Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 06/02/2020

Par : **Amal Elfassihi**

Titre : **Méthodes de prédiction et de correction automatique de données de mortalité, et applications**

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

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

Membres présents du jury de la filière

Caroline Hillairet

Entreprise : Milliman
Nom : Alexandre Boumezoued
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Alexandre Boumezoued
Signature :

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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

TABLE DES MATIÈRES

Résumé **iv**

Abstract **v**

Note de synthèse **vi**

Executive Summary **x**

Remerciements **xiv**

Introduction **1**

1 Modélisation stochastique de la mortalité **3**
1.1 Notations 3
1.2 La structure en âge et en temps de la mortalité 5
1.3 La famille des modèles de mortalité GAPC 8
1.3.1 Le modèle de Lee-Carter 9
1.3.2 Le modèle de Renshaw-Haberman 10
1.3.3 Le modèle Age-Period-Cohort 11
1.3.4 Le modèle de Cairns-Blake-Dowd 12
1.3.5 Le modèle de Plat 12
1.4 Calibration des modèles de mortalité GAPC 14

2 Approches par apprentissage automatique **15**
2.1 L'apprentissage automatique 15
2.1.1 Introduction 15
2.1.2 Apprentissage supervisé : régression et classification binaire 16
2.1.3 Méthode de minimisation du risque empirique 17
2.1.4 Applications de l'apprentissage automatique en modélisation de la mortalité 19
2.2 Les réseaux de neurones 19
2.2.1 Fondements mathématiques des réseaux de neurones 19
2.2.2 Fonctions d'activation 21
2.2.3 Optimisation des paramètres d'un réseau de neurones 22
2.2.4 Optimisation des hyperparamètres d'un réseau de neurones 24

2.2.5 Méthodes de régularisation 24
2.2.6 Raccourcis 26
2.3 Les forêts aléatoires 26
2.3.1 Les arbres de régression 26
2.3.2 Les forêts aléatoires 29

3 Une nouvelle méthode de correction de données de mortalité 30
3.1 Contexte et structure du chapitre 30
3.2 La correction des estimations de mortalité 31
3.2.1 Le problème de correction 31
3.2.2 Observables dans le diagramme de Lexis 33
3.3 La méthode de correction des estimations de mortalité 37
3.3.1 Séparation de l'échantillon de données en un échantillon d'entraînement et un échantillon de test 37
3.3.2 Variables d'entrée et *feature engineering* 38
3.3.3 L'approche par réseau de neurones 39
3.3.4 Le modèle de régression linéaire comme baseline 41
3.4 Résultats 42
3.5 Application de la méthode de correction 43
3.5.1 La matrice corrigée des taux d'amélioration de la mortalité 43
3.5.2 Impact de la méthode de correction sur des modèles stochastiques de mortalité 49
3.6 Bilan de la méthode de correction et ouverture 54

4 Modélisation prédictive de la mortalité future 55
4.1 Introduction 55
4.2 Données et définition de l'échantillon de données 55
4.2.1 Données 55
4.2.2 Définition de l'échantillon de données 57
4.2.3 Séparation de l'échantillon de données en un échantillon d'entraînement et un échantillon de test 58
4.3 Comparaison des approches de prédiction de la mortalité 59
4.3.1 Modèles stochastiques de mortalité 59
4.3.2 Forêt aléatoire 60
4.3.3 Réseau de neurones 60
4.4 Résultats comparatifs intermédiaires 62
4.5 Discussion 63
4.5.1 Choix de métrique et critères de comparaison 63
4.5.2 Discussion sur les âges très élevés retenus dans l'échantillon de données 64
4.5.3 Nouvelle comparaison 65
4.5.4 Bilan comparatif des approches 67
4.6 Application : étude d'un portefeuille de rentes viagères 68
4.6.1 Objectifs 68
4.6.2 Analyse descriptive du portefeuille fictif de rentes 68
4.6.3 Calculs des projections de mortalité 71
4.6.4 Le bilan économique d'un assureur 72
4.6.5 Première approche 72
4.6.6 Deuxième approche 74
4.6.7 Troisième approche 76

4.7 Bilan 80

Conclusion **81**

Bibliographie **83**

Annexes **88**
A Acronymes et symboles 88
B Extrait de code R de l'architecture du réseau de neurones 89
C Liste des pays disponibles dans la HMD 90
D Early Stopping 91
E Monte Carlo Dropout 93
F Indicateurs de correction prédits pour chaque échantillon de test, moyennés par âge au sein de chaque cohorte 94
G Paramètres du modèle M1, calibré sur des données de mortalité brutes et corrigées 98
G.1 Calibration sur les données de l'Allemagne de l'Ouest 98
G.2 Calibration sur les données de la Biélorussie 99
G.3 Calibration sur les données de l'Ukraine 100
G.4 Calibration sur les données de la Pologne 101
H Discussion : Détails du calcul des estimations de mortalité corrigées dans la version 6 de la HMD 102
I Tables réglementaires TGF 05 et TGH 05 103

Table des figures **105**

Liste des tableaux **108**

RÉSUMÉ

Ce mémoire d'Actuariat porte sur les dernières avancées en matière de correction d'anomalies dans les tables de mortalité nationales ainsi que la modélisation prédictive de la mortalité future.

Dans un premier temps, on se propose de présenter une approche par apprentissage automatique afin de corriger des anomalies de données de mortalité pour des pays pour lesquels on ne dispose pas de données mensuelles de fertilité. Ce travail permet ainsi de produire des tables de mortalité dites "corrigées" et s'inscrit dans la continuité des travaux de Cairns *et al.* (2016) et Boumezoued (2016). Comme résultat principal de la première partie du mémoire, il est démontré que cette nouvelle méthode de correction permet de fournir une correction de haute qualité, ainsi que d'améliorer la calibration de modèles stochastiques de mortalité sans composante de cohorte, comme c'est le cas pour les méthodes de correction existantes qui utilisent des données mensuelles de fertilité.

Dans un deuxième temps, on se propose de comparer des approches de prédiction de mortalité par apprentissage automatique (réseaux de neurones, forêts aléatoires) avec des approches plus classiques, dans la continuité des travaux de Richman et Wüthrich (2019). Par ailleurs, l'impact du choix de modèle de mortalité sur le calcul de projections d'un portefeuille fictif de rentes est également discuté. En plus du critère de la précision des prédictions obtenues par les différentes approches, d'autres critères comparatifs sont examinés. Tant dans la partie s'intéressant à la prédiction de la mortalité future que dans celle abordant la modélisation d'un portefeuille fictif de rentes dans son ensemble, les résultats obtenus montrent qu'il n'existe pas de modèle qui se dégage de manière claire au regard des critères de comparaison considérés.

Mots-clefs — *Human Mortality Database*, effet cohorte, anomalies de données de mortalité, modèles stochastiques de mortalité, apprentissage automatique, réseau de neurones, portefeuille de rentes

ABSTRACT

This thesis in Actuarial Science focuses on the latest advances in the correction of anomalies in national mortality tables, as well as the predictive modelling of future mortality.

First, we propose to present a machine learning approach in order to correct anomalous mortality data for countries for which monthly fertility data is not available. This work thus makes it possible to produce so-called "corrected" mortality tables and follows on from the work of Cairns *et al.* (2016) and Boumezoued (2016). As a main result, it is demonstrated that the new correction method proposed in this first part allows to provide a high quality correction and to improve the fitting of stochastic mortality models without a cohort component, as it is the case for the existing correction methods based on monthly fertility data.

Secondly, and following on from the work of Richman and Wüthrich (2019), machine learning approaches (neural networks, random forests) and more classical approaches to predict future mortality are compared. We also propose to illustrate the impact of the choice of the mortality model on the calculation of projections of a fictitious annuity portfolio. In addition to a predictive accuracy criterion, other comparative criteria are examined. Both in the part dealing with the future mortality prediction and in that considering the modelling of a fictitious annuity portfolio, the results obtained show that there is no model that emerges clearly regarding the comparison criteria that are discussed.

Keywords — Human Mortality Database, cohort effect, anomalous mortality data, stochastic mortality models, machine learning, neural network, annuity portfolio

NOTE DE SYNTHÈSE

Contexte de la correction des données de mortalité

Depuis son lancement en 2000, la *Human Mortality Database* (HMD) est devenue le fournisseur de référence des estimations de mortalité de plus de quarante pays ou régions. Ces indicateurs nationaux sont largement utilisés par les démographes, ainsi que par les compagnies d'assurance, afin d'obtenir des informations sur la mortalité future. Si la littérature consacrée à la sophistication des modèles stochastiques de mortalité est abondante au cours des dernières décennies, peu de contributions se sont concentrées sur la fiabilité des données de mortalité utilisées dans l'analyse historique et prédictive de la mortalité.

La prise de conscience d'anomalies de données de mortalité est apparue suite à la conjecture de Richards (2008), qui avait suggéré que certains effets cohorte en Angleterre pouvaient être dus à la sensibilité de la méthode de calcul des taux de mortalité aux chocs de fertilité. Cette conjecture a été confirmée par Cairns *et al.* (2016) qui a proposé une méthode pour corriger les données de mortalité pour l'Angleterre et le pays de Galles sur la base des données trimestrielles du taux de natalité. Le *Convexity Adjustment Ratio* introduit dans leurs travaux a ensuite été adapté par Boumezoued (2016) qui s'est concentré sur les données HMD et a montré que ces anomalies sont universelles. Pour construire des tables de mortalité corrigées, un lien avec la *Human Fertility Database* (HFD, 2018), l'équivalent de la HMD pour la fertilité, a été établi, rendant possible la correction de ces anomalies de manière systématique. La dernière version du protocole de méthodes HMD (Wilmoth *et al.*, 2019) introduit un changement dans la façon dont les taux de mortalité sont construits dans la base de données, en améliorant leur estimation par l'utilisation de données mensuelles de fertilité.

Néanmoins, ces méthodes de correction de données de mortalité ne s'appliquent pas aux pays pour lesquels les données de naissance par mois ne sont pas disponibles. C'est dans ce contexte que l'on s'intéresse dans une première partie de ce mémoire à la construction d'une méthode de production de tables de mortalité corrigées pour des pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles.

Résultats de notre nouvelle méthode de correction

Notre nouvelle méthode de correction a été testée sur des pays pour lesquels des données mensuelles de fertilité sont disponibles et les performances de la méthode ont été déterminées en calculant les erreurs MSE (pour *Mean-Squared Errors*) moyennes des prédictions de l'indicateur de correction, l'indicateur de correction observé correspondant à celui obtenu par la

méthode de correction de la HMD utilisant des données mensuelles de fertilité. L'échantillon de données utilisées comprend les pays européens suivants : Autriche, Danemark, Espagne, Finlande, France, Italie, Norvège, Suède, pour la période temporelle $[\![1935, 2016]\!]$, les cohortes $[\![1915, 1950]\!]$ et les âges $[\![18, 70]\!]$. Le tableau 1 consigne les résultats obtenus et montre que le modèle de base de régression linéaire donne des résultats très satisfaisants en matière d'erreur de prédiction MSE moyenne, les résultats moyens étant améliorés par l'approche par réseau de neurones.

Échantillon de test		**Régression linéaire**	**Réseau de neurones**	
Pays	Code HMD	MSE	MSE moyen	MSE médian
Autriche	AUT	$1,11 \cdot 10^{-4}$	$1,21 \cdot 10^{-4}$	$1,20 \cdot 10^{-4}$
Danemark	DNK	$1,28 \cdot 10^{-4}$	$7,66 \cdot 10^{-5}$	$7,56 \cdot 10^{-5}$
Espagne	ESP	$6,34 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$
Finlande	FIN	$1,15 \cdot 10^{-4}$	$8,48 \cdot 10^{-5}$	$8,41 \cdot 10^{-5}$
France	FRATNP	$1,44 \cdot 10^{-4}$	$8,56 \cdot 10^{-5}$	$8,54 \cdot 10^{-5}$
Italie	ITA	$7,42 \cdot 10^{-5}$	$6,55 \cdot 10^{-5}$	$6,59 \cdot 10^{-5}$
Norvège	NOR	$9,81 \cdot 10^{-5}$	$6,19 \cdot 10^{-5}$	$6,09 \cdot 10^{-5}$
Suède	SWE	$6,14 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$

TABLEAU 1. Erreurs MSE obtenues sur les échantillons de test par la régression linéaire et l'approche par réseau de neurones

La méthode a ensuite été appliquée à l'Allemagne de l'Ouest, l'Ukraine, la Biélorussie et la Pologne, pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles. Il s'est avéré que la volatilité des taux d'amélioration de la mortalité pour ces pays est considérablement réduite lors de l'utilisation de données de mortalité corrigées produites par notre méthode de correction.

L'impact de la méthode de correction a également été étudié sur trois modèles stochastiques de mortalité : le modèle M1, le modèle M3 et le modèle de Plat simplifié. Il apparaît que le modèle M1 est significativement amélioré lorsqu'il est calibré avec des données de mortalité corrigées, tandis que la log-vraisemblance demeure inchangée pour les modèles stochastiques de mortalité ayant une composante de cohorte. Ainsi, notre méthode de correction a le même effet que les méthodes de correction fondées sur les données mensuelles de fertilité.

La dynamique des paramètres de cohorte des modèles M3 et Plat simplifié est considérablement plus régulière lorsque ces modèles sont calibrés sur des données de mortalité corrigées. Il convient de souligner que la hiérarchie des modèles en matière de qualité d'ajustement peut même être inversée dans certains cas lorsque des données de mortalité corrigées sont utilisées dans la calibration. Ainsi, d'un point de vue statistique, il peut être dans certains cas davantage pertinent d'intégrer dans un modèle stochastique de mortalité un paramètre de sensibilité spécifique à l'âge plutôt qu'un paramètre de sensibilité spécifique à la cohorte.

Pour aller plus loin, notons que même pour le modèle de Plat simplifié qui présente une qualité d'ajustement particulièrement élevée quand il est calibré sur des données brutes, aucun changement n'est observé en matière de log-vraisemblance lorsqu'il est calibré sur des données de mortalité corrigées, alors que la dynamique du paramètre de co-

horte est considérablement modifiée. À l'inverse, le modèle M1 est instable en matière de log-vraisemblance mais ses paramètres demeurent stables. Le modèle M3 et le modèle de Plat simplifié sont stables en matière de log-vraisemblance mais leurs paramètres de cohorte demeurent instables.

Enfin, il convient de noter que notre méthode de correction peut être améliorée en ajoutant dans l'échantillon d'entraînement d'autres pays pour lesquels on dispose de données mensuelles de fertilité. De plus, des recherches plus approfondies sur l'architecture du réseau de neurones pourraient mener à des résultats plus optimaux que ceux présentés ici. D'autres améliorations possibles de notre méthode comprennent la considération de variables d'entrée supplémentaires ainsi que l'intégration des données de mortalité des femmes.

Contexte de la modélisation prédictive de la mortalité

Les prévisions de mortalité sont utilisées dans une grande variété de domaines : pour l'élaboration des politiques de santé, pour orienter la recherche pharmaceutique, la sécurité sociale, la planification des fonds de retraite et l'assurance-vie, pour ne citer que quelques-uns. Au cours du vingtième siècle, il est maintenant bien connu que la mortalité humaine a globalement diminué : dans la plupart des pays industrialisés, la mortalité chez les adultes et les personnes âgées révèle une probabilité de décès annuelle décroissante.

Par ailleurs, dans la littérature sur la modélisation de la mortalité, les approches par apprentissage automatique sont peu présentes. Parmi ces approches, Richman et Wüthrich (2019) ont étendu le modèle de Lee-Carter à plusieurs populations grâce au *Deep Learning*.

On se propose en deuxième partie de ce mémoire de discuter leur approche par réseau de neurones en comparant sa capacité prédictive avec celles de modèles stochastiques de mortalité ainsi qu'une approche par forêt aléatoire. En outre, l'impact du choix de modèle de mortalité sur le calcul de projections d'un portefeuille fictif de rentes est également discuté.

Comparaison des approches de prédiction de la mortalité

Des approches classiques de prédiction de la mortalité ont été comparées avec des approches par apprentissage automatique. La métrique classique de l'erreur MSE sur la force de mortalité μ utilisée notamment par Richman et Wüthrich (2019) n'a pas semblé être satisfaisante, car elle pénalise beaucoup plus les erreurs sur les âges élevés. Il existe dans notre échantillon de données des points qui correspondent à des forces de mortalité estimées supérieures à 1 (voir la figure 1) et qui, en raison des fluctuations d'échantillonnage, présentent des coefficients de variations énormes. Ceci nous a orienté vers le choix d'une nouvelle métrique d'erreur (que l'on nomme LMSE), ainsi que sur le choix de retirer les points correspondant aux âges supérieurs à 95 ans dans l'échantillon de données.

En retirant ces âges très élevés du calcul d'erreur, l'approche par réseau de neurones donne pour l'un des échantillons testés l'erreur MSE la plus élevée parmi l'ensemble des modèles considérés. Ainsi, la qualité des prédictions obtenues par l'approche de Richman et Wüthrich (2019) est dans ce cas moins bonne que celle d'un modèle de Lee-Carter à une population.

Tant dans la partie s'intéressant à la prédiction de la mortalité future que dans celle discutant la modélisation d'un portefeuille fictif de rentes dans son ensemble, les résultats obtenus montrent qu'il n'existe pas de modèle qui se dégage de manière claire. Les modèles de mortalité classiques restent pertinents et permettent par exemple d'obtenir facilement

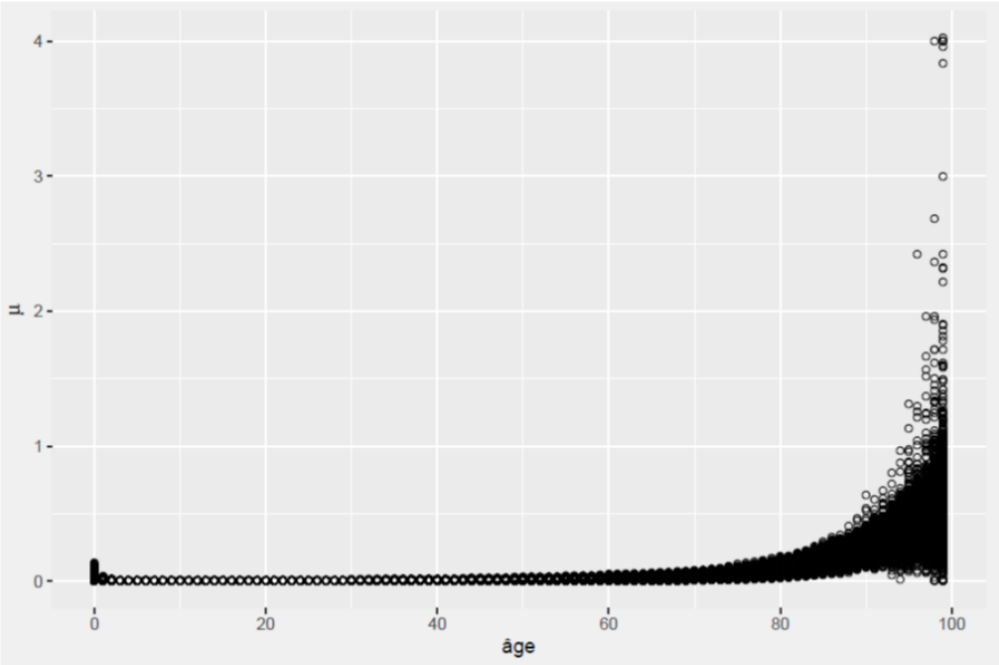


FIGURE 1. Nuage de points des données de force de mortalité μ de l'échantillon de données en fonction de l'âge

des intervalles de confiance, les approches par apprentissage automatique étant notamment pertinentes dans un cadre multi-populations (voir le tableau 2 pour plus de détails).

Critère	M1	M3	NN	RF
Précision des prédictions	+	-	+	++
Parcimonie	++	++	-	+
Simplicité à mettre en œuvre	+	+	-	+
Application à toute la tranche d'âges	-	-	+	+
Capacité à générer des intervalles de confiance	+	+	-	-
Capacité à gérer des populations multiples	-	-	+	+
Interprétabilité	+	+	-	-

TABLEAU 2. Bilan comparatif des approches. Le tableau met en évidence les critères satisfaits par chacun des modèles. Les dénominations NN et RF désignent respectivement l'approche par réseau de neurones et par forêt aléatoire.

EXECUTIVE SUMMARY

Background to the correction problem of mortality data

Since its launch in 2000, the Human Mortality Database (HMD) has become the benchmark supplier of mortality estimates for more than forty countries or regions. These national indicators are widely used by demographers, as well as by insurance companies, to obtain information on future mortality. While the literature on the sophistication of stochastic models of mortality has been abundant in recent decades, few contributions have focused on the reliability of the mortality data used in the historical and predictive analysis of mortality.

The awareness of anomalies in mortality data appeared following the conjecture of Richards (2008), who had suggested that some cohort effects in England could be due to the sensitivity of the method of calculating mortality rates to fertility shocks. This conjecture has been confirmed by Cairns *et al.* (2016) who proposed a method to correct mortality data for England & Wales based on quarterly birth rate data. The Convexity Adjustment Ratio introduced into their work was then adapted by Boumezoued (2016) who focused on HMD data and showed that these anomalies are universal. To build corrected mortality tables, a link with the Human Fertility Database (HFD, 2018), the counterpart of HMD for fertility data, has been established to systematically correct these anomalies. The latest version of the HMD method protocol (Wilmoth *et al.*, 2019) introduces a change in the way mortality rates are constructed in the database, improving their estimation through the use of monthly fertility data.

However, these methods of correcting mortality data do not apply to countries for which birth-by-month data is not available. It is in this context that we are interested in the first part of this thesis in the construction of a method allowing to produce corrected mortality tables for countries for which monthly fertility data is not available.

Results of our new correction method

Our new correction method has been tested for countries for which monthly fertility data is available and the performance of the method has been determined by computing the mean-squared errors (MSE) of the correction indicator predictions, the observed correction indicator corresponding to that obtained by the HMD correction method using monthly fertility data. The considered dataset includes the following European countries : Austria, Denmark, Finland, France, Italy, Norway, Spain, Sweden, for the time period $[\![1935, 2016]\!]$, cohorts $[\![1915, 1950]\!]$ and ages $[\![18, 70]\!]$. Table 3 presents the results obtained and shows that the baseline linear regression model gives very good results in terms of mean MSE prediction error, the results being improved by the neural network approach.

Test set		**Linear regression**	**Neural network**	
Country	HMD code	MSE	Mean MSE	Median MSE
Austria	AUT	$1,11 \cdot 10^{-4}$	$1,21 \cdot 10^{-4}$	$1,20 \cdot 10^{-4}$
Denmark	DNK	$1,28 \cdot 10^{-4}$	$7,66 \cdot 10^{-5}$	$7,56 \cdot 10^{-5}$
Spain	ESP	$6,34 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$
Finland	FIN	$1,15 \cdot 10^{-4}$	$8,48 \cdot 10^{-5}$	$8,41 \cdot 10^{-5}$
France	FRATNP	$1,44 \cdot 10^{-4}$	$8,56 \cdot 10^{-5}$	$8,54 \cdot 10^{-5}$
Italy	ITA	$7,42 \cdot 10^{-5}$	$6,55 \cdot 10^{-5}$	$6,59 \cdot 10^{-5}$
Norway	NOR	$9,81 \cdot 10^{-5}$	$6,19 \cdot 10^{-5}$	$6,09 \cdot 10^{-5}$
Sweden	SWE	$6,14 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$

TABLE 3. Test set MSEs of the baseline linear regression and the neural network

The method was then applied to West Germany, Ukraine, Belarus and Poland, countries for which monthly fertility data is not available. It turned out that the volatility of mortality improvement rates for these countries is considerably reduced when using corrected mortality data produced by the correction method.

