès attendus.
- $A/E > 100$ indique qu'il y a eu un excès de décès.

Cependant, pour déterminer si le A/E est significativement différent de 100%, les intervalles de confiance (IC) à 95% doivent être calculés.

Une difficulté à manipuler le ratio A/E concerne le choix de la population "standard". Il est impossible de comparer les ratios sans utiliser la même population standard. Outre cette limite, le A/E reste une quantité facile à calculer et à interpréter.

2.4.3 Les intervalles de confiance

Comme mentionné auparavant, les intervalles de confiance sont nécessaires afin de déterminer si le ratio calculé est significativement différent de 100% ou non. Pour capturer les déviations en fonction du nombre de décès, Liddell (1984) fournit des formules exactes pour les bornes inférieure et supérieure de notre intervalle de confiance. L'IC à 95% est alors définit par :

$$IC = \left[\frac{A}{E} \times (1 - (9 \times A)^{-1} \pm z_{1-\frac{\alpha}{2}} \times (9 \times A)^{-\frac{1}{2}})^3\right].$$

Avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite. Pour $\alpha = 5\%$, les tables indiquent $z_{1-\frac{\alpha}{2}} = z_{97.5\%} = 1{,}96$.

Il est également possible de calculer le ratio A/E dans d'autres quantités que le nombre de décès. En effet, celui-ci peut être basé sur le montant assuré par les individus. Dans ce cas, les quantiles de la loi de Tweedie fournissent une approximation pour la construction d'un intervalle de confiance à 95%, Tweedie (1984).

Les distributions de Tweedie sont une famille de distributions incluant les lois normale, gamma, de poisson et poisson-gamma, agrémentées d'une masse en zéro. Pour toute variable aléatoire Y suivant une distribution de la famille de Tweedie, on a la relation :

$$\mathbb{V}(Y) = \phi\ \mathbb{E}(Y)^p.$$

Avec ϕ et p des constantes positives telles que, ϕ est le paramètre de dispersion et p l'indice de la distribution.

Ainsi, pour p = 0, on obtient une distribution normale, pour p = 1, une distribution de poisson, pour p = 2, une distribution gamma, etc.

Les bornes de l'intervalle de confiance sont obtenues à l'aide des expressions suivantes :

$$Borne\_Inf = qtweedie_{0.025}(\mu = A, p, \phi)/E,$$

$$Borne\_Sup = qtweedie_{0.975}(\mu = A, p, \phi)/E.$$

Où μ, la moyenne correspond ici au montant observé A. p et ϕ sont respectivement les paramètres de puissance et de dispersion.

2.4.4 Descriptif du modèle appliqué

Concernant les deux prochaines applications, il sera question d'utiliser une extension de l'algorithme MOB, la fonction `glmtree()`. Cette approche combine des modèles paramétriques tels que les modèles linéaires généralisés aux arbres de décision. Cette méthode des GLM a été choisie pour faciliter les applications. En particulier, la régression logistique sera utiliséée pour modéliser les décès observés par rapport à ceux prédits. Cela permettra de modéliser le ratio A/E (Actual/Expected) en %, pour mesurer les déviations, ce qui facilitera les calculs.

Pour la première étape de l'algorithme où un modèle paramétrique est ajusté à un ensemble de données, le nombre de décès observé A (Actual) est supposé être distribué de manière binomiale.

Chaque cellule est déterminée par une combinaison unique de covariables :

$$A \sim \mathcal{B}(n, q)$$

où nos paramètres n et q représentent respectivement l'exposition totale et la probabilité de décès. L'utilisation de la fonction de lien *logit()* pour notre modèle conduit à l'équation suivante :

$$logit(q) = \ln(\frac{q}{1-q}) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

Cependant, dans notre modèle, les variables explicatives ne seront pas utilisées. Seules les variables de partitionnement interviendront dans l'équation. La régression se fera donc uniquement selon l'intercept β_0. Il est courant d'utiliser la loi de poisson pour modéliser le nombre d'occurences d'un certain événement. En effet, le nombre de décès étant une variable aléatoire de comptage selon Brillinger (1986), l'hypothèse de poisson semble plausible. Cependant, cette distribution ne permet pas d'intégrer les poids de manière adéquate afin de définir la taille des groupes. Dans la version actuelle de l'algortihme, l'utilisation de la loi de poisson utilise le nombre de lignes comme poids plutôt que l'exposition.

Cela explique le choix de la loi de distribution binomiale. L'utilisation de cette dernière nécessitera une approximation qui impactera la précision des déviations calculées sans toutefois remettre en cause l'analyse. En définissant l'exposition comme les poids, l'équation suivante est obtenue :

$$\ln\left(\frac{q \times n}{(1-q) \times n}\right) = \ln\left(\frac{\mathbb{E}(A)}{n - \mathbb{E}(A)}\right) = \beta_0.$$

L'idée étant de trouver les variations de β_0 et de capturer des déviations, une information supplémentaire sera rajoutée au travers de l'offset et la fonction *ln()*. Cette quantité correspondra ici au taux de mortalité espéré μ^{ref}.
En outre, l'approximation générée par l'utilisation de la loi binomiale est la suivante : $1 - q \approx 1$. Cette approximation tire sa source des taux de décès q très faibles, de l'ordre de 10^{-3}. Cela revient à supposer que $n - A \approx n$, car il survient très peu de décès dans notre ensemble de données. On obtient donc l'équation suivante :

$$\ln(\frac{\mathbb{E}(A)}{n}) = \beta_0 + log(\mu^{ref}),$$

La quantité β_0 correspond donc à la quantité suivante :

$$\begin{aligned}\beta_0 &= \ln(\tfrac{\mathbb{E}(A)}{n}) - log(\mu^{ref}) \\ &= \ln(\tfrac{\mathbb{E}(A)}{n \times \mu^{ref}}) \\ &= \ln\left(\tfrac{\mathbb{E}(A)}{E}\right)\end{aligned}$$

C'est à dire :

$$\beta_0 = \ln\left(\frac{\mathbb{E}(A)}{E}\right).$$

Où $E = n \times \mu^{ref}$ représente le nombre de décès espéré (Expected). Nos déviations seront capturées par l'ntercept β_0.

Ce modèle binomial sera appliqué à nos deux jeux de données afin de capturer les déviations observées via la quantité β_0. Cette première étape commune à nos applications prochaines consitute la première phase de l'algorithme du `glmtree()`.

2.5 L'algorithme CART

En comparaison au GLM Tree, et dans le but d'évaluer la pertinence de celui-ci, le CART sera également appliqué à nos données.

Les arbres de décision sont des techniques d'apprentissage automatique qui servent à construire des modèles prédictifs à partir d'un échantillon de données, Timofeev (2004). Ces techniques consistent en un partitionnement récursif de l'échantillon de données initial et la construction d'un modèle prédictif sur chaque partition. Ces techniques peuvent être classées en deux familles : les arbres de classification et les arbres de régression. Les arbres de classification ont pour but de prédire une variable qualitative. Les arbres de régression permettent quant à eux de prédire une variable quantitative.

Le CART présente un intérêt actuariel dès lors qu'il est question de construire des classes de risques, Barbaste (2017). Plusieurs applications sont disponibles dans ce domaine. Notamment, un arbre de régression peut être construit pour l'étude du risque de maintien en arrêt de travail, Lopez et al. (2015). La variable expliquée est alors la durée moyenne dans l'arrêt (en jours). Les variables explicatives quant à elles peuvent être la classe d'âge à l'entrée, le sexe de l'assuré et la cause de l'arrêt.

L'algorithme CART est un arbre de décision binaire. À chaque fois que l'on divise l'échantillon, celui-ci est séparé en deux sous-échantillons grâce à une condition sur les variables explicatives. L'algorithme CART procède en trois étapes :

- Création de l'arbre maximal,
- Élagage de l'arbre maximal,
- Choix de l'arbre optimal parmi les arbres élagués.

La base de données est découpée comme suit :

– un échantillon d'apprentissage pour construire le modèle prédictif,

– un échantillon de validation pour calibrer le modèle,

– un échantillon de test pour tester le modèle.

L'échantillon d'apprentissage est composé de n observations. Pour chaque observation i, $1 \leq i \leq n$, le but est d'expliquer la variable réponse Y avec les variables explicatives X_j, $1 \leq j \leq p$. Pour expliquer la variable Y , les modèles par arbre de décision partitionnent récursivement l'échantillon d'apprentissage grâce aux variables explicatives X_j, $1 \leq j \leq p$.

2.5.1 Création de l'arbre maximal (1)

Abre de classification

Dans le cas d'un arbre de classification, la variable à expliquer Y est une variable qualitative. Autrement dit, on cherche à connaître la classe de Y. L'exemple d'arbre de classification le plus connu est celui de l'étude sur la survie des passagers à bord du Titanic. En effet, grâce aux caractéristiques d'un individu (âge, sexe, prix du ticket, etc), il est possible d'estimer avec précision si celui-ci est décédé ou a survécu, Milla (2017).

On appelle couramment :

– La racine : l'ensemble de la population. C'est l'échantillon initial qui n'a pas encore été segmenté.

– Les branches : les règles qui permettent de segmenter la population en deux.

– Les nœuds : les sous-échantillons qui sont créés dans l'arbre.

– Les feuilles : les sous-populations homogènes créées qui donnent un estimateur de ce que l'on souhaite estimer.

Le but de l'arbre est de partitionner la population initiale en sous-populations. L'idée est de couper la racine en deux feuilles plus homogènes selon une règle. Cette dernière est construite en choisissant une seule variable explicative X_i, ainsi qu'un seuil j comme le montre la figure 2.1.

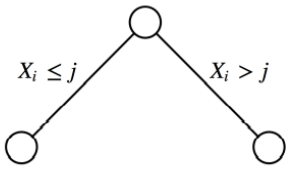


Figure 2.1: Partitionnement de la racine par rapport à la variable X_i au seuil j

Cependant,la variable explicative ainsi que son seuil ne sont pas choisis au hasard. La force de l'algorithme de CART repose sur le choix optimal de la variable explicative ainsi que de son seuil pour segmenter la population en deux sous-populations les plus homogènes possibles. Un critère est donc nécessaire pour mesurer l'homogénéité d'une population. En classification, les fonctions d'impureté sont utilisées. On note :

– t : le nœud en question,

- t_{gauche} : le nœud fils gauche,
- t_{droite} : le nœud fils droite,
- P_{droite} : la proportion d'individus dans le nœud droit,
- P_{gauche} : la proportion d'individus dans le nœud gauche,
- i(.) : la fonction d'impureté choisie

Le principe est de segmenter chaque nœud en deux, en minimisant l'impureté des deux nouveaux nœuds. Pour trouver la meilleure variable explicative ainsi que son meilleur seuil, on résout le problème de maximisation suivant :

$$\min_{X_i \leq j} P_{gauche} \times i(t_{gauche}) + P_{droite} \times i(t_{droite}).$$

On note :

- t : le nœud dont on veut connaître l'impureté,
- K : le nombre de classes possibles,
- p(k—t) : la proportion d'individus de classe k dans le nœud t.

La fonction d'impureté la plus utilisée pour les arbres CART en classification est l'indice de Gini. Il se définit comme suit :

$$i(t) = 1 - \sum_{k=1}^{K} p(k|t)^2.$$

À chaque nœud, pour séparer la population en deux sous-populations, on teste chaque variable explicative avec chaque seuil possible et on retient le couple qui minimise l'impureté des deux nouveaux nœuds. Ceci entraîne la création de deux nouvelles populations. Le processus est répété sur chacun des nouveaux nœuds crées.

Le processus s'arrête lorsque l'on ne peut plus segmenter les nœuds obtenus. Ceci se produit quand il n y a plus qu'une seule observation dans la population ou lorsque les observations de la population ne peuvent plus être différenciées avec les covariables du modèle.

Abre de régression

Dans le cas où l'arbre de décision est un arbre de régression, la variable à expliquer est une variable continue. On cherche ainsi à connaître la valeur d'une quantité d'intérêt.

On note :

- Y : la variable réponse,
- p : le nombre de covariables,
- X_j, $1 \leq j \leq p$: les covariables,

– π_0 : la quantité d'intérêt que l'on veut prédire.

Dans la plupart des cas, la quantité d'intérêt que l'on veut prédire est :

$$\pi_0 = \mathbb{E}[Y|X = x].$$

Un quantile peut être choisi comme quantité d'intérêt. Pour cela, il suffit de choisir le bon critère pour mesurer l'homogénéité des nœuds.

Dans le cas où la quantité d'intérêt est l'espérance, la fonction de perte utilisée sera l'erreur de généralisation des moindres carrés ou mean squared error (MSE) : $\mathbb{E}[(\pi(x) - Y)^2]$.

L'espérance étant la solution de la minimisation de l'erreur quadratique ; en pratique, les calculs d'espérance sont faits de façon empirique. Ainsi, on cherche :

$$\pi_n(x) = \arg\min_{\pi(x)} \mathbb{E}_n[\phi(Y, \pi(x))|X = x],$$

avec $\phi(Y, \pi(x)) = (Y - \pi(x))^2$.

On note :

– $V_n(.)$: la variance empirique,

– $E_n(.)$: l'espérance empirique.

$$E_n = \frac{1}{n}\sum_{i=1}^{n} y_i \quad et \quad V_n = \frac{1}{n}\sum_{i=1}^{n}(y_i - E_n[Y])^2.$$

Ayant $E[(Y-E[Y])^2] = Var[Y]$, cela revient à choisir la variance empirique comme critère de sélection. Pour construire l'arbre, le but est de segmenter chaque nœud en deux nœuds fils, en minimisant la variance des deux nouveaux nœuds. À chaque nœud construit, le nouvel estimateur de $E[Y]$ devient l'espérance empirique de l'ensemble des observations du nœud en question.

Pour trouver la meilleure variable explicative ainsi que son meilleur seuil, on résout le problème de maximisation suivant :

$$\min_{X_i \leq j} P_{gauche} \times V(t_{gauche}) + P_{droite} \times V(t_{droite}).$$

De la même manière que pour les arbres de classification, pour séparer l'ensemble en deux sous-ensembles plus homogènes, on teste chaque variable explicative et chaque seuil et on retient le couple qui minimise la variance des deux nouveaux nœuds. Le processus est répété sur chaque nouveau nœud créé.

L'algorithme de la création de l'arbre maximal est alors le suivant :
1. On se place au niveau de la racine de l'arbre.
2. On minimise la somme pondérée de l'impureté (la variance) des deux nœuds fils en testant chaque covariable et chaque seuil pour la classification (pour la régression).
3. Si on ne peut plus segmenter, on s'arrête. Sinon, on recommence l'étape 2 sur chaque nœud fils.

2.5.2 Élagage de l'arbre maximal (2)

Le principe de l'élagage

Élaguer un arbre signifie extraire un sous-arbre à partir de celui-ci. Prenons l'arbre de la figure 2.2 qui n'est pas encore élagué et supposons que l'on veuille élaguer les branches qui suivent le premier nœud fils gauche.

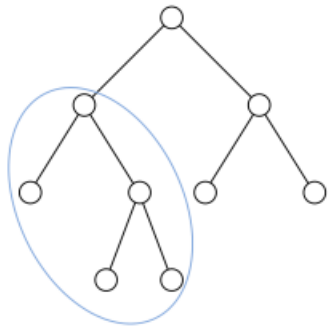


FIGURE 2.2: Arbre avant élagage

On souhaite donc élaguer les branches qui sont dans la partie entourée. La figure 2.3 présente le nouvel arbre élagué.

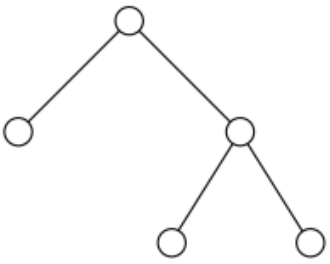


FIGURE 2.3: Arbre après élagage

On est donc passé de 5 feuilles à 3 feuilles. Soit un sous-arbre T qui contient $|T|$ feuilles. On définit le taux d'erreur relatif de ce sous-arbre :

$$R(T) = E_n[\phi(Y, \pi(x))|X = x]$$

Dans le cas d'un arbre de classification, cette erreur correspond à la moyenne pondérée des taux de mauvaise classification des feuilles (avec l'indice de Gini). Dans le cas d'un arbre de régression pour lequel la quantité d'intérêt est l'espérance, cela correspond à la moyenne pondérée des variances empiriques des feuilles.

La fonction coût-complexité

Soit un arbre T composé de $|T|$ feuilles. Sa fonction coût-complexité est égale à :

$$\begin{aligned} R_\alpha(T) &= R(T) + \alpha \times |T| \\ &= E_n[\phi(Y, \pi(x))|X = x] + \alpha \times |T|. \end{aligned}$$

On note :

$$\begin{aligned} T(\alpha) &= \arg \min_{T \leq T_{max}} R_\alpha(T) \\ &= \arg \min_{T \leq T_{max}} E_n[\phi(Y, \pi(x))|X = x] + \alpha \times |T|. \end{aligned}$$

$T(\alpha)$ correspond au sous-arbre de l'arbre maximal T_{max} qui minimise la fonction de complexité en α. Bien évidemment, $T(0) = T_{max}$.

La procédure est assez simple : on fait croître itérativement α de 0 à $+\infty$ et on sélectionne pour chaque α, le sous-arbre de T_{max} qui minimise la fonction de coût-complexité en α.

Avec cette procédure, nous créons une suite de sous-arbres optimaux de T_{max}, ainsi qu'une suite de nombres $0 < \alpha_1 < ... < \alpha_{np}$ où n_p désigne le nombre de sous-arbres créés avant de tomber sur la racine. L'existence de cette séquence a été démontrée par Breiman et al. (1984).

D'après Breiman et al. (1984), nous avons :

$$\forall i \in [0, n_p], T(\alpha_i) \subset T(\alpha_{i-1}).$$

et :

$$T(\alpha) = T(\alpha_i), \forall \alpha_i < \alpha < \alpha_{i+1}.$$

Cette procédure permet d'obtenir une suite de sous-arbres emboîtés optimaux de l'arbre maximal.

Quelques résultats importants de la procédure d'élagage

- Le cas où $\alpha = 0$ correspond à l'arbre maximal et le cas où $\alpha = \alpha_{n_p}$ correspond à la racine.
- Pour α fixé, T(α) est unique.
- Pour $i = 1, 2, ..., (n_p - 1)$, on pose $I_i = [\alpha_i; \alpha_{i+1}]. \forall \alpha \in I_i$, T($\alpha$) est unique.

La dernière étape de l'algorithme de CART consiste à choisir le meilleur arbre optimal parmi la suite des sous-arbres emboîtés de l'arbre maximal.

2.5.3 Choix de l'arbre optimal parmi les arbres élagués (3)

La méthode la plus simple pour choisir le meilleur arbre est d'utiliser un échantillon de validation. En effet, il suffit de prédire la réalisation de la variable réponse de chaque individu de l'échantillon de validation et de la comparer avec la réalisation réelle. Le sous-arbre optimal qui est le plus proche de la réalité sera alors choisi. La méthode de ré-échantillonage la plus efficace dans le choix du meilleur sous arbre optimal est la validation croisée.

Le principe de l'utilisation de la validation croisée dans le choix du meilleur sous-arbre est le suivant :

- Construction de l'arbre maximal sur toute la population.

- Obtention des paramètres de complexité $0 < \alpha_1 < ... < \alpha_{n_p} < +\infty$ par le critère coût-complexité de la méthode d'élagage.

- Construction d'une nouvelle suite de paramètres de complexité : $\beta_1 = 0$, $\beta_2 = \sqrt{\alpha_1.\alpha_2}$, ..., $\beta_{n_p-1} = \sqrt{\alpha_{n_p-2}.\alpha_{n_p-1}}$, $\beta_{n_p} = +\infty$.

- Division de la population totale en k sous-groupes $G_1, G_2, ..., G_k$ de même taille.

- $\forall i \in [[1; k]]$:
  1. Sur la population totale privée de G_i, on détermine un arbre à partir des règles de l'arbre maximal.
  2. On construit les sous-arbres $T(\beta_1), .., T(\beta_{n_p})$ et on prédit les réalisations de chaque observation de G_i sur chaque sous-arbre.
  3. On calcule l'erreur sur chaque sous-arbre.

- Pour chaque paramètre de complexité β, on somme les erreurs des sous-groupes G_i. On trouve alors le paramètre de complexité β_{min} qui minimise cette somme et on retient $T(\beta_{min})$ comme meilleur sous-arbre.

2.5.4 La validation croisée

Le principe de la validation croisée (ou k-Fold Cross Validation) est de découper la base de données en k échantillons de même taille. Cela consiste à découper le dataset en K sous-ensemble (ou K folds) puis prendre un des K sous-ensemble comme dataset de validation (validation set) et les K-1 restants comme dataset d'entrainement (training set). La procédure est repétée sur toutes les combinaisons possibles. Les modèles ainsi construits disposent chacun d'un score de performance. La moyenne des scores permet d'obtenir un score de performance moyen. La figure 2.4 illustre cette technique :

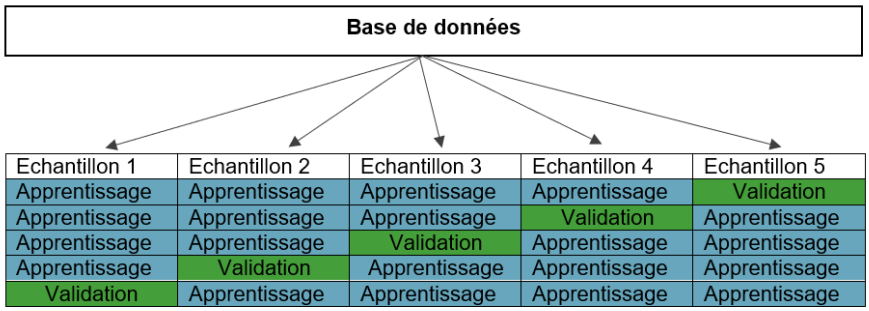


FIGURE 2.4: Validation croisée avec k=5

Le cas particulier où nous prenons $k = n^1$ s'appelle la validation croisée leave-one-out. C'est la technique qui permet d'avoir une mesure de la performance de notre modèle avec le biais le plus faible mais en échange d'un temps d'exécution très long. Cette technique est intéressante lorsque la base de données est petite.

Comparaison modèles linéaires VS CART

Les deux algorithmes présentés plus haut ont tous deux leurs points fort et faible comme le suggère le tableau 2.2 ci dessous.

Caractéristiques	Modèles Linéaires	CART
Hypothèse de linéarité	✓	✗
Hypothèse de distribution	✓	✗
Données manquantes	✗	✓
Intervalles de confiance, p-valeur	✓	✗
Validation croisée intégrée	✗	✓

✗: non; ✓: oui

TABLE 2.2: Comparaison modèle linéaire et CART

Par la suite, nous verrons comment ces derniers se comportent avec nos données au delà de ces informations. Au vu des résultats, il sera plus facile de discerner l'algorithme le plus performant, facile à implémenter et à interpréter.

Chapitre 3

Étude des déviations de mortalité de la population assurée aux USA

3.1 Objectif de l'étude

Après avoir décrit les algorithmes et modèle qui seront appliqués aux données, vient l'étape des applications. La première concerne les données de l'industrie américaine où les déviations de mortalité sont capturées par rapport à la table réglementaire 2015 VBT, SOA (2018). La Valuation Basic Table capture déjà certains effets tels que l'âge, le sexe, la durée passée dans le contrat (duration), et le statut fumeur/non-fumeur.

L'idée est donc de construire un arbre afin de regrouper les segments d'individus où on observe un comportement similaire, en d'autres termes où on détecte les mêmes déviations de mortalité par rapport à la table réglementaire. Cette approche permettra également d'identifier les variables qui discriminent au mieux la différence entre mortalité observée et prédite.

3.2 Source des données et définition de l'échantillon

Les données de l'industrie américaine proviennent de la SOA (Society Of Actuaries), l'équivalent de l'IA (Institut des Actuaires) en France. Elles sont publiques[†]. La SOA a fait appel à l'ILEC (Individual Life Experience Committee) par ses 18 membres pour concevoir le projet Committee (2017). La SOA a passé un contrat avec le MIB's Actuarial and Statistical Research Group pour collecter, valider et compiler les données de ce rapport.

Les données ont été collectées pour la période de 2003 à 2013. Les données de 2003 à 2009 proviennent des études volontaires de l'ILEC. Celles pour les périodes d'observation 2009 à 2013 proviennent de la collecte obligatoire de la Life Statistical Services (LSS) pour l'État de New York de 2009 à 2013 et de l'État du Kansas de 2011 à 2013.

À partir de la base de données originale contenant près de 26 millions de lignes et 33 variables, le traitement de la base a conduit aux variables contenues dans le tableau 4.1.

[†]https ://www.soa.org/resources/experience-studies/2017/2009-13-indiv-life-ins-mort-exp/

Variables	Description	Type	Modalités
Insurance Plan	Le type de produit souscrit	Qualitative	Perm Term UL VL ULSG VLSG
Gender	Le sexe de l'individu	Qualitative	F M
Duration	La durée passée dans le contrat	Quantitative	1-2 3-7 8-12 13-19 20-24 25+
Face amount band	Le montant assuré	Quantitative	1-49999 50000-99999 100000-249999 250000-499999 500000-999999 1000000$-10M$+
Attained age	L'âge de l'individu à la souscription du contrat (par tranche de 5 ans)	Quantitative	18-22 23-27 28-32 33-37 ... 113-117
Smoker status	Cette variable informe du statut fumeur ou non fumeur de l'individu	Qualitative	Smoker Non Smoker
Risk class	Une variable plus granulaire que le statut fumeur/non-fumeur avec 6 modalités. Cette variable indique la classe de risque dans laquelle se trouve l'individu avec SS qui est la classe avec les individus en moins bonne santé et SPNS qui est celle avec les individus en meilleure santé.	Qualitative	SS PS SNS S+NS PNS SPNS
Expected Count (EC)	Le nombre de décès espéré et prédit par la table standard	Quantitative	-
Number of Deaths (A)	Le nombre de décés réellement observé dans les données de l'industrie	Quantitative	-

TABLE 3.1: Descriptif des variables utilisées

Concernant le traitement des variables, il y a d'abord eu la suppression des lignes de la base contenant des valeurs manquantes. Puis, certaines variables qui possédaient un nombre élevé de modalités ont été retraitrées. Il s'agit des variables : Duration, Face amount band, attained age et risk class.

Les variables Duration et Face amount band sont passées de respectivement 33 et 11 modalités à des variables de 6 modalités chacune. Ensuite, la variable attained age a été regroupée par groupe d'âge quinquennal. Enfin, la variable risk class provient de la combinaison de 3 variables de la base originale : preferred class, number of preferred classes et smoker status. Les modalités suivantes sont obtenues : SS (Standard Smoker), PS (Preferred Smoker), SNS (Standard Non Smoker), S+NS (Standard + Non Smoker), PNS (Preferred Non Smoker), SPNS (Super Preferred Non Smoker) telles que mentionnées dans le tableau 4.1.

En aggrégant ces données avec nos 9 variables, cela permet de réduire considérablement la taille de notre base qui passe désormais à 14 millions de ligne.

Les sept premières variables (à l'exception de smoker status) font partie des variables de partitionnement Z_j evoquées précédemment. Les autres sont celles qui serviront à modéliser les déviations de mortalité en termes de A/E (Number of Deaths A / Exposure Count EC)

3.2.1 Statistiques descriptives sur les données

Nous allons à présent voir certaines statistiques descriptives sur nos données afin de cerner la base que nous allons manipuler. Les variables étudiées dans cette partie vont se révéler pertinentes dans la suite.

Commençons par la distribution par âge et par statut fumeur/non-fumeur telle que représentée sur la figure 3.1.

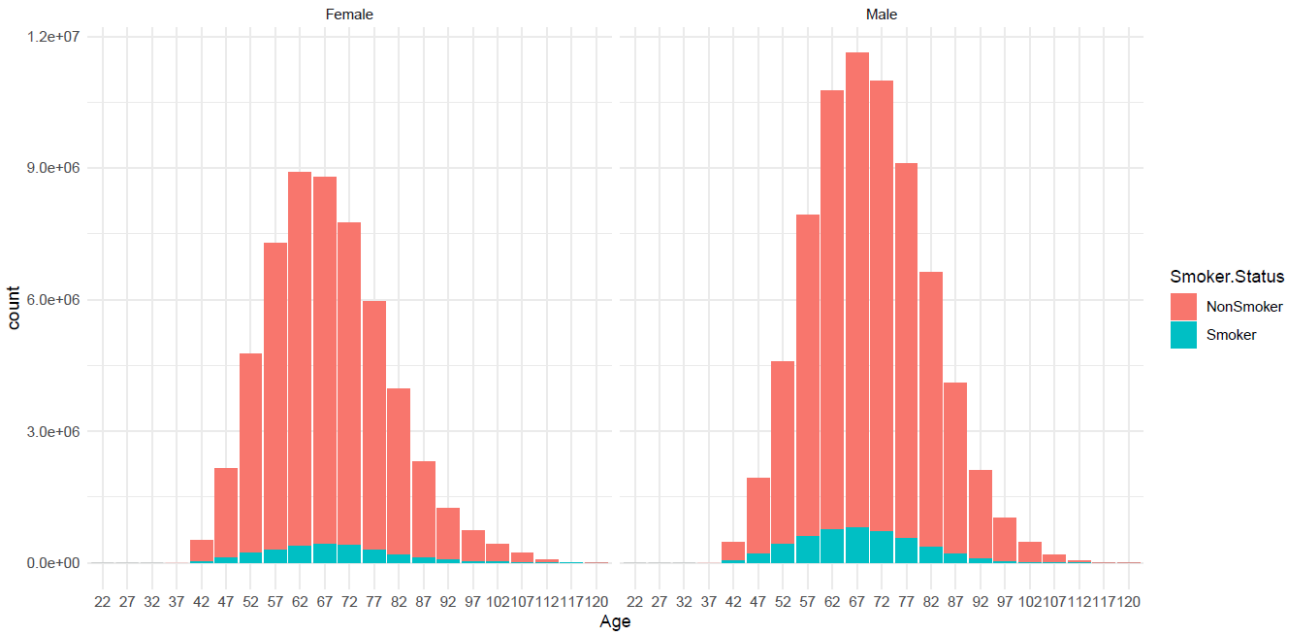


FIGURE 3.1: Distribution de la population par âge et par statut fumeur/non-fumeur.

La distribution de l'exposition pour les deux sexes indique un nombre plus élevé d'hommes que de femmes dans notre base. Pour les deux sexes, la population globale est assez âgée : l'âge moyen du portefeuille est de 68 ans. On s'aperçoit de la faible population présente aux âges extrêmes et ce pour les deux sexes. Cela semble logique car il est rare pour les plus jeunes de souscrire un contrat d'assurance et il y a naturellement moins de personnes aux âges élevés.

Concernant le statut fumeur/non fumeur, il a été décidé de manipuler la variable la moins granulaire (2 modalités) pour ces statistiques afin de faciliter l'interprétation. L'observation est claire : il y a moins de personnes fumeuses que non-fumeuses dans les données globales et ce pour les deux sexes.

La figure 3.2 présente la distribution de la population totale sans distinction de sexe par rapport à l'âge et au montant assuré.

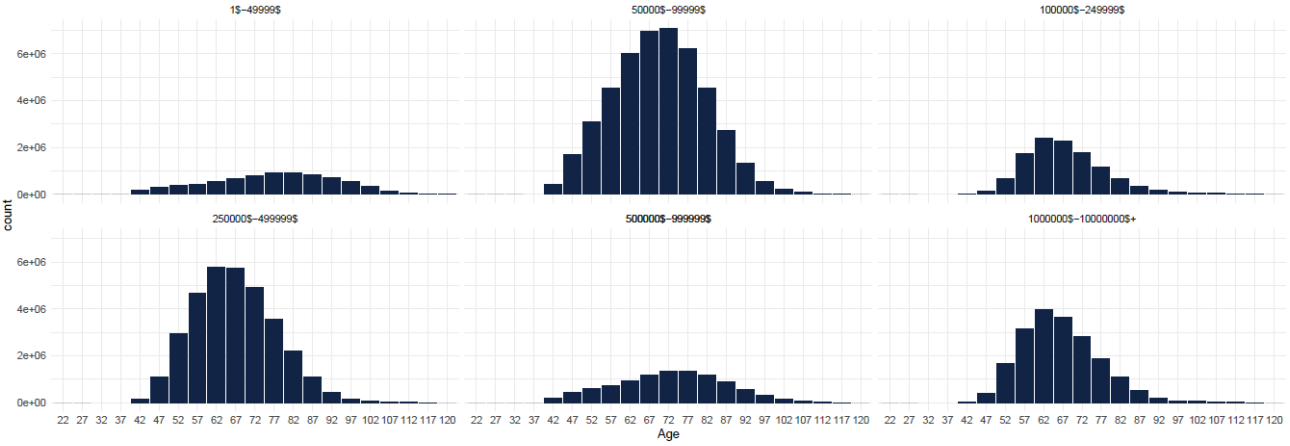


FIGURE 3.2: Distribution de la population par âge et par montant assuré.

Pour tout montant, se dégagent des formes de distribution relativement différentes. De plus, les plus faibles expositions sont observées pour les montants les plus faibles (1-49999) et pour certains des montants les plus élevés (100000-249999 et 500000-999999). La plus grande exposition est constatée pour des montants moyennement faibles compris entre 50000$ et 99999$.

3.3 Arbre obtenu à l'issu du partitionnement avec GLM Tree

Après l'ajustement d'un modèle paramétrique à nos données (ici, le modèle binomial définit dans la sous-section 2.4.4), vient la seconde étape de l'algorithme qui concerne les tests d'instabilité pour les variables de partitionnement. Parmi toutes les variables mentionnées dans le tableau 4.1, seules six d'entre elles (Attained age, Face amount band, Gender, Risk class, Duration et Insurance plan) seront utilisées comme variables de partitionnement. Le test de $\lambda_{sup_{LM}}$ est utilisé pour évaluer les variables numériques et celui du χ^2 pour les variables catégorielles. Au final, la variable qui montrera le plus d'instabilité sera utilisé comme première variable de partitionnement et ainsi de suite dans les nœuds résultants de l'arbre.

Enfin, concernant la dernière étape de l'algorithme, un partitionnement binaire des variables (qu'elles soient catégorielles ou numériques) est mis en œuvre tout en choisissant le découpage qui optimise la fonction objective.

Ainsi, après avoir appliqué la fonction `glmtree()` du package **partykit**, il en résulte l'arbre de

la figure 3.3. Les hyperparamètres de la fonction ont été fixés de sorte à avoir une taille raisonnable et conséquente pour chacun de nos groupes. Dans une seconde partie, un test de sensibilité de ces hyperparamètres sera réalisé afin d'analyser le comportement de l'algorithme.

Dans cet arbre, les déviations (en nombre) qui correspondent au rapport des nombres de décès observés sur ceux prédits seront analysés. Notons qu'en plus des déviations mentionnées dans les nœuds terminaux, est aussi affiché le nombre d'observations contenu dans chaque feuille.

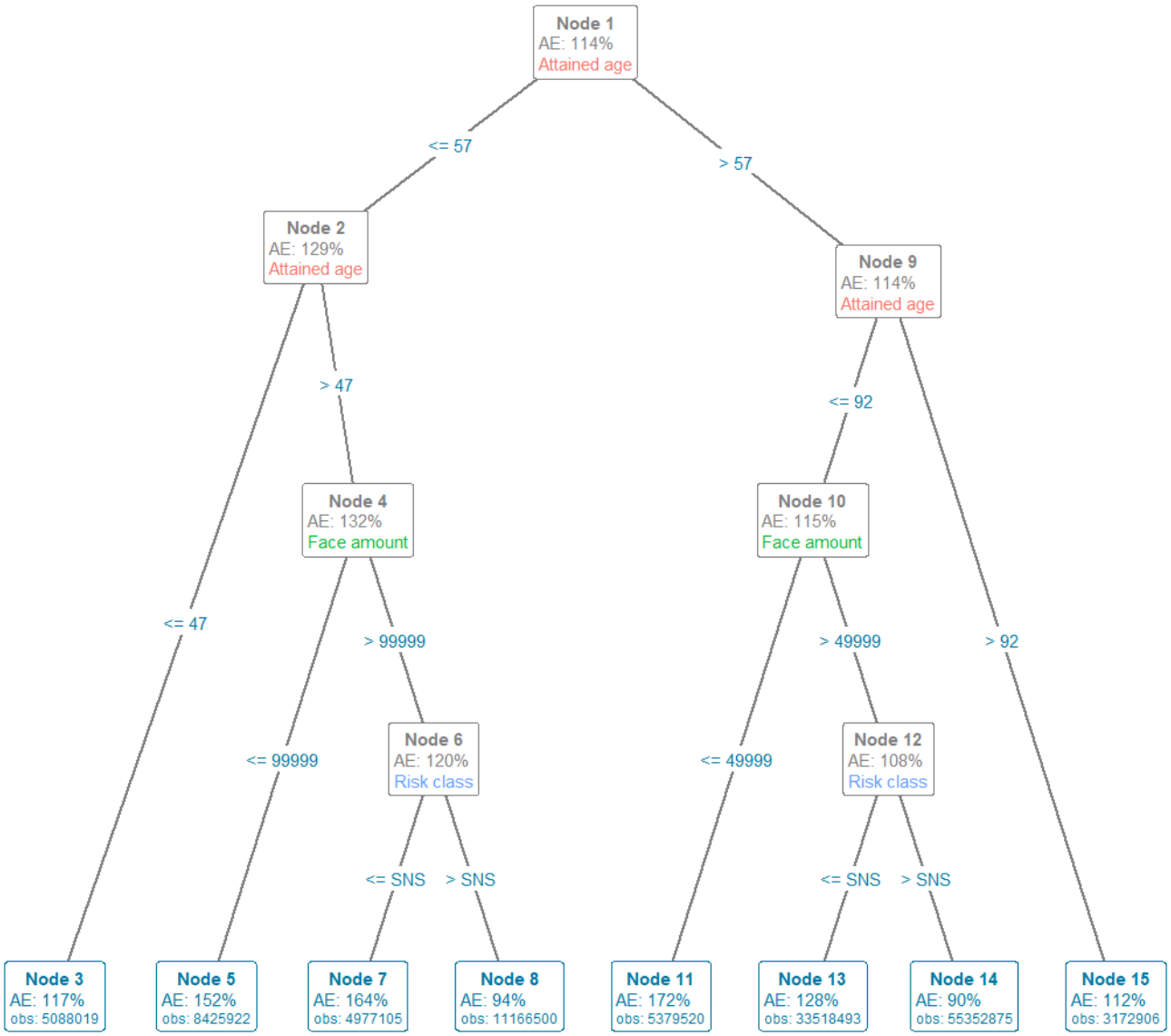


FIGURE 3.3: Arbre issu du partitionnement des données

Notons que les paramètres de cet arbre ont été choisis de sorte à avoir une taille assez conséquente dans chaque groupe. Cependant, un test de sensibilité sur les paramètres sera tout de même effectué afin de déterminer la stabilité de cet arbre.

Au niveau global, la première déviation obtenue s'élève à 114% : la table de référence sous estime et ne capture donc pas de façon adéquate le comportement de mortalité des assurés qui sont représentés dans cette base.

Il existe une symétrie parfaite des variables utilisées pour le partitionnement pour les individus d'âge inférieur et supérieur à 57 ans. En effet, dès le second niveau de l'arbre, les mêmes variables servent au partitionnement pour les nœuds résultants.
L'ordre des variables utilisées pour le découpage montre quel différentiel de mortalité est le plus important. Parmi les six variables de partitionnement utilisées dans la fonction (Attained age, Face amount band, Gender, Risk class, Duration et Insurance plan), seules trois sont utilisées pour le partitionnement. Il s'agit de :

- Attained age (x2),
- Face amount,
- Risk class.

L'âge est la principale variable de partitionnement. Elle est utilisée deux fois de suite en tant que telle. Cela montre que c'est la variable où on observe le plus de fluctuations par rapport à la table standard. Pour la variable âge, il y a clairement une différence de forme entre les deux tables car c'est celle qui présente le plus d'instabilité. Cela revient à dire qu'il y a une différence d'évolution de la mortalité par âge en comparant nos deux tables. La table 2015 VBT était censée capturer les effets liés à l'âge mais on s'aperçoit que ce n'est pas le cas avec nos données. Le premier découpage en fonction de l'âge sépare les individus les plus jeunes des plus âgés. Ensuite, survient le second partitionnement par rapport à l'âge. Ce second découpage permet une plus grande granularité des groupes en fonction de l'âge.