The impact of the correction method was also discussed on three stochastic mortality models : the M1 model, the M3 model and the simplified Plat model. It appears that the M1 model is significantly improved when calibrated with corrected mortality data, while the log-likelihood is unchanged for the stochastic mortality models with a cohort component. Thus, our correction method has the same effect as correction methods based on monthly fertility data.

The dynamics of the cohort parameters of the M3 and simplified Plat models are much more regular when these models are calibrated on corrected mortality data. It should be noted that the ranking of models in terms of goodness-of-fit can even be reversed in some cases when corrected mortality data is used in the calibration. Thus, from a statistical point of view, it can be more relevant to integrate into a mortality model an age-specific sensitivity parameter rather than a cohort-related parameter.

To go further, note that even for a model which has a particularly high goodness-of-fit with raw data like the simplified Plat model, no change is observed in terms of log-likelihood when using retreated mortality data, whereas the dynamics of the cohort parameter is greatly changed. On the contrary, the M1 model in unstable in terms of log-likelihood but its parameters are stable. The M3 model and the simplified Plat model are stable in terms of log-likelihood but their cohort-related parameters are unstable

Finally, one should note that the correction method may be improved by adding other countries with available birth-by-month data in the training set. Moreover, a more comprehensive search over the architecture of the neural network, including networks with more layers and different hyperparameters settings, may produce results that are more optimal than those presented here. Other improvements to the model are the inclusion of additional input variables and the inclusion of Females data.

Background to the predictive mortality modelling

Mortality forecasting is used in a wide variety of areas : for health policy-making, to guide pharmaceutical research, social security, pension planning and life insurance, to name a few. During the twentieth century, it is now a well-known fact that human mortality has decreased overall : in most industrialised countries, mortality among adults and the elderly reveals a decreasing annual probability of death.

Furthermore, in the literature on mortality modelling, machine learning approaches are rarely present. Among these approaches, Richman and Wüthrich (2019) extended the Lee-Carter model to multiple populations using Deep Learning.

In the second part of this thesis, we propose to discuss their neural network approach by comparing its predictive capacity with those of stochastic mortality models as well as a random forest approach. In addition, the impact of the choice of mortality model on the calculation of annuity portfolio projections will be discussed.

Comparison of mortality prediction approaches

Classical approaches to predict mortality have been compared with machine learning approaches. The classical metric of the MSE error on the force of mortality μ used for instance by Richman and Wüthrich (2019) did not seem to be satisfactory, as it penalises errors at older ages much more. There are points in our dataset that correspond to estimated forces of mortality greater than 1 (see Figure 2) and which, due to sampling fluctuations, present enormous variation coefficients. This guided us towards the choice of a new error metric (that we called LMSE), as well as the choice to remove the points corresponding to ages greater than 95 from the dataset.

By removing these very high ages from the error calculation, the neural network approach gives in one case the highest MSE error among all the considered models. Thus, the quality of the predictions obtained by the approach of Richman and Wüthrich (2019) is in this case worse than that of a single population Lee-Carter model.

Both in the mortality prediction part and in the fictitious annuity portfolio modelling part, the results obtained show that there is no model that emerges clearly ; unfortunately, there is no miracle solution. Classical mortality models remain relevant and allow e.g. to easily generate forecast percentiles, machine learning approaches being particularly relevant in a multi-population framework (see Table 4 for more details).

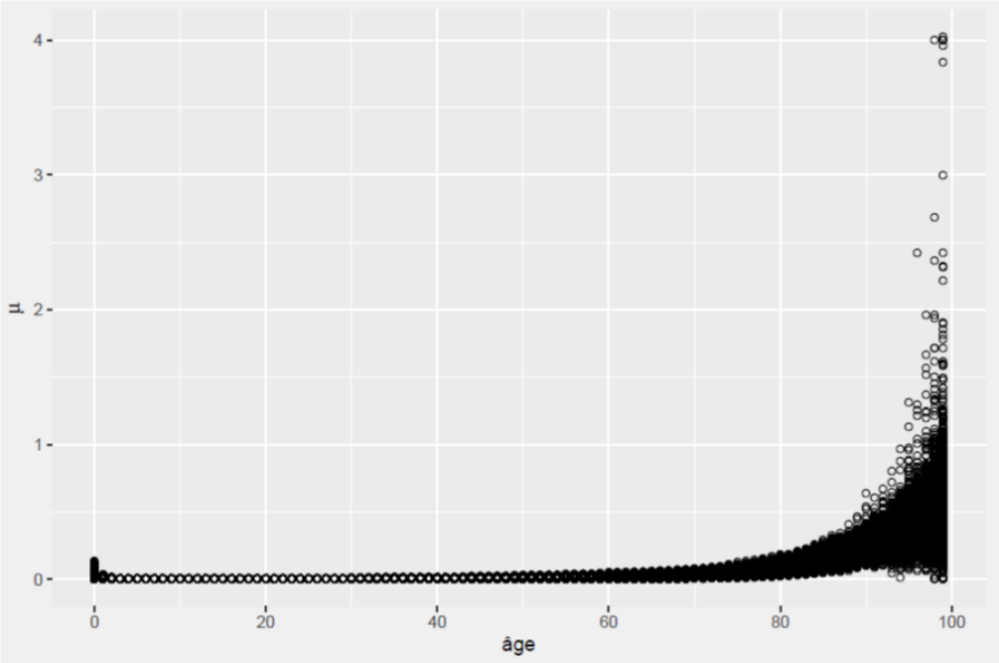


FIGURE 2. Scatterplot representing the estimated force of mortality μ as a function of the age in the dataset

Criterion	M1	M3	NN	RF
Predictions accuracy	+	-	+	++
Parsimony	++	++	-	+
Ease of implementation	+	+	-	+
The model is applicable for a full age range	-	-	+	+
Ability to generate forecast percentiles	+	+	-	-
Ability to manage multiple populations	-	-	+	+
Transparency	+	+	-	-

TABLE 4. Comparative assessment of approaches. The table highlights the criteria satisfied by each of the models. NN and RF respectively denote the neural network and the random forest approach.

REMERCIEMENTS

Je tiens à remercier le cabinet de conseil en actuariat Milliman pour l'opportunité qui m'a été donnée de travailler dans un cadre agréable et motivant.

Plus particulièrement, je remercie Alexandre Boumezoued, responsable du département Recherche et Développement de Milliman Paris, pour m'avoir supervisé le long de ce mémoire. J'ai pu apprécier sa disponibilité, ses précieux conseils et son sens aigu de la pédagogie.

Je suis également reconnaissant envers Caroline Hillairet pour son accompagnement depuis l'ENSAE Paris.

J'adresse également mes remerciements à mes collègues de Milliman pour les encouragements et les conseils qu'ils m'ont apportés.

Je remercie l'ensemble de l'équipe pédagogique de l'ENSAE Paris pour la qualité de la formation et de l'enseignement dispensés.

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

INTRODUCTION

Depuis son lancement en 2000, la *Human Mortality Database* (HMD) est devenue le fournisseur de référence des estimations de mortalité de plus de quarante pays ou régions. Ces indicateurs nationaux sont largement utilisés par les démographes, ainsi que par les compagnies d'assurance, afin d'obtenir des informations sur la mortalité future. Si la littérature consacrée à la sophistication des modèles stochastiques de mortalité est abondante au cours des dernières décennies, peu de contributions se sont concentrées sur la fiabilité des données de mortalité utilisées dans l'analyse historique et prédictive de la mortalité.

La prise de conscience d'anomalies de données de mortalité est apparue suite à la conjecture de Richards (2008), qui avait suggéré que certains effets cohorte en Angleterre pouvaient être dus à la sensibilité de la méthode de calcul des taux de mortalité aux chocs de fertilité. Cette conjecture a été confirmée par Cairns *et al.* (2016) qui a proposé une méthode pour corriger les données de mortalité pour l'Angleterre et le pays de Galles sur la base des données trimestrielles du taux de natalité. Le *Convexity Adjustment Ratio* introduit dans leurs travaux a ensuite été adapté par Boumezoued (2016) qui s'est concentré sur les données HMD et a montré que ces anomalies sont universelles. Pour construire des tables de mortalité corrigées, un lien avec la Human Fertility Database (HFD, 2018), l'équivalent de la HMD pour la fertilité, a été établi, rendant possible la correction de ces anomalies de manière systématique. La dernière version du protocole de méthodes HMD (Wilmoth *et al.*, 2019) introduit un changement dans la façon dont les taux de mortalité sont construits dans la base de données, en améliorant leur estimation par l'utilisation de données mensuelles de fertilité.

Néanmoins, ces méthodes de correction de données de mortalité ne s'appliquent pas aux pays pour lesquels les données de naissance par mois ne sont pas disponibles. C'est dans ce contexte que l'on s'intéresse dans une première partie de ce mémoire à la construction d'une méthode de production de tables de mortalité corrigées pour des pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles.

Les prévisions de mortalité sont utilisées dans une grande variété de domaines : pour l'élaboration des politiques de santé, pour orienter la recherche pharmaceutique, la sécurité sociale, la planification des fonds de retraite et l'assurance-vie, pour ne citer que quelques-uns. Au cours du vingtième siècle, il est maintenant bien connu que la mortalité humaine a globalement diminué : dans la plupart des pays industrialisés, la mortalité chez les adultes et les personnes âgées révèle une probabilité de décès annuelle décroissante.

Par ailleurs, dans la littérature sur la modélisation de la mortalité, les approches par apprentissage automatique sont peu présentes. Parmi ces approches, Richman et Wüthrich

(2019) ont étendu le modèle de Lee-Carter à plusieurs populations grâce au *Deep Learning*.

On se propose en deuxième partie de ce mémoire de discuter leur approche par réseau de neurones en comparant sa capacité prédictive avec celles de modèles stochastiques de mortalité ainsi qu'une approche par forêt aléatoire. En outre, l'impact du choix de modèle de mortalité sur le calcul de projections d'un portefeuille fictif de rentes sera discuté.

Ainsi, ce mémoire est organisé comme suit. Le chapitre 1 présente des approches de modélisation stochastique de la mortalité tandis que le chapitre 2 expose les approches par apprentissage automatique utilisées dans la suite. Le chapitre 3 présente et applique à des pays européens notre nouvelle méthode de correction des données de mortalité dans le cas où les données mensuelles de fertilité ne sont pas disponibles. Enfin, le chapitre 4 se propose de comparer des approches de prédiction de mortalité par apprentissage automatique avec des approches plus classiques, dans la continuité des travaux de Richman et Wüthrich (2019). Dans une partie plus pratique, l'impact du choix de modèle de mortalité sur le calcul de projections d'un portefeuille fictif de rentes est également discuté.

CHAPITRE 1

MODÉLISATION STOCHASTIQUE DE LA MORTALITÉ

1.1 Notations

Plusieurs modèles de mortalité sont considérés dans ce mémoire, il est donc important d'utiliser des notations claires et cohérentes.

On désigne par année civile t l'année qui s'étend du temps exact t au temps exact $t+1$. Soit $d(a,t)$ le nombre de décès enregistrés d'individus d'âge exact a au dernier anniversaire au cours de l'année civile t. Soit de plus $E(a,t)$ la somme des durées de vie de chaque individu d'âge exact a au dernier anniversaire au cours de l'année civile t, appelée exposition centrale au risque.

Soit T la variable aléatoire qui représente la durée de vie d'un individu donné. Alors la force de mortalité pour cet individu à l'âge a, qui est également la probabilité instantanée de décès, est définie par :

$$\mu(a) = \lim_{\mathrm{d}a \to 0^+} \frac{\mathbb{P}(T \in [a, a+\mathrm{d}a] \mid T > a)}{\mathrm{d}a}. \tag{1.1}$$

Néanmoins, la force de mortalité évolue dans le temps. Il est donc nécessaire d'ajouter une variable temporelle pour mieux décrire cette grandeur. La probabilité de décès de l'individu d'âge a au temps $t = a + \upsilon$, où υ désigne sa date de naissance, est donnée par :

$$\mu(a,t) = \mu(a, a+\upsilon) = \lim_{\mathrm{d}a \to 0^+} \frac{\mathbb{P}(T \in [a, a+\mathrm{d}a] \mid T > a, \upsilon)}{\mathrm{d}a}. \tag{1.2}$$

On définit également le taux de mortalité $q(a,t)$ comme la probabilité qu'un individu d'âge a au temps exact t décède dans l'année civile suivante, *i.e.* entre t et $t+1$:

$$q(a,t) = \mathbb{P}(a \leq T < a+1 \mid T \geq a, \upsilon). \tag{1.3}$$

Ainsi, la fonction ou probabilité de survie à n années pour cet individu s'obtient par :

$$S(n) = \exp\left(-\int_0^n \mu(\alpha, \alpha+\upsilon)\,\mathrm{d}\alpha\right) = \exp\left(-\sum_{k=0}^{n-1} \mu(k, k+\upsilon)\right). \tag{1.4}$$

Dans la suite, l'hypothèse standard (Cairns *et al.*, 2009; Currie, 2016) suivante est adoptée :

$$\forall a, \forall t, \forall \alpha, \tau \in [0,1[,\ \ \mu(a+\alpha, t+\tau) = \mu(a,t). \tag{1.5}$$

Ceci est mieux illustré à l'aide d'un système de coordonnées (t, a). Une telle représentation s'appelle un diagramme de Lexis (Lexis, 1875). Les deux échelles de temps sont divisées en bandes annuelles, qui divisent le diagramme de Lexis :

- des horizontales qui correspondent aux âges exacts entiers,
- des verticales qui délimitent les années civiles,
- des diagonales qui délimitent les cohortes ou générations.

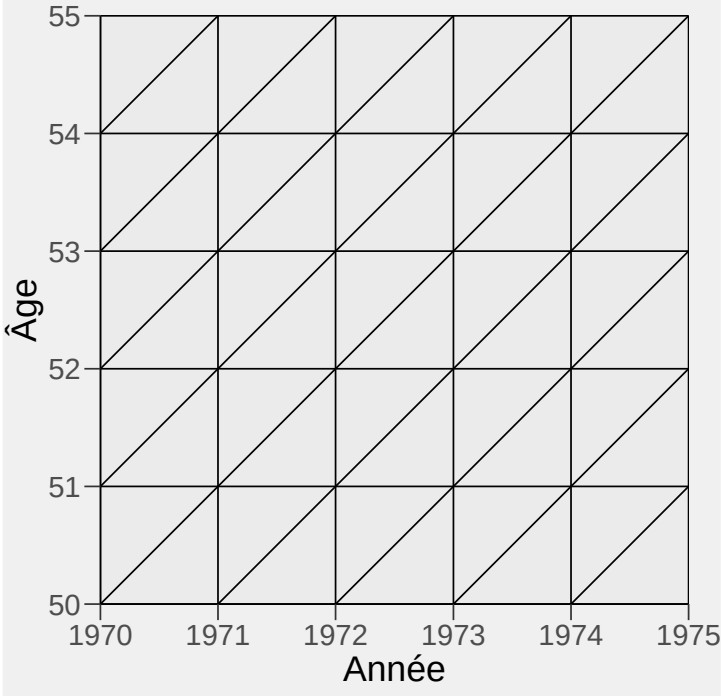


FIGURE 1.1. Diagramme de Lexis

L'hypothèse (1.5) suppose donc que la force de mortalité est constante dans chaque carré du diagramme de Lexis, mais lui permet de varier d'un carré à l'autre. Ainsi, il est par exemple supposé que la force de mortalité à l'âge exact $a + \frac{1}{2}$ et au temps exact $t + \frac{1}{2}$ est simplement $\mu(a, t)$.

Par ailleurs, cette hypothèse (1.5) a deux conséquences importantes :

- Il existe une relation simple entre la force de mortalité $\mu(a, t)$ et le taux de mortalité $q(a, t)$:
$$q(a, t) \approx 1 - \exp(-\mu(a, t)). \tag{1.6}$$
- Si $d(a, t)$ est traitée comme une variable aléatoire $D(a, t)$, et l'exposition centrale $E(a, t)$ comme fixée, alors $D(a, t)$ a une distribution de Poisson :
$$D(a, t) \sim \mathcal{P}(E(a, t) \cdot \mu(a, t)). \tag{1.7}$$

Les actuaires sont souvent davantage intéressés par le taux de mortalité $q(a, t)$ que par la force de mortalité $\mu(a, t)$. Il existe un résultat homologue à (1.7) pour le taux de mortalité $q(a, t)$. Si l'on approche l'exposition initiale au risque $E^0(a, t) \approx E(a, t) + \frac{1}{2}d(a, t)$, alors $D(a, t)$ a une distribution binomiale :

$$D(a, t) \sim \mathcal{B}(E^0(a, t), q(a, t)). \tag{1.8}$$

Les équations (1.7) et (1.8) mènent directement à des estimateurs de maximum de vraisemblance de $\mu(a,t)$ et $q(a,t)$:

$$\hat{\mu}(a,t) = \frac{d(a,t)}{E(a,t)}, \tag{1.9}$$

$$\hat{q}(a,t) = \frac{d(a,t)}{E^0(a,t)}. \tag{1.10}$$

En outre, une application directe du théorème de Taylor montre que :

$$-\log(1-\hat{q}(a,t)) - \hat{\mu}(a,t) \approx \frac{1}{12}\hat{\mu}(a,t)^3, \tag{1.11}$$

cette quantité étant suffisamment petite pour que les estimateurs de maximum de vraisemblance de $\mu(a,t)$ sous (1.7) et $q(a,t)$ sous (1.8) vérifient l'équation (1.6).

1.2 La structure en âge et en temps de la mortalité

En 1825, Gompertz propose que pour les populations humaines, la force de mortalité μ augmente de manière exponentielle avec l'âge a pour des âges suffisamment grands, de telle sorte qu'il existe deux constantes positives B et C telles que l'on ait :

$$\mu(a) = B \cdot C^a. \tag{1.12}$$

En 1860, Makeham propose d'ajouter une constante D à la loi de Gompertz afin de tenir compte des causes de décès indépendantes de l'âge. On parle alors du modèle de Gompertz-Makeham qui s'écrit :

$$\mu(a) = B \cdot C^a + D. \tag{1.13}$$

En démographie, la forme de cette fonction s'interprète de la manière suivante : le paramètre D représente un taux de décès accidentel et indépendant de l'âge, tandis que le terme $B \cdot C^a$ modélise un vieillissement exponentiel (si $C > 1$). Ce modèle a ainsi une ambition "explicative", en intégrant explicitement deux causes de décès clairement identifiées.

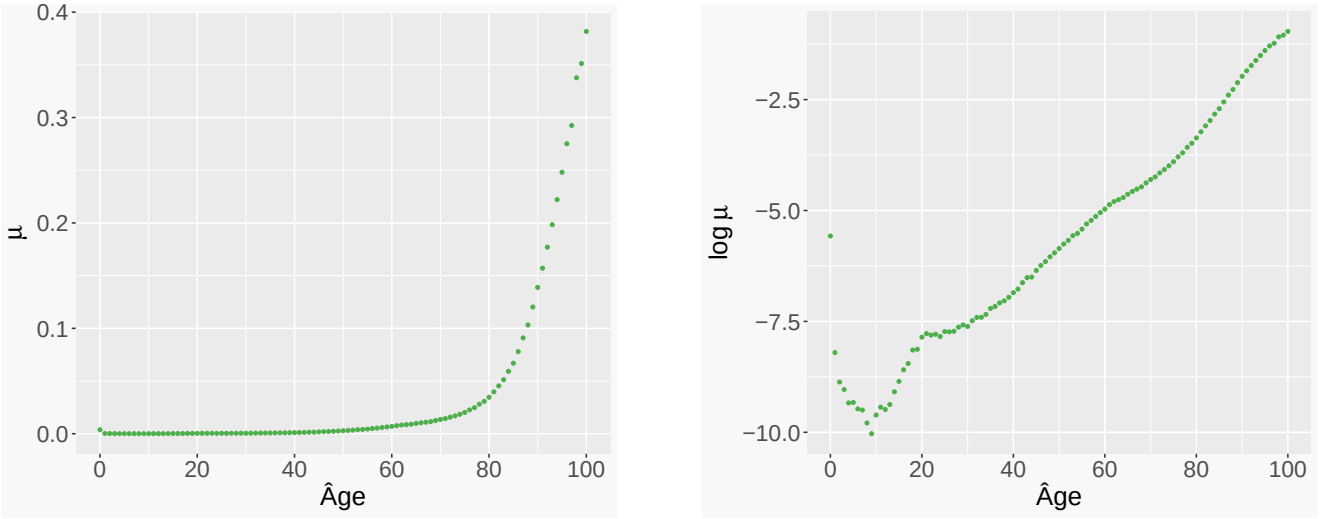
La figure 1.2 représente la force de mortalité μ de la population française et son logarithme $\log \mu$ en fonction de l'âge pour l'année 2017. On observe dans ce cas qu'aux âges élevés, le logarithme de la force de mortalité est bien une fonction linéaire de l'âge. Néanmoins, la force de mortalité présente une structure en fonction de l'âge bien particulière. On distingue en effet :

- la mortalité infantile : la force de mortalité des enfants de moins d'un an est significativement supérieure à celle des enfants âgés entre deux et dix ans,
- un creux de mortalité pour les âges au voisinage de dix ans,
- la célèbre "bosse des accidents", attribuée aux accidents de la route et aux suicides, et qui concerne surtout les individus âgés entre 15 et 25 ans,
- à partir de quarante ans, une croissance approximativement linéaire de la force de mortalité.

La figure 1.3 représente la force de mortalité μ de la population française et son logarithme $\log \mu$ en fonction de l'âge et du sexe pour l'année 2017. Elle permet d'observer que les femmes présentent un niveau de force de mortalité inférieur à celui des hommes, cet écart étant

surtout présent pour les âges supérieurs à 15 ans, tandis que la mortalité infantile semble du même ordre de grandeur entre les deux sexes. En outre, la structure en âge de la mortalité entre les deux sexes est différente. On remarque notamment que la "bosse des accidents", très marquée chez les hommes, l'est moins chez les femmes.

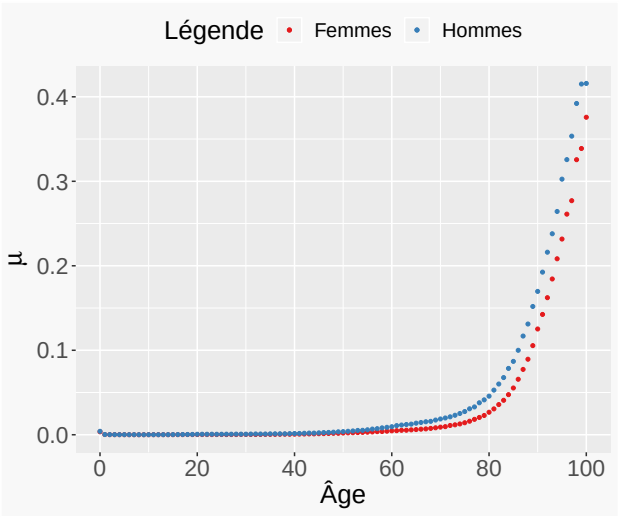
En plus d'une structure en âge, la force de mortalité possède également une structure en temps. Au cours du vingtième siècle, il est maintenant bien connu que la mortalité humaine a globalement diminué au cours du temps. Dans la plupart des pays industrialisés, la mortalité chez les adultes et les personnes âgées révèle en effet une probabilité de décès annuelle décroissante (Benjamin et Soliman, 1993; Macdonald *et al.*, 1998). La figure 1.4 représente le logarithme de la force de mortalité des hommes et des femmes en France en fonction de l'âge sur la période temporelle 1900-2017. Elle permet d'observer une diminution de la mortalité au cours du temps, en particulier de la mortalité infantile et de celle des âges élevés. Cette figure met par ailleurs en évidence l'impact de certains événements catastrophiques comme l'effet conjoint de la Première Guerre mondiale et de la pandémie de la grippe de 1918 ("grippe espagnole"), ainsi que celui de la Seconde Guerre mondiale. On constate que les niveaux de force de mortalité sont différents entre les hommes et les femmes, les événements catastrophiques ayant également un impact différent en fonction du sexe considéré.