Après la variable âge, vient le partitionnement selon le montant assuré entre les petits et les grands montants. Ce type de partitionnement, entre petits et grands montants, mais à des niveaux différents est retrouvé dans plusieurs nœuds de cet arbre, en particulier dans les nœuds 4 et 10. Pour les individus dont l'âge est compris entre 48 et 57 ans, le découpage par montant assuré se fait à 99999$ tandis que pour ceux dont l'âge est dans la tranche 58-92 ans, le découpage en fonction du montant est à un niveau inférieur de 49000$. Les effets de cette variable ne sont pas censés être capturés par la table standard, il est donc logique de la voir apparaître dans l'arbre comme variable de partitionnement.

Quant à la variable faisant état de la classe de risque, il semble également logique de partitionner les données en fonction de celle-ci. En effet, il s'agit d'une variable plus granulaire que la variable smoker status (statut fumeur/non fumeur). Or, c'est seulement les effets de cette dernière qui sont censés être capturés par la table réglementaire.
Dans les nœuds 6 et 12, contenant les personnes ayant des montants élevés, l'algorithme partitionne entre les fumeurs et la classe la plus saine. Pour cette dernière qui peut être considérée comme la classe des non-fumeurs, le ratio est inférieur 100% dans les nœuds 8 et 14. Ce sont les seules feuilles de l'arbre où on observe une mortalité observée plus faible que celle attendue et prédite par la table réglementaire. Ce sont également les feuilles dont les ratios sont les plus proches de 100%, il semble donc que la table standard capture plus ou moins bien la mortalité sur ce segment de la population. Au sujet de la classe contenant les individus les moins sains (les fumeurs) avec de grands montants assurés, les déviations les plus importantes sont observées. Il s'agit des nœuds 7 et 13.

Pour les autres feuilles (les individus aux âges extêmes), nous observons une déviation relativement élevée qui varie de 112% pour certaines individus à 117% pour d'autres.

Les plus grandes déviations sont observées pour les nœuds 5 et 11 contenants les personnes ayant de faibles montants assurés.

Pour récapituler ces résultats, les plus fortes déviations sont observées pour tout âge en regardant les segments suivants :

- Celui des faibles montants assurés,
- Celui des personnes en moins bonne santé (SS + PS + SNS) ayant de grands montants assurés.

En ce qui concerne les âges extrêmes (< 47 et > 97), on observe des déviations moyennement élevées ainsi que les expositions les plus faibles.

3.3.1 Focus sur les feuilles de l'arbre

Nous allons nous concentrer sur les 8 feuilles de l'arbre qui sont les 8 groupes formés à l'issu du partionnement de nos données. Sur la figure 3.4, sont affichés les groupes formés ainsi que leur caractéristique. Le ratio présent sur la figure demeure celui indiqué précédemment sur l'arbre. Les nœuds seront étudiés deux par deux en prenant les symétries par rapport à la première variable de partitionnement.

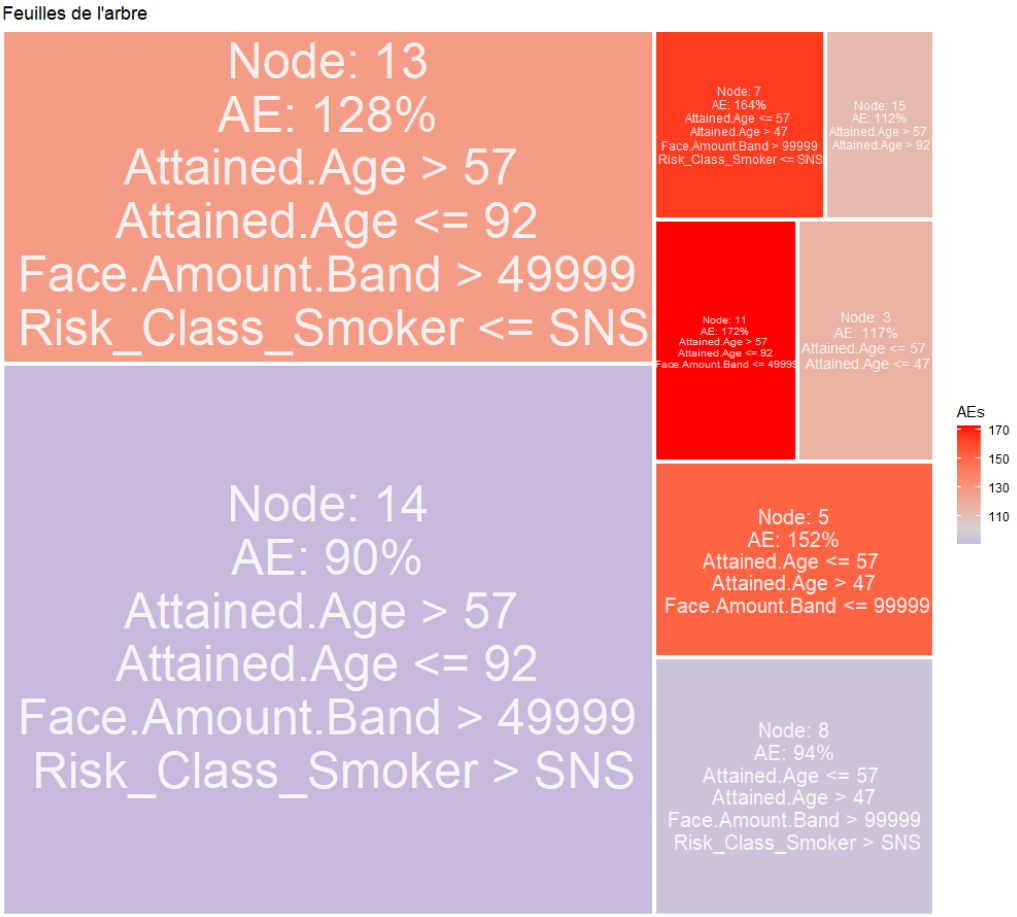


FIGURE 3.4: Les nœuds terminaux de l'arbre

La taille de la cellule est fonction du nombre d'assurés dans le nœud. La couleur de la cellule met en évidence à quel point nos données sont en adéquation ou non par rapport à la table standard 2015 VBT : plus la cellule est rouge, plus la mortalité est sous-estimée.
En outre, plus le ratio est proche de 100%, plus nos données sont en adéquation avec la table standard et dans le cas contraire, cela signifie que nos données ne s'ajustent pas correctement. Les déviations obtenues pourront être analysées plus en détail à l'aide des figures suivantes et des intervalles de confiance.

La plus grande cellule, c'est-à-dire le nœud 14 et le nœud 8 de taille moyenne affichent des déviations proches de 100%. Ces segments contiennent les individus dont l'âge est compris entre 47 et 92 ans, qui ont un montant assuré supérieur 49999$, et font partie des classes de risque les plus saines (c'est à dire non-fumeurs). Dans notre cas, le ratio étant inférieur à 100%, la mortalité observée est inférieure à celle prédite par la table. Il y a moins de décès observés que prévus : la table surestime alors la mortalité pour ce segment de la population. Dans le contexte de l'assurance vie et dans le souci de faire preuve de prudence, ce genre de situation n'est pas problématique pour les compagnies.

Les nœuds 7 et 13 sont moyennement importants en termes de volume et les ratios sont tout de même élevés : 128% et 164% pour respectivement les nœuds 13 et 7. Pour ces segments, les écarts par rapport à la table standard concernent les personnes âgées entre 47 et 92 ans, ayant un montant assuré supérieur à 49999$ et en moins bonne santé (SS + PS + SNS). La mortalité n'est clairement pas bien capturée pour ces deux groupes. L'écart des déviations pour ces deux segments s'élève à 36%, ce qui n'est pas négligeable. En effet, le ratio augmente lorsqu'on s'intéresse aux individus de 48 à 57 ans dont les montants assurés sont plus grands et atteignent plus de 100000$.

La situation devient assez problématique lorsque le ratio tend vers de larges valeurs au dessus de 100%. Pour les personnes aux petits montants c'est à dire les nœuds 5 et 11, les plus grands écarts sont observés : la mortalité pour ces segments ne coïncide pas à celle prédite par la table 2015 VBT. L'écart le plus important dans cette étude est observé dans le nœud 11 contenant les personnes âgées de 58 à 92 ans avec les montants assurés les plus faibles ($<$ 49999$) : la table standard ne capture pas du tout la mortalité pour ce segment de la population. La mortalité pour les personnes aux faibles montants assurés est différente des autres et cela peu importe le niveau de montant où est effectuée le découpage (49999$ ou 99999$). La mortalité observée est plus élevée que celle prédite : ceci est sans doute lié à la classe économique et sociale de ces personnes. De plus, il peut être aussi question d'une faible sélection médicale avant d'entrer dans le portefeuille d'assurance pour les individus aux montants les plus faibles.

Pour les cellules restantes, c'est à dire les nœuds 3 et 15 contenant les individus aux âges extrêmes ($<$ 47 ans et $>$ 92 ans),on constate moins d'observations car moins de personnes à ces âges dans notre portefeuille et des sinistres relativement plus élevés que prévus avec des déviations supérieures à 100%.

3.3.2 Stratégie de post-élagage

Nous aurions souhaité adopter une stratégie de post-élagage sur l'arbre 3.3. Rappelons que cette technique vise à réduire la taille de l'arbre. Il s'agit d'abord de construire un arbre très grand, en utilisant une valeur assez élevée du paramètre α, variable qui indique le niveau de confiance. Il sera ensuite question d'élaguer l'arbre construit à l'aide d'un critère d'information (AIC/BIC). Toutefois, dans notre étude, l'utilisation d'un critère d'information pour l'élagage ne sera pas possible. En effet, il est question de déviations moyennes. La régression se fait donc selon l'intercept, sans prise en compte des variables explicatives. Il n'y a donc que les variables de partionnement qui entrent en jeu.

L'alternative proposée par Zeileis et al. (2008) qui consiste à utiliser une plus petite valeur du paramètre α ne fonctionne pas non plus. En effet, l'arbre construit est stable selon les valeurs de ce paramètre. Notons que certains paramètres de la fonction `glmtree()` feront l'objet d'une étude plus approfondie par la suite.

3.4 Analyse des déviations

Dans cette partie, 2 types de déviations calculées seront étudiées. D'une part, les déviations en fonction du nombre de décès, qui correspondent à celles mentionnées sur les graphes précédents (figures 3.3 et 3.4). D'autre part, les déviations en fonction des montants assurés, qui seront plus explicitées dans la suite.

3.4.1 Déviations en nombre

Le graphique 3.5 apporte des informations supplémentaires à propos des feuilles de l'arbre. Outre les déviations observées et leurs intervalles de confiance, on y retrouve pour chacun des nœuds terminaux, la distribution des montants assurés et du nombre de décès. La ligne grise en pointillé illustre le ratio

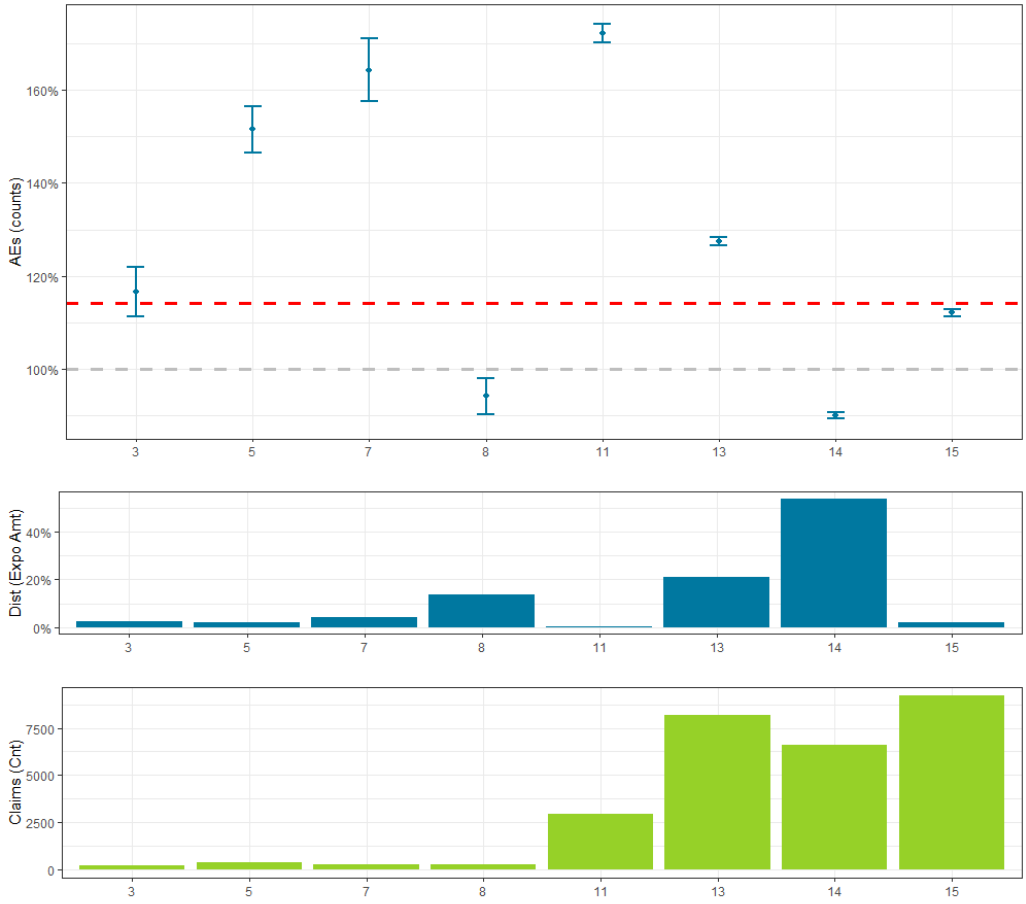


FIGURE 3.5: Ratio des décès observés et prédits par rapport à la table 2015 VBT

A/E de 100% qui indique l'adéquation parfaite entre les tables de l'industrie et la 2015 VBT. La ligne rouge en pointillé est le ratio A/E à 114% qui illustre la déviation globale de la table par rapport à la table réglementaire. Dans ce cas, les intervalles de confiance proposés par LIDDELL (1984) sont construits non à partir d'approximation mais à l'aide d'une formule exacte explicitée à la sous-section 2.4.3. Ces derniers permettront d'évaluer le caractère significatif ou non des déviations obtenues.

Le premier panel présentant les déviations déjà observées ainsi que les IC associés, révèle que le

ratio à 100% n'est inclus dans aucun des IC de nos déviations. Il est donc possible d'affirmer que la table standard ne capture pas de manière adéquate le schéma de mortalité pour ces différents segments de la population.

Le second panel quant à lui indique la distribution des montants assurés pour chaque feuille de l'arbre. Plusieurs nœuds (3, 5, 7, 11 et 15) affichent les proportions les plus faibles des montants assurés.

En ce qui concerne le dernier panel, les premiers nœuds terminaux (3, 5, 7 et 8) expriment les nombres de sinistres les plus faibles de notre base. Il s'agit en réalité des individus les plus jeunes ($<= 57$ ans) comme l'indique l'arbre 3.3 qui décèdent donc moins que le reste de la population (> 57 ans).

Nous pouvons voir que les nœuds 5 et 7 (contenant respectivement les personnes avec un faible montant et les personnes les moins saines déclarant un montant élevé) présentent des écarts largement plus élevés que la moyenne ; Cependant, en regardant le second panel de la figure 3.5, celui de la distribution des montants, on s'aperçoit que les montants exposés pour ces deux segments sont très faibles. De plus, le dernier panel qui met en évidence la distribution des sinistres révèle un nombre de sinistres extrêmement faible. Dans le cas où ces données proviendraient d'un client souhaitant étudier son portefeuille, nous ne nous focaliserons pas sur ces segments de la population. En effet, avec une proportion des montants et des sinistres faible, ces segments ne sont pas alarmants et ne représentent pas réellement un risque pour le résultat d'une compagnie.

Le nœud 11 quant à lui, présente la déviation la plus élevée. Cependant, bien qu'affichant une sinistralité relativement forte (> 2500), la distribution des montants est extrêmement faible. Dès lors, le nœud 11 rejoint le lot des nœuds 5 et 7 précédents qui sont jugés moins inquiétants pour une compagnie.

En observant le nœud 13 (avec les individus fumeurs aux montants élevés), il se trouve que la déviation affichée (128%) est supérieure au ratio global de 114% et encore plus au ratio à 100%. En outre, les montants pour ce nœud représentent plus de 20% des montants assurés et la sinistralité correspondante est également très forte (> 7500). Ce type de segment est préoccupant car assez coûteux pour une compagnie. Il doit donc être surveillé au plus près.

En ce qui concerne les nœuds 8 et 14, où nous observons les déviations les plus faibles par rapport à la table standard, la mortalité semble être plus faible que prévue. Dans un contexte où l'on souhaite rester prudent, ce type de segments n'est pas problématique et ne représente pas un danger pour le résultat d'une compagnie.

Enfin les nœuds 3 et 15 représentent les individus aux âges extrêmes. Pour ces segments, les déviations sont moyennement élevées et la distribution des montants est quant à elle faible. La proportion des sinistres est naturellement moins forte dans le nœud 3 (qui contient des individus jeunes) que dans le 15 (contenant les individus plus âgés avec plus de chances de décès). Concernant le nœud 15, bien que le nombre de sinistres soit très élevé, une faible exposition des montants assurés conduit à placer ces nœuds dans le lot des nœuds 5, 7 et 11 qui présentent moins de risques pour la compagnie.

Au final, la plupart des segments étudiés ne sont pas les plus inquiétants pour un portefeuille d'assureur. Cependant, certains profils de la population, c'est à dire des individus en moins bonne santé, assez âgés, avec des montants élevés (nœud 13), doivent être contrôlés.

3.4.2 Déviations en montant

Pour affiner les conclusions et les mesures à prendre à la suite de cette étude, nous pouvons avoir une perspective actuarielle en analysant le ratio A/E en montant assuré sur la figure 3.6. Pour ce faire, nous utiliserons les mêmes groupes créés précédemment à l'issu du partitionnement. Pour chacune des feuilles de l'arbre, les ratios en montant seront calculés à partir de 2 nouvelles variables. On a :

$$A/E = \frac{Claim\ Amount}{EA}.$$

où Claim Amount représente le nombre de décès observés par le montant assuré et EA quant à lui exprime le nombre de décès prédit par la table standard toujours par le montant assuré. L'idée de cette approche par montant est de faciliter l'interprétation pour les compagnies d'assurance en mettant en avant les segments où une déviation est observée, car il est plus logique pour ces dernières de se référer à des montants et à leurs déviations pour prendre des décisions adéquates.

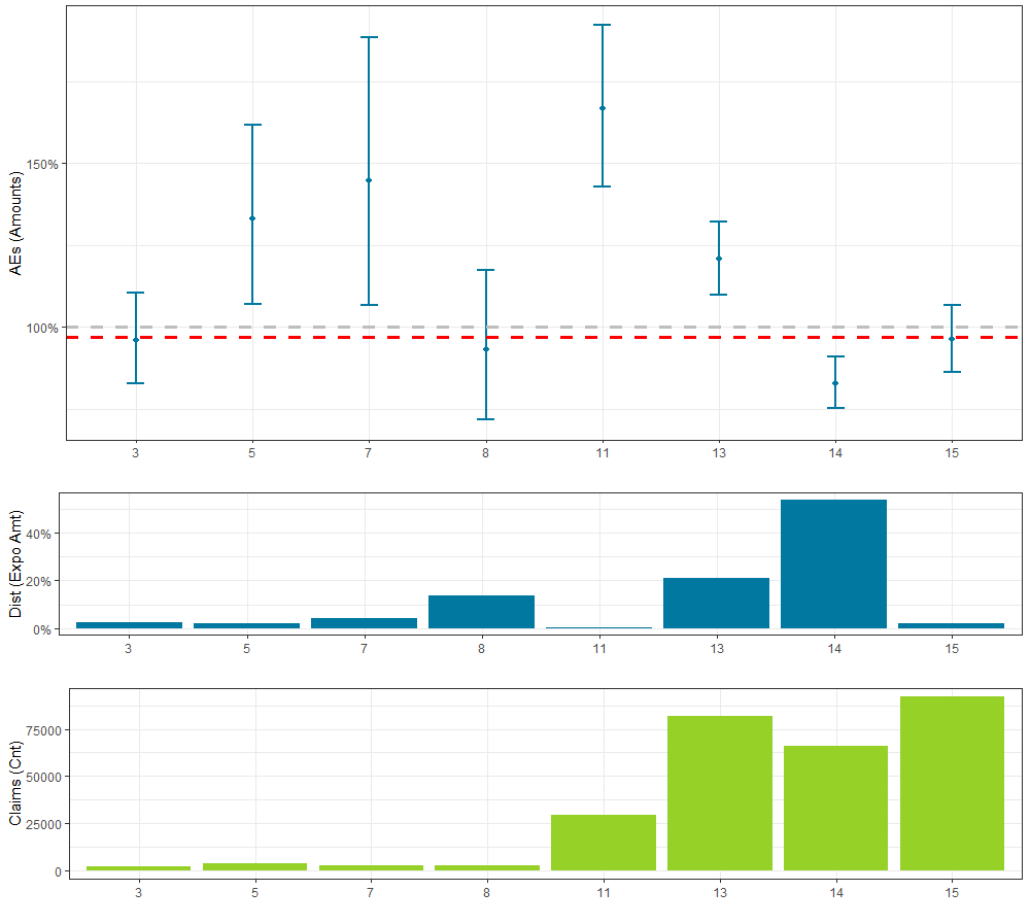


FIGURE 3.6: Ratio des montants assurés observés et prédits par rapport à la table 2015 VBT

Pour plusieurs nœuds où une déviation en nombre est observée, la table en montant capture bien le modèle de mortalité. La table standard étant construite sur la base des montants, il apparaît alors évident que les lignes en pointillé grise et rouge soient très proches.

L'intervalle de confiance construit ici à l'aide des quantiles de la loi de Tweedie et explicité à la sous-section 2.4.3 est plus grand que précédemment car il est plus difficile d'étudier la mortalité en utilisant un raisonnement par montant : il y a plus d'incertitudes.

On remarque en premier lieu, les nœuds 3, 8 et 15 pour lesquels le ratio à 100% est à l'intérieur de l'IC des déviations : il n'existe donc pas une déviation significative entre les montants observés et prédits pour ces segments. Nous nous concentrons donc uniquement sur les nœuds où il n'y a pas d'intersection entre l'intervalle de confiance et la ligne du ratio à 100%. Il s'agit des nœuds 5, 7, 11, 13 et 14.

Commençons avec le nœud 14 dont la déviation observée est en dessous de celle globale et du ratio à 100%. Ce segment représente plus de 50% des montants assurés totaux. Comme évoqué auparavant, dans une optique de prudence, ce segment n'est pas problématique.

Au sujet des nœuds 5, 7 qui ont des écarts importants mais une proportion de montant exposé et de sinistres très faible : ils ne semblent toujours pas alarmants et donc ne posent pas non plus de problème pour notre portefeuille. Ces résultats sont valables dans une moindre mesure pour le nœud 11. En effet, ce dernier, bien qu'affichant un nombre de sinistres relativement élevé, la distribution des montants indique que ce n'est pas le plus coûteux des segments.

Le nœud 13 quant à lui a son intervalle de confiance juste au-dessus du ratio à 100%, mais avec un grand montant assuré et un nombre important de sinistres. Comme vu antérieurement, il y a plus de 20% du montant assuré pour ce nœud et les sinistres dépassent le compte des 75000. Cela confirme les résultats déjà obtenus par l'analyse des déviations en nombre de décès pour ce nœud à la sous-section 3.4.1. Il est important de mettre en place des mesures de gestion (actions de management) pour faire face à ce type de segments.

Cette première application concerne les données de l'industrie américaine et n'est donc pas représentative d'une compagnie en particulier. Les données, typiques de données contenues dans un portefeuille de compagnie d'assurance donnent une idée des réalisations et résultats possibles en utilisant cet algorithme. Néanmoins, les résultats obtenus auraient pu être totalement différents au vue de la pluralité de compagnies participantes et donc de la diversité des données contenues dans cette base.

3.4.3 Focus sur le nœud 13

Après avoir étudié notre ensemble de données, un certain segment de la population se démarque des autres. Il s'agit du nœud 13 contenant des individus assez agés (> 57 ans et ≤ 92 ans), faisant partie de la classe de risque la moins saine et ayant des montants assurés élevés. Nous allons donc étudier en détail ce profil de la population.

Commençons par examiner les taux de mortalité tels qu'observés et prédits en fonction de l'âge comme on peut le voir sur la figure 3.7.

Avec une déviation à 128% et donc une surmortalité dans nos données par rapport à la table standard, il apparaît évident d'avoir une courbe des taux observés généralement au dessus de celle des taux prédits.
Pour les âges les plus jeunes, 62 et 67 ans, les taux observé et prédit sont assez fluctuants. Cependant, dès l'âge de 72 ans, on observe clairement que le taux observé est au dessus du taux prédit. Notons

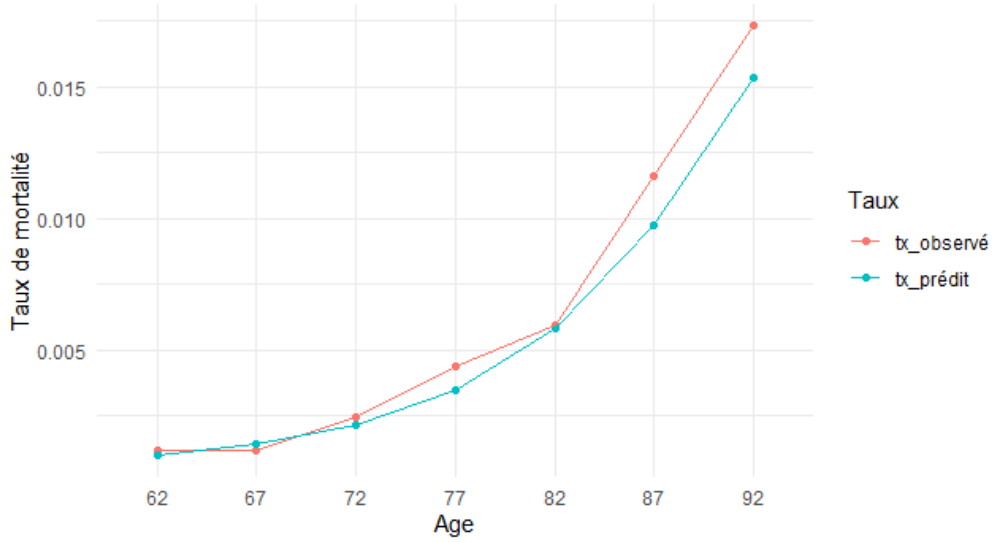


FIGURE 3.7: Comparaison des taux de mortalité observé et prédit dans le nœud 13

également que l'écart entre les taux se creuse en allant vers les âges les plus élevés.

Nous allons ensuite tenter de comparer les flux observés et ceux prédits. Des hypothèses seront émises afin de calculer les flux que devrait débourser une compagnie d'assurance selon les taux observés et prédits. Nous supposons donc que tous les individus présents dans ce segment de la population, souscrivent un contrat permanent d'assurance décès vie entière. L'assurance en cas de décès de type vie entière prévoit le versement d'un capital décès ou d'une rente au(x) bénéficiaire(s) désigné(s) par l'assuré. Dans notre cas, nous supposons un versement en capital au décès de l'assuré. Nous allons donc calculer les flux potentiels que la compagnie devra payer aux assurés de notre segment. Il paraît évident que des écarts de flux surviendront en réalisant ce calcul avec les taux observés puis prédits. Ayant un taux observé supérieur à celui prédit, les individus de notre portfeuille auront tendance à décéder plus rapidement et donc les versements aux bénéficiaires de l'assuré interviendront également plus vite.

Pour réaliser le calcul des flux (observés et prédits) à payer, il nous faut les taux de décès des individus associés à chaque âge, les montants pour chaque individu ainsi qu'un taux d'actualisation. Pour ce qui est de l'actualisation en fonction du nombre d'années, nous utilisons la courbe des taux sans risque de 2019 publiée par le trésor public américain.

Après avoir obtenu nos trois quantités qui serviront à calculer les flux sur 30 ans pour chaque âge disponible, nous réalisons le calcul suivant.
Pour un individu d'âge x et de montant $montant_x$, les flux sont les suivants :

$$A\ t = i,\ Flux_i\ =\ montant_x \times q_{x+i} \times Discount\ Year_i.$$

Chacun de ces flux sera calculé d'une part en utilisant les taux de décès observés et d'autre part les taux de décès prédits. Nous obtenons ainsi deux matrices contenant les flux par année et par âge pour nos deux taux de décès. L'idée maintenant sera de sommer les flux par année afin d'obtenir une distribution de flux en fonction de la maturité comme on peut le constater sur la figure 3.8.

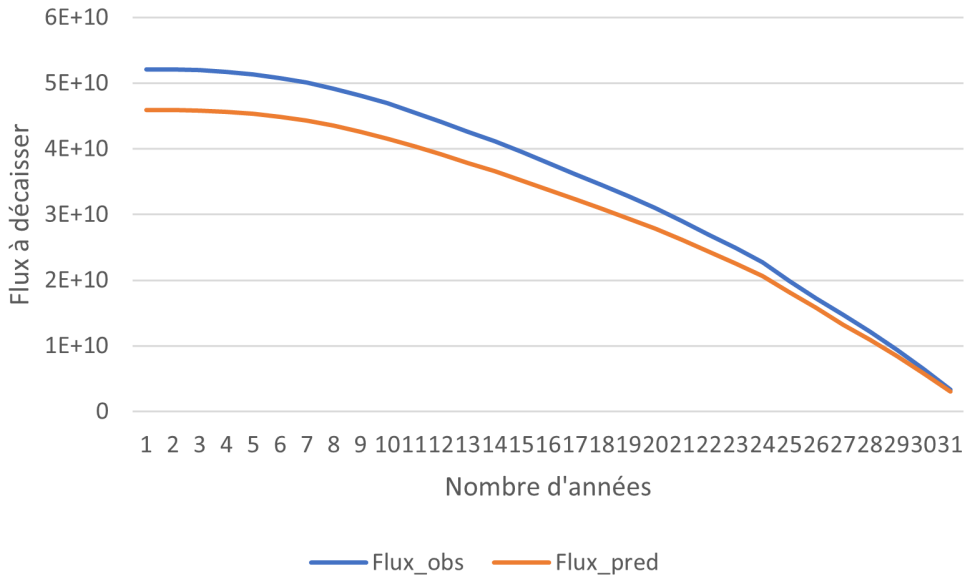


FIGURE 3.8: Comparaison des flux observés et prédits en fonction du nombre d'années projetées

Il apparaît dès lors une décroissance des flux en fonction de la maturité. On s'aperçoit qu'avec une déviation de mortalité à 128% dans ce nœud, les flux observés, c'est à dire, ceux à réellement décaisser pour couvrir les bénéficiaires des assurés sont nettement plus élevés que ceux qu'on aurait prédit avec la table standard. La somme des flux prédits et observés à décaisser pour les 30 années de projection sont respectivement de 957,99 Md$ et 1076,53 Md$. L'écart entre les 2 types de flux s'élève donc à 11% et ne peut être considéré comme négligeable vu les montants en jeu.

Un écart de mortalité de presque 30% peut donc avoir de graves conséquences sur le résultat d'une compagnie d'assurance. Il est donc important de calibrer au mieux nos taux de mortalité, ce qui pourrait préserver une compagnie de pertes considérables.

3.5 Test de sensibilité des hyperparamètres

Dans cette partie, trois paramètres de la fonction `glmtree()` seront soumis à des tests. Il s'agit de :

- *alpha* : Le niveau de confiance.
- *minsize* : Le minimum d'observations dans un nœud.
- *maxdepth* : La profondeur maximale de l'arbre.

Pour analyser la sensibilité de l'arbre construit précédemment aux différents paramètres définis plus haut, nous allons à tour de rôle faire varier l'un d'entre eux en gardant les deux autres fixes.

3.5.1 Variation de *alpha*

Pour ce qui est de ce paramètre, plusieurs tests ont été faits pour les valeurs suivantes : 0.001, 0.01, 0.05, 0.1 et 1. L'utilisation d'une valeur assez faible pour ce paramètre était suggérée afin de réaliser une stratégie de post-élagage. Cependant, pour chacune de ces valeurs, la structure de l'arbre est inchangée. En effet, le même ordre des variables de partionnement ainsi que les mêmes seuils de découpage étaient obtenus. Pour un très grand volume de données, le niveau de confiance n'est pas un paramètre utile. La raison étant que pour des échantillons de cette taille, même les plus petites instabilités de paramètres peuvent être détectées.

3.5.2 Variation de *minsize*

Étant le minimum d'observations dans un nœud, ce paramètre permet de construire des groupes plus ou moins grand par rapport à ceux obtenus avec l'arbre précédent. Notons que les observations ici sont très grandes, étant aggrégées sur les dix années d'étude. Dans l'arbre précédent, cette variable était fixée à 3000000. Les deux autres paramètres, *alpha* et *maxdepth* sont fixés à respectivement 0.01 et 5.

Le paramètre *minsize* = 300000 conduit à la figure 3.9.

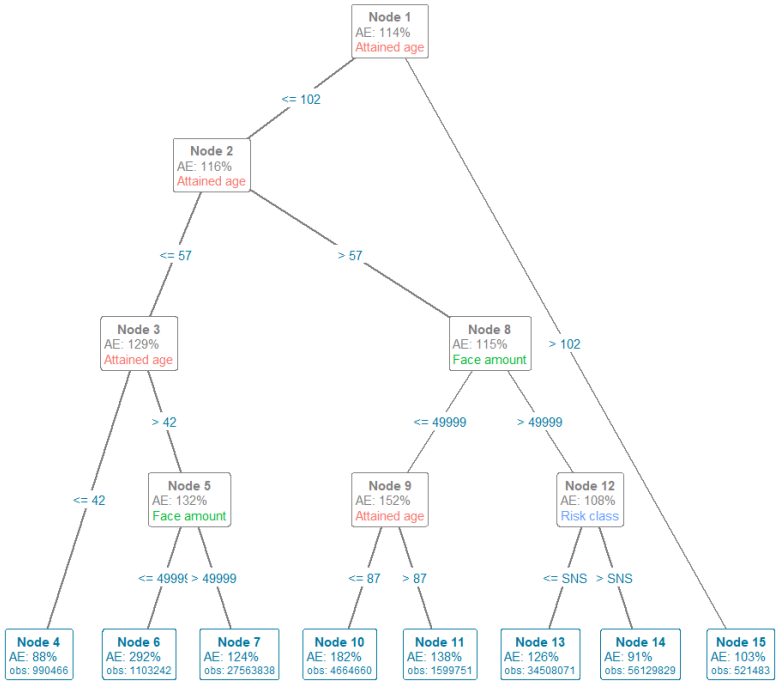


FIGURE 3.9: Abre obtenu à l'issu du partitionnement pour *alpha* = 0.01, *maxdepth* = 5 et *minsize* = 300000

De prime abord, les variables utilisées pour le partitionnement dans cet arbre sont les mêmes présentes dans l'arbre de la figure 3.3. La structure reste sensiblement la même. On retrouve des découpages et déviations similaires et donc les mêmes conclusions pour certains groupes d'individus. Pour les individus d'âge inférieur ou égale à 57 ans, il n'y a plus de découpage selon la classe de risque.

L'âge revient plusieurs fois dans cet arbre, ce qui confirme encore une fois qu'il y a bel et bien une

différence d'évolution de la mortalité par âge entres nos deux tables. Le découpage selon le montant assuré se fait toujours au même seuil dans cet arbre contrairement à ce qu'on obtient dans l'arbre précédent, car la taille des groupes est plus petite.

On retrouve le résultat obtenu précédemment par rapport à la mortalité des individus aux plus faibles montants assurés. En effet, les nœuds 6 et 9 en comparaison respective des nœuds 7 et 12 montrent les déviations les plus fortes pour les individus aux plus petits montants assurés. La mortalité de ces derniers est largement plus forte que celle prédite. Ce résultat s'accentue pour les individus âgés de 43 à 57 ans pour lesquels les déviations affichées sont les plus élevées dans ce second arbre.

En outre, le résultat pour les individus appartenant à la classe de risque la moins saine reste le même. Comme le montre le nœud 13, la déviation de ces derniers est plus élevée que celle des individus les plus sains (nœud 14).

Le fait de diminuer la taille des nœuds et de garder la même profondeur ne permet pas de retrouver la structure exacte de l'arbre initial. En effet, dans le soucis de respecter la contrainte de taille, la structure entière de l'arbre est modifiée. Des découpages sont ajoutés ou retirés mais les conclusions restent les mêmes.

Le paramètre *minsize* = 1000000 conduit à la figure 3.10.

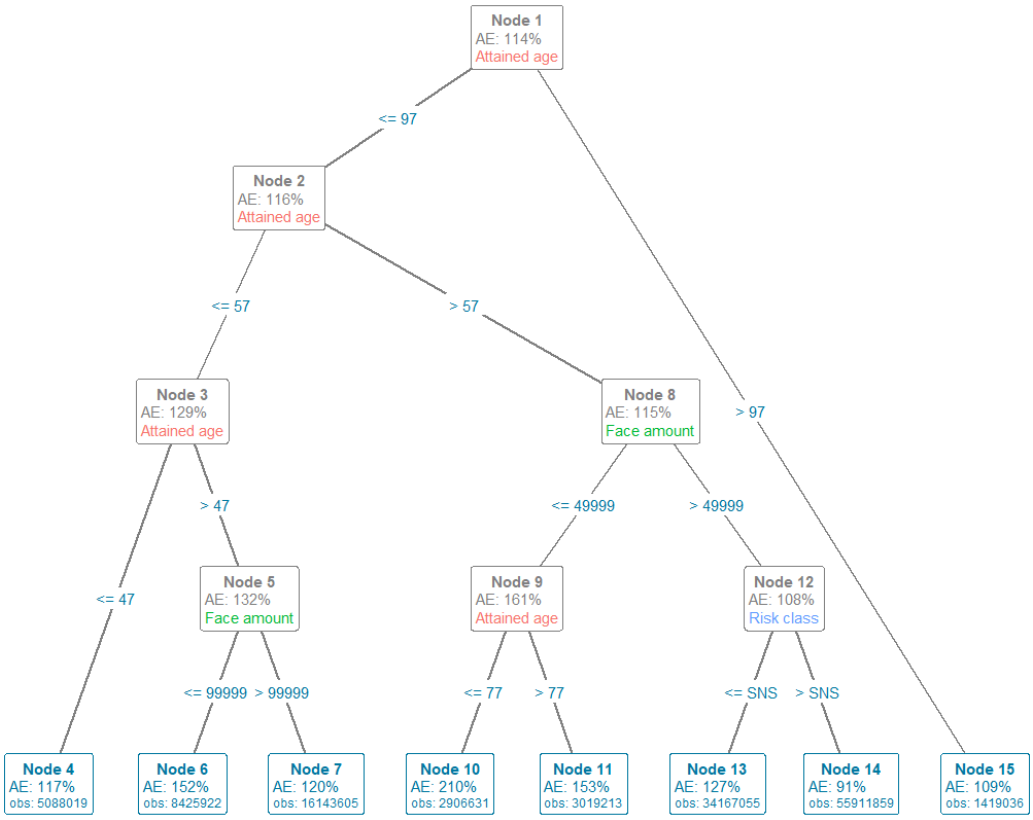


FIGURE 3.10: Abre obtenu à l'issu du partitionnement pour *alpha* = 0.01, *maxdepth* = 5 et *minsize* = 1000000

La structure de ce nouvel arbre est très similaire à celle de l'arbre de la figure ci-dessus, 3.9. Le même nombre de groupes formés est obtenu à l'issu du partitionnement. L'ordre des variables utilisées pour le partitionnement reste le même. Les seuils de découpage sont les seuls qui varient (à la hausse

ou à la baisse), toujours dans le but de respecter la nouvelle contrainte des tailles fixée. L'âge et le montant assuré sont les seules variables dont le seuil de découpage change. Ces changements non structurels ne changent pas les résultats déjà obtenus.

Le paramètre *minsize* = 5000000 conduit à la figure 3.11.