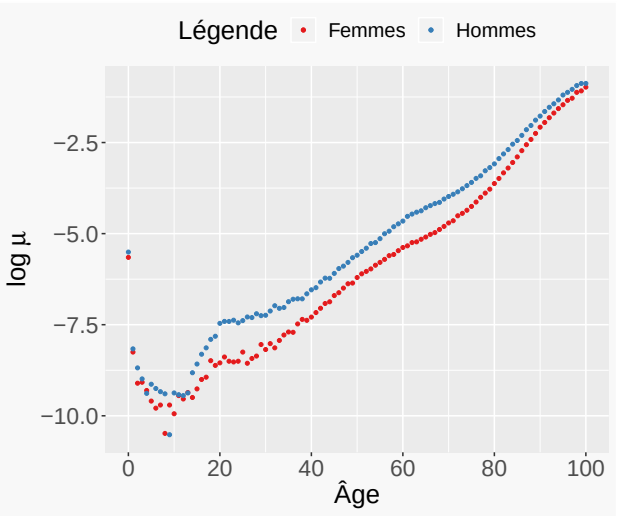
(a) Force de mortalité

(b) Logarithme de la force de mortalité

FIGURE 1.2. Force de mortalité de la population totale française (a) et son logarithme (b) en fonction de l'âge en 2017 — données provenant de la *Human Mortality Database*

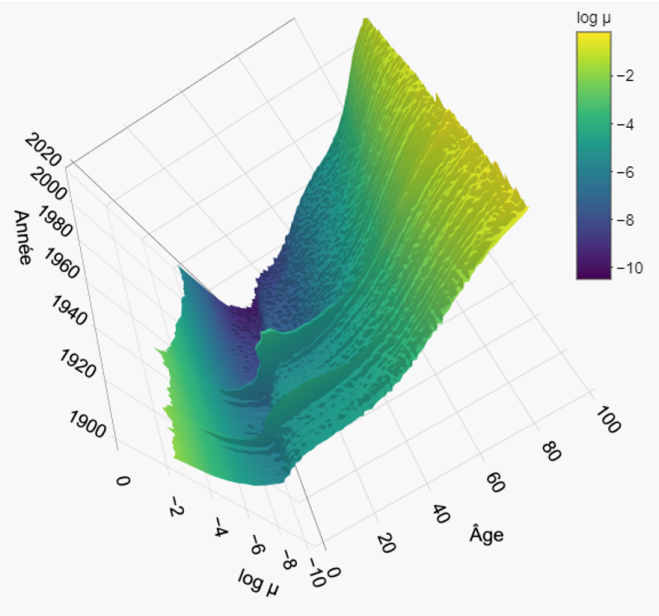


(a) Force de mortalité

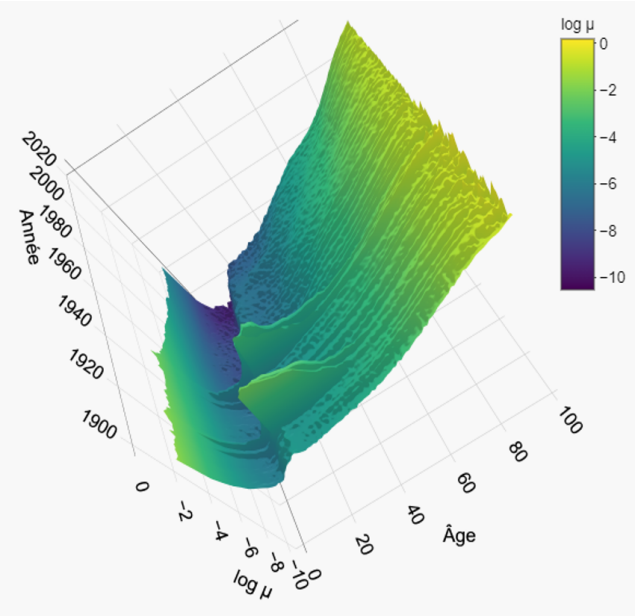


(b) Logarithme de la force de mortalité

FIGURE 1.3. Force de mortalité (a) et son logarithme (b) en fonction de l'âge et du sexe en 2017 — données de la population française provenant de la *Human Mortality Database*



(a) Données des femmes



(b) Données des hommes

FIGURE 1.4. Logarithme de la force de mortalité des femmes (a) et des hommes (b) en fonction de l'âge sur la période 1900-2017 — données de la population française provenant de la *Human Mortality Database*

1.3 La famille des modèles de mortalité GAPC

On considère à présent la famille GAPC (*Generalized Age-Period-Cohort*) des modèles stochastiques généralisés de mortalité âge-période-cohorte (Villegas *et al.*, 2015). Cette famille est composée de quatre éléments :

1. **La composante aléatoire :** la variable aléatoire $D(a,t)$ suit une distribution de Poisson ou une distribution binomiale, si bien que l'une des équations (1.7) ou (1.8) soit vérifiée.
2. **La composante systématique :** les effets de l'âge a, l'année civile t et l'année de naissance (cohorte) $c := t-a$ sont pris en compte par le prédicteur $\eta(a,t)$ donné par :

$$\eta(a,t) = \alpha_a + \sum_{i=1}^{p} \beta_a^{(i)} \kappa_t^{(i)} + \beta_a^{(0)} \gamma_{t-a}, \tag{1.14}$$

   où :
   - α_a désigne le paramètre spécifique à l'âge a,
   - $p \geq 0$ est un entier indiquant le nombre de termes âge-période décrivant les tendances de la mortalité, où chaque $\kappa_t^{(i)}$, $i \in [\![1,p]\!]$, décrit l'évolution générale de la mortalité et $\beta_a^{(i)}$ modifie l'effet de $\kappa_t^{(i)}$ à travers les âges,
   - $\beta_a^{(0)}\gamma_{t-a}$ représente l'effet cohorte, où γ_{t-a} est le paramètre de cohorte et $\beta_a^{(0)}$ modifie son effet à travers les âges.

   $\kappa_t^{(i)}$, $i \in [\![1,p]\!]$, et le paramètre de cohorte γ_{t-a} sont des processus stochastiques. Il s'agit de la caractéristique-clef qui permet la projection stochastique des modèles GAPC et ainsi la génération de prévisions probabilistes des taux de mortalité futurs.
3. **La fonction de lien** ℓ, qui est associée à la composante aléatoire et à la composante systématique selon la relation suivante :

$$\eta(a,t) = \ell\left(\mathbb{E}\left(\frac{D(a,t)}{E(a,t)}\right)\right). \tag{1.15}$$

   Bien qu'un certain nombre de fonctions de lien soient possibles, il est standard d'utiliser des fonctions de lien dites "canoniques". La distribution de Poisson est ainsi associée à la fonction de lien log et la distribution binomiale est associée à la fonction de lien logit [1] (voir Currie (2016) pour une discussion de ce choix dans le cadre des modèles de mortalité).
4. **L'ensemble de contraintes sur les paramètres :** la plupart des modèles stochastiques de mortalité ne sont identifiables qu'à une transformation près et nécessitent donc des contraintes sur les paramètres afin de garantir des estimations de paramètres uniques. Ces contraintes sur les paramètres sont appliquées via une fonction de contrainte v qui transforme un vecteur arbitraire de paramètres

$$\theta := (\alpha_a, \beta_a^{(1)}, \ldots, \beta_a^{(p)}, \kappa_t^{(1)}, \ldots, \kappa_t^{(p)}, \beta_a^{(0)}, \gamma_{t-a}) \tag{1.16}$$

   en un vecteur transformé

$$\tilde{\theta} := (\tilde{\alpha}_a, \tilde{\beta}_a^{(1)}, \ldots, \tilde{\beta}_a^{(p)}, \tilde{\kappa}_t^{(1)}, \ldots, \tilde{\kappa}_t^{(p)}, \tilde{\beta}_a^{(0)}, \tilde{\gamma}_{t-a}) \tag{1.17}$$

   qui satisfait les contraintes du modèle. Cette transformation n'a pas d'effet sur le prédicteur $\eta(a,t)$ (*i.e.* que θ et $\tilde{\theta}$ donnent le même $\eta(a,t)$).

1. $\forall x \in]0,1[, \text{logit}(x) = \log(\frac{x}{1-x})$

Dans la suite, cinq modèles de mortalité appartenant à la famille GAPC seront présentés de manière détaillée.

1.3.1 Le modèle de Lee-Carter

En 1992, Lee et Carter ont mis au point une approche puissante et élégante des prévisions de mortalité. Ces auteurs ont proposé un modèle remarquablement simple pour les projections de mortalité, spécifiant une forme log-bilinéaire pour le taux de mortalité. Au lieu de recourir comme dans leur approche originelle à une décomposition en valeurs singulières de la matrice des taux de mortalité, Brouhns *et al.* (2002) ont implémenté le modèle de Lee-Carter, également connu sous le nom de modèle M1, en supposant une distribution Poissonnienne du nombre de décès et en utilisant la fonction de lien logarithmique (log-link) appliquée à la force de mortalité $\mu(a,t)$. La structure prédictive proposée par Lee et Carter (1992) suppose qu'il existe un paramètre d'âge statique, α_a, un unique terme âge-période ($p=1$), et aucun effet cohorte. Ainsi, le prédicteur s'écrit :

$$\eta(a,t) = \alpha_a + \beta_a^{(1)} \cdot \kappa_t^{(1)}. \tag{1.18}$$

Afin de projeter la mortalité future, la série des $\kappa_t^{(1)}$ est modélisée par un processus ARIMA (pour *Auto-Regressive Integrated Moving Average*). En règle générale, une marche aléatoire avec dérive permet d'obtenir des résultats satisfaisants, c'est-à-dire que l'on écrit :

$$\kappa_t^{(1)} = \delta + \kappa_{t-1}^{(1)} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\kappa^2) \text{ i.i.d.}, \tag{1.19}$$

où δ est le paramètre de dérive (*drift*) et ε_t est un bruit blanc Gaussien de variance σ_κ^2.

Le modèle de Lee-Carter n'est identifiable qu'à une transformation près, car comme pour toutes constantes réelles arbitraires c_1 et $c_2 \neq 0$, les paramètres de l'équation (1.18) peuvent être transformés de la manière suivante :

$$(\alpha_a, \beta_a^{(1)}, \kappa_t^{(1)}) \to \left(\alpha_a + c_1\beta_a^{(1)}, \frac{1}{c_2}\beta_a^{(1)}, c_2(\kappa_t^{(1)} - c_1)\right), \tag{1.20}$$

aboutissant à un prédicteur $\eta(a,t)$ inchangé. Afin d'assurer l'identifiabilité du modèle, Lee et Carter (1992) suggèrent de considérer l'ensemble suivant de contraintes sur les paramètres :

$$\sum_a \beta_a^{(1)} = 1, \tag{1.21}$$

$$\sum_t \kappa_t^{(1)} = 0, \tag{1.22}$$

qui peut être imposé en choisissant de poser :

$$c_1 = \frac{1}{n}\sum_t \kappa_t^{(1)}, \tag{1.23}$$

$$c_2 = \sum_a \beta_a^{(1)} \tag{1.24}$$

dans la transformation (1.20).

1.3.2 Le modèle de Renshaw-Haberman

Renshaw et Haberman (2006) généralisent le modèle de Lee-Carter et incluent un effet cohorte. Ce modèle est également connu sous le nom de modèle M2 et s'écrit :

$$\eta(a,t) = \alpha_a + \beta_a^{(1)} \cdot \kappa_t^{(1)} + \beta_a^{(0)} \cdot \gamma_{t-a}, \tag{1.25}$$

où γ_c est un paramètre additionnel de sensibilité spécifique à la cohorte, permettant d'ajuster les taux de mortalité pour une cohorte née l'année civile $c = t - a$.

Les projections de mortalité pour ce modèle sont dérivées en modélisant les séries $\kappa_t^{(1)}$ et γ_{t-a} par des processus ARIMA, en faisant une hypothèse d'indépendance entre les deux effets période et cohorte.

Afin d'estimer le modèle, Renshaw et Haberman (2006) supposent une distribution Poissonnienne du nombre de décès et utilisent la fonction de lien log appliquée à la force de mortalité $\mu(a,t)$. Comme dans le modèle de Lee-Carter, le prédicteur $\eta(a,t)$ est invariant par la transformation suivante :

$$\begin{aligned}(\alpha_a, \beta_a^{(1)}, \kappa_t^{(1)}, \beta_a^{(0)}, \gamma_{t-a}) \rightarrow \Bigg(& \alpha_a + c_1\beta_a^{(1)} + c_2\beta_a^{(0)}, \frac{1}{c_3}\beta_a^{(1)}, \\ & c_3(\kappa_t^{(1)} - c_1), \frac{1}{c_4}\beta_a^{(0)}, c_4(\gamma_{t-a} - c_2)\Bigg)\end{aligned} \tag{1.26}$$

où $c_1, c_2, c_3 \neq 0$, et $c_4 \neq 0$ sont des constantes réelles. L'identifiabilité du modèle peut être assurée en considérant l'ensemble suivant de contraintes sur les paramètres :

$$\sum_a \beta_a^{(1)} = 1, \tag{1.27}$$

$$\sum_t \kappa_t^{(1)} = 0, \tag{1.28}$$

$$\sum_a \beta_a^{(0)} = 1, \tag{1.29}$$

$$\sum_c \gamma_c = 0, \tag{1.30}$$

qui peuvent être imposées en choisissant de poser :

$$c_1 = \frac{1}{n}\sum_t \kappa_t^{(1)}, \tag{1.31}$$

$$c_2 = \frac{1}{n+k-1}\sum_c \gamma_c\ , \tag{1.32}$$

$$c_3 = \sum_a \beta_a^{(1)}, \tag{1.33}$$

$$c_4 = \sum_a \beta_a^{(0)}, \tag{1.34}$$

dans la transformation (1.26).

1.3.3 Le modèle Age-Period-Cohort

Comme le modèle M2 est connu pour être difficile à calibrer, Currie (2006) propose d'introduire une sous-structure couramment utilisée du modèle de Renshaw-Haberman, en posant simplement $\beta_a^{(1)} = 1$ et $\beta_a^{(0)} = 1$. Le modèle en résultant est nommé modèle Age-Period-Cohort (APC) ou modèle M3. Il s'écrit sous la forme :

$$\eta(a,t) = \alpha_a + \kappa_t^{(1)} + \gamma_{t-a}. \tag{1.35}$$

Currie (2006) utilise une approche par splines pénalisées à α_a, $\kappa_t^{(1)}$ et γ_{t-a} afin d'assurer leur lissage. Dans notre analyse du modèle APC, nous n'imposons aucune condition de lissage. Le modèle APC est connu pour être invariant par rapport aux deux transformations suivantes :

$$\begin{aligned}
(\alpha_a, \kappa_t^{(1)}, \gamma_{t-a}) &\to \left(\alpha_a + \phi_1 - \phi_2 a, \kappa_t^{(1)} + \phi_2 t, \gamma_{t-a} - \phi_1 - \phi_2(t-a)\right), \\
(\alpha_a, \kappa_t^{(1)}, \gamma_{t-a}) &\to \left(\alpha_a + c_1, \kappa_t^{(1)} - c_1, \gamma_{t-a}\right),
\end{aligned} \tag{1.36}$$

où c_1, ϕ_1 et ϕ_2 désignent des constantes réelles. Néanmoins, l'identifiabilité du modèle peut être assurée en considérant l'ensemble suivant de contraintes sur les paramètres :

$$\sum_t \kappa_t^{(1)} = 0, \tag{1.37}$$

$$\sum_c \gamma_c = 0, \tag{1.38}$$

$$\sum_c c\gamma_c = 0, \tag{1.39}$$

les deux dernières contraintes impliquant que le paramètre de cohorte oscille autour de zéro sans tendance linéaire perceptible. Suite à Haberman et Renshaw (2011), les contraintes sur l'effet cohorte peuvent être imposées en appliquant la transformation (1.36), les constantes ϕ_1 et ϕ_2 étant obtenues en régressant γ_{t-a} sur $t-a$, de sorte que :

$$\gamma_{t-a} = \phi_1 + \phi_2(t-a) + \varepsilon_{t-a}, \quad \varepsilon_{t-a} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.} \tag{1.40}$$

La contrainte sur le paramètre de période peut alors être imposée en appliquant la transformation (1.36) avec :

$$c_1 = \frac{1}{n}\sum_t \kappa_t^{(1)}. \tag{1.41}$$

1.3.4 Le modèle de Cairns-Blake-Dowd

Cairns *et al.* (2006) ont proposé une structure à deux termes âge-période ($p = 2$), avec des paramètres de modulation de l'âge prédéfinis aux valeurs $\beta_a^{(1)} = 1$ et $\beta_a^{(2)} = a - \bar{a}$, sans paramètre spécifique à l'âge ($\alpha_a = 0$), et sans paramètre de cohorte ($\gamma_{t-a} = 0$). Le modèle en résultant est nommé modèle de Cairns-Blake-Dowd (CBD) ou modèle M5. Il s'écrit sous la forme :

$$\eta(a,t) = \kappa_t^{(1)} + (a - \bar{a})\kappa_t^{(2)}, \tag{1.42}$$

où $\bar{a}$ est l'âge moyen de la plage d'âges considérée. Cairns *et al.* (2006) obtiennent des projections de mortalité en projetant les séries $\kappa_t^{(1)}$ et $\kappa_t^{(2)}$ en les modélisant par des marches aléatoires bivariées avec dérive.

Le modèle CBD n'a pas de problème d'identification, l'ensemble de contraintes sur les paramètres est donc réduit à l'ensemble vide. Afin d'estimer les paramètres du modèle CBD, Haberman et Renshaw (2011) supposent une distribution binomiale du nombre de décès et utilisent la fonction de lien logit appliquée au taux de mortalité $q(a,t)$.

1.3.5 Le modèle de Plat

Plat (2009) a proposé de combiner le modèle CBD avec le modèle de Lee-Carter afin de créer un modèle qui capte l'effet cohorte et qui est adapté à la plage d'âges complète. La structure proposée comprend un paramètre spécifique à l'âge α_a, trois termes âge-période ($p = 3$) avec des paramètres de modulation de l'âge prédéfinis aux valeurs $\beta_a^{(1)} = 1$, $\beta_a^{(2)} = \bar{a} - a$, $\beta_a^{(3)} = (\bar{a} - a)^+ = \max(0, \bar{a} - a)$, et un paramètre de cohorte γ_{t-a}. Le modèle en résultant est nommé modèle de Plat et s'écrit sous la forme :

$$\eta(a,t) = \alpha_a + \kappa_t^{(1)} + (\bar{a} - a)\kappa_t^{(2)} + (\bar{a} - a)^+\kappa_t^{(3)} + \gamma_{t-a}. \tag{1.43}$$

Afin d'estimer le modèle, Plat (2009) suppose une distribution Poissonnienne du nombre de décès et utilisent la fonction de lien log appliquée à la force de mortalité $\mu(a,t)$. Le prédicteur $\eta(a,t)$ est invariant par les transformations suivantes :

$$\begin{aligned}(\alpha_a, \kappa_t^{(1)}, \kappa_t^{(2)}, \kappa_t^{(3)}, \gamma_{t-a}) \to \Big(&\alpha_a + \phi_1 - \phi_2 a + \phi_3 a^2, \kappa_t^{(1)} + \phi_2 t + \phi_3(t^2 - 2\bar{a}t), \\ &\kappa_t^{(2)} + 2\phi_3 t, \kappa_t^{(3)}, \gamma_{t-a} - \phi_1 - \phi_2(t-a) - \phi_3(t-a)^2\Big),\end{aligned} \tag{1.44}$$

$$\begin{aligned}(\alpha_a, \kappa_t^{(1)}, \kappa_t^{(2)}, \kappa_t^{(3)}, \gamma_{t-a}) \to \Big(&\alpha_a + c_1 + c_2(\bar{a} - a) + c_3(\bar{a} - a)^+, \\ &\kappa_t^{(1)} - c_1, \kappa_t^{(2)} - c_2, \kappa_t^{(3)} - c_3, \gamma_{t-a}\Big),\end{aligned} \tag{1.45}$$

où $c_1, c_2, c_3, \phi_1, \phi_2$, et ϕ_3 désignent des constantes réelles. L'identifiabilité du modèle peut

être assurée en considérant l'ensemble suivant de contraintes sur les paramètres :

$$\sum_t \kappa_t^{(1)} = 0, \tag{1.46}$$

$$\sum_t \kappa_t^{(2)} = 0, \tag{1.47}$$

$$\sum_t \kappa_t^{(3)} = 0, \tag{1.48}$$

$$\sum_c \gamma_c = 0, \tag{1.49}$$

$$\sum_c c\gamma_c = 0, \tag{1.50}$$

$$\sum_c c^2\gamma_c = 0. \tag{1.51}$$

Les trois premières contraintes garantissent que les paramètres de période sont centrés autour de zéro, tandis que les trois dernières contraintes garantissent que le paramètre de cohorte oscille autour de zéro sans tendance linéaire ou quadratique perceptible. Suite à Haberman et Renshaw (2011), les contraintes sur le paramètre de cohorte peuvent être imposées en appliquant la transformation (1.44), les constantes ϕ_1, ϕ_2, et ϕ_3 étant obtenues en régressant γ_{t-a} sur $t-a$ et $(t-a)^2$, de telle sorte que :

$$\gamma_{t-a} = \phi_1 + \phi_2(t-a) + \phi_3(t-a)^2 + \varepsilon_{t-a}, \quad \varepsilon_{t-a} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.} \tag{1.52}$$

Les contraintes sur les paramètres de période peuvent alors être imposées en appliquant la transformation (1.45) avec :

$$c_i = \frac{1}{n}\sum_t \kappa_t^{(i)}, i \in \{1, 2, 3\}. \tag{1.53}$$

Le modèle de Plat simplifié Dans les cas où seuls les âges élevés sont d'intérêt, Plat (2009) suggère d'omettre le quatrième terme de l'équation (1.43) qui devient :

$$\eta(a,t) = \alpha_a + \kappa_t^{(1)} + (\bar{a} - a)\kappa_t^{(2)} + \gamma_{t-a}. \tag{1.54}$$

On note que ce modèle de Plat simplifié a les mêmes problèmes d'identification que le modèle de Plat complet, à l'exception des transformations et contraintes impliquant les paramètres $\kappa_t^{(3)}$ et c_3.

1.4 Calibration des modèles de mortalité GAPC

L'estimation des paramètres d'un modèle stochastique de mortalité GAPC peut être obtenue en maximisant la log-vraisemblance $\log \mathscr{L}$ du modèle, qui est donnée par

$$\log \mathscr{L}(d(a,t), \hat{d}(a,t)) = \sum_a \sum_t \omega(a,t) \cdot \Big(d(a,t) \log \hat{d}(a,t) - \hat{d}(a,t) - \log d(a,t)! \Big) \tag{1.55}$$

dans le cas d'une distribution Poissonnienne du nombre de décès, et par

$$\begin{aligned} \log \mathscr{L}(d(a,t), \hat{d}(a,t)) = \sum_a \sum_t \omega(a,t) \cdot \Bigg(& d(a,t) \log \left(\frac{\hat{d}(a,t)}{E^0(a,t)} \right) \\ & + (E^0(a,t) - d(a,t)) \log \left(\frac{E^0(a,t) - \hat{d}(a,t)}{E^0(a,t)} \right) + \log \binom{E^0(a,t)}{d(a,t)} \Bigg) \end{aligned} \tag{1.56}$$

dans le cas d'une distribution binomiale du nombre de décès. On note que l'équation (1.56) fait intervenir le coefficient binomial $\binom{\mathscr{V}}{\mathscr{W}}$, défini par :

$$\binom{\mathscr{V}}{\mathscr{W}} = C_{\mathscr{V}}^{\mathscr{W}} = \frac{\mathscr{V}!}{\mathscr{W}!(\mathscr{V} - \mathscr{W})!}. \tag{1.57}$$

Dans les deux cas, les quantités $\omega(a,t)$ sont des poids qui prennent la valeur de 0 si un point de données (a,t) est omis ou 1 si ce point de données est inclus. En outre, la grandeur

$$\hat{d}(a,t) = E(a,t)\ \ell^{-1} \left(\alpha_a + \sum_{i=1}^{p} \beta_a^{(i)} \kappa_t^{(i)} + \beta_a^{(0)} \gamma_{t-a} \right) \tag{1.58}$$

est le nombre moyen de décès estimés par le modèle, où ℓ^{-1} désigne la fonction réciproque de la fonction de lien ℓ.