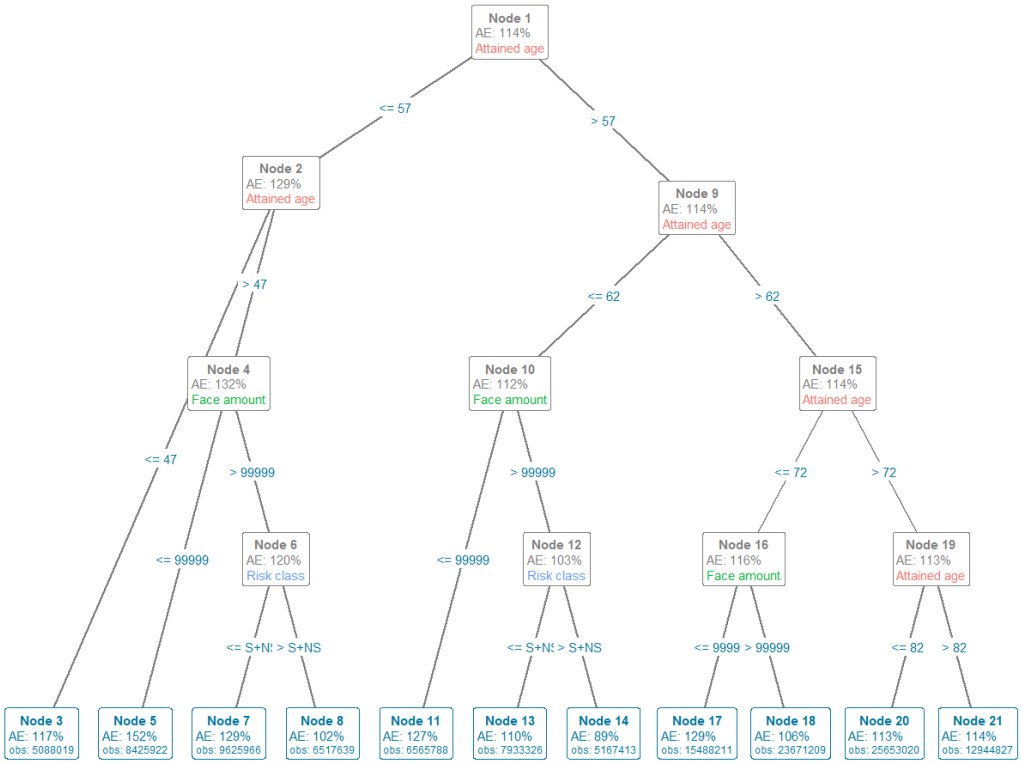


FIGURE 3.11: Abre obtenu à l'issu du partitionnement pour *alpha* = 0.01, *maxdepth* = 5 et *minsize* = 5000000

Bien que cet arbre utilise les mêmes variables de partitionnement que les précédents, il induit beaucoup plus de découpages selon celles-ci. On note à ce titre l'utilisation des variables âge et montant assuré à respectivement 5 et 3 reprises. Nous avons à présent 11 groupes formés à l'issu du partitionnement contre 8 pour les deux arbres précédents. Contrairement à ce qui aurait été attendu, l'augmentation du nombre minimal d'observations dans le nœud augmente le nombre de feuilles de l'arbre. Cela induit de nouveaux seuils de découpage pour les variables du montant assuré et de la classe de risque et davantage d'informations sur les individus à un certain âge. Toutefois, les principales conclusions ne sont pas inchangées.

Le paramètre *minsize* = 10000000, conduit à la figure 3.12.

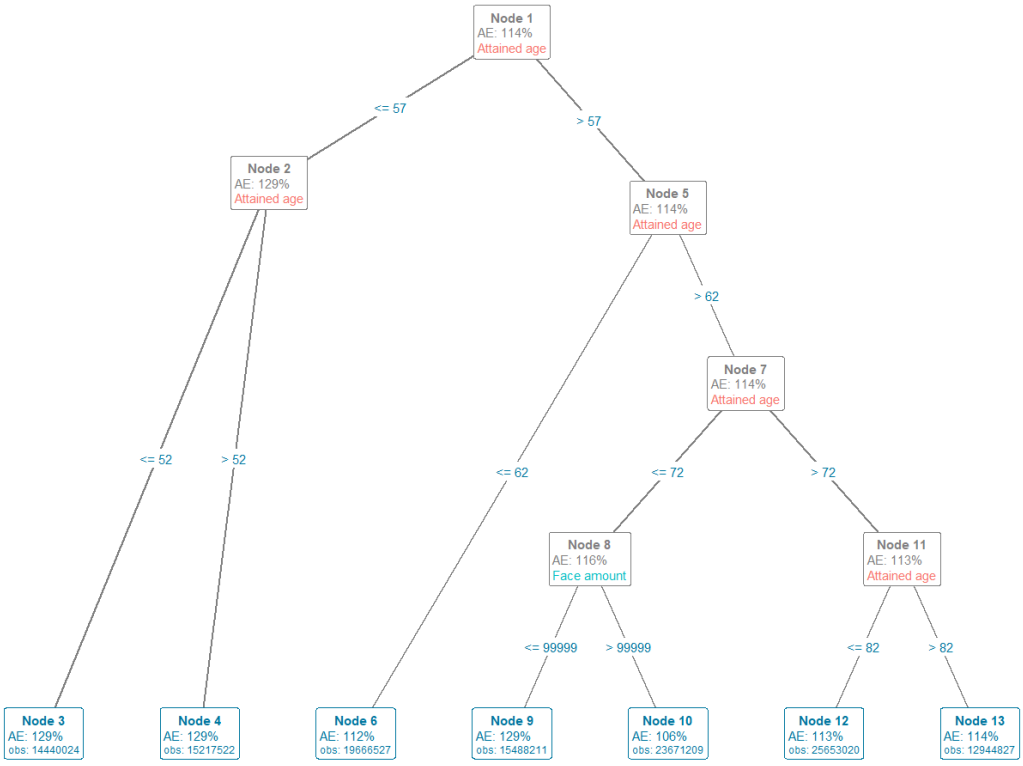


FIGURE 3.12: Abre obtenu à l'issu du partitionnement pour *alpha* = 0.01, *maxdepth* = 5 et *minsize* = 10000000

Pour la plus grande valeur du paramètre *minsize* testée, le nombre de groupes baisse et tombe à 7. On remarque qu'il y a principalement des découpages selon l'âge (qu'un seul pour le montant). On comprend aisément que la containte choisie du nombre minimal dans le nœud est devenue trop forte pour nos données et pour la construction de l'arbre. Pour répondre à cette exigence de taille, l'algortihme se doit d'éviter certains découpages qui ne respecteront pas la taille demandée.

3.5.3 Variation de *maxdepth*

Étant la profondeur maximale de l'arbre, ce paramètre permet d'utiliser plus ou moins de variables de partitionnement dans le découpage de l'arbre. Évidemment, une variable peut être utilisée plusieurs fois. Dans l'abre précédent, cette variable était fixée à 5. Les deux autres paramètres, *alpha* et *minsize* sont fixés à respectivement 0.01 et 3000000.

Le paramètre *maxdepth* = 4 conduit à la figure 3.13.

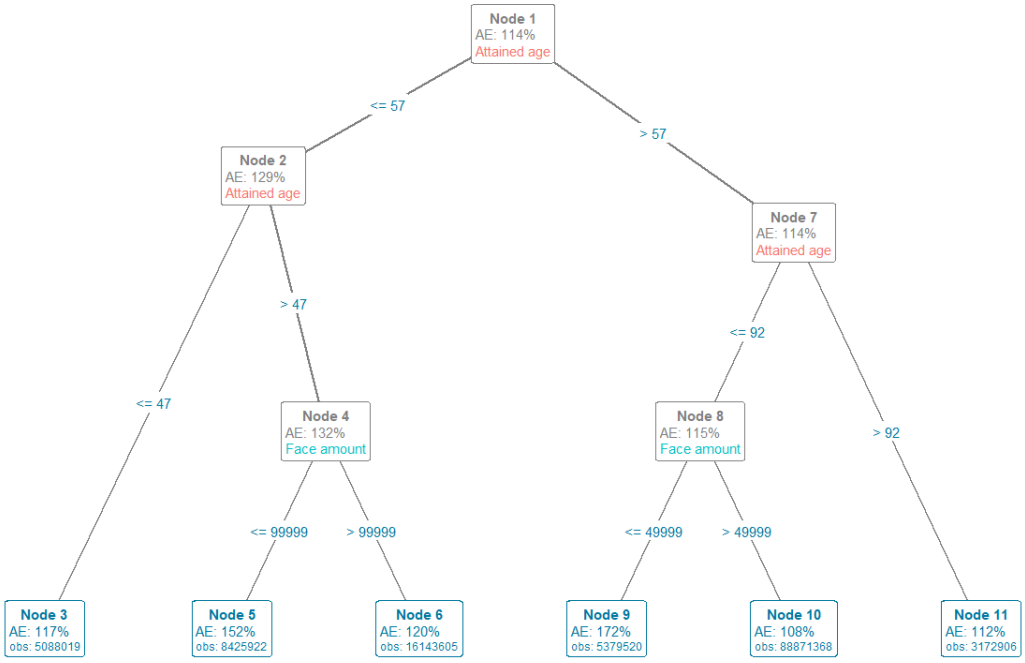


FIGURE 3.13: Abre obtenu à l'issu du partitionnement pour *alpha* = 0.01, *maxdepth* = 4 et *minsize* = 3000000

Il est cohérent de retrouver exactement la même structure que l'arbre initial de la figure 3.3. En effet, diminuer la profondeur n'a pour conséquence que de supprimer le dernier niveau de l'arbre qui induisait le découpage selon la classe de risque.

Le paramètre *maxdepth* = 6 conduit à la figure 3.10.

Toujours en gardant la même structure que l'arbre de la figure 3.3, un niveau se rajoute. Pour ce dernier niveau, les variables utilisées pour le partitionnement ne sont plus les mêmes de part et d'autre des âges inférieur et supérieur à 57 ans comme c'était le cas.

Pour les individus les plus jeunes (<57 ans), une distinction en fonction du sexe est introduite. Notons que cette variable n'était apparue dans aucun des arbres construits jusque là. Il se trouve que parmi les individus en meilleure santé et ayant un montant assuré élevé, la déviation est plus importante pour les hommes. Ces derniers ont une mortalité supérieure à celle prédite par la table règlementaire 2015 VBT. En ce qui concerne les femmes, leur mortalité se trouve inférieure aux prédictions.

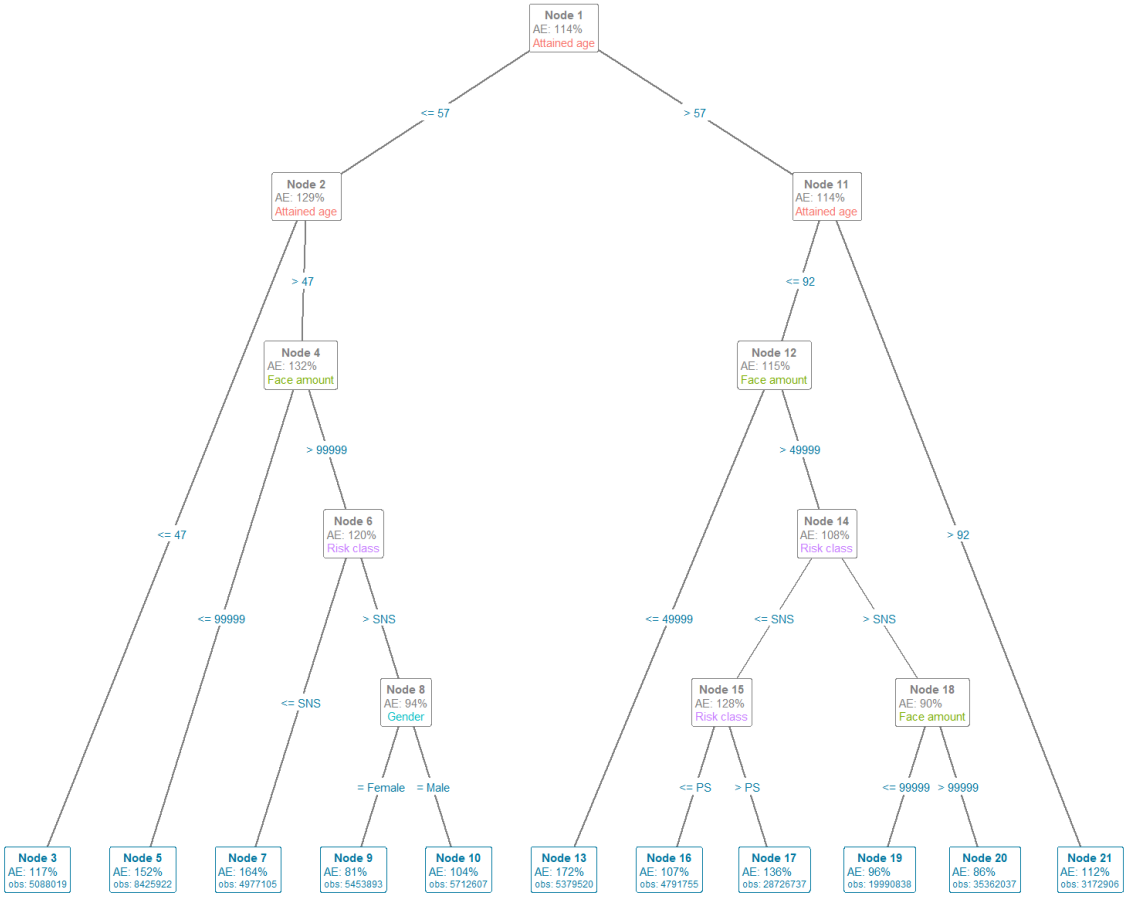


FIGURE 3.14: Abre obtenu à l'issu du partitionnement pour $alpha = 0.01$, $maxdepth = 6$ et $minsize = 3000000$

Pour les individus les plus âgés (> 57 ans), on remarque un découpage supplémentaire selon la classe de risque et le montant assuré.
Ce nouveau découpage selon la classe de risque indique une déviation étonnamment plus faible pour les fumeurs (SS + PS) par rapport aux non-fumeurs (SNS). Ceci explique la déviation observée dans le nœud 13 de l'arbre de la figure 3.3. La table réglementaire a du mal à capturer le comportement des non-fumeurs standars (SNS). Une profondeur supplémentaire permet donc d'apporter une informaiton plus fine sur ce groupe.
Pour ce qui est du nouveau découpage selon le montant assuré, les conclusions sont inchangées.

3.6 Arbre obtenu à l'issu du partitionnement avec CART

Dans cette partie, l'algorithme CART sera utilisé dans un but de régression. Ainsi, la variable réponse correspond à la déviation de chaque profil d'individus. Les poids sont spécifiés par l'exposition. L'arbre maximal obtenu n'est pas exploitable et interprétable.

Nous passons donc à l'étape de l'élagage. La validation croisée avec le paramètre $k = 10$ sera utilisé pour trouver le paramètre de complexité qui maximise la précision du modèle. L'idée est de minimiser la moyenne des taux d'erreur sur les 10 échantillons de validation. Celui-ci est minimal avec

un paramètre de complexité égale à $6.4.10^{-3}$. Le sous-arbre correspondant est présenté à la figure 3.15.

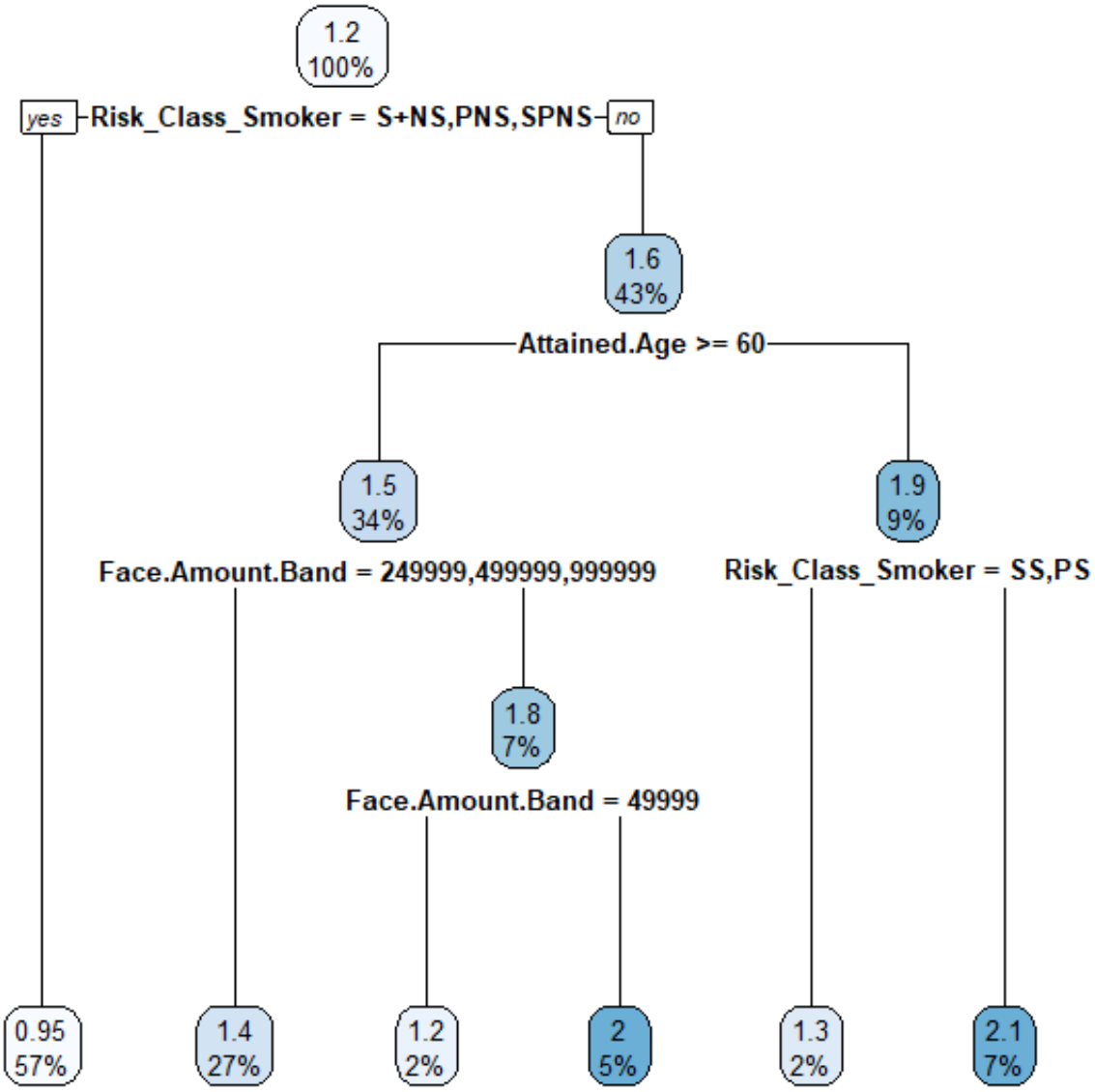


FIGURE 3.15: Abre obtenu à l'issu de l'élagage pour CART

La déviation globale de l'arbre (1.2) est supérieure à 1. Cela indique que d'une manière générale, la table de référence sous-estime le comportement de mortalité des assurés qui sont représentés dans cette base.

Pour chacun des nœuds dans cet arbre sont affichés la déviation en nombre de décès ainsi qu'un pourcentage qui représente la proportion de la population contenue dans ce nœud. Notons que la déviation exprimée ici n'est pas multipliée par 100 comme précédemment.

Afin de comparer cet arbre à celui obtenu avec l'approche GLM Tree, des hyperparamètres com-

parables ont étés appliqués. En effet, la même profondeur de l'arbre (=5) a été choisie. Pour ce qui est du minimum d'observations dans le nœud, il n'a pas été possible de le fixer à la même valeur que celle du GLM Tree. Il existe un biais dans cet algorithme. En effet, les proportions en pourcentage affichées dans l'arbre ne correspondent pas à l'exposition mais plutôt au nombre de lignes que contient notre base. L'algorithme CART ne semble pas fonctionner avec des données agrégées.

Cet arbre sera tout de même analyser afin de déterminer si les résultats concordent ou non à ceux obtenus à l'issu du partitionnement avec le GLM Tree.

On s'apercoit en premier lieu de l'utilisation de la variable indiquant la classe de risque pour entamer le partitionnement des données. Notons que celle-ci est utilisée le même nombre de fois (=2) que dans l'arbre précédent. Par contre, elle était utilisée auparavant au tout dernier niveau de l'arbre. Le GLM Tree détectait moins de variations selon cette variable. Dans l'arbre actuel, cette dernière semble être plus importante. Notons que le seuil du découpage reste le même. La majorité des profils d'individus font partie de la classe de risque la plus saine. Ils sont en effet plus nombreux que les individus fumeurs. La déviation calculée est plus forte pour la classe la moins saine (1.6) par rapport à la classe la plus saine (0.95). Ce résultat est cohérent avec l'approche GLM Tree. Après ce premier découpage, il ne reste que 43% de la population à répartir pour les deux niveaux restants de l'arbre.

Après le découpage selon la classe de risque, intervient celui selon l'âge. La variable âge n'est utilisée qu'une seule fois dans cet arbre contre 3 fois dans l'arbre précédent.

Pour les individus les plus âgés, cet arbre indique que la déviation est plus forte pour ceux dont les montants assurés sont les plus faibles (1.8). Ce résultat est rapidement contredit par le nœud suivant. Il se trouve que les individus avec des montants assurés de 49999$ ont une déviation plus faible que ceux ayant des montants plus élevés.

Pour les individus les plus jeunes, la déviation obtenue est plus faible pour les fumeurs (SS + PS à 1.3) que pour les non fumeurs (SNS à 2.1). Ce dernier résultat pour les fumeurs SNS est retrouvé dans l'arbre de la figure 3.14 où la profondeur de l'arbre GLM Tree était revu à la hausse.

Conclusion partielle

Pour conclure ce chapitre, au niveau du GLM Tree, il existe des segments de portefeuilles d'assureurs qui représentent moins de risques que d'autres. En effet, bien qu'affichant des déviations, ceux-ci sont considérés moins inquiétants. Soit en se plaçant dans un contexte de prudence, soit parce que les montants assurés sont assez faibles pour impacter négativement le résultat d'un (ré)assureur. Toutefois, il existe des groupes pour lesquels le risque de mauvaise calibration des taux de décès peut générer des pertes énormes à la compagnie. Dans ce dernier cas, différents types d'actions s'offrent à la compagnie afin de gérer ce risque :

- L'augmentation de la prime,
- Une réduction du risque (par rétrocession).

L'arbre crée avec le GLM Tree et soumis aux tests de sensibilité des paramètres est relativement stable. En effet, peu de changements de structure s'opèrent. Les différents arbres créés mènent aux mêmes conclusions.
Certains résultats du CART se contredisent et même ne concordent pas avec le GLM Tree. Notons

que les conclusions tirées du CART sont à prendre avec des pincettes. La construction de l'algorithme en tenant compte des lignes de la base comme les poids est la seule raison plausible à ces écarts de conclusion. Il faut tout de même noter que les variables utilisées dans les deux arbres sont exactement les mêmes.

Le GLM Tree permet de répondre à notre problématique, celle d'identifier les groupes d'individus dont la mortalité s'écarte des hypothèses. Contrairement au GLM Tree, le CART n'intègre pas convenablement les poids dans l'analyse pour définir la taille des groupes. De plus, l'utilisation de la validation croisée avec des données agrégées est discutable et peut être remise en cause. La validation croisée fonctionne mieux sur les données individuelles. Au vu des éléments qui précèdent, l'utilisation de l'algorithme CART ne permet pas d'affirmer de la pertinence du GLM Tree.

Une seconde application sera alors réalisée pour tester le GLM Tree d'une manière différente. En effet, il sera question de l'appliquer sur les données de la population nationale américaine dont les résultats général en termes de mortalité sont connus et rappelés à la section 1.3.

Chapitre 4

Étude des déviations de mortalité de la population nationale américaine

4.1 Objectif de l'étude

La variable *ethnie* légalement renseignée dans cette base n'est utilisée qu'à des fins illustratives dans cette seconde application. En tarification, son utilisation poserait un réel problème d'éthique. Les résultats obtenus dans cette partie ne sauraient donc servir de base afin de proposer un quelconque contrat d'assurance. Cette seconde application à but démographique permettra d'illustrer la pertinence de l'algorithme GLM Tree.

Après avoir étudié les déviations de mortalité dans l'industrie américaine, nous allons à présent voir ce qu'il en est généralement de la mortalité aux USA en étudiant la population nationale. Dans ce nouveau contexte, il n'existe pas de table standard qu'on pourrait utiliser et considérer comme la mortalité de référence. Des hypothèses seront donc émises sur cette mortalité standard. Contrairement à la première application, il sera question ici des déviations en nombre de décès uniquement. Les données récoltées ne permettent pas d'avoir cette seconde vision des déviations en montant assuré.

L'idée de cette nouvelle application reste la même que précédemment, tout en intégrant de nouvelles variables. Ces dernières nous ouvrent donc de nouvelles perspectives en permettant d'obtenir des résultats totalement différents de la première application mais tout aussi intéressants.

Toujours à l'aide de la fonction `glmtree()`, le partitionnement des données se fera en fonction de l'âge, du sexe, de l'ethnie, du niveau d'éducation et des causes de décès. Le but est de former des groupes d'individus qui présentent un comportement similaire. Les résultats de cette seconde application seront comparés à ceux mentionnés dans la partie sur l'étude de la mortalité américaine, section 1.3. Dans le cas où les résultats concordent, cela indiquerait que l'algortihme GLM Tree au travers du modèle binomial est pertinent dans le cadre de cette étude.

4.2 Source des données et définition de l'échantillon

Les données utilisées dans cette seconde application proviennent de différentes sources. Les données concernant les nombres et les causes de décès proviennent de la CDC (Centers for Disease and Control),

NCHS (2010) et sont observées annuellement. Un rapport correspondant est également disponible pour chaque année. Au sujet de l'exposition, on utilise une base différente qui provient du Census Bureau *. Les données ont été collectées pour la période allant de 2000 à 2015 puis aggrégées. Après traitement de la base, les variables et modalités inscrites dans le tableau 4.1 ont été retenues.

Variables	Description	Type	Modalités
Gender	Le sexe de l'individu	Qualitative	F M
Ethnie	L'origine de l'individu	Qualitative	Asian Black Hispanic White
Educ	Le niveau d'éducation de l'individu	Qualitative	High School HS not grad Graduate College no bach Bach
Age	L'âge de l'individu à la souscription du contrat	Quantitative	15-24 25-34 35-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79 80-84 85-89 90-94 95-99 100+
COD (Cause of Deaths)	Les causes de décès	Qualitative	Circulatory Nervous External Cancer Other
Deaths (A)	Le nombre de décés observé dans les données	Quantitative	-

TABLE 4.1: Descriptif des variables utilisées

*https ://www.census.gov/programs-surveys/popest/data/tables.html

Nous avons décidé de travailler avec les individus de la tranche d'âge 15-24 et plus en faisant fi des personnes d'âge inférieur.

Dans le but de faciliter les interprétations et figures dans la suite,les causes de décès (COD) ont été regroupées en 5 grandes causes. La modalité External correspond ici aux causes non médicales (suicides, chutes, accidents de la route). La modalité Other regroupe quant à elle, les 11 autres causes non mentionnées dans ce tableau mais qu'on peut retrouver aux annexes A.3, A.4 et A.5.

4.2.1 Statistiques descriptives sur les données

Nous commençons notre étude en examinant quelques statisques de notre base.

La première concerne la distribution de la population en fonction de l'âge et de l'ethnie comme le suggère la figure 4.1.

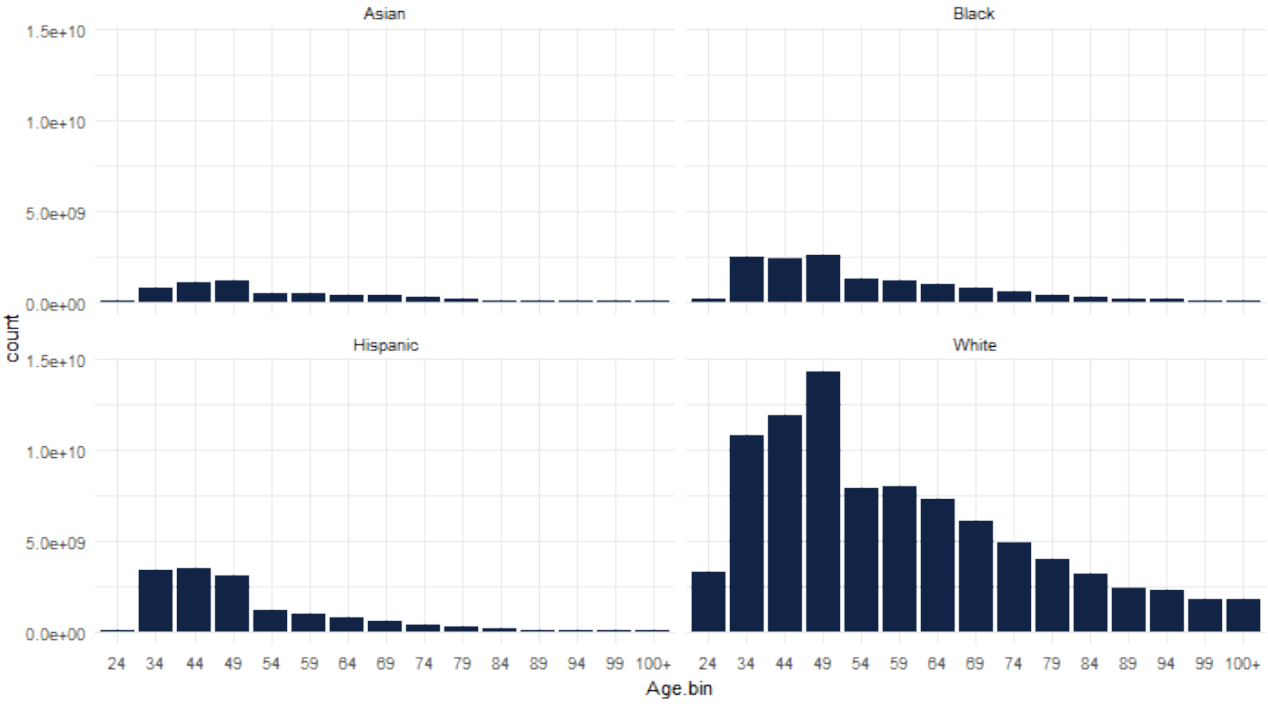


FIGURE 4.1: Distribution de la population par âge et par ethnie

Il se trouve que la population étudiée est à forte majorité blanche. Ce résultat n'est pas surprenant lorsqu'on sait qu'en 2015, les américains blancs représentaient environ 60% de la population américaine totale. Les asiatiques présentent l'exposition la plus faible dans notre jeu de données. En 2015, ils représentaient 5.6% de la population américaine. Au niveau des âges, on observe pour toutes les ethnies une distribution dissymétrique à gauche. En effet, les individus de cette base sont assez jeunes avec un âge moyen de 50 ans.

La même étude avec les causes de décès en lieu et place de l'ethnie conduit à la figure 4.2.

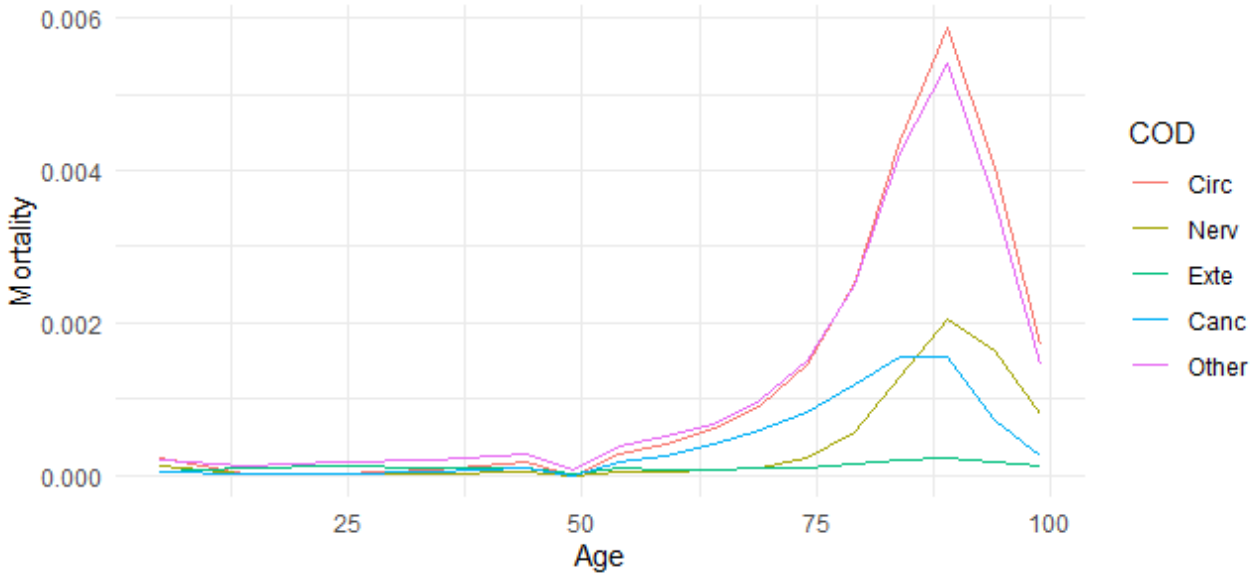


FIGURE 4.2: Distribution de la mortalité par âge et par causes de décès

Les individus âgés d'au moins 50 ans décèdent plus des maladies circulatoires et des autres causes de maladies. Concernant la modalité Other regroupant les 11 autres variables, il est naturel d'avoir des taux de décès si élevés. On s'aperçoit que la population étudiée décède moins de causes externes. Les causes de décès affichent cette fois une dissymétrie à droite selon l'âge : il y a plus de décès, toutes causes confondues, aux âges les plus élevés.

4.3 Arbre obtenu à l'issu du partitionnement avec GLM Tree

En ne prenant pas en compte la variable causes de décès dans cette première partie, nous allons étudier les déviations de mortalité par rapport à la mortalité moyenne. Le schéma suivit ici reste le même que celui décrit dans la sous-section 2.4.4 pour capturer les déviations. Cependant, dans notre cas, la mortalité observée correspond au nombre de décès par âge, sexe, ethnie, niveau d'education et la mortalité moyenne est celle obtenue par âge et par sexe.

4.3.1 Sans prise en compte de la variable COD

Comme précédemment, l'algorithme GLM Tree sera appliqué à nos données et les déviations seront capturées à l'aide de l'intercept β_0. Ensuite, les tests d'instabilité sur les variables de partitionnement (âge, sexe, ethnie, niveau d'éducation) révèleront les variables qui différencient au mieux la mortalité observée de la mortalité moyenne par âge et par sexe. Nous obtenons la figure 4.3.

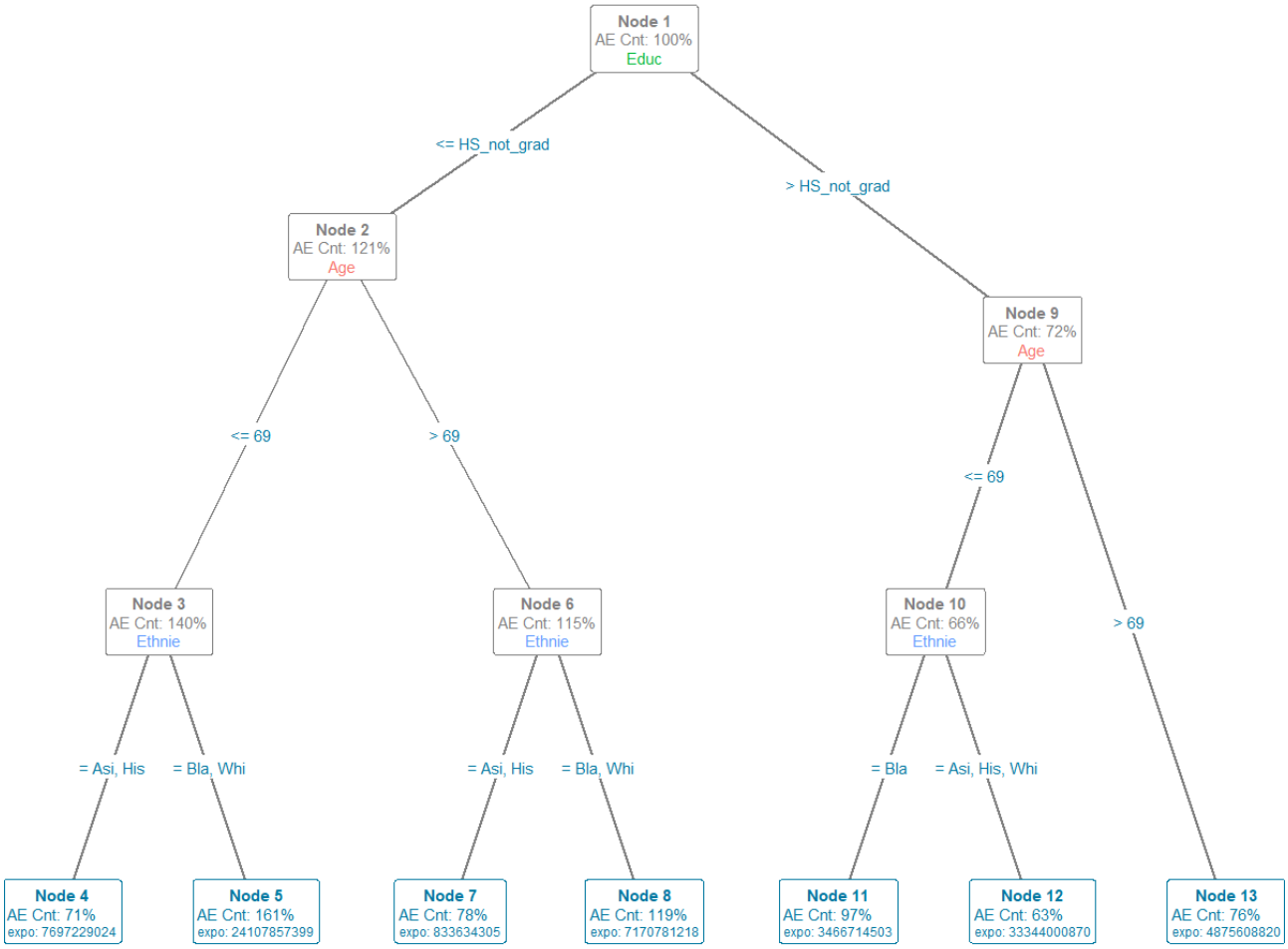


FIGURE 4.3: Arbre issu du partitionnement des données sans les causes de décès

La première déviation affichée est à 100%. C'est la quantité attendue vu que nous capturons les déviations par rapport à la mortalité moyenne. Sur les quatre variables de partionnement, seulement trois sont utilisées dans cet arbre. La variable ethnie, bien qu'utilisée au 3^e niveau de l'arbre est utilisée trois fois pour partitionner les données.

La première variable de partitionnement, Education, est celle pour laquelle on capture le plus d'instabilités. En effet, ces effets ne sont pas capturés par la mortalité moyenne utilisée dans ce cas. Ce premier découpage sépare les lycéens de ceux qui ont obtenu l'équivalent du Baccalauréat. On observe une mortalité plus faible par rapport à la mortalité moyenne pour les individus ayant au moins le Bac. Ce résultat est valable quelque soit l'âge et l'ethnie des individus.

Pour les individus ayant un niveau d'instruction inférieur ou égal au lycée, la mortalité observée est au dessus de la moyenne par âge et par sexe. En considérant pour tout âge confondu les américains asiatiques et hispaniques, on observe une mortalité largement inférieure à celle des américains noirs et blancs. Cela confirme les résultats révélés dans la littérature par la CDC (2002) : les asiatiques et hispaniques ont un taux de mortalité beaucoup plus faible que le reste de la population aux USA.

Bien que les effets de la variable âge soient capturés, cette dernière apparaît tout de même dans l'arbre. Avec un âge moyen à 50 ans observé dans les données issues de la population nationale, les américains blancs et noirs dont l'âge est inférieur à 69 ans représentent le segment de la population

pour lequel la déviation de mortalité par rapport à la moyenne est la plus élevée.

En somme, les déviations les plus faibles par rapport à la mortalité moyenne sont observées dans les segments suivants :

- Les individus les plus éduqués ayant au moins obtenu l'équivalent du Baccalauréat,
- Les américains asiatiques et hispaniques tout âge confondu.

Les américains noirs et blancs se démarquent dès lors des autres ethnies en affichant une mortalité largement supérieure à la moyenne.

Analyse des déviations

Encore une fois, les intervalles de confiance vont servir à analyser les déviations obtenues. Le premier panel de la figure 4.4 montre que le ratio de mortalité à 100% n'est compris dans aucun des IC de nos déviations. Le second panel apporte l'information concernant le nombre de décès observé dans chaque nœud.