Dans la littérature sur la mortalité, la maximisation de la log-vraisemblance est généralement utilisée en utilisant par exemple l'algorithme de Newton-Raphson (voir *e.g.* Renshaw et Haberman (2006) et Cairns *et al.* (2009)). Néanmoins, comme discuté en détail par Currie (2016), de nombreux modèles stochastiques de mortalité sont des exemples de modèles linéaires généralisés (GLM) ou de modèles non linéaires généralisés (GNM), ce qui permet de simplifier leur calibration par l'utilisation de librairies standard sur R. Currie (2016) illustre ce fait en calibrant plusieurs modèles stochastiques de mortalité dans R en utilisant la fonction **glm** ou la fonction **gnm** de la librairie R `gnm` (Turner et Firth, 2007). Dans ce mémoire, la librairie R `StMoMo` sera utilisée. Elle comprend la fonction **fit**, qui repose sur la fonction **gnm** de la librairie `gnm` afin d'estimer les paramètres des modèles stochastiques de mortalité GAPC.

CHAPITRE 2

APPROCHES PAR APPRENTISSAGE AUTOMATIQUE

2.1 L'apprentissage automatique

2.1.1 Introduction

Ces dernières années, les techniques d'apprentissage automatique (*Machine Learning*) ont pris une place de plus en plus importante dans de nombreux domaines de la recherche, de l'informatique à la médecine en passant par la science actuarielle. L'un des pionniers dans ce domaine est Vapnik, notamment grâce à ses apports fournis dans l'ouvrage Vapnik (2013).

L'objectif général de l'apprentissage automatique est d'élaborer des procédures automatiques qui permettent de mettre en évidence des règles générales à partir d'exemples. Il s'agit donc d'imiter le fonctionnement inductif du cerveau humain dans le but de développer des systèmes d'intelligence artificielle. Ceci est considéré aujourd'hui comme une branche de l'informatique. Cependant, les liens avec la statistique sont très étroits, notamment avec la théorie non paramétrique. Pour schématiser, la différence principale entre la statistique et l'apprentissage automatique est que le concept central en statistique est le modèle, tandis qu'en apprentissage automatique, c'est l'algorithme. Bien que l'apprentissage automatique ne puisse pas tout expliquer, il demeure très utile pour détecter des régularités (*patterns*), même inconnues et non identifiables, ainsi que des corrélations cachées. Ainsi, ces techniques permettent de mieux comprendre des processus, de prédire l'avenir à partir de données historiques ou de classer automatiquement des ensembles de données (Dalalyan, 2018).

Tout comme en statistique, le point de départ en apprentissage automatique est l'échantillon $(Z_1, \ldots, Z_n)$, que l'on veut utiliser, par exemple, pour faire des prévisions. Il existe différentes branches de l'apprentissage automatique caractérisées par la nature de l'échantillon et l'objectif poursuivi.
Dans les méthodes d'apprentissage supervisé, les observations $Z_i = (X_i, Y_i)$ sont composées d'une variable d'entrée $X_i \in \mathbb{R}^d$, souvent appelée prédicteur ou *feature*, et d'une variable de sortie Y_i appelée étiquette ou *label*, appartenant soit à un ensemble fini, soit à $\mathbb{R}$. L'objectif est de prévoir, pour un $x \in \mathbb{R}^d$ donné (choisi de façon déterministe ou aléatoire), la valeur de l'étiquette correspondante.
Dans la méthode non supervisée, les observations Z_i ne sont pas étiquetées. Le plus souvent, $Z_i \in \mathbb{R}^d$ pour un $d \in \mathbb{N}$ assez grand, et le but est de caractériser la loi de probabilité ayant engendré ces observations. Le clustering ou encore l'estimation de la densité sont les

problèmes les plus étudiés en apprentissage non supervisé.
Dans la méthode semi-supervisée, seule une faible proportion n_1 des observations est étiquetée. On a donc $Z_i = (X_i, Y_i)$ pour $i \in [\![1, n_1]\!]$ et $Z_i = X_i$ pour $i \in [\![n_1 + 1, n]\!]$. Le but est le même qu'en apprentissage supervisé. Ce cadre est intéressant pour de nombreuses applications où le coût d'étiquetage est très élevé.

2.1.2 Apprentissage supervisé : régression et classification binaire

Supposons que l'on observe une base de données composée de n couples $Z_i = (X_i, Y_i)$ que nous supposons être des réalisations indépendantes d'une même loi P inconnue. On écrira :

$$Z_i = (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P. \tag{2.1}$$

Les X_i appartiennent à un espace $\mathcal{X}$ et s'appellent les entrées ou les *features*. Typiquement, $\mathcal{X} = \mathbb{R}^d$ pour un grand entier d. Les Y_i appartiennent à un espace $\mathcal{Y}$, et s'appellent les sorties ou les étiquettes. Typiquement, $\mathcal{Y}$ est fini ou $\mathcal{Y}$ est un sous-ensemble de $\mathbb{R}$.

L'objectif de l'apprentissage supervisé est de prévoir l'étiquette Y associée à toute nouvelle entrée X, où il est sous-entendu que la paire (X, Y) est une nouvelle réalisation de la loi P, cette réalisation étant indépendante des réalisations précédemment observées.

Une fonction de prédiction est une fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$. Dans ce qui suit, nous supposons que toutes les quantités que nous manipulons sont mesurables. L'ensemble de toutes les fonctions de prédiction est noté $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. La base de données $(Z_1, \dots, Z_n)$ est appelée ensemble d'apprentissage.

Un algorithme d'apprentissage est une fonction qui à tout ensemble d'apprentissage renvoie une fonction de prédiction, c'est-à-dire une fonction de la réunion :

$$\bigcup_{k=1}^{+\infty} \mathcal{Z}^k$$

dans l'ensemble $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, où $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. C'est un estimateur de la "meilleure" fonction de prédiction, où le terme "meilleure" sera précisé ultérieurement.

Soit $\rho(y, y')$ la perte encourue lorsque la sortie réelle est y et la sortie prédite est y'. La fonction $\rho : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ est appelée fonction de perte. Les deux exemples les plus fréquemment utilisés sont :

- La classification : $\rho(y, y') = \mathbb{1}_{\{y \neq y'\}}$, *i.e.* $\rho(y, y') = 1$ si $y \neq y'$ et $\rho(y, y') = 0$ sinon. Un problème d'apprentissage pour lequel cette fonction de perte est utilisée est appelé problème de classification. L'ensemble $\mathcal{Y}$ considéré en classification est le plus souvent fini, voire même de cardinal deux en classification binaire.
- La régression L_p : $\mathcal{Y} = \mathbb{R}$ et $\rho(y, y') = |y - y'|^p$ où $p \geq 1$ est un réel fixé. Dans ce cas, on parle de régression L_p. La tâche d'apprentissage lorsque $p = 2$ est aussi appelée régression aux moindres carrés.

La qualité d'une fonction de prédiction $g : \mathcal{X} \to \mathcal{Y}$ est mesurée par son risque ou erreur de généralisation :

$$\mathscr{R}_P(g) = \mathbb{E}_P[\rho(Y, g(X))]. \tag{2.2}$$

Le risque est donc l'espérance par rapport à loi P de la perte encourue sur la donnée (X, Y) par la fonction de prédiction g. La "meilleure" fonction de prédiction est une fonction de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ minimisant le risque $\mathscr{R}_P$:

$$g_P^* \in \operatorname*{arg\,min}_{g \in \mathcal{F}(\mathcal{X}, \mathcal{Y})} \mathscr{R}_P(g). \tag{2.3}$$

Une telle fonction g_P^* n'existe pas nécessairement mais existe pour les fonctions de pertes usuelles. Cette "meilleure" fonction est appelée fonction oracle ou prédicteur de Bayes. Elle dépend de la probabilité inconnue P et, par conséquent, est inconnue.

2.1.3 Méthode de minimisation du risque empirique

Le but d'un algorithme d'apprentissage est de trouver une fonction de prédiction dont le risque est aussi faible que possible (autrement dit aussi proche que possible du risque des prédicteurs oracles).

La distribution P générant les données étant inconnue, le risque $\mathscr{R}_P$ et les prédicteurs oracles sont inconnus. Néanmoins, le risque $\mathscr{R}_P(g)$ peut être estimé par son équivalent empirique :

$$\widehat{\mathscr{R}}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, g(X_i)). \tag{2.4}$$

Si l'on suppose que $\mathbb{E}_P[\rho(Y, g(X))^2] < +\infty$, alors d'après la loi forte des grands nombres et le théorème central limite, il en résulte que :

$$\widehat{\mathscr{R}}_n(g) \xrightarrow[n\to+\infty]{p.s.} \mathscr{R}_P(g), \tag{2.5}$$

$$\sqrt{n}\cdot(\widehat{\mathscr{R}}_n(g) - \mathscr{R}_P(g)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mathbb{V}(\rho(Y, g(X)))). \tag{2.6}$$

Pour toute fonction de prédiction g, la variable aléatoire $\widehat{\mathscr{R}}_n(g)$ effectue donc des déviations en $\mathcal{O}(\frac{1}{\sqrt{n}})$ autour de sa moyenne $\mathscr{R}_P(g)$. Puisque nous cherchons une fonction qui minimise le risque $\mathscr{R}_P$ et puisque ce risque est approché par le risque empirique $\widehat{\mathscr{R}}_n$, il est naturel de considérer l'algorithme d'apprentissage, dit de minimisation du risque empirique, défini par :

$$\widehat{g}_{n,\mathcal{G}} = \arg\min_{g\in\mathcal{G}} \widehat{\mathscr{R}}_n(g), \tag{2.7}$$

où $\mathcal{G}$ est un sous-ensemble de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$. Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ n'est pas une bonne idée. Tout d'abord, cela entraîne un problème de choix puisqu'en général, pour tout ensemble d'apprentissage, il existe une infinité de fonctions de prédiction minimisant le risque empirique. D'autre part, si l'on prend l'algorithme du plus proche voisin comme minimiseur du risque empirique, alors on peut montrer que cet algorithme est loin d'être universellement consistant (Dalalyan, 2018).

Prendre $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$ mène en général à un sur-apprentissage (*overfitting*) dans la mesure où l'algorithme qui en résulte a un risque empirique qui peut être très inférieur à son risque réel. En pratique, il faut prendre $\mathcal{G}$ suffisamment grand pour pouvoir raisonnablement approcher toute fonction, tout en ne le prenant pas trop grand afin d'éviter que l'algorithme sur-apprenne. La "grandeur" de l'ensemble $\mathcal{G}$ est appelée capacité ou complexité. Un autre point de vue consiste à rajouter à $\widehat{\mathscr{R}}_n(g)$ une pénalisation, quand par exemple, la fonction g est trop irrégulière. Ces deux approches sont en fait proches l'une de l'autre.

Soit g une fonction minimisant le risque sur $\mathcal{G}$:

$$g_{P,\mathcal{G}}^* = \arg\min_{g\in\mathcal{G}} \mathscr{R}_P(g). \tag{2.8}$$

On suppose le minimum atteint pour simplifier l'explication. La fonction $g_{P,\mathcal{G}}^*$ est appelée prédicteur oracle sur $\mathcal{G}$. D'après l'inégalité :

$$\mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) \geq \mathscr{R}_P(g_{P,\mathcal{G}}^*) \geq \mathscr{R}_P(g_P^*), \tag{2.9}$$

l'excès de risque de $\widehat{g}_{n,\mathcal{G}}$ se décompose en deux termes positifs, appelés erreur stochastique (ou erreur d'estimation) et erreur systématique (ou erreur d'approximation ou biais) :

$$\mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) - \mathscr{R}_P(g_P^*) = \underbrace{\mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) - \mathscr{R}_P(g_{P,\mathcal{G}}^*)}_{\text{erreur stochastique}} + \underbrace{\mathscr{R}_P(g_{P,\mathcal{G}}^*) - \mathscr{R}_P(g_P^*)}_{\text{erreur systématique}} \tag{2.10}$$

Plus $\mathcal{G}$ est grand, plus l'erreur d'approximation est faible mais plus l'erreur d'estimation est en général grande. Il y a donc un compromis à trouver dans le choix de $\mathcal{G}$. Ce compromis est souvent appelé "dilemme biais-variance", où le terme variance provient du lien entre l'erreur stochastique et la variabilité de l'ensemble d'apprentissage que nous avons supposé dans notre formalisme être une réalisation de variables aléatoires indépendantes et identiquement distribuées (i.i.d.).

L'erreur stochastique peut être bornée par deux fois le supremum du processus empirique : $g \mapsto |\mathscr{R}_P(g) - \widehat{\mathscr{R}}_n(g)|$ défini sur $\mathcal{G}$. En effet, nous avons :

$$\begin{aligned}
\mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) - \mathscr{R}_P(g_{P,\mathcal{G}}^*) &= \mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) - \widehat{\mathscr{R}}_n(\widehat{g}_{n,\mathcal{G}}) + \underbrace{\widehat{\mathscr{R}}_n(\widehat{g}_{n,\mathcal{G}}) - \widehat{\mathscr{R}}_n(g_{P,\mathcal{G}}^*)}_{\leq 0} + \widehat{\mathscr{R}}_n(g_{P,\mathcal{G}}^*) - \mathscr{R}_P(g_{P,\mathcal{G}}^*) \\
&\leq \sup_{g\in\mathcal{G}} |\mathscr{R}_P(g) - \widehat{\mathscr{R}}_n(g)| + \sup_{g\in\mathcal{G}} |\mathscr{R}_P(g) - \widehat{\mathscr{R}}_n(g)| \\
&\leq 2 \cdot \sup_{g\in\mathcal{G}} |\mathscr{R}_P(g) - \widehat{\mathscr{R}}_n(g)|.
\end{aligned}$$

Des bornes plus élaborées sont données par le résultat suivant, qui découle de l'inégalité de Hoeffding (1963).

Théorème Supposons qu'il existe deux nombres réels $a < b$ tels que la fonction de perte ρ vérifie $a \leq \rho(y, y') \leq b$ pour tout $y, y' \in \mathcal{Y}$. Alors pour tout $\varepsilon > 0$, avec probabilité au moins $1 - \varepsilon$:

$$\mathscr{R}_P(\widehat{g}_{n,\mathcal{G}}) - \mathscr{R}_P(g_{P,\mathcal{G}}^*) \leq (b-a)\sqrt{\frac{2\log(2\ \mathrm{Card}(\mathcal{G})\ \varepsilon^{-1})}{n}} \tag{2.11}$$

Ainsi, lorsque l'ensemble $\mathcal{G}$ de fonctions considérées est fini, le cardinal de $\mathcal{G}$ est une première mesure de complexité (ou capacité) de $\mathcal{G}$.

2.1.4 Applications de l'apprentissage automatique en modélisation de la mortalité

Dans la littérature sur la modélisation de la mortalité, les approches par apprentissage automatique sont très peu présentes. Les travaux de Deprez *et al.* (2017) ont montré que les algorithmes d'apprentissage automatique sont utiles pour évaluer la qualité de l'ajustement des estimations de mortalité fournies par les modèles de mortalité stochastiques standard (ils ont considéré les modèles de Lee-Carter et de Renshaw-Haberman). Les auteurs ont appliqué un modèle de *gradient boosting* pour analyser comment la modélisation devrait être améliorée en fonction des caractéristiques d'un individu, tels que son âge ou sa cohorte de naissance. Cette approche par régression (non paramétrique) permet ensuite de détecter les faiblesses de différents modèles stochastiques de mortalité. Les travaux menés par Hainaut (2018) ont utilisé des réseaux de neurones pour trouver les facteurs de mortalité latents d'un modèle de Lee-Carter et les prévoir en fonction d'une marche aléatoire avec dérive. Richman et Wüthrich (2019) ont étendu le modèle de Lee-Carter à plusieurs populations à l'aide de réseaux de neurones profonds *feedforward*. Ludkovski *et al.* (2018) considèrent des processus Gaussiens dans un cadre uni-population sur les données de mortalité des États-Unis, du Japon, et de l'Angleterre et du pays de Galles. Les auteurs considèrent les individus d'âges 50-84, leur objectif principal étant d'obtenir des projections court-terme de mortalité. Enfin, les travaux de Levantesi et Pizzorusso (2019) étudient la capacité de l'apprentissage automatique à améliorer la précision de certains modèles stochastiques de mortalité standard, tant dans l'estimation que dans la prévision des taux de mortalité. Les auteurs utilisent des techniques d'apprentissage automatique basées sur des arbres pour calibrer un paramètre (l'estimateur d'apprentissage automatique) à appliquer aux taux de mortalité ajustés par le modèle de mortalité standard.

Pour d'autres applications de l'apprentissage automatique en actuariat vie et non-vie, le lecteur pourra se référer à Richman (2018).

2.2 Les réseaux de neurones

2.2.1 Fondements mathématiques des réseaux de neurones

Le terme de réseau de neurones désigne un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones du cerveau humain (McCulloch et Pitts, 1943; Wiener, 1948). Son architecture comprend notamment des neurones, des connexions synaptiques qui relient les neurones entre eux, et des algorithmes d'apprentissage. Un réseau de neurones est composé de trois types de couches, appelées respectivement couches d'entrée, couches cachées et couches de sortie, chacune d'entre elles comprenant un ou plusieurs neurones. Chaque unité d'un réseau obtient des informations "pondérées", via les liaisons synaptiques des autres unités connectées et renvoie une sortie à l'aide d'une fonction d'activation transformant la somme pondérée des signaux d'entrée.

Les réseaux de neurones visent à approximer des fonctions complexes et non linéaires définies dans un espace à dimension finie, en se reposant sur la composition de plusieurs couches de fonctions simples. La base des réseaux de neurones provient du théorème de représentation de Kolmogorov-Arnold, solution au treizième problème de Hilbert. (Kolmogorov, 1956; Arnold, 1957). Ces auteurs ont montré que toute fonction continue de n variables a une représentation exacte en termes de superpositions finies et de compositions d'un nombre

finie de fonctions continues d'une variable, *i.e.* :

$$\begin{aligned}\forall f \in \mathcal{C}(\mathbb{R}^n, \mathbb{R}), \exists(\Phi_q) \in (\mathcal{C}(\mathbb{R}, \mathbb{R}))^{2n+1}, \exists(\varphi_{q,p}) \in (\mathcal{C}(\mathbb{R}, \mathbb{R}))^{n(2n+1)} / \\ \forall (x_1, \ldots, x_n) \in \mathbb{R}^n, f(x_1, \ldots, x_n) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^{n} \varphi_{q,p}(x_p) \right),\end{aligned} \tag{2.12}$$

où $\mathcal{C}(\mathscr{A}, \mathscr{B})$ désigne l'ensemble des fonctions continues de $\mathscr{A}$ dans $\mathscr{B}$. Dans la suite, un réseau de neurones *feedforward* (par opposition aux réseaux de neurones récurrents) est considéré — aussi appelé perceptron multi-couches. Le terme *feedforward* signifie que l'entrée dans un neurone ne peut jamais dépendre de la sortie de ce neurone. Les flux de données sont à sens unique, des entrées aux sorties, et la fonction calculée par le réseau est une composition des fonctions calculées par les neurones individuels. Mathématiquement, un tel réseau de neurones peut être modélisé par une fonction non linéaire :

$$x \in \mathcal{X} \subset \mathbb{R}^r \mapsto \Xi(x, \Theta) \in \mathbb{R},$$

où Ξ s'écrit comme une composition finie de fonctions. Soit $L \geq 2$ le nombre de couches du réseau de neurones et soit $\sigma : \mathbb{R} \to \mathbb{R}$ une fonction non constante, bornée et continue, appelée fonction d'activation. On a :

$$\Xi = Z_L \circ \sigma \circ Z_{L-1} \circ \cdots \circ \sigma \circ Z_1,$$

où

$$\forall l \in \{1, \ldots, L\}, \forall x \in \mathbb{R}^{d_{l-1}}, Z_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}, x \mapsto W_l x + \beta_l.$$

La fonction σ est appliquée composante par composante, β_l est le biais relatif à la couche l, d_l est le nombre de lignes de la matrice de poids W_l et correspond au nombre de neurones de la couche l. La grandeur Θ intègre les paramètres de toutes les couches et se définit par $\Theta = (W_l, \beta_l)_{l=1,\ldots,L} \in \mathbb{R}^{N_d}$ avec $N_d = \sum_{l=1}^{L} d_l(1 + d_{l-1})$.

Ces paramètres à optimiser sont l'ensemble des poids et des biais qui ont été utilisés dans les différentes couches du réseau de neurones. Plus le nombre de neurones par couche est élevé, plus le nombre de paramètres à optimiser est important.

Soient I_m l'hypercube unité $[0, 1]^m$ de dimension m et $\mathcal{C}(I_m, \mathbb{R})$ l'espace des fonctions continues sur I_m à valeurs réelles. Soit de plus $\varsigma : \mathbb{R} \to \mathbb{R}$ une fonction sigmoïdale, *i.e.* telle que ς est une fonction continue vérifiant les conditions :

$$\begin{cases} \lim\limits_{x \to -\infty} \varsigma(x) = 0 \\ \lim\limits_{x \to +\infty} \varsigma(x) = 1 \end{cases} \tag{2.13}$$

Alors Cybenko (1989) a montré que :

$$\begin{aligned}\forall \varepsilon > 0, \forall f \in \mathcal{C}(I_m, \mathbb{R}), \exists N \in \mathbb{N}, \exists(v_i), (b_i) \in \mathbb{R}^N, \exists(w_i) \in (\mathbb{R}^m)^N / \\ \forall x \in I_m, \begin{cases} F(x) = \sum\limits_{i=1}^{N} v_i \cdot \varsigma(w_i^T x + b_i) \\ |F(x) - f(x)| < \varepsilon \end{cases}\end{aligned} \tag{2.14}$$

Autrement dit, l'ensemble des fonctions de la forme $F(x)$ est dense dans $\mathcal{C}(I_m, \mathbb{R})$.

Hornik (1991) a montré que ce n'est pas le choix spécifique de la fonction d'activation, mais plutôt l'architecture *feedforward* multi-couches elle-même qui donne aux réseaux de

neurones le potentiel d'approximateurs universels. Il prouve également qu'un réseau de neurones à trois couches avec n_1 neurones d'entrée, une fonction d'activation tanh en deuxième couche et une fonction d'activation linéaire en troisième couche de n_2 neurones peuvent approcher avec une précision arbitraire toute fonction continue de $\mathbb{R}^{n_1}$ dans $\mathbb{R}^{n_2}$ à condition que le nombre de neurones dans la deuxième couche soit suffisamment grand.

Ces résultats fondamentaux justifient notre approche qui consiste à considérer des réseaux de neurones *feedforward*.