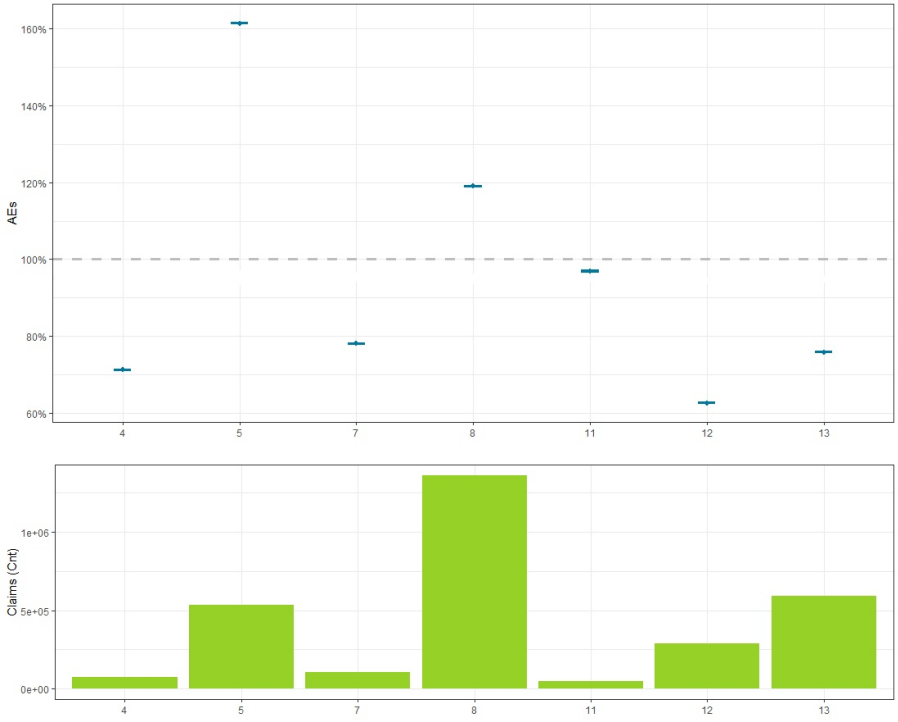


FIGURE 4.4: Ratio des décès observés et moyens par âge et par sexe

On obtient que les nœuds pour lesquels la mortalité est en dessous de la moyenne affichent un nombre de sinistres assez faible. Il s'agit des nœuds 4, 7, 11 et dans la moindre mesure les nœuds 12 et 13.

Pour les nœuds 4, 7 contenant les américains hispaniques et asiatiques et le nœud 11 contenant les américains noirs, le second panel de la figure 4.4 indique qu'ils présentent les nombres de décès les plus

faibles. Rappelons aussi les expositions de ces trois ethnies lors de l'étude des statistiques descriptives qui étaient les plus faibles de la base.

Les nœuds 12 et 13 des personnes ayant au moins obtenues le baccalauréat affichent également une mortalité plus faible que la moyenne. Cela suggère que les individus avec un niveau d'études élevé décèdent moins que la population moyenne comme le mentionne les études réalisées par Olshansky et al. (2012). Le niveau d'éducation accentue l'écart de mortalité entre les personnes ayant au moins le Bac et les autres.

Les nœuds restants (5 et 8), renfermant les américains noirs et blancs peu éduqués affichent les déviations ainsi que les nombres de décès les plus élevés. On remarque que le plus grand nombre de décès se produit pour les américains noirs et blancs (nœud 8) pour lesquels l'âge est supérieur à 69 ans. Dans ce nœud, l'écart de déviation de la mortalité par rapport à la moyenne s'élève à 19%. Finalement, parmi les individus peu éduqués, les hispaniques et asiatiques ont une mortalité plus faible que la moyenne alors que les blancs et noirs ont une surmortalité.

Dans la suite, nous intégrerons les causes de décès afin de rajouter une information supplémentaire et déduire de nouveaux résultats concernant les types de causes de décès qui influenceraient plus certains segments que d'autres.

4.3.2 Intégration de la variable cause de décès dans notre fonction

Nous répétons encore une fois la même procédure que précédemment. Pour cette partie, la population moyenne sera calibrée en fonction de l'âge, du sexe et des causes de décès. Nous obtenons dès lors l'arbre construit à l'issu du partionnement des données et les déviations associées à chaque segment. Dans un souci d'affichage, ne seront affichées sur la figure 4.5 que les premières lettres de nos différentes causes de décès.

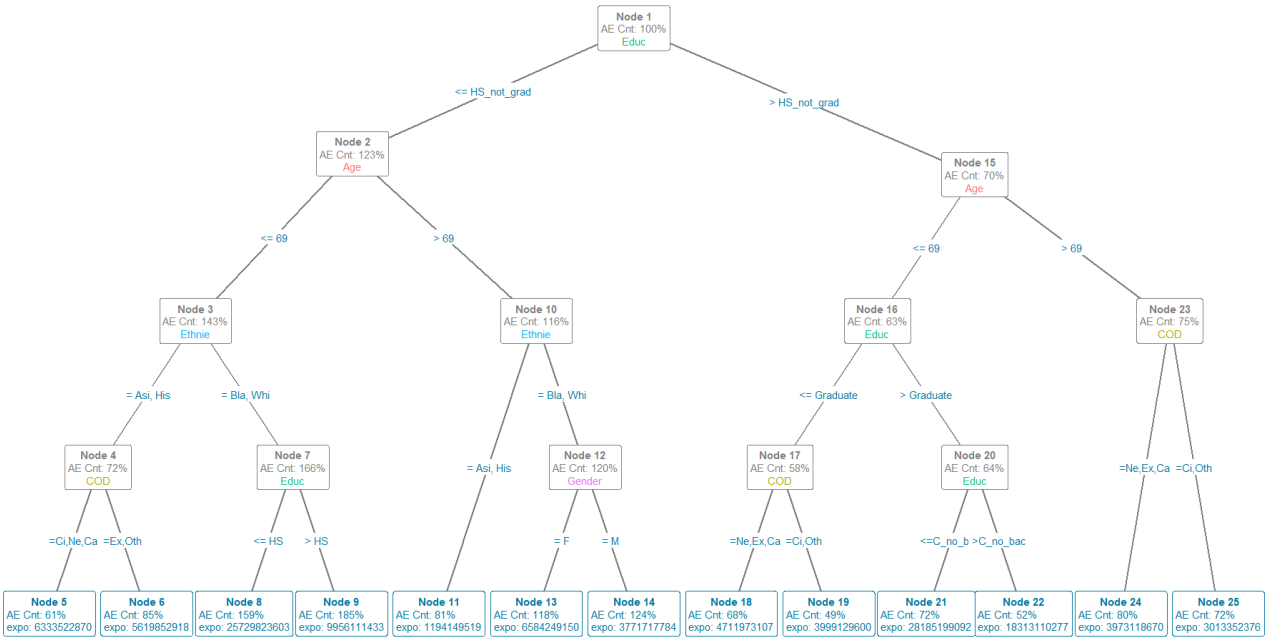


Figure 4.5: Arbre issu du partitionnement des données avec les causes de décès

Nous avons décidé de rajouter une profondeur par rapport à l'arbre de la figure 4.3 afin de capturer

plus d'informations. Pour ce qui concerne les deux premiers niveaux de l'arbre, il n'est pas aberrant de retrouver le même partitionnement que précédemment. Chacune des cinq variables (age, sexe, COD, niveau d'éducation et ethnie) utilisées comme variables de partionnement apparaît au moins une fois dans l'arbre.

Dès le premier nœud, on remarque que la mortalité est différente pour les individus ayant fait l'école primaire et ceux ayant au moins obtenus l'équivalent du Bac. La mortalité pour ces derniers est plus faible que la mortalité moyenne de la population.

Dès l'utilisation de l'âge comme variable de partitionnement, on observe un changement de comportement entre les personnes les moins éduquées et les autres. En considérant les individus les moins éduqués, on remarque encore une fois que les asiatiques et hispaniques se démarquent toujours autant des autres ethnies en affichant une mortalité plus faible que la moyenne. Cependant parmi les cinq causes de décès étudiées, la sous mortalité observée provient surtout des maladies circulatoires, nerveuses et cancéreuses (déviation à 61%) que des autres maladies (déviation à 85%). Résultat qui fait écho aux bonnes pratiques et hygiènes de vie de ces populations et déjà mentionné à la sous-section 1.3.1. Concernant les américains noirs et blancs, un niveau d'éducation supérieur au collège et inférieur à la dernière année de lycée (nœud 9) ne semble pas réduire le niveau de mortalité de ces populations. Bien au contraire, on observe une déviation à 185% pour ces individus contre 159% pour ceux encore moins éduqués. Notons par ailleurs que pour les individus d'âge supérieur à 69 ans, il apparaît une surmortalité des américains noirs et blancs de sexe masculin (nœud 14) par rapport au sexe opposé (nœud 13).

Voyons maintenant le cas des individus les plus éduqués, ayant au moins obtenu l'équivalent du Bac. Pour ceux dont l'âge est supérieur à 69 ans, la sous mortalité observée provient plus des maladies nerveuses, externes et cancéreuses que des maladies circulatoires et des autres causes. De même, on retrouve que la faible mortalité due aux maladies circulatoires et des autres causes par rapport à la population nationale explique la sous mortalité des individus les plus jeunes (< 69 ans) ayant un niveau d'études équivalent au lycée. Toutefois, ceux d'entre eux ayant au moins 1 diplôme post Bac décèdent moins que les autres n'ayant que le Bac.

Conclusion partielle

Au terme de cette seconde application, il ressort que les américains asiatiques et hispaniques se démarquent en affichant une mortalité plus faible que le reste de la population. La mortalité la plus forte concerne donc les américains noirs et blancs. Le niveau d'éducation est un facteur important à prendre en jeu car il creuse l'écart de mortalité lorsqu'on compare les individus les plus éduqués des autres. Pour ce qui touche les causes de décès, on remarque une sous mortalité des asiatiques et hispaniques par rapport à la population nationale qui s'explique par la faible mortalité due aux maladies cancéreuses, nerveuses et circulatoires. Les résultats de cette étude confirment ceux déjà présents dans la littérature au sujet de la mortalité américaine.

Outre la comparaison avec l'algorithme CART dans la première étude, cette seconde application est un moyen d'attester de la pertinence du GLM Tree et du modèle choisit dans le cadre de cette étude. En effet, retrouver des résutats déjà présents dans la littérature confortent dans l'idée que cet algorithme et le modèle implémenté pour les déviations (bien qu'approximé) est adapté à nos données.

Toutefois, comme mentionné au début de cette seconde application, l'utilisation de la variable *ethnie* rend inutilisable les résultats de cette étude dans un but de tarification de contrats d'assurance. Il aurait été possible de remplacer cette variable par une autre faisant état de la classe socio-économique,

comme les revenus annuels des individus par exemple.

Au final, ces deux applications permettent de mettre en évidence les réalisations possibles avec ce type d'algorithme. La fonction `glmtree()`, au delà de faire ressortir les segments de la population qui présentent des déviations par rapport à une population de référence, ne permet pas de tirer l'information concernant les évolutions de mortalité dans les différents segments.

Conclusion

Les risques biométriques sont au cœur du métier des compagnies d'assurance vie. Calibrer au mieux les tables de mortalité en particulier permet de tarifer au plus juste les contrats d'assurance. Dans ce contexte, une analyse comparative de la mortalité observée et prédite est indispensable afin d'assurer un suivi de la mortalité du portefeuille.

Ainsi, pour une première application, les données publiques de la SOA fournies par différentes compagnies d'assurance américaines ont fait l'objet d'une analyse approfondie. Nous avons comparé la mortalité de cette base par rapport à la table réglementaire 2015 VBT dans le but de capturer les déviations de mortalité en termes de A/E (Actual/Expected) en nombre de décès et en montant assuré. Cette étude a été menée sachant que les résultats obtenus ne seraient repésentatifs d'aucune des compagnies d'assurance volontaires.
Une seconde application, plus démographique, a été menée dans le but de capturer les déviations de mortalité de la population nationale américaine par rapport à une mortalité moyenne (par âge et sexe dans un premier temps, puis en intégrant les causes de décès).

L'enjeu de ce mémoire était de déceler les segments clés de la population qui présentent une déviation de mortalité significative par rapport aux hypothèses définies. L'approche choisie via la fonction `glmtree()` pour segmenter nos données et capturer les écarts de déviations a permis de comprendre les facteurs de mortalité sous-jacents, de surveiller les segments de la population présentants de larges déviations, mais également de mettre en place des actions de management pour faire face à ce genre de segments.

La pertinence de cette approche GLM Tree a été testée en utilisant d'abord un algorithme concurrent qu'est le CART. Puis, les résultats de l'application de l'algorithme GLM Tree à la population générale américaine ont été comparés à ceux issus de la littérature sur la démographie américaine.

Pour la première application, ce mémoire démontre d'une part le poids de la variable âge dans les déviations observées. Elle est suivie de la variable indiquant le montant assuré, variable repésentative de la classe socio-économique de l'individu et de la classe de risque, faisant état du degré du statut fumeur/non fumeur de la personne. D'autre part, ce mémoire indique que les plus grands écarts sont obtenus pour les personnes ayant de faibles montants assurés. Il se trouve que les individus dont les montants sont inférieurs à un certain seuil (100000$) ne suivent pas le même protocole de souscription que les autres, notamment en matière de sélection médicale ce qui pourrait expliquer leur différence de mortalité. En outre, l'analyse d'un segment de la population qui présente un écart de près de 30% par rapport à la mortalité standard, a révélé que les flux réels à débourser par la compagnie dans ce cas pourraient représenter environ 11% de plus que ce qui avait été prédit, d'où l'importance de calibrer de façon minutieuse les taux de mortalité pour une compagnie. Une des management actions qu'il est possible de mettre en place pour gérer ce genre de segments serait de se séparer de celui-ci ou sinon d'augmenter la prime demandée. Toutefois, dans un souci de réduire la volatilité du résultat, le transfert du risque porté par ce segment peut être envisagé. L'arbre construit avec le `glmtree()` n'est

que très peu sensible aux variations des différents hyperparamètres.

En ce qui concerne l'application de l'algorithme CART à ce jeu de données, les résultats sont à prendre avec précaution. En effet, les groupes sont définis selon le nombre de lignes de la base en lieu et place de l'exposition. Cela impacte les résultats au sein de l'arbre qui se contredisent parfois et ne concordent pas à ceux du GLM Tree. Il faut tout de même noter que les résultats de déviations obtenues dans cet arbre pour les classes de risque les plus ou moins saines sont les mêmes qu'auparavant. De plus, les variables utilisées pour le partitionnement dans chacun des arbres sont exactement les mêmes. Nous comprenons dès lors que certaines variables comme la classe de risque, l'âge et le montant assuré ont plus d'influence sur la mortalité et ses déviations que les autres.

Pour la seconde application plus démographique, nos résultats confirment ceux présents dans la littérature et permettent d'affirmer la pertinence du GLM Tree. En effet, ces derniers indiquent une mortalité pour les hispaniques et asiatiques plus faible que le reste de la population américaine. Le niveau d'éducation influence également la mortalité des individus. Étant un marqueur de la catégorie socio-économique, la mortalité des personnes les moins éduquées se trouve supérieure à celle des personnes les plus instruites.

En définitif, l'analyse de ces déviations a permis de mettre en lumière les segments les plus risqués pour une compagnie pour ainsi proposer des solutions adaptées.

De par sa robustesse, sa simplicité d'utilisation et d'interprétabilité, cette méthode des GLM Tree peut être intégrée à la boîte à outils de tout actuaire.

Les travaux réalisés ont permis de souligner un certain nombre de pistes d'améliorations. En particulier, il aurait été possible d'utiliser un autre modèle pour capturer les déviations. Une idée serait de construire un modèle plus complexe en intégrant l'âge parmi les variables de régression. De plus, l'inclusion de variables supplémentaires telles que la localisation géographique aurait permis de rajouter plus d'informations pour le partitionnement. En outre, il est évident que l'utilisation de la loi de poisson aurait été plus adaptée pour nos données. Il convient alors de retravailler la fonction `glmtree()` afin d'utiliser cette loi de distribution et intégrer les poids de façon convenable : c'est à dire, utiliser le nombre d'assurés comme paramètre du modèle et plus particulièrement pour construire la taille des nœuds finaux. Enfin, une analyse des tendances de mortalité dans un portefeuille aurait été possible en intégrant les années d'observation.

Bibliographie

Andrews, D. et Ploberger, W. (1994). Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative. *Econometrica* 62, 1383–1414.

Andrews, D. (1993). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica* 61, 821–856.

Arias, E. (2011). United States life tables: 2007.

Barbaste, M. (2017). Une méthode de provisionnement individuel par apprentissage automatique.

Beltrán-Sánchez, H., Finch, C. et Crimmins, E. (2015). *Proceedings of the National Academy of Sciences* 112.29, p. 8993-8998.

Benjamin, B. et Soliman, A. (1993). Mortality on the Move. *Actuarial Education Service, Oxford*, p. 105-130.

Bound, J., Geronimus, A., Rodriguez, J. et Waidmann, T. (2014). The Implications of Differential Trends in Mortality for Social Security Policy.

Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984). Classification And Regression Trees.

Brillinger, D. (1986). The natural variability of vital rates and associated statistics. *Biometrics* 42, 693–734.

CDC (2002). 2002 With Chartbook on Trends in the Health of Americans. url : www.cdc.gov/nchs/.

Committee, I. L. E. (2017).

Curtin, L. et Klein, R. (1995). Direct Standardisation. Centre for disease control and prevention. url : https://www.cdc.gov/nchs/data/statnt/statnt06rv.pdf.

Dowd, J. et Hamoudi, A. (2014). Is life expectancy really falling for groups of low socioeconomic status? Lagged selection bias and artefactual trends in mortality. *International Journal of Epidemiology* 43.4.

Harper, S., Lynch, J., Burris, S. et Davey-Smith, G. (2007). Trends in the black-white life expectancy gap in the United States, 1983-2003. *Journal of the American Medical Association* 11.297, p. 1224-1232.

Liddell, F. (1984). Simple exact analysis of the standardised mortality ratio. *Journal of Epidemiology and Community Health* 38, p. 85-88.

Loh, W. (2002). Regression Trees with Unbiased Variable Selection and Interaction Detection. *Statistica Sinica* 12, 361–386.

Lopez, O., Milhaud, X. et Thérond, P.-E. (jan. 2015). Arbres de régression et de classification (CART).

Macdonald, A., Cairns, A., Gwilt, P. et Miller, K. (1998). An International Comparison of Recent Trends in Population Mortality. *British Actuarial Journal* 4.1, p. 3-141.

Manchester, J. et Topoleski, J. (2008). Growing Disparities in Life Expectancy. Congressional Budget Office Economic and Budget Issue Brief. url : http://www.cbo.gov/publication/41681.

McCullagh, P. et Nelder, J. (1989). Generalized Linear Models. *Monographs on Statistics and Applied Probability* 37.

Milla, D. (2017). Introduction to Decision Trees Titanic dataset. URL : https://www.kaggle.com/dmilla/introduction-to-decision-trees-titanic-dataset.

National Academies of Sciences, E. et Medicine (2015). The Growing Gap in Life Expectancy by Income: Implications for Federal Programs and Policy Responses.

NCHS (2010).

Nelder, J. et Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society* 135, 370–384.

Nyblom, J. (1989). Testing for the Constancy of Parameters Over Time. *Journal of the American Statistical Association* 84, 223–230.

Olshansky, S., Antonucci, T., Berkman, L. et Binstock, R. (2012). Differences in life expectancy due to race and educational differences are widening, and many may not catch up. 31.8, p. 1803-1813.

Planchet, F. (2020). Modèle de durée - Tables de mortalité.

Planchet, F. et Théron, P. (2011). Modélisation Statistique Des Phénomènes de Durée - Applications Actuarielles. Economica.

Ploberger, W. et Krämer, W. (1992). The CUSUM Test with OLS Residuals. *Econometrica* 60, 271–285.

Potts, D. et Sammut, C. (2005). Incremental Learning of Linear Model Trees. *Machine Learning* 61, 5–48.

Preston, S. et Elo, I. (1995). Are educational differentials in adult mortality increasing in the United States? *Journal of Aging and Health* 7.4, p. 476-496.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : https://www.R-project.org/.

Scott, W. (1981). Some applications of the poisson distribution in mortality studies. *Transactions of the Faculty of Actuaries* 38, p. 255-263.

Singh, G. et Yu, S. (2019). Infant Mortality in the United States, 1915-2017: Large Social Inequalities have Persisted for Over a Century. *International journal of MCH and AIDS* 1.8.

SOA (2018). 2015 Valuation Basic Table Report.

Su, X., Wang, M. et Fan, J. (2004). Maximum Likelihood Regression Trees. *Journal of Computational and Graphica Statistics* 13, 586–598.

Therneau, T. et Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. URL : https://CRAN.R-project.org/package=rpart.

Timofeev, R. (2004). Classification and Regression Trees (CART) Theory and Applications.

Tweedie, M. (1984). An index which distinguishes between some important exponential families. *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute, Golden Jubilee International Conference.*

White, H. (1994). Estimation, Inference and Specification Analysis.

Zeileis, A. (2005). A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals. *Econometric Reviews* 24, 445–466.

Zeileis, A. et Croissant, Y. (2010). Extended Model Formulas in R: Multiple Parts and Multiple Responses. *Journal of Statistical Software* 34.1, 1–13.

Zeileis, A., Hothorn, T. et Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics* 17.2, 492–514.

Annexe A

Compléments des chapitres 3 et 4

A.1 Focus sur le nœud 13

Pour calculer les différents flux, il nous faut réaliser une interpolation de nos taux de décès ainsi que des montants.

Pour ce qui est des taux de décès, nous commençons par une interpolation de nos 7 points marqués sur la figure 3.7 afin d'avoir les taux associés à chaque âge individuellement.
Pour ce faire, nous utilisons une interpolation exponentielle pour chacun de nos deux taux. On veut avoir une fonction de la forme exponentielle : $y = A\ r^x$ qui représente bien la relation entre les âges et les taux observés et prédits. Pour le taux observé, l'équation est la suivante : $y = 3.10^{-6}e^{0,0894x}$. Concernant le taux prédit, l'équation de la courbe d'intepolation est donnée par : $y = 2.10^{-6}e^{0,0927x}$.

L'interpolation pour les deux taux avec les différentes équations permet d'obtenir la figure A.1.

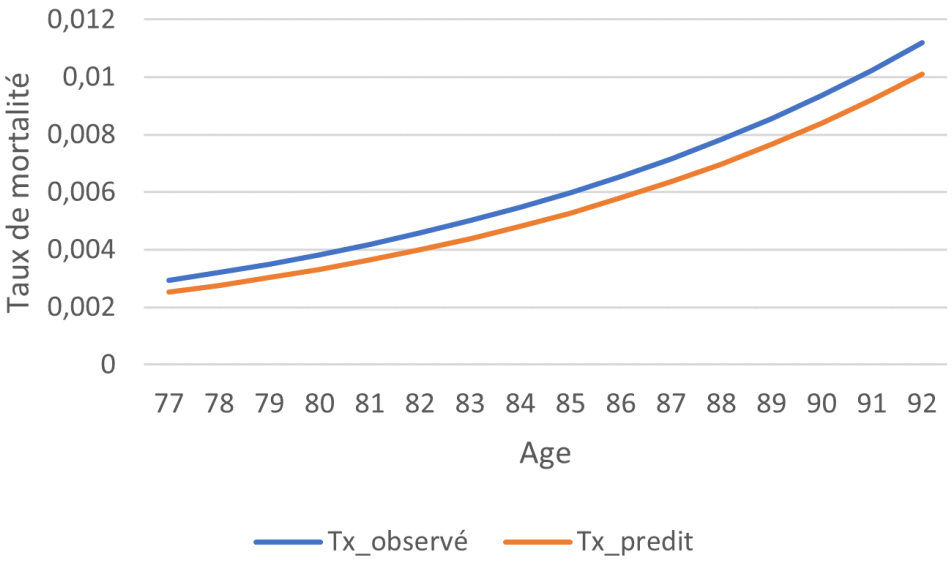


FIGURE A.1: Comparaison des taux de mortalité ajusté observé et prédit dans le nœud 13

On remarque que le taux observé est toujours au dessus de celui prédit. Cependant, il demeure

assez de différences par rapport à la figure 3.7. L'utilisation de l'interpolation exponentielle nous assure un R^2 de 0,9781 et 0,9888 pour respectivement les taux observés et prédits. Le R^2 étant proche de 1, la qualité de l'ajustement est jugée bonne.

Pour ce qui concerne les montants, nous partons du tableau A.1, qui contient la somme des montants assurés pour chaque point d'âge.

Age	Montant en $
62	$1,724818.10^{12}$
67	$1,940815.10^{12}$
72	$2,043877.10^{12}$
77	$2,154555.10^{12}$
82	$2,206264.10^{12}$
87	$2,161528.10^{12}$
92	$2,125034.10^{12}$

TABLE A.1: Tableau des montants assurés par âge

Il est important de noter que les données contenues dans ce tableau ne représentent pas exactement les données de départ. En effet, ayant travaillé avec des tranches de montant, il est contenu dans ce tableau la somme de la borne maximale des tranches pour chaque âge.

Comme pour les taux de décès, il nous faut trouver les montants associés à chaque âge individuellement. Pour ce faire, sachant que le montant affiché dans le tableau pour un âge donné est la somme des montants pour les 4 âges précédents, on répartit équitablement les montants par tranche de 5 en divisant les montants contenus dans le tableau par 5. À titre d'exemple, pour l'âge 67 ans, la somme de $3,88163.10^{11} (= \frac{1,940815.10^{12}}{5})$ correspondra aux montants pour les âges de 63 à 67 ans inclus et ainsi de suite.

Le tableau A.2 est ainsi obtenu à l'issu de cette procédure. Nous avons donc de cette manière une certaine interpolation pour nos montants en fonction de l'âge.

Age	Montant en $
62	2156791,8
63	2159320,7
64	2161849,6
65	2164378,5
66	2166907,4
67	2169436,3
68	2171965,2
69	2174494,1
70	2177023
71	2179551,9
72	2182080,8
73	2184609,7
74	2187138,6
75	2189667,5
76	2192196,4
77	2194725,3
78	2197254,2
79	2199783,1
80	2202312
81	2204840,9
82	2207369,8
83	2209898,7
84	2212427,6
85	2214956,5
86	2217485,4
87	2220014,3
88	2222543,2
89	2225072,1
90	2227601
91	2230129,9
92	2232658,8

TABLE A.2: Tableau des montants assurés ajustés par âge

A.2 Comparaison des taux de mortalité observés et prédits pour les différentes feuilles de l'arbre 3.3

La figure A.2 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 3.

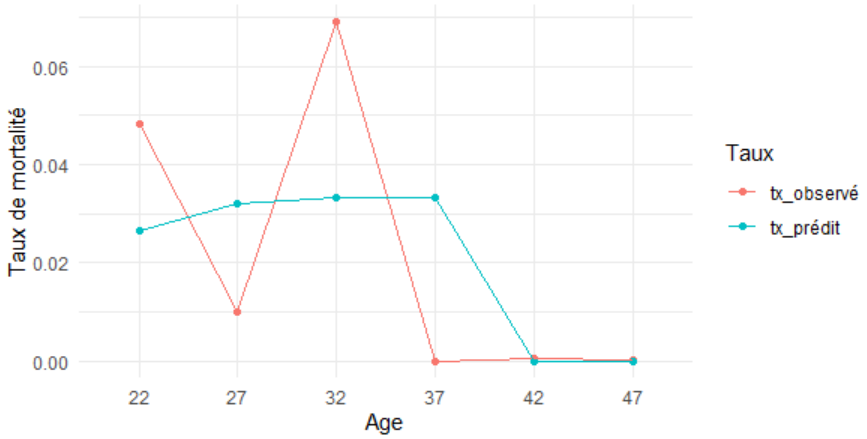


FIGURE A.2: Comparaison des taux de mortalité observé et prédit dans le nœud 3

Rappelons que ce nœud affiche une déviation s'élevant à 117%. Cela indique qu'il y a généralement plus de décès observés que prédits.

La figure A.3 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 5.

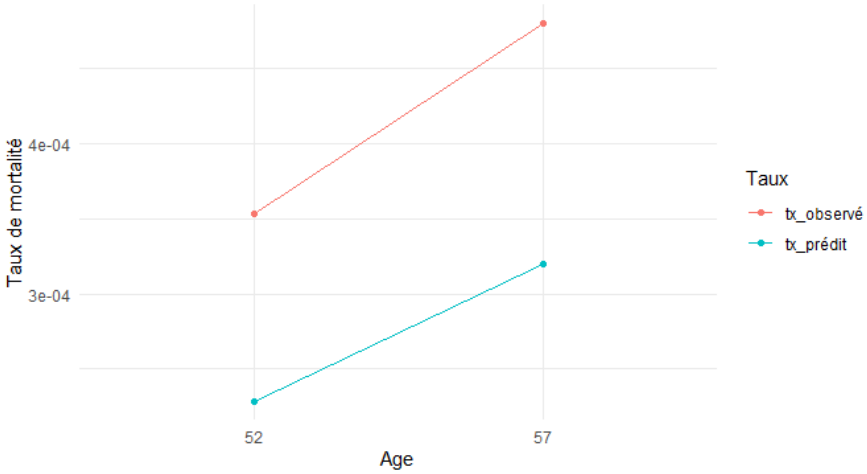


FIGURE A.3: Comparaison des taux de mortalité observé et prédit dans le nœud 5

Rappelons que ce nœud affiche une déviation s'élevant à 152%. Cela indique qu'il y a généralement plus de décès observés que prédits. L'écart de mortalité est pratiquement constant entre les deux âges de ce groupe de la population.

La figure A.4 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table

règlementaire dans le nœud 7.

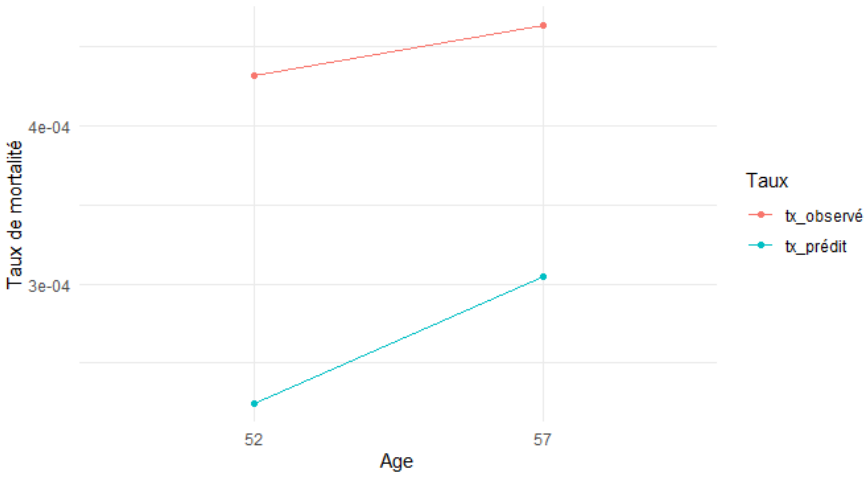


FIGURE A.4: Comparaison des taux de mortalité observé et prédit dans le nœud 7

Rappelons que ce nœud affiche une déviation s'élevant à 164%. Cela indique qu'il y a généralement plus de décès observés que prédits.

La figure A.5 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 8.

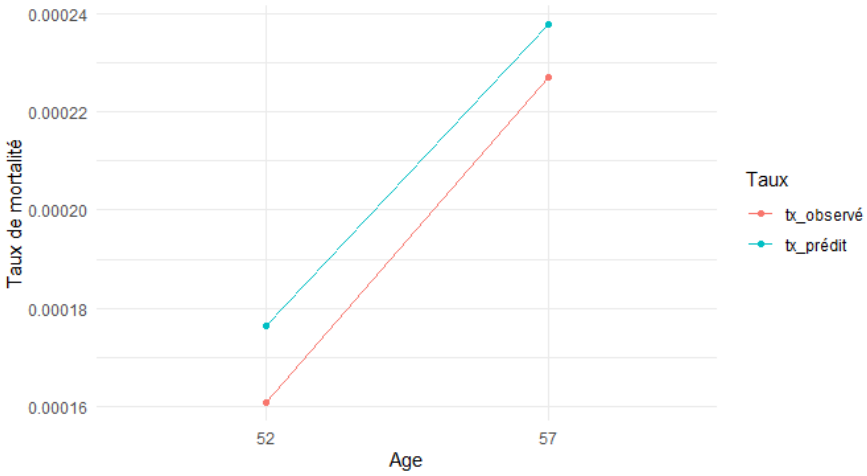


FIGURE A.5: Comparaison des taux de mortalité observé et prédit dans le nœud 8

Rappelons que ce nœud affiche une déviation s'élevant à 94%. Cela indique qu'il y a généralement plus de décès prédits qu'observés dans ce groupe.

La figure A.6 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 11.

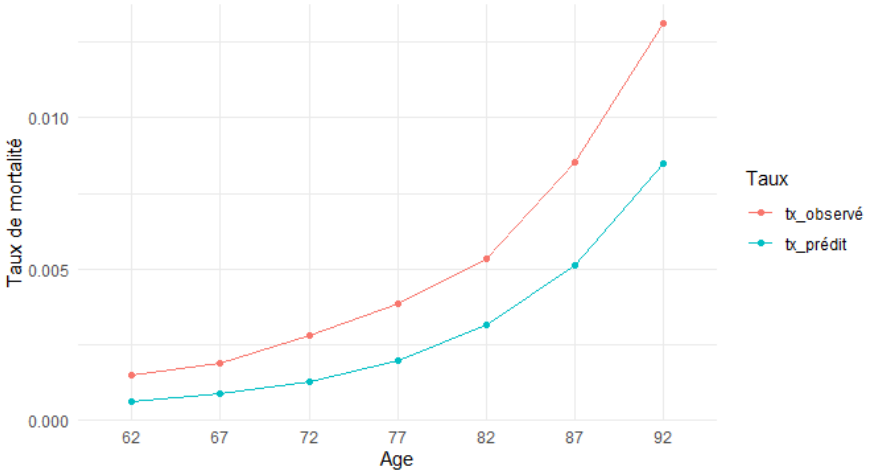


FIGURE A.6: Comparaison des taux de mortalité observé et prédit dans le nœud 11

Rappelons que ce nœud affiche une déviation s'élevant à 172%. Cela indique qu'il y a généralement plus de décès observés que prédits.

La figure A.7 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 14.

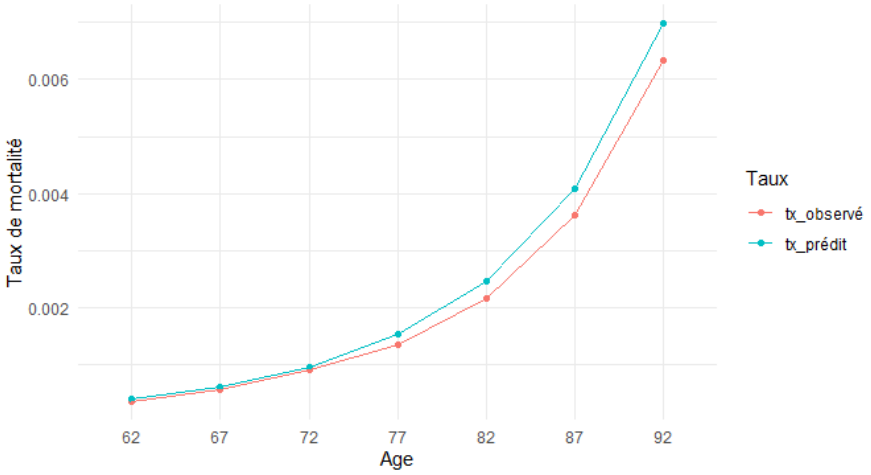


FIGURE A.7: Comparaison des taux de mortalité observé et prédit dans le nœud 14

Rappelons que ce nœud affiche une déviation s'élevant à 90%. Cela indique qu'il y a généralement plus de décès prédits qu'observés.

La figure A.8 ci-dessous permet d'observer les taux de mortalité observé et prédit par la table règlementaire dans le nœud 15.

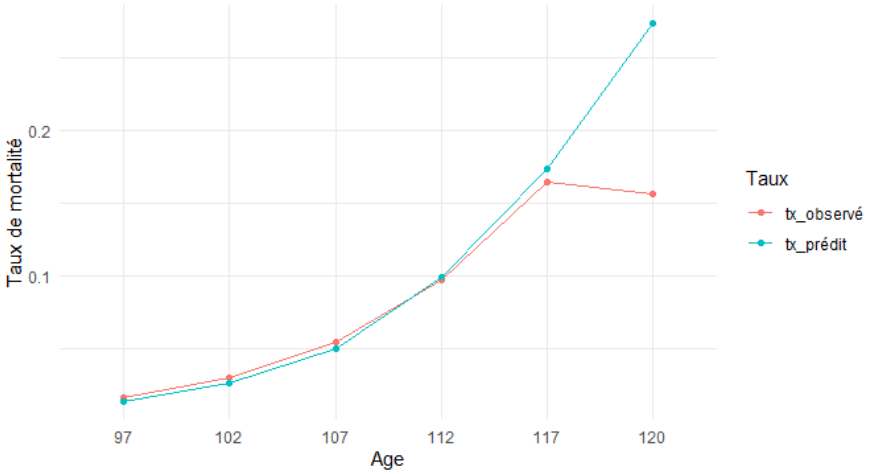


FIGURE A.8: Comparaison des taux de mortalité observé et prédit dans le nœud 15

Rappelons que ce nœud affiche une déviation s'élevant à 112%. Cela indique qu'il y a généralement plus de décès observés que prédits. Pour ces individus aux âges élevés, les taux observé et prédit fluctuent assez.

A.3 Les causes de décès

Les tableaux A.3, A.4 et A.5 ci-dessous donnent le détail des 5 grandes causes de maladies utilisées et mentionnées dans la seconde application de ce mémoire. On retrouve pour chaque grande cause, les maladies sous-jacentes ainsi que le codage associé.

	COD (Cause of death)	Coding
1. Circulatory	1. Diseases of the circulatory system	
	1 Heart diseases 1: Acute myocardial infarction	I20-I25
	2 Heart diseases 2: Due to heart failure (post-AMI, other diseases)	I00-I09, I30
	3 Heart diseases 3: Pulmonary embolism, night ventricle dysfunction	I26-I28, I80
	4 Stroke 1: Ischemic stroke (Atherosclerosis, etc.)	I10-I15, I63
	5 Stroke 2: Haemorrhage stroke (Aneurysm, etc)	I60-I62, I71
2. Nervous	2.Diseases of the nervous system	
	6 Parkinson's disease	G20-G21
	7 Alzheimer's disease	G30
	8 Other diseases of nervous system	G00-G19, G32
	3.Mental and behavioral disorders	
	9 Organic dementia	F01, F03
	10 Other mental and behavioral disorders	F02, F04-F09
3.External	4.External causes of mortality	
	11 Motor vehicle accidents	V02-V04, V09
	12 Suicide	U03, X60
	13 Homicide	U01, U02
	14 Other external causes of mortality	V01, V05-V09

TABLE A.3: Classification internationale des maladies

4.Cancer		
	5.Neoplasms	
	15 Malignant neoplasms of colon and rectum	C18-C20
	16 Malignant neoplasms of pancreas	C25
	17 Malignant neoplasms of trachea, bronchus and lung	C33-C34
	18 Malignant neoplasms of breast	C50
	19 Malignant neoplasms of prostate	C61
	20 Leukemia	C91-C95
	21 Malignant melanoma of skin	C43
	22 Malignant neoplasms of cervix uteri, corpus uteri and ovary	C52-C56
	23 Malignant neoplasms of lip, oral cavity and pharynx	C00-C14
	24 Other neoplasms	C15, C17
5.Other		
	6.Certain infectious and parasitic diseases	
	25 Tuberculosis	A16-A19
	26 Malaria	B50-B54
	27 Human immunodeficiency virus (HIV) disease	B20-B24
	28 Other infectious diseases	A00-A015
	7. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	
	8. Endocrine, nutritional and metabolic diseases	
	29 Diabetes mellitus	E10-E14
	30 Obesity and other hyperalimentation	E65-E68
	31 Other endocrine, nutritional and metabolic diseases	E00-E09

TABLE A.4: Classification internationale des maladies

9. Diseases of the eye and adnexa & diseases of the ear and mastoid process	
10. Diseases of the respiratory system	
32 Influenza and pneumonia	J09-J18
33 Chronic lower respiratory diseases	J40-J47
34 Other diseases of respiratory system	J00-J08, J19
11. Diseases of the digestive system	
35 Diseases of the digestive system	K00-K92
12. Diseases of the skin and subcutaneous tissue & diseases of the musculoskeletal system and connective tissue	
13. Diseases of the genitourinary system	
14. Pregnancy, childbirth and the pueperium & Certain conditions originating in the perinatal period	
15. Congenital malformations, deformations and chromosomal abnormalities	
16. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	

TABLE A.5: Classification internationale des maladies