2.2.2 Fonctions d'activation

Beaucoup de choix de fonctions d'activation existent (James *et al.*, 2013). Nous commencerons par nous demander s'il est vraiment nécessaire de considérer une fonction d'activation non linéaire. Supposons que la fonction d'activation soit égale à l'identité : $\sigma = Id$. Si le réseau est composé de L couches (on suppose les biais β_l nuls par souci de simplicité, mais les garder non nuls ne changerait pas l'argument), alors si $Z = W^T \cdot X$ est la fonction de pré-activation, on a :

$$A_L = f(Z) = (W_L)^T \cdot A_{L-1} + 0 = (W_L)^T \cdot (W_{L-1})^T \cdots (W_1)^T \cdot X = (W_1)^T_{TOTAL} \cdot X, \quad (2.15)$$

qui est une fonction linéaire en X. Ainsi, disposer d'un nombre de L couches ne change pas la capacité de représentation du réseau : la non-linéarité de la fonction d'activation est donc fondamentale et il est donc nécessaire que la fonction d'activation soit non linéaire. Les fonctions d'activation non linéaires les plus utilisées en pratique sont présentées dans la suite.

Fonction de Heaviside La fonction de Heaviside (ou fonction step) s'écrit :

$$\text{step}(z) = \begin{cases} 0 & \text{si } z < 0 \\ 1 & \text{sinon} \end{cases} \quad (2.16)$$

Fonction ReLU La fonction unité de rectification linéaire ou ReLU s'écrit :

$$\text{ReLU}(z) = \begin{cases} 0 & \text{si } z < 0 \\ z & \text{sinon} \end{cases} \quad (2.17)$$

Fonction sigmoïde La fonction sigmoïde s'écrit :

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (2.18)$$

Elle représente la fonction de répartition de la loi logistique et est la fonction réciproque de la fonction logit.

Fonction tangente hyperbolique La fonction tangente hyperbolique s'écrit :

$$\forall z \in [-1, 1], \tanh(z) = \frac{\sinh(z)}{\cosh(z)} = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)} \quad (2.19)$$

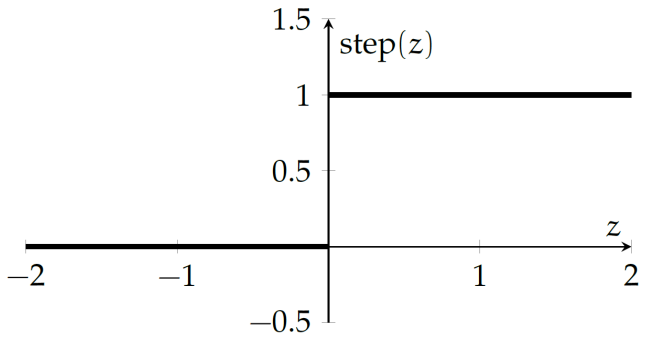


FIGURE 2.1. Représentation graphique de la fonction de Heaviside

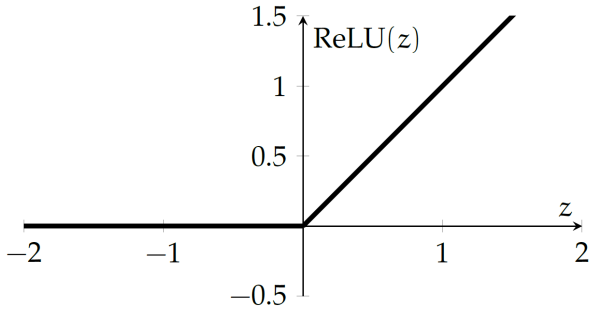


FIGURE 2.2. Représentation graphique de la fonction ReLU

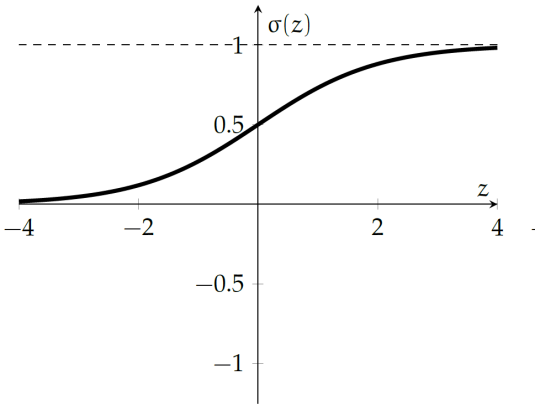


FIGURE 2.3. Représentation graphique de la fonction sigmoïde

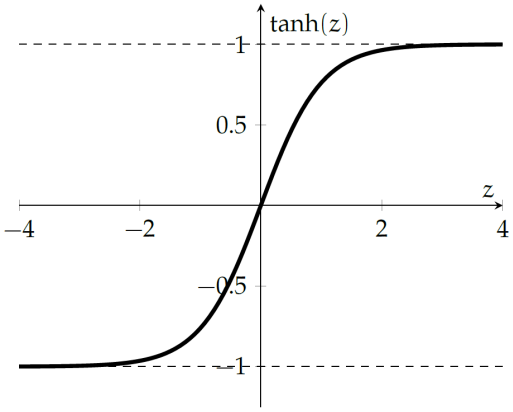


FIGURE 2.4. Représentation graphique de la fonction tanh

Fonction Leaky ReLU La fonction Leaky ReLU est une variante de la fonction ReLU qui introduit une légère pente négative α sur $\mathbb{R}_-$ afin de maintenir les mises à jour de poids en vie tout au long du processus de propagation. Ce paramètre α a été introduit comme solution aux problèmes du neurone mort (*dead neuron problem*) concernant les fonctions d'activation ReLU. Cela permet aux gradients de ne jamais être nuls pendant la période d'entraînement. La fonction Leaky ReLU calcule le gradient avec une valeur constante très faible pour le gradient négatif (ordre de grandeur de 0,01). Ainsi, la fonction Leaky ReLU est calculée comme suit :

$$\text{Leaky ReLU}(z) = \begin{cases} \alpha \cdot z & \text{si } z < 0 \\ z & \text{sinon} \end{cases} \tag{2.20}$$

Cette fonction d'activation donne des résultats identiques par rapport à la fonction ReLU standard, à l'exception du fait qu'elle a des gradients non nuls, ce qui suggère qu'il n'y a pas d'amélioration significative des résultats sauf en ce qui concerne la faible densité et la dispersion par rapport aux fonctions standard ReLU et tanh.

2.2.3 Optimisation des paramètres d'un réseau de neurones

Les paramètres d'un réseau de neurones sont optimisés grâce à une fonction de coût qui analyse la valeur en sortie de la couche de sortie et la compare aux valeurs réelles (fournis

lors de la phase d'entraînement) et calcule un coût.

2.2.3.1 Fonction de coût

Le choix de la fonction de coût est important. Trois fonctions de coût sont généralement utilisées dans les problèmes de régression et classification par réseau de neurones : la MSE (ou erreur quadratique moyenne ou perte L2), la MAE (ou erreur absolue moyenne ou perte L1), et la fonction log-loss $\mathcal{L}$ (ou *binary cross-entropy*). Elles sont définies par les relations suivantes :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.21}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2.22}$$

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \tag{2.23}$$

où les y_i sont les valeurs observées et les $\hat{y}_i$ sont les valeurs prédites de la variable de sortie du réseau de neurones. La fonction log-loss $\mathcal{L}$ est généralement utilisée dans les problèmes de classification.

Remarque Le choix d'une fonction de coût particulière est intéressant si l'on souhaite par exemple pénaliser plus les sur-estimations par rapport aux sous-estimations de la variable de sortie.

2.2.3.2 Algorithme de rétro-propagation du gradient

La fonction de coût permet de définir une descente de gradient sur l'ensemble des poids du réseau de neurones, c'est ce que l'on appelle la rétro-propagation du gradient. La descente de gradient est un algorithme classique d'optimisation qui définit aléatoirement un point de départ x puis va calculer à chaque étape le gradient de la fonction à optimiser pour modifier la valeur de x de manière à aller dans le sens opposé du gradient. Mathématiquement, si on note f la fonction à optimiser, on a à chaque étape de l'algorithme le calcul suivant :

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \tag{2.24}$$

avec x_0 le point initial à choisir, de manière aléatoire ou non, η_k le taux d'apprentissage (*learning rate*) et $\nabla f(x_k)$ est le gradient de la fonction f au point x_k. De manière imagée, l'idée derrière la descente de gradient est similaire à la "descente d'une colline" jusqu'à ce qu'un minimum global ou local soit atteint. À chaque mise à jour, la recherche se déplace dans la direction opposée du gradient ; la pente du gradient et le taux d'apprentissage déterminent l'amplitude de ce mouvement. De plus, le choix du taux η_k est un élément important. En effet, une trop faible valeur de η_k implique un nombre élevé d'itérations et une convergence lente de l'algorithme. À l'inverse, si la valeur de η_k est trop élevée, la descente de gradient peut dépasser le minimum, cela entraînant une non-convergence de l'algorithme vers le minimum global ou même une divergence. (voir figure 2.5).

Une description complète de l'algorithme de rétro-propagation du gradient figure par exemple dans Goodfellow *et al.* (2016) qui donnent également un aperçu détaillé des algorithmes d'optimisation par descente de gradient.

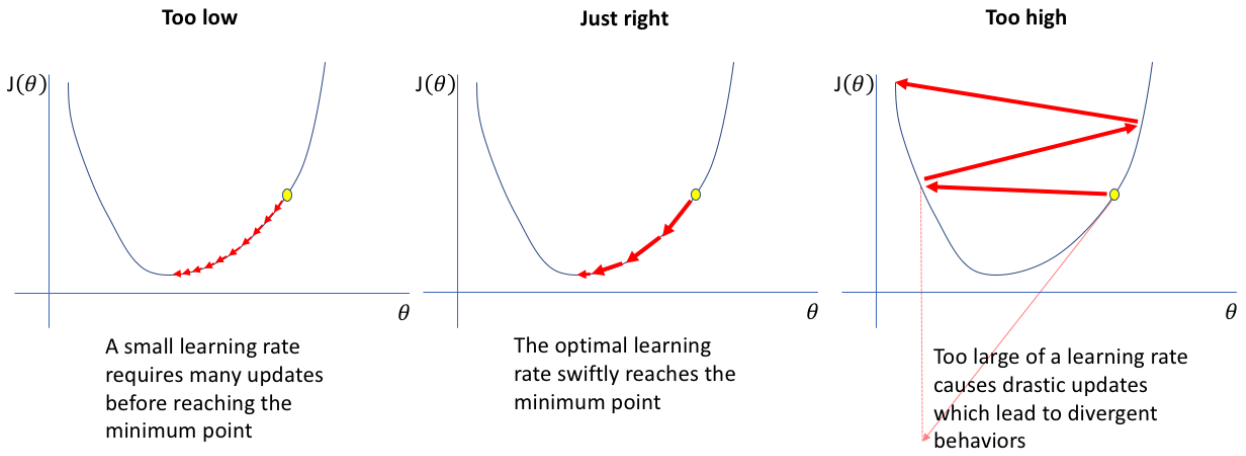


FIGURE 2.5. Taux d'apprentissage et descente de gradient. Cas d'une fonction de coût à minimiser $J(\theta)$ à une variable θ. Source : `https://www.jeremyjordan.me/nn-learning-rate`

2.2.4 Optimisation des hyperparamètres d'un réseau de neurones

Les choix concernant le type d'architecture d'un réseau de neurones (par exemple, le nombre de couches cachées, le nombre de neurones pour chaque couche) et les valeurs d'autres hyperparamètres (le taux d'apprentissage, les fonctions d'activation, le nombre d'époques, le nombre de batches, etc.) restent un problème heuristique et expérimental (Thome, 2019). Leur choix dépend souvent du type de données et cette étape est généralement aussi chronophage que fondamentale.

Une description détaillée et centrée sur la pratique du réglage des hyperparamètres d'un réseau de neurones (*hyperparameters tuning*) figure dans Ng (2017).

2.2.5 Méthodes de régularisation

Si l'optimisation de la fonction de perte sur les données d'entraînement constitue un objectif important pour entraîner un réseau de neurones, il convient de prendre garde au risque de sur-apprentissage (*overfitting*). Il existe plusieurs stratégies importantes de régularisation d'un réseau de neurones.

2.2.5.1 Dropout

Le Dropout est une méthode de régularisation conçue pour fonctionner avec des réseaux de neurones profonds (Srivastava *et al.*, 2014; Hinton *et al.*, 2012). Son principe est, plutôt que de perturber les données à chaque étape d'entraînement, de perturber le réseau de neurones lui-même. De manière aléatoire, à chaque étape d'entraînement, un ensemble d'unités est sélectionné dans chaque couche et est supprimé du réseau de neurones. Ainsi, toutes les unités devront assurer une sorte de responsabilité "collective" pour obtenir la bonne prédiction et ne pourront pas compter sur un petit sous-ensemble de poids pour effectuer tous les calculs nécessaires. Cela tend également à rendre le réseau plus robuste aux perturbations de données.

Pendant la phase d'entraînement, pour chaque exemple d'entraînement, et pour chaque unité, aléatoirement et avec une probabilité p, on définit temporairement $a_j^l := 0$. Il n'y aura aucune contribution à la sortie ni aucune mise à jour du gradient pour l'unité associée.

Après la période d'entraînement, quand le réseau est utilisé pour faire des prédictions, tous les poids sont multipliés par p pour obtenir les mêmes niveaux moyens d'activation. Pour implémenter la méthode de Dropout, on définit dans le passage forward pendant la période d'entraînement :

$$a^l = f(z^l) \odot d^l, \tag{2.25}$$

où $\odot$ désigne le produit composante par composante et d^l est un vecteur de 0 et de 1 construit aléatoirement avec une probabilité p. Le passage backward dépend de a^l, donc il n'est pas nécessaire de modifier l'algorithme.

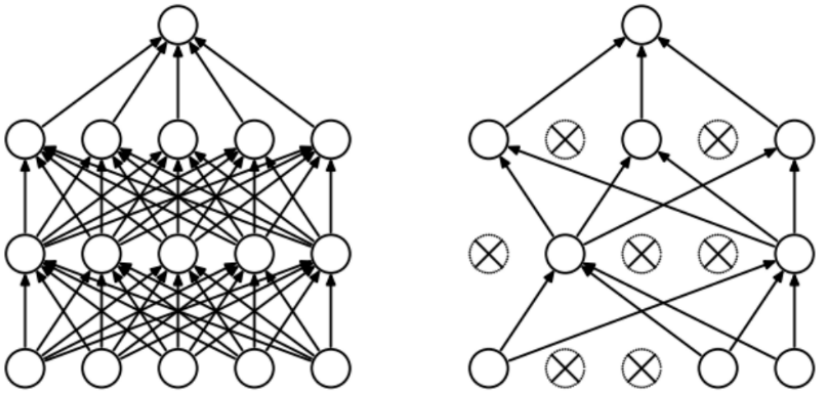


FIGURE 2.6. Régularisation par Dropout (Srivastava *et al.*, 2014). **Gauche** : Un réseau de neurones standard avec deux couches cachées. **Droite** : Le même réseau de neurones auquel est appliquée la méthode de Dropout, les neurones marqués d'une croix ayant été supprimés.

2.2.5.2 Normalisation par batch (BatchNorm)

Une autre technique de régularisation est la normalisation par batch (ou couches BatchNorm) qui a été développée à l'origine pour traiter un problème de décalage de covariable (Ioffe et Szegedy, 2015). L'idée est qu'il est très difficile pour un réseau de neurones d'apprendre lorsque la distribution des entrées est en train de changer : les poids doivent être modifiés pour améliorer les prédictions, mais aussi simplement pour compenser un changement des entrées. Ainsi, lors de l'entraînement avec des mini-batches, l'idée est de normaliser les valeurs d'entrée pour chaque mini-batch, en soustrayant la moyenne et en divisant par l'écart-type de chaque entrée. Ainsi, pour une couche cachée avec une entrée d-dimensionnelle $x = (x^1, \ldots, x^d)$, chaque dimension est normalisée selon l'équation de normalisation :

$$\widetilde{x}^k = \frac{x^k - \mathbb{E}(x^k)}{\sqrt{\mathbb{V}(x^k)}}. \tag{2.26}$$

Cela signifie que l'échelle des entrées de chaque couche reste la même, peu importe l'évolution des poids dans les couches précédentes. La normalisation par batch finit par avoir un effet de régularisation pour des raisons similaires à celles de l'ajout de bruit et du Dropout : chaque mini-batch de données finit par être légèrement perturbé, ce qui empêche le réseau d'exploiter des valeurs très particulières de l'échantillon de données.

2.2.6 Raccourcis

Des raccourcis (appelées aussi "skip connections") sont utilisés en établissant des connexions entre la couche de sortie et la dernière couche cachée intermédiaire, ainsi que la première couche cachée. Ces connexions supplémentaires ont été utilisées dans la littérature sur la vision afin d'entraîner avec succès des réseaux de neurones profonds (He *et al.*, 2016; Huang *et al.*, 2017). Elles sont censées permettre de résoudre le problème d'évanouissement du gradient ("vanishing gradient"), en raccourcissant le chemin que l'algorithme de rétro-propagation du gradient a besoin de parcourir pour revenir à la première couche du réseau de neurones.

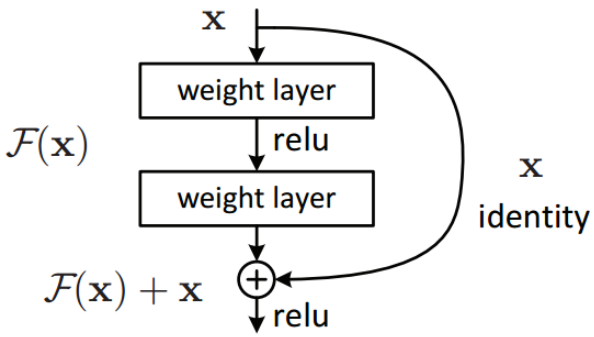


FIGURE 2.7. Représentation d'un raccourci (He *et al.*, 2016)

2.3 Les forêts aléatoires

2.3.1 Les arbres de régression

Les arbres de classification et de régression (CART ou *Classification And Regression Trees*) sont devenus des alternatives populaires à la régression linéaire (Breiman *et al.*, 1984). Elles sont basées sur la partition de l'espace des prédicteurs $\mathcal{X}$, via une séquence de scissions binaires, dont l'ensemble des règles peut être résumé dans une arborescence. Une fois que l'espace entier est scindé de manière récursive en un certain nombre de régions simples, la réponse pour une observation donnée peut être prédite à l'aide de la moyenne des observations d'entraînement de la région à laquelle appartient cette observation (James *et al.*, 2017).

Nous discutons maintenant du processus de construction d'un arbre de régression qui est composé de deux grandes étapes.

- Division de l'espace des prédicteurs $\mathcal{X} = \{X_1, \ldots, X_p\}$ en J régions distinctes, disjointes deux à deux : $R_1, \ldots, R_J$.
- Pour toute nouvelle observation des prédicteurs $X = x_0$, on regarde dans quelle région on tombe, disons R_j. La prédiction correspondant à cette observation est la moyenne des valeurs observées dans la partie de l'ensemble d'entraînement qui tombent dans R_j.

Comment construire les régions $R_1, \ldots, R_J$? En théorie, les régions pourraient avoir n'importe quelle forme. Néanmoins, il est choisi de diviser l'espace des prédicteurs en rectangles ou *boîtes* de grande dimension, pour des raisons de simplicité et de facilité d'interprétation

du modèle prédictif résultant. Le but est de trouver les *boîtes* $R_1, \ldots, R_J$ qui minimisent la somme des résidus au carré (RSS ou *Residual Sum of Squares*), donnée par :

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{2.27}$$

où $\hat{y}_{R_j}$ est la réponse moyenne sur les observations d'entraînement qui tombent dans la *boîte* R_j.

Hélas, considérer l'ensemble des partitionnements possibles de l'espace des prédicteurs en J *boîtes* nécessiterait des calculs excessivement lourds. Pour cette raison, un algorithme glouton et *top-down* (appelé *recursive binary splitting*) est utilisé afin de construire l'arbre de façon récursive. On parle d'algorithme glouton car à chaque étape de la construction de l'arbre, on construit la meilleure division possible du nœud en deux sous-nœuds. Par ailleurs, on parle d'algorithme *top-down* car l'algorithme démarre à la racine de l'arbre (au point où toutes les observations appartiennent à une seule région) et sépare ensuite l'espace des prédicteurs en ajoutant progressivement des nœuds.

Afin de construire l'arbre de façon récursive, on sélectionne d'abord le prédicteur X_j et le point de coupure s de sorte que le fractionnement de l'espace du prédicteur dans les régions $\{X \mid X_j < s\}$ et $\{X \mid X_j \geq s\}$ entraîne la plus grande réduction possible du RSS. La notation $\{X \mid X_j < s\}$ désigne la région de l'espace des prédicteurs dans laquelle X_j prend une valeur strictement inférieure à s. Autrement dit, on considère tous les prédicteurs $X_1, \ldots, X_p$ et toutes les valeurs possibles des points de coupure s pour chacun des prédicteurs, puis on choisit le prédicteur et le point de coupure de telle sorte que l'arbre résultant ait le RSS le plus petit. Autrement dit, pour tout j et s, on définit la paire de demi-plans

$$R_1(j, s) = \{X \mid X_j < s\} \tag{2.28}$$
$$R_2(j, s) = \{X \mid X_j \geq s\}, \tag{2.29}$$

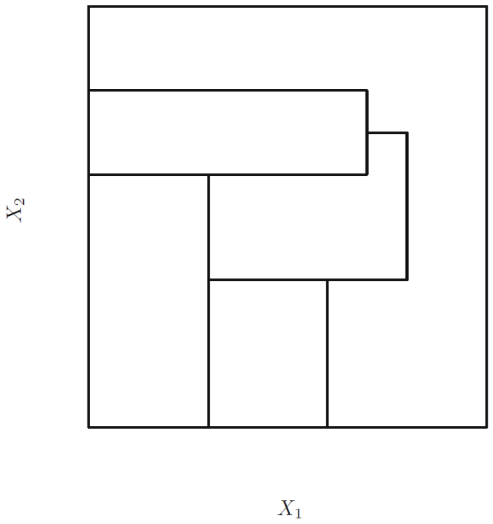
et on cherche les valeurs de j et s qui minimisent la quantité suivante

$$\sum_{i \mid x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i \mid x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{2.30}$$

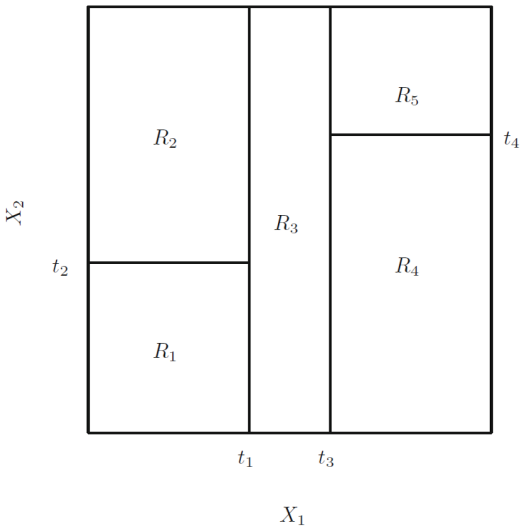
où $\hat{y}_{R_1}$ est la réponse moyenne pour les observations d'entraînement dans $R_1(j, s)$ et $\hat{y}_{R_2}$ est la réponse moyenne pour les observations d'entraînement dans $R_2(j, s)$. Trouver les valeurs de j et s qui minimisent l'équation (2.30) peut être fait assez rapidement, surtout lorsque le nombre d'entités p n'est pas trop grand.

Le processus est ensuite répété sur chaque nœud, en recherchant le meilleur prédicteur et le meilleur point de coupure afin de diviser davantage les données et minimiser le RSS dans chacune des régions résultantes. Le processus se poursuit jusqu'à ce qu'un critère d'arrêt soit atteint. Par exemple, nous pouvons continuer jusqu'à ce qu'aucune région ne contienne plus de cinq observations. Une fois les régions $R_1, \ldots, R_J$ créées, il s'agit de prédire la réponse pour une observation de test donnée en utilisant la moyenne des observations d'entraînement dans la région à laquelle appartient cette observation de test.

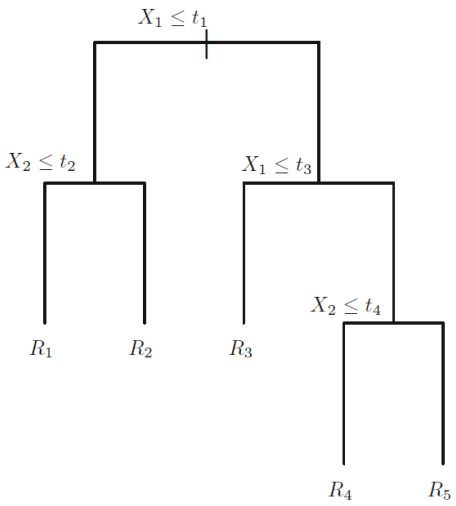
La figure 2.8 présente un exemple avec cinq régions. Pour une description complète de l'algorithme CART, le lecteur pourra se référer à Breiman *et al.* (1984) et James *et al.* (2017).



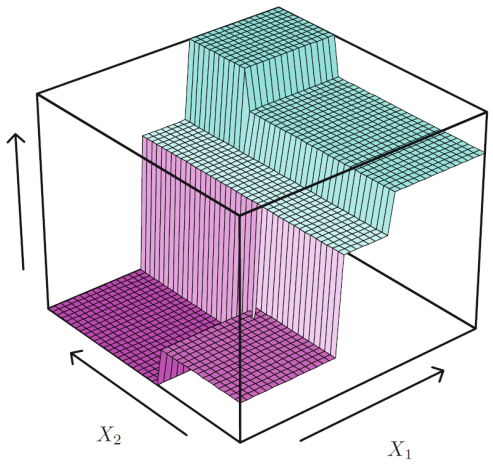
(a) Partition d'un espace de prédicteurs bidimensionnel qui ne peut être obtenue par l'algorithme *recursive binary splitting*



(b) Partition d'un espace de prédicteurs bidimensionnel par l'algorithme *recursive binary splitting*



(c) Arbre de régression qui correspond à la partition obtenue en figure (b)



(d) Surface de prédiction qui correspond à l'arbre de régression présenté en figure (c)

FIGURE 2.8. Illustration de l'approche dans le cas de deux prédicteurs X_1, X_2 et cinq régions $R_1, \ldots, R_5$. Source : James *et al.* (2017)

2.3.2 Les forêts aléatoires

Les méthodes de *bagging* consistent en une agrégation aléatoire par *bootstrap* d'un grand nombre de modèles. Supposons avoir à notre disposition un grand nombre de modèles donnant une estimation d'une variable de sortie y (taux de mortalité par exemple), estimés sur des échantillons indépendants. Alors l'espérance de ces estimateurs est un estimateur sans biais et de faible variance de y. Cela signifie qu'il suffit de calculer la moyenne des prédictions de tous les modèles pour obtenir un nouvel estimateur plus précis de y. Cependant, il n'est pas envisageable d'obtenir un grand nombre d'échantillons d'observations indépendants. C'est alors que l'algorithme de *bootstrap* intervient. Les échantillons sont remplacés par des tirages aléatoires, avec remises et de taille fixe au sein de la base de données initiale. Ce type de modèles est particulièrement bien adapté aux modèles dont les résultats sont instables (et donc de variance importante dans les prédictions). C'est pourquoi cette méthode est particulièrement adaptée à l'utilisation de modèles d'arbres de décision comme modèle élémentaire.

Les forêts aléatoires ou *random forests* (Breiman, 2001) sont issues d'une méthode de *bagging* appliquée aux arbres. Celle-ci part d'une constatation simple : les arbres ont de fortes chances d'être très semblables entre eux, car ce sont les variables les plus explicatives qui vont ressortir au détriment des autres. L'idée du modèle est donc, au-delà de la sélection aléatoire des échantillons sur lesquels sont estimés les arbres, de sélectionner également aléatoirement les variables intervenant dans le modèle à chaque itération. En particulier, à chaque partition, l'algorithme ne considère que m variables explicatives de manière aléatoire parmi les p variables explicatives du problème. Souvent, l'ordre de grandeur de m est donné par la relation $m \approx \sqrt{p}$. Ainsi, chaque modèle de base est moins performant, mais l'agrégation de modèles très diversifiés produit des résultats pertinents.

Remarque Une forêt aléatoire est équivalente à un *bagging* si l'on a : $m = p$.

L'estimateur de l'indicateur par forêt aléatoire s'écrit de la manière suivante :

$$\hat{y}_{RF} = \frac{1}{n_{\text{a}}} \sum_{i=1}^{n_{\text{a}}} \hat{y}_i,$$

où n_{a} est le nombre d'arbres de régression. Cet estimateur peut être obtenu en appliquant l'algorithme du package R `randomForest` (Liaw et Wiener, 2002). Comme cette procédure a un coût calculatoire élevé, le nombre d'arbres doit être choisi avec soin : il ne doit pas être trop grand, mais il doit également être capable de produire un pourcentage adéquat de variance expliqué et une faible moyenne de résidus au carré.

CHAPITRE 3

UNE NOUVELLE MÉTHODE DE CORRECTION DE DONNÉES DE MORTALITÉ

3.1 Contexte et structure du chapitre

La construction des tables de mortalité reste un problème non résolu, depuis les premiers développements de Lexis (1875) et ses contemporains. Des travaux récents ont montré l'importance des interactions entre la fertilité et la mortalité, ainsi que leur impact sur les taux de mortalité agrégés.

Les études portant sur la modélisation de la mortalité supposent généralement que les données sous-jacentes (par exemple, les décès et les expositions) sont exactes. Néanmoins, il existe des anomalies dans les données de mortalité pour la plupart des pays de la *Human Mortality Database* (HMD, 2018), une base de données qui contient des tables de mortalité pour plus de quarante pays ou régions, ainsi que toutes les données brutes utilisées pour construire ces tables. La prise de conscience de telles anomalies de données est apparue suite à la conjecture de Richards (Richards, 2008), qui avait suggéré que certains effets cohorte en Angleterre pouvaient être dus à la sensibilité de la méthode de calcul des taux de mortalité aux chocs de fertilité. Cette conjecture a été confirmée par Cairns *et al.* (2016) qui a proposé une méthode pour corriger les données de mortalité pour l'Angleterre et le pays de Galles sur la base des données trimestrielles du taux de natalité. Le *Convexity Adjustment Ratio* introduit dans leurs travaux a ensuite été adapté par Boumezoued (2016) qui s'est concentré sur les données HMD et a montré que ces anomalies sont universelles. Pour construire des tables de mortalité corrigées, un lien avec la *Human Fertility Database* (HFD, 2018), l'équivalent de la HMD pour la fertilité, a été établi, rendant possible la correction de ces anomalies de manière systématique. Boumezoued *et al.* (2018) ont introduit une stratégie d'inférence à partir d'un modèle de dynamique de population déterministe afin de calculer les taux de mortalité sur la base d'informations extraites des recensements. Leurs travaux ont confirmé d'un point de vue mathématique l'efficacité de l'utilisation de données mensuelles de fertilité supplémentaires pour calculer correctement les tables de mortalité annuelles.

La dernière version du protocole de méthodes HMD (Wilmoth *et al.*, 2019) introduit un changement dans la façon dont les taux de mortalité sont construits dans la base de données. Les données de naissance par mois qui ont été recueillies sont maintenant utilisées pour estimer plus précisément les expositions de la population.

Cependant, ces méthodes ne s'appliquent pas aux pays pour lesquels les données de naissance par mois ne sont pas disponibles. L'objectif de ce chapitre est donc de construire une méthode afin de produire des tables de mortalité corrigées pour des pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles.

À notre connaissance, la seule tentative de correction automatique a été proposée par Cairns *et al.* (2016) qui ont développé un cadre bayésien pour récupérer des expositions corrigées. À notre connaissance également, aucune autre étude prédictive n'a porté sur la correction des estimations de mortalité en l'absence de données mensuelles de fertilité.

Le reste de ce chapitre est organisé comme suit. La section 3.2 spécifie le problème de correction et les observables dans le diagramme Lexis. La section 3.3 présente la méthode de correction des estimations d'exposition. La section 3.4 expose les résultats de la méthode et la section 3.5 est une application de la méthode à quatre pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles. Enfin, la section 3.6 se termine par une discussion et des pistes pour de futures recherches.

3.2 La correction des estimations de mortalité

3.2.1 Le problème de correction

Un effet cohorte apparaît lorsqu'une génération a des caractéristiques de longévité très différentes des générations adjacentes, par exemple lorsque différentes distributions d'une maladie découlent d'une cause environnementale changeante ou nouvelle et affectent différemment l'ensemble des groupes d'âge. En épidémiologie, un effet cohorte désigne un effet période qui est vécu de manière différente par une exposition spécifique à l'âge ou une sensibilité à un événement ou à une cause. En d'autres termes, un effet cohorte peut survenir lorsqu'une cause environnementale qui affecte la population est inégalement répartie dans la population (Keyes *et al.*, 2010).

À l'inverse, des variations soudaines de naissances peuvent produire de faux effets cohorte. En fait, la force de mortalité $\mu(a,t)$ à l'âge a et à l'année t est estimée en comparant le nombre de décès $d(a,t)$ survenus au cours de l'année t, avec l'exposition correspondante au risque $E(a,t)$. L'exposition au risque représente la quantité d'individus à risque de décéder, *i.e.* le temps total vécu par la population au cours de la période considérée (voir la section 1.1 pour plus de détails). L'estimateur de la force de mortalité est donc donné par :

$$\hat{\mu}(a,t) = \frac{d(a,t)}{E(a,t)}. \tag{3.1}$$

Définissons les taux d'amélioration :

$$r(a,t) = \frac{\mu(a,t+1) - \mu(a,t)}{\mu(a,t)}, \tag{3.2}$$

qui sont souvent utilisés pour observer des *patterns* particuliers. Si l'on représente la matrice des taux d'amélioration de la mortalité des hommes pour l'Allemagne de l'Ouest (voir la figure 3.1), des effets cohorte peuvent être observés pour des générations spécifiques (nées vers 1915, 1920 et 1940).

Comme l'exposition au risque par âge est approximée par la moyenne des populations en début et fin d'année, les fluctuations à la naissance à ces moments rendent l'approximation

inexacte (voir la figure 3.2 qui représente le nombre de naissances par mois observés en France de 1861 à 2017).

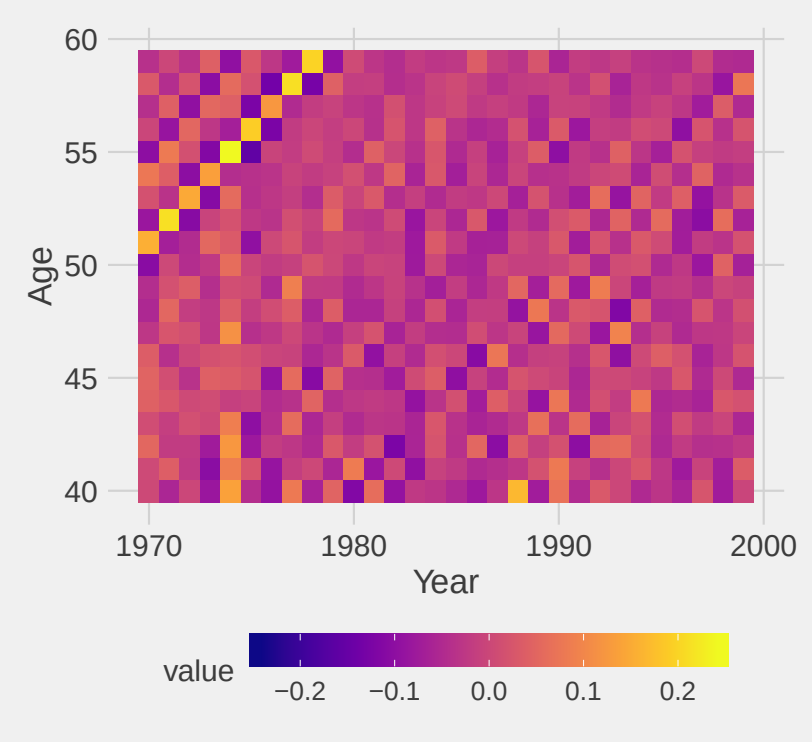


FIGURE 3.1. Taux d'amélioration de la mortalité des hommes — données brutes de l'Allemagne de l'Ouest

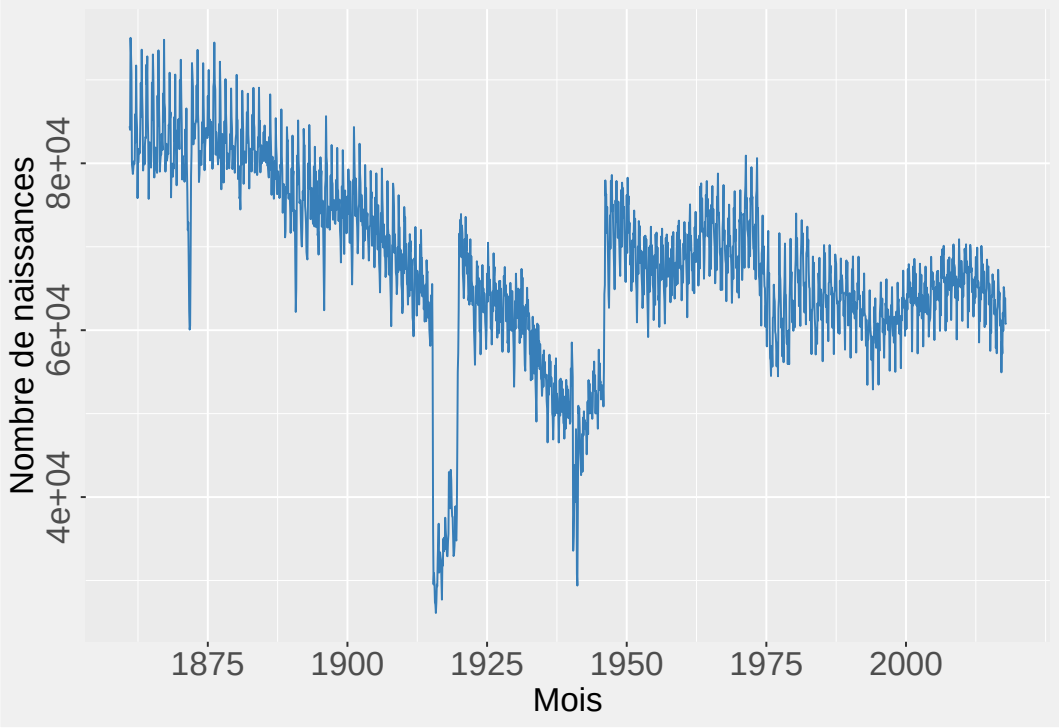


FIGURE 3.2. Nombre de naissances par mois observées en France de 1861 à 2017 — données provenant de la *Human Fertility Database*

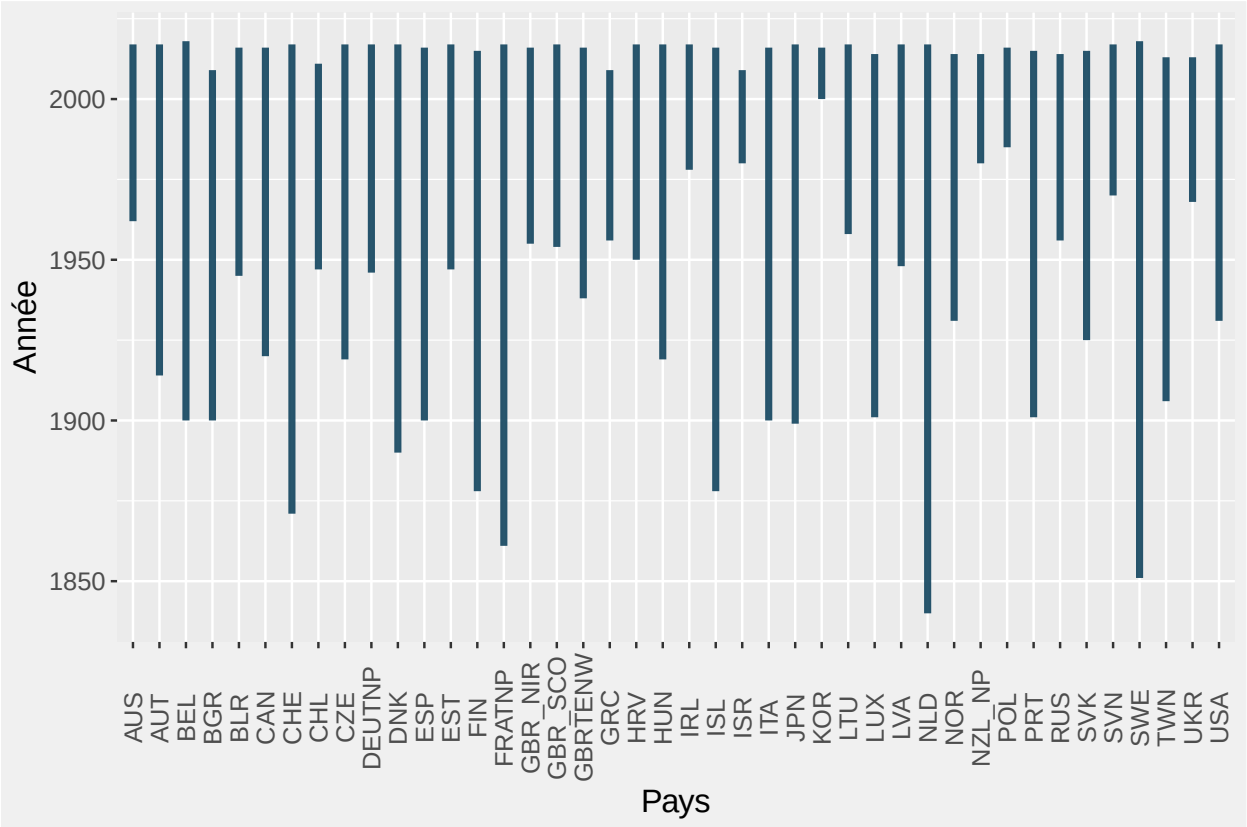


FIGURE 3.3. Disponibilité des données mensuelles de fertilité dans la HMD et la HFD

Des erreurs ont été constatées dans la HMD dès 2007 par Wilmoth *et al.* (2007). Sur la base des données mensuelles de fertilité, il est possible d'améliorer l'approximation du niveau d'exposition au risque, et donc de corriger les estimations du taux de mortalité (Wilmoth *et al.*, 2019; Boumezoued, 2016). Ce processus élimine les faux effets cohorte et produit des tables de mortalité plus précises, appelées dans la suite "tables de mortalité corrigées". Cependant, les données de naissance par mois ne sont pas disponibles pour de nombreux pays pour lesquels les relevés mensuels de fertilité sont fournis à partir des années postérieures à 1918 (voir la figure 3.3). Elle limite ainsi l'ensemble des cohortes pour que la méthode de correction soit applicable. D'où la nécessité de trouver une nouvelle méthode de correction pour corriger les tables de mortalité pour un plus grand nombre de pays.

3.2.2 Observables dans le diagramme de Lexis

Considérons le diagramme de Lexis (voir la figure 3.4), constitué d'un système de coordonnées cartésiennes (t, a), où t désigne l'année à laquelle un individu est observé et a désigne l'âge de l'individu. Notons également par $g(a, t)$ la densité de population au point (a, t).

Les estimations suivantes dans le diagramme de Lexis peuvent alors être observées :

- La population au temps t, d'âge a au dernier anniversaire : $P(a,t) = \int_a^{a+1} g(x,t)\,\mathrm{d}x$,
- Le nombre d'individus ayant atteint l'âge exact a au cours de l'année civile $[t, t+1)$: $N(a,t) = \int_t^{t+1} g(a,s)\,\mathrm{d}s$,
- Le nombre de naissances annuelles $B(t) = \int_t^{t+1} g(0,s)\,\mathrm{d}s$.

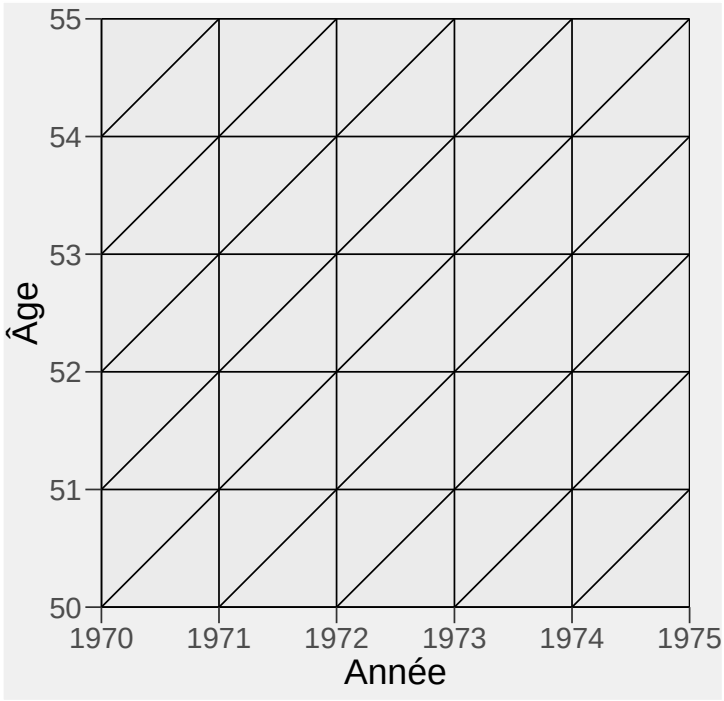


FIGURE 3.4. Diagramme de Lexis

De plus, le nombre de décès est fourni sur les triangles supérieur (T_U) et inférieur (T_L) du diagramme de Lexis :

$$d_U(a,t) = \iint_{T_U(a,t)} \mu(x,s)g(x,s)\,\mathrm{d}x\,\mathrm{d}s \text{ et } d_L(a,t) = \iint_{T_L(a,t)} \mu(x,s)g(x,s)\,\mathrm{d}x\,\mathrm{d}s, \tag{3.3}$$

où :

$$T_U(a,t) = \{(x,s) \mid x \in [a,a+1) \text{ et } s \in [t,t-a+x)\}, \text{ et}$$
$$T_L(a,t) = \{(x,s) \mid x \in [a,a+1) \text{ et } s \in [t-a+x,t+1)\}.$$

Ainsi, le nombre de décès pour les carrés de Lexis de dimension 1×1 est donné par $d(a,t) = d_U(a,t) + d_L(a,t)$.

On considère les données pour les individus de genre masculin de la version 5 de la HMD (voir `https://v5.mortality.org`) et de la version 6 de la HMD (voir `https://www.mortality.org`). Soit k la variable de pays, $\mu^{(5)}$ et $\mu^{(6)}$ les estimations des forces de mortalité provenant respectivement de la version 5 et de la version 6 de la HMD, définis par :

$$\mu^{(v)}(a,t,k) = \frac{d^{(v)}(a,t,k)}{E^{(v)}(a,t,k)}, v \in \{5,6\},$$

avec $E^{(v)}(a,t,k)$ et $d^{(v)}(a,t,k)$ respectivement l'exposition au risque et le nombre de décès par rapport à la version $v \in \{5,6\}$ de la HMD pour le pays k. $I(a,t,k)$ désigne l'indicateur de correction des données de mortalité, défini comme suit :

$$I(a,t,k) := \frac{\mu^{(5)}(a,t,k)}{\mu^{(6)}(a,t,k)}. \tag{3.4}$$

Comme on a : $\forall a, t, d^{(5)}(a,t,k) = d^{(6)}(a,t,k)$ pour un pays fixé k, il en découle l'égalité suivante :

$$I = \frac{\mu^{(5)}(a,t,k)}{\mu^{(6)}(a,t,k)} = \frac{E^{(6)}(a,t,k)}{E^{(5)}(a,t,k)}. \tag{3.5}$$

Le protocole des méthodes HMD correspondant à la version 6 de la HMD (Wilmoth *et al.*, 2019) utilise les données de naissance mensuelles lorsqu'elles sont disponibles afin d'estimer plus précisément les expositions. Par conséquent, il est supposé dans la suite du chapitre que si $I \neq 1$ pour un triplet fixé (a,t,k), la quantité $\mu^{(5)}(a,t,k)$ correspond à une estimation de mortalité brute qui n'a pas été corrigée à l'aide des données de naissance par mois, tandis que $\mu^{(6)}(a,t,k)$ correspond à sa version corrigée.

Boumezoued (2016) suggèrent que le rapport entre les estimations de mortalité corrigées et brutes est le même pour chaque âge au sein d'une cohorte donnée. La valeur de la variable I est analysée pour les âges 18-70, les cohortes 1915, 1919, 1940 et 1941 et les pays suivants : Espagne, Italie, France, Autriche, Danemark, Suède, Finlande et Norvège. Ces résultats sont illustrés par les figures 3.5 à 3.8. À première vue, ces figures montrent que la variable I est assez stable au fil des âges au sein d'une cohorte donnée. La stabilité globale est observée pour la plupart des pays, bien que cette remarque générale ne soit pas valable pour les cohortes 1940 et 1941, qui montrent quelques variations de I pour l'Espagne et l'Italie. L'analyse de I en matière de formules et de variabilité en âge est discutée en annexe H.

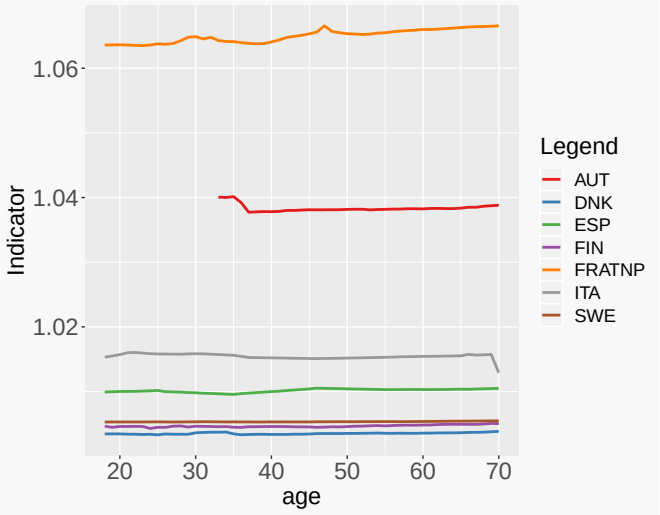


FIGURE 3.5. Valeurs de I pour la cohorte 1915 (hommes)

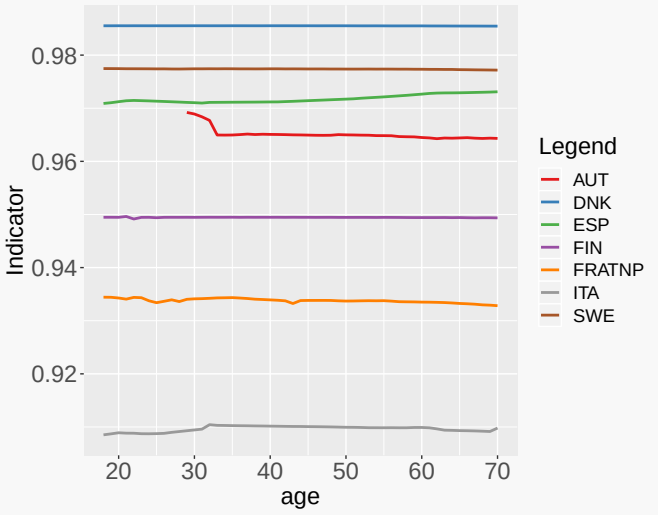


FIGURE 3.6. Valeurs de I pour la cohorte 1919 (hommes)

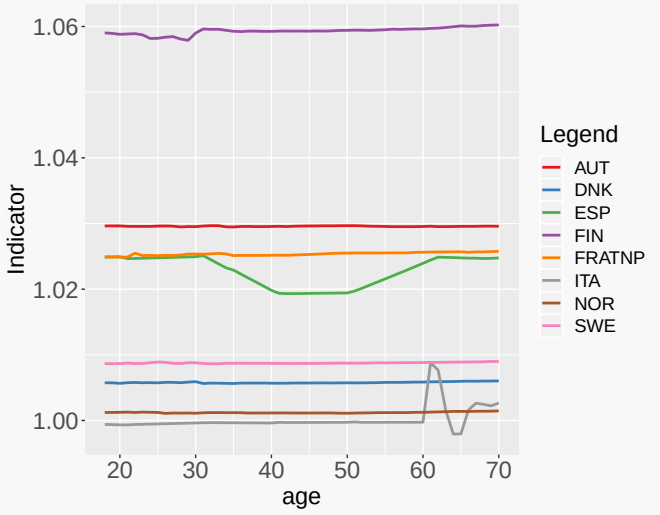


FIGURE 3.7. Valeurs de I pour la cohorte 1940 (hommes)

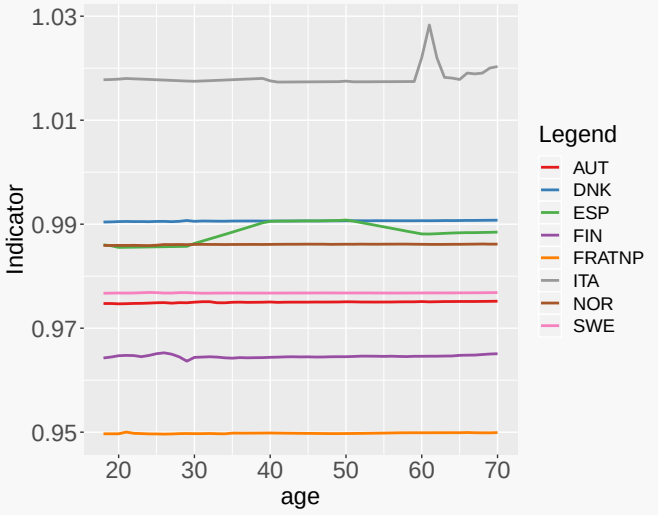


FIGURE 3.8. Valeurs de I pour la cohorte 1941 (hommes)

3.3 La méthode de correction des estimations de mortalité

3.3.1 Séparation de l'échantillon de données en un échantillon d'entraînement et un échantillon de test

L'échantillon de données $\mathcal{D}$ comprend N_{obs} observations :

$$\mathcal{D} = \{(\mathbf{x}_i, I_i) \mid I_i \neq 1 \text{ et } i \in [\![1, N_{\text{obs}}]\!]\}, \tag{3.6}$$

où $\mathbf{x}$ désigne le vecteur de variables d'entrée de dimension D et l'indicateur de mortalité I est la variable de sortie ou cible.

Les vecteurs-colonnes des entrées pour toutes les N_{obs} observations sont agrégés dans une matrice X de dimension $D \times N_{\text{obs}}$, appelée *design matrix*. Les cibles sont collectées dans le vecteur $\mathbf{I}$, si bien que l'on puisse écrire $\mathcal{D} = (X, \mathbf{I})$. Dans notre problème de régression, les cibles sont des valeurs réelles. On s'intéresse à inférer la relation entre les entrées et les cibles, *i.e.* la distribution conditionnelle des cibles étant données les entrées (mais nous ne sommes pas intéressés de modéliser la distribution des entrées elle-même).

Dans un scénario de correction réaliste où nous aimerions corriger les données de mortalité d'un pays $\mathcal{C}$ pour lequel les données mensuelles de fertilité ne sont pas disponibles, nous utiliserions un ensemble de pays pour lesquels ces données mensuelles de fertilité sont disponibles en tant qu'échantillon d'entraînement. Nous entraînerions ensuite notre modèle sur cet échantillon d'entraînement et calculerions les prédictions de l'indicateur de correction pour ce pays $\mathcal{C}$.

Suite à cette observation, au lieu de séparer l'échantillon de données $\mathcal{D}$ de façon aléatoire entre un échantillon d'entraînement $\mathcal{D}_{\text{train}}$ et un échantillon de test $\mathcal{D}_{\text{test}}$ afin de déterminer les performances de notre modèle, nous décidons donc d'utiliser les données de chaque pays comme échantillon de test. Plus précisément, notons K le nombre de pays considérés. L'échantillon de données $\mathcal{D}$ est divisé en K sous-échantillons, chaque sous-échantillon correspondant aux données disponibles d'un pays fixé. L'un des sous-échantillons est ensuite sélectionné comme échantillon de test $\mathcal{D}_{\text{test}}$ et la réunion des $K-1$ autres sous-échantillons est l'échantillon d'entraînement $\mathcal{D}_{\text{train}}$. Le processus est ensuite répété en sélectionnant un autre sous-échantillon parmi les $K-1$ sous-échantillons qui n'ont pas été utilisés. Ainsi, le processus est répété K fois afin que chaque sous-échantillon soit utilisé exactement une fois comme échantillon de test $\mathcal{D}_{\text{test}}$. On note que l'ensemble des K échantillons de test est une partition de l'échantillon de données d'origine.

On considère le cas où les informations disponibles sont similaires pour tous les pays considérés dans l'échantillon de données. Les données communes disponibles pour ces pays sont donc constituées d'un ensemble d'estimations de forces de mortalité et d'expositions au risque pour les années comprises entre $t_{\min}$ et $t_{\max}$, les âges compris entre $a_{\min}$ et $a_{\max}$, et les cohortes entre $c_{\min}$ et $c_{\max}$.

Finalement, le modèle est entraîné et testé K fois sur les K échantillons d'entraînement et les K échantillons de test. La moyenne des K erreurs quadratiques moyennes (MSE pour *Mean Squared Error*) de la variable cible I est ensuite calculée pour estimer l'erreur de prédiction du modèle. La MSE est une mesure fréquemment utilisée pour évaluer la précision des prédictions obtenues par un modèle. Elle mesure les différences ou résidus entre les valeurs

observées et prédites, respectivement $\hat{I}_i$ et I_i. La formule de calcul de la MSE est la suivante :

$$\text{MSE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (\hat{I}_i - I_i)^2, \tag{3.7}$$

où N_{test} désigne le nombre de points de données dans l'échantillon de test considéré, donné par :

$$\begin{aligned} N_{\text{test}} &= (t_{\max} - t_{\min} + 1)(a_{\max} - a_{\min} + 1) \\ &\quad - \left(\mathbb{1}_{\{c_{\min} \geq t_{\min} - a_{\max} + 1\}} \sum_{j=1}^{c_{\min} - t_{\min} + a_{\max}} j + \mathbb{1}_{\{c_{\max} \leq t_{\max} - a_{\min} - 1\}} \sum_{j=1}^{t_{\max} - a_{\min} - c_{\max}} j \right). \end{aligned}$$

3.3.2 Variables d'entrée et *feature engineering*

En apprentissage automatique, le *feature engineering* désigne le processus informel mais fondamental qui consiste à créer de nouvelles variables définies à partir de combinaisons de variables déjà existantes et disponibles dans l'échantillon de données. Ce processus permet grâce à l'introduction de ces nouvelles variables, d'améliorer parfois considérablement la capacité prédictive d'un modèle donné.

Soit k la variable de pays et c la variable de cohorte, avec $c = t - a$. Dans la suite, pour simplifier les notations, il est supposé que k est fixé. On considère les dérivées secondes discrètes des estimées de la force de mortalité, par rapport à l'année civile :

$$\Phi(a,t) = \frac{\mu(a,t+2)}{\mu(a,t+1)} - \frac{\mu(a,t+1)}{\mu(a,t)}, \tag{3.8}$$

$\Phi(a,t)$ est donc la dérivée première discrète par rapport à t du taux d'amélioration de la mortalité $r(a,t)$ défini par l'équation (3.2). De même, on considère également les dérivées secondes discrètes de la force de mortalité, par rapport à l'âge :

$$\Psi(a,t) = \frac{\mu(a+2,t)}{\mu(a+1,t)} - \frac{\mu(a+1,t)}{\mu(a,t)}. \tag{3.9}$$

Comme le *Convexity Adjustment Ratio* de Cairns *et al.* (2016), les quantités $\Phi(a,t)$ et $\Psi(a,t)$ peuvent être interprétées comme des matrices de convexité de la force de mortalité $\mu(a,t)$. Comme ces variables ne sont pas de bons prédicteurs de l'indicateur de correction I, l'idée est de calculer une moyenne de ces variables pour chaque cohorte fixée. Dans ce qui suit, un opérateur $\mathcal{F}$ est défini pour formaliser cette idée.

On définit l'opérateur $\mathcal{F}$ par :

$$\mathcal{F} : V(a,t) \mapsto \mathcal{F}_V(c) = \frac{1}{N_c} \sum_{t-a=c} V(a,t), \text{ avec } N_c = \mid \{V(a,t) \mid t - a = c\} \mid, \tag{3.10}$$

où V désigne une variable qui dépend de l'âge a et l'année civile t, et $|\mathcal{E}|$ désigne le nombre d'éléments d'un ensemble donné $\mathcal{E}$. On a par exemple que $\mathcal{F}_\Psi(c) = \frac{1}{N_c} \sum_{t-a=c} \Psi(a,t)$ avec $N_c = \mid \{\Psi(a,t) \mid t - a = c\} \mid$. Dans ce cas, N_c est le nombre de quantités $\Psi(a,t)$ observées à partir des données pour la cohorte fixée c. Comme le montre la figure 3.9, $\mathcal{F}_\Psi(c)$ correspond donc à la moyenne arithmétique des dérivées secondes discrètes de la quantité $\mu(a,t)$ par rapport à l'âge a pour la cohorte fixée $c = t - a$. Comme les données de mortalité sont disponibles pour plusieurs pays, les nouvelles variables considérées dépendent aussi de la variable pays k.

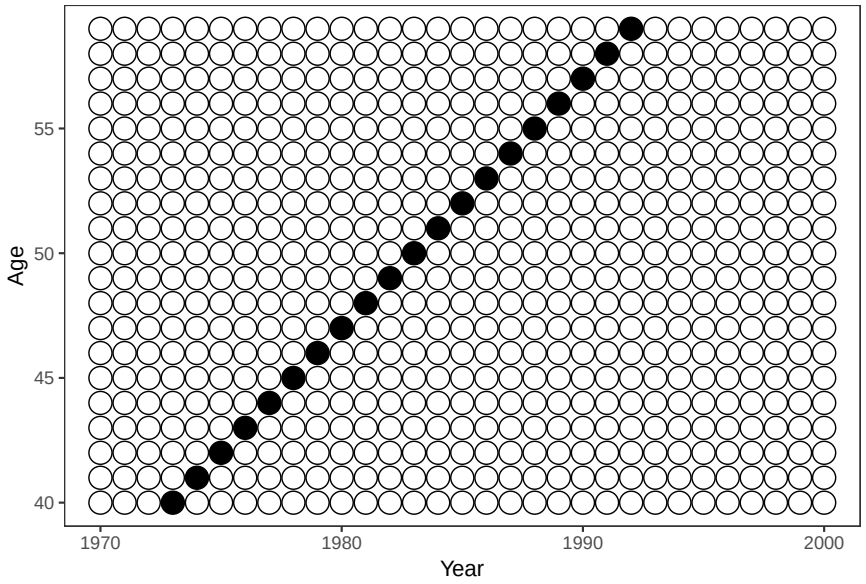


FIGURE 3.9. Dans cet exemple illustratif simple, les points de données représentent les valeurs prises par la variable V pour les âges 40-59 et les années civiles 1970-2000. Les points noirs représentent les valeurs correspondant à la cohorte de 1933 qui sont moyennées pour calculer la quantité $\mathcal{F}_V(1933)$. Dans ce cas, on a $N_{1933} = 20$, ce qui est le nombre total de points noirs.

Variables d'entrée retenues Plusieurs variables d'entrée ont été testées dans le modèle baseline et l'approche par réseau de neurones. Ces variables testées incluent notamment les dérivées premières et secondes des variables $B(k,c)$, $P(k,c)$, a, $\mathcal{F}_{\Phi_k}(k,c)$, et $\mathcal{F}_{\Psi_k}(k,c)$. Finalement, les variables $\mathcal{F}_{\Phi_k}(k,c)$ et $\mathcal{F}_{\Psi_k}(k,c)$ sont retenus pour le modèle baseline et la variable d'âge a est ajoutée dans l'approche par réseau de neurones. Ainsi, le vecteur d'entrée est $\mathbf{x} = (x_1, x_2, x_3) = (\mathcal{F}_{\Phi_k}(k,c), \mathcal{F}_{\Psi_k}(k,c), a)$.

3.3.3 L'approche par réseau de neurones

Les variables d'entrée n'ont pas des valeurs du même ordre de grandeur. Des écarts importants peuvent nuire aux algorithmes d'apprentissage en les rendant inefficaces. Ainsi, on effectue un changement de variable en remplaçant chacune d'entre elles par leur transformée *Mean-Std* $\breve{\xi}_i$ (*Mean-Std transformation*). Soit ξ une variable d'entrée continue du modèle. La transformation pour la variable ξ de l'exemple i, notée ξ_i, s'écrit :

$$\begin{cases} \mathcal{T} : \xi_i \mapsto \breve{\xi}_i = \dfrac{\xi_i - \bar{\xi}}{\sigma(\xi)} \\ \mathcal{T}^{-1} : \breve{\xi}_i \mapsto \xi_i = \sigma(\xi) \cdot \breve{\xi}_i + \bar{\xi} \end{cases} \tag{3.11}$$

avec

$$\begin{cases} \bar{\xi} = \dfrac{1}{N_{\text{train}}} \displaystyle\sum_{i=1}^{N_{\text{train}}} \xi_i \\ \sigma(\xi) = \sqrt{\dfrac{1}{N_{\text{train}} - 1} \displaystyle\sum_{i=1}^{N_{\text{train}}} (\xi_i - \bar{\xi})^2} \end{cases} \tag{3.12}$$

où N_{train} désigne le nombre de points de données dans l'échantillon d'entraînement.

Les valeurs initiales des poids et des biais du réseau de neurones ont besoin d'être définies afin d'éviter des problèmes relatifs à la vitesse d'apprentissage et la capacité de généralisation. L'initialisation retenue est l'initialisation de Glorot uniforme (Glorot et Bengio, 2010).

Parmi toutes les architectures ayant été testées, celle finalement retenue comprend trois couches cachées, contenant respectivement 64, 32 et 16 neurones. Une fonction d'activation ReLU est choisie pour les 3 couches cachées, et une fonction linéaire est choisie comme fonction d'activation à la sortie. Afin d'éviter le phénomène de sur-apprentissage et d'obtenir une meilleure capacité de généralisation du réseau de neurones (Hinton *et al.*, 2012), une couche de Dropout (paramètre $p = 0.15$) est ajoutée entre chaque couche cachée.

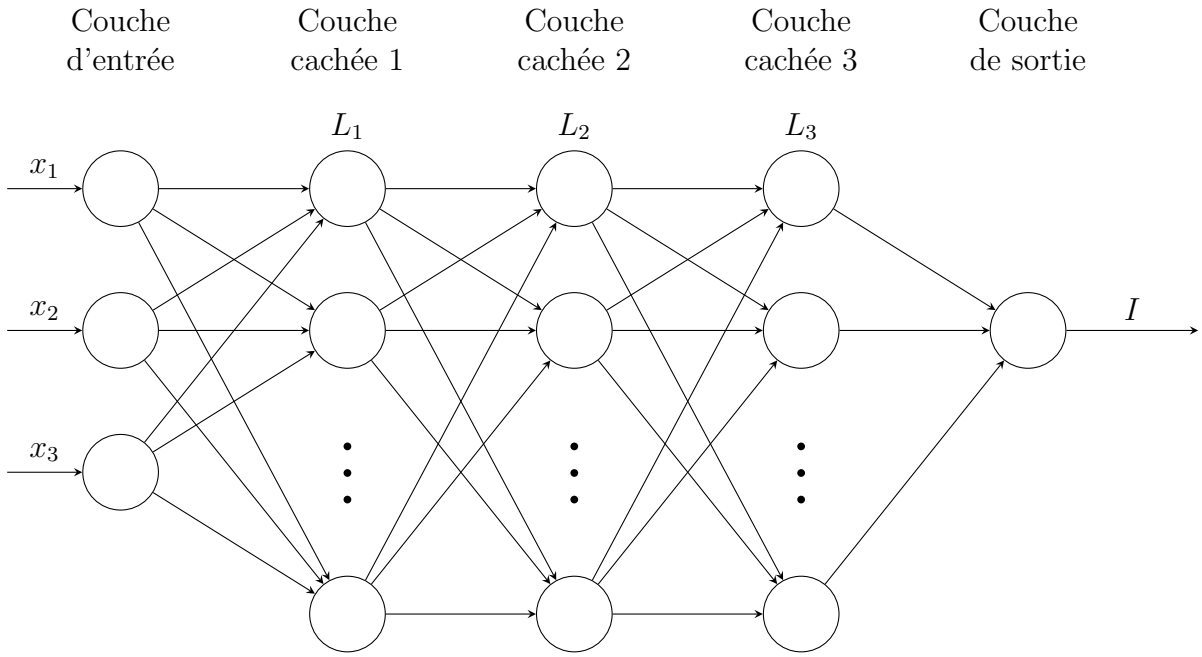


FIGURE 3.10. Architecture du réseau de neurones. Pour des raisons de clarté, toutes les connexions entre neurones ne sont pas représentées ; de plus, les couches de Dropout ne sont pas représentées.

Le nombre de fois où l'échantillon d'entraînement est utilisé pour entraîner le modèle (nombre d'époques) est de 30 et l'on suppose que le meilleur modèle est celui qui minimise l'erreur de validation. Un échantillon aléatoire de 10% de l'échantillon d'entraînement est utilisé comme échantillon de validation. En d'autres termes, le réseau de neurones est calibré sur 90% des données d'entraînement, et ses performances sont évaluées sur un échantillon de validation correspondant à 10% de l'échantillon d'entraînement.

Pour calibrer le modèle, la librairie `Keras` (Chollet et Allaire, 2018) est utilisé sur R (Team, 2018), l'optimiseur choisi étant `Adam` (Kingma et Ba, 2014). Par ailleurs, l'erreur MSE est la fonction de coût à minimiser.

3.3.4 Le modèle de régression linéaire comme baseline

Afin de comparer les performances de l'approche par réseau de neurones, nous nous appuyons également sur modèle de régression linéaire, définie pour chaque pays k et cohorte c par :

$$I(k,c) = \alpha_1 \mathcal{F}_{\Phi_k}(k,c) + \alpha_2 \mathcal{F}_{\Psi_k}(k,c) + \alpha_3, \tag{3.13}$$

où $\alpha = (\alpha_1, \alpha_2, \alpha_2)$ est le vecteur de paramètres. Il est à noter que cette régression linéaire fournit une correction qui ne dépend que de la variable de cohorte c et du pays k, contrairement à l'approche par réseau de neurones pour laquelle l'âge a est une variable d'entrée supplémentaire.

3.4 Résultats

Notre échantillon de données comprend les pays européens suivants : Espagne, Italie, France, Autriche, Danemark, Suède, Finlande, Norvège, pour la période temporelle 1935-2016, les cohortes 1915-1950 et les âges 18-70.

Il convient de souligner que les prédictions produites par un réseau de neurones sont variables. Ceci est une conséquence de l'algorithme d'optimisation qui initialise les poids du réseau de neurones à des valeurs aléatoires, de la sélection aléatoire des batches pendant l'entraînement pour calculer les gradients utilisés dans l'algorithme de rétro-propagation du gradient, et de la régularisation par Dropout qui est appliquée aléatoirement à certains neurones du réseau. Ainsi, les erreurs MSE d'un réseau de neurones sur l'ensemble de test varient d'un cycle d'entraînement à un autre. Nos résultats peuvent être reproductibles en définissant la valeur de départ du générateur de nombres aléatoires utilisé lors de la calibration du modèle. Néanmoins, entraîner un réseau de neurones une seule fois peut en général produire des résultats non représentatifs des performances moyennes du modèle.

Ainsi, dans la suite, chaque réseau de neurones est calibré dix fois et le MSE moyen est considéré pour indiquer la performance du modèle pour chaque échantillon de test. Les résultats sont présentés dans le tableau 3.1, où l'on peut voir que le réseau de neurones est le plus performant en matière de MSE moyen sur six des huit échantillons de test. Cependant, les résultats indiqués dans le tableau 3.1 sont les moyennes de dix exécutions d'un processus d'entraînement aléatoire. Comme indiqué précédemment, entraîner un réseau de neurones une seule fois au lieu de dix peut produire des résultats différents, comme le montrent les *outliers* de la boîte à moustaches sur la figure 3.11.

Échantillon de test		**Régression linéaire**	**Réseau de neurones**	
Pays	Code HMD	MSE	MSE moyen	MSE médian
Autriche	AUT	$1,11 \cdot 10^{-4}$	$1,21 \cdot 10^{-4}$	$1,20 \cdot 10^{-4}$
Danemark	DNK	$1,28 \cdot 10^{-4}$	$7,66 \cdot 10^{-5}$	$7,56 \cdot 10^{-5}$
Espagne	ESP	$6,34 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$	$7,45 \cdot 10^{-5}$
Finlande	FIN	$1,15 \cdot 10^{-4}$	$8,48 \cdot 10^{-5}$	$8,41 \cdot 10^{-5}$
France	FRATNP	$1,44 \cdot 10^{-4}$	$8,56 \cdot 10^{-5}$	$8,54 \cdot 10^{-5}$
Italie	ITA	$7,42 \cdot 10^{-5}$	$6,55 \cdot 10^{-5}$	$6,59 \cdot 10^{-5}$
Norvège	NOR	$9,81 \cdot 10^{-5}$	$6,19 \cdot 10^{-5}$	$6,09 \cdot 10^{-5}$
Suède	SWE	$6,14 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$	$4,52 \cdot 10^{-5}$

TABLEAU 3.1. Erreurs MSE obtenues sur les échantillons de test par la régression linéaire et l'approche par réseau de neurones

Afin de comparer les performances globales du réseau de neurones et de la régression linéaire, la moyenne des erreurs quadratiques moyennes correspondant à chaque échantillon de test est calculée.

On obtient une erreur MSE de prédiction moyenne de $9,95.10^{-5}$ pour la régression linéaire et $7,69.10^{-5}$ pour le réseau de neurones. Plus précisément, l'approche par réseau de neurones a permis d'améliorer l'erreur de prédiction MSE de 22.7% en moyenne, par rapport à l'approche par régression linéaire.

Une comparaison graphique des prédictions obtenues par les deux approches est disponible en annexe F.

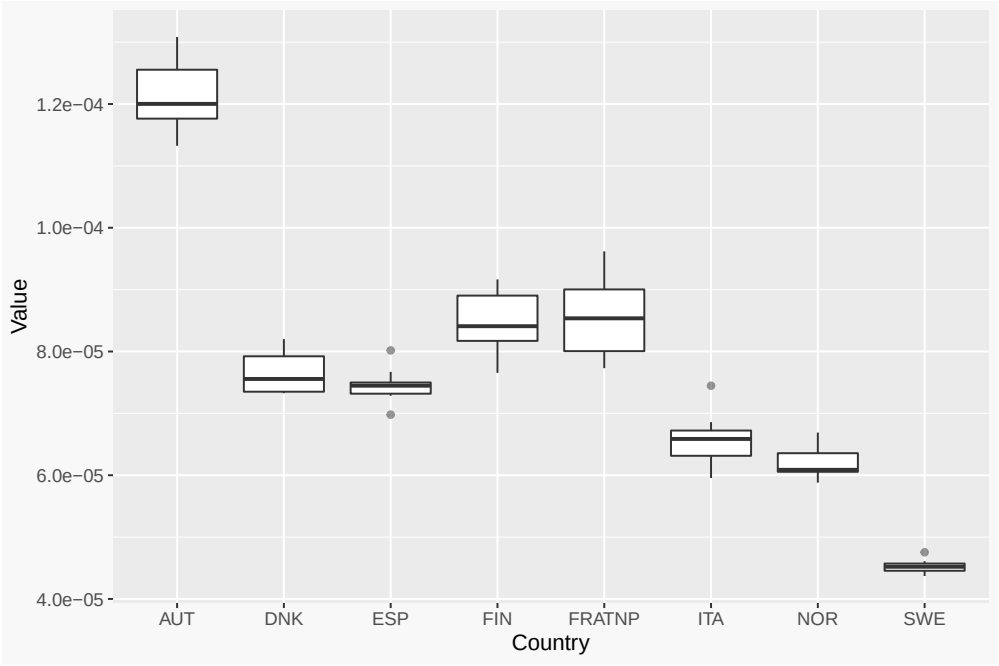


FIGURE 3.11. Boîtes à moustaches des erreurs MSE obtenues sur les échantillons de test. Le réseau de neurones est entraîné et testé à dix reprises pour chacun des échantillons de test.

3.5 Application de la méthode de correction

3.5.1 La matrice corrigée des taux d'amélioration de la mortalité

Comme mentionné en section 3.2, la méthodologie de correction de la HMD ne peut être appliquée qu'aux années de naissance pour lesquelles les données mensuelles de fertilité sont fournies par la HFD. Malheureusement, ce n'est pas le cas pour certains pays comme l'Allemagne de l'Ouest pour lesquels les données mensuelles de fertilité HFD sont fournies à partir de l'année 1946, limitant par conséquent l'ensemble des années pour lesquelles la méthodologie de correction est applicable.

La suite de cette section présente un exemple d'application de notre méthode de correction à des pays pour lesquels les données de naissance par mois ne sont pas disponibles. Les données de l'Allemagne de l'Ouest, de l'Ukraine, de la Biélorussie et de la Pologne (population des hommes) sont utilisées à partir de la version 5 de la HMD pour la période temporelle 1970-2000 et pour les âges 40-59. Notre méthode de correction est ensuite appliquée pour prédire les indicateurs de correction I_{LR} et I_{NN}, respectivement par régression linéaire et par approche par réseau de neurones (voir les figures 3.12 à 3.15). Les forces de mortalité estimées après correction μ_{LR} et μ_{NN} sont ensuite calculées, selon l'équation (3.4). Ainsi, la matrice des taux d'amélioration de mortalité de chaque pays $r_{\text{raw}}(a,t)$ (voir *e.g.* la figure 3.16 pour l'Allemagne de l'Ouest) peut être comparée avec les matrices corrigées par chacune des approches, $r_{\text{LR}}(a,t)$ et $r_{\text{NN}}(a,t)$ (voir les figures 3.17 à 3.18), définies par les relations suivantes :

$$r_{\text{LR}}(a,t) = \frac{\mu_{\text{LR}}(a,t+1) - \mu_{\text{LR}}(a,t)}{\mu_{\text{LR}}(a,t)} \text{ et } r_{\text{NN}}(a,t) = \frac{\mu_{\text{NN}}(a,t+1) - \mu_{\text{NN}}(a,t)}{\mu_{\text{NN}}(a,t)}. \tag{3.14}$$

La méthode de correction semble être très efficace concernant les faux effets cohorte isolés. En effet, sur les graphes des matrices de taux d'amélioration de la mortalité, les

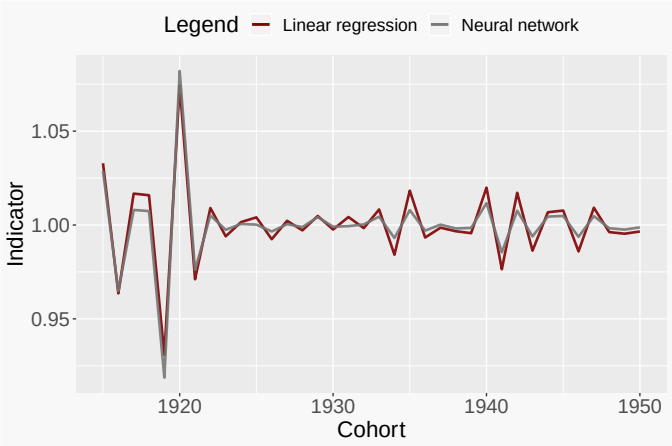


FIGURE 3.12. Indicateurs de correction prédits — Allemagne de l'Ouest

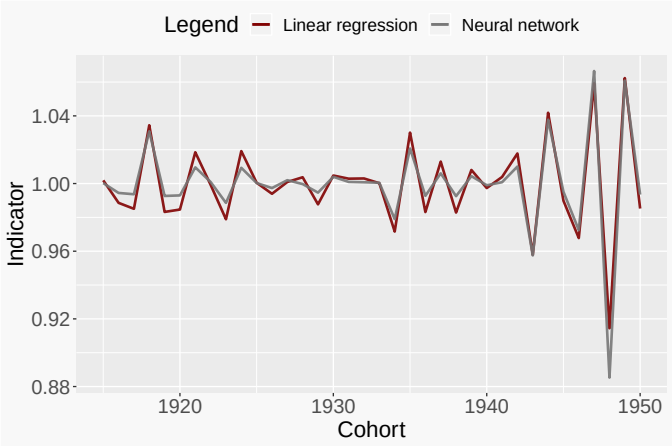


FIGURE 3.13. Indicateurs de correction prédits — Ukraine

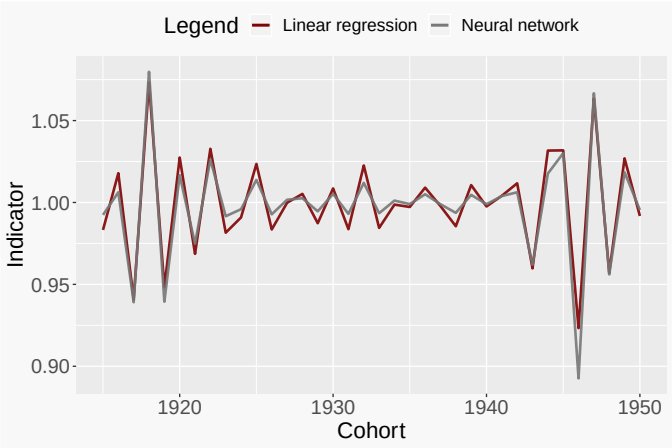


FIGURE 3.14. Indicateurs de correction prédits — Biélorussie

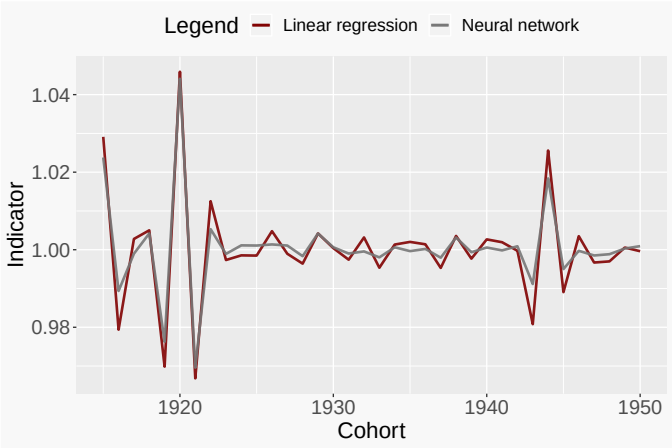


FIGURE 3.15. Indicateurs de correction prédits — Pologne

diagonales correspondant à de faux effets cohorte, présentes autour des années 1915 et 1920, ainsi que celles présentes autour de l'année 1940, disparaissent. Il s'avère que la volatilité des taux d'amélioration de la mortalité est considérablement réduite après retraitement des données de mortalité par la méthode de correction. En outre, à partir de la matrice des taux d'amélioration de la mortalité, on relève que les caractéristiques en période et en d'âge sont préservées dans les versions avec données corrigées. Cette propriété de la méthode de correction est intéressante, car elle permet de retrouver les principales caractéristiques de mortalité sans distorsion dans les zones où les diagonales correspondant à de faux effets cohorte ne sont pas présentes.

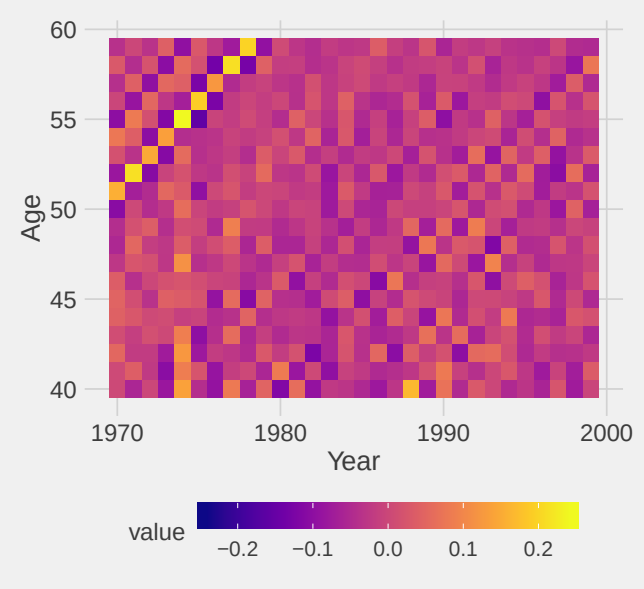


FIGURE 3.16. Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données brutes

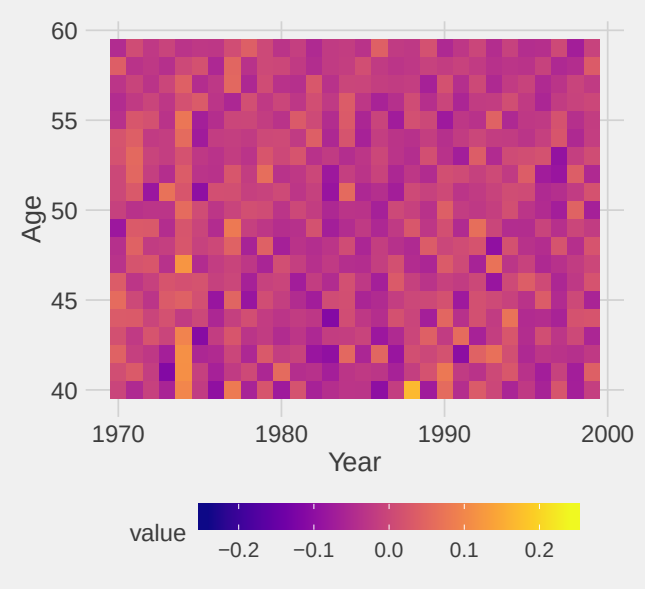


FIGURE 3.17. Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données corrigées par régression linéaire

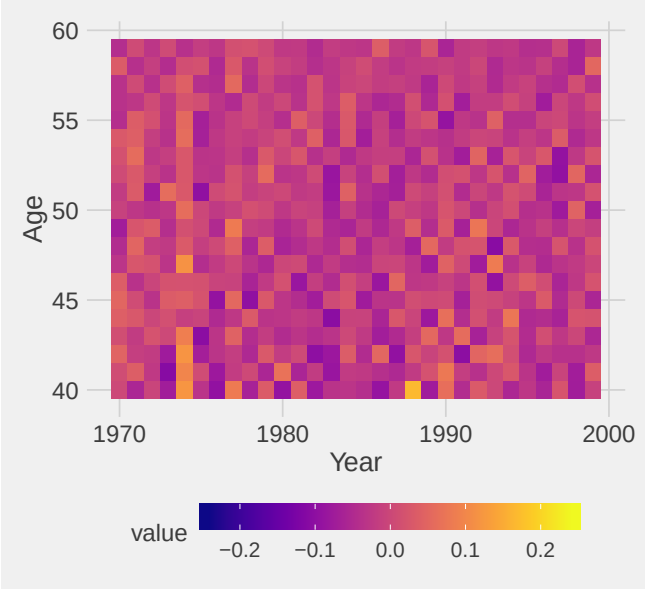


FIGURE 3.18. Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données corrigées par réseau de neurones

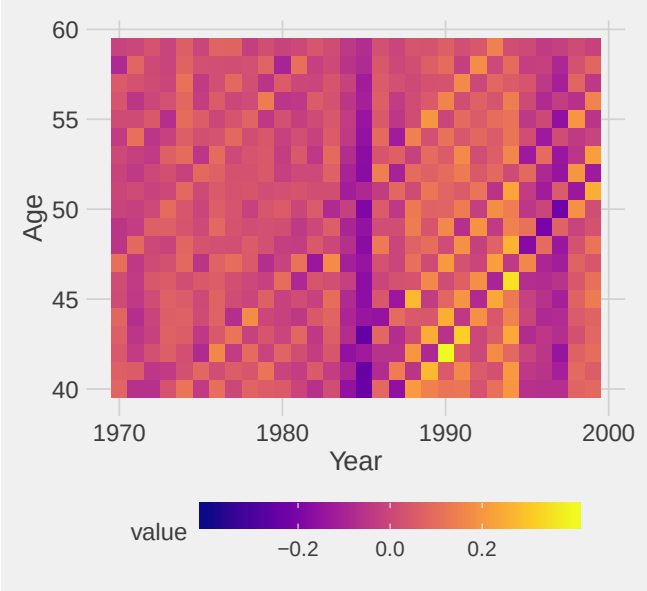


FIGURE 3.19. Taux d'amélioration de mortalité pour l'Ukraine — Données brutes

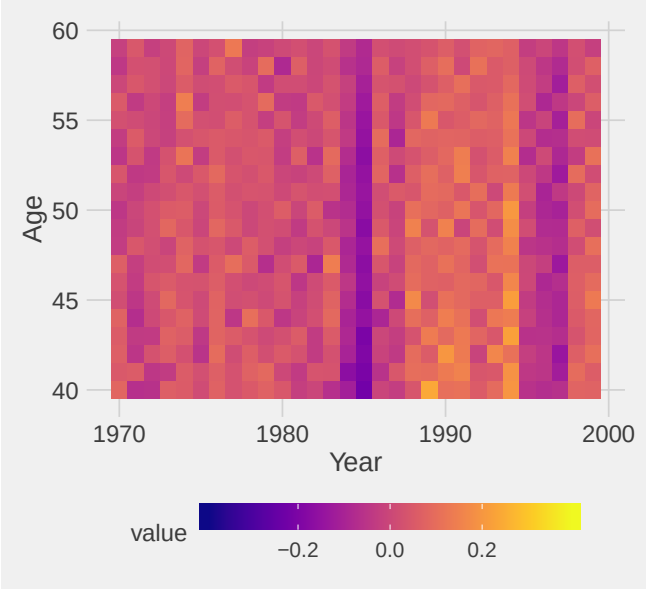


FIGURE 3.20. Taux d'amélioration de mortalité pour l'Ukraine — Données corrigées par régression linéaire

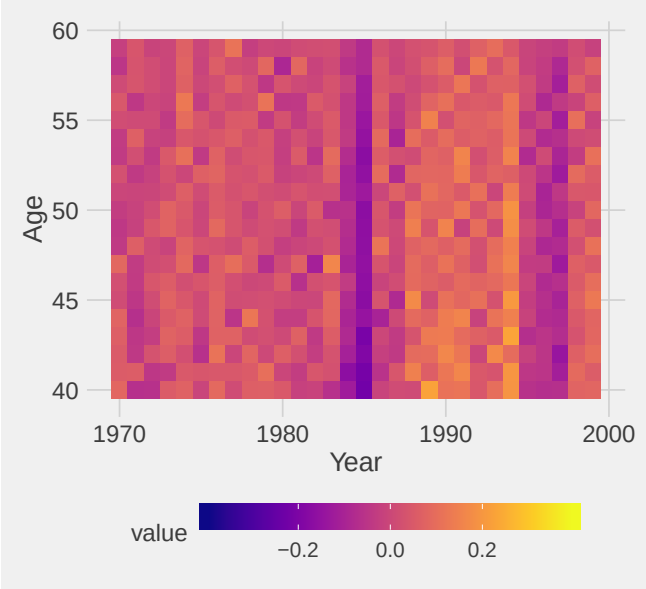


FIGURE 3.21. Taux d'amélioration de mortalité pour l'Ukraine — Données corrigées par réseau de neurones

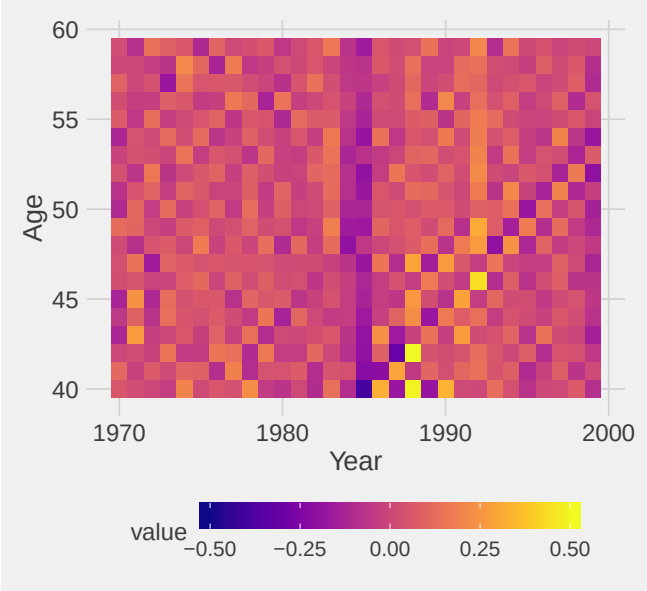


FIGURE 3.22. Taux d'amélioration de mortalité pour la Biélorussie — Données brutes

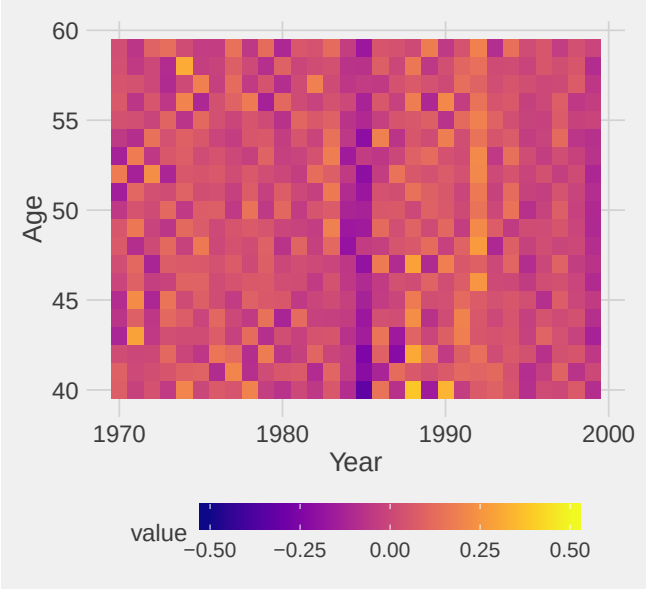


FIGURE 3.23. Taux d'amélioration de mortalité pour la Biélorussie — Données corrigées par régression linéaire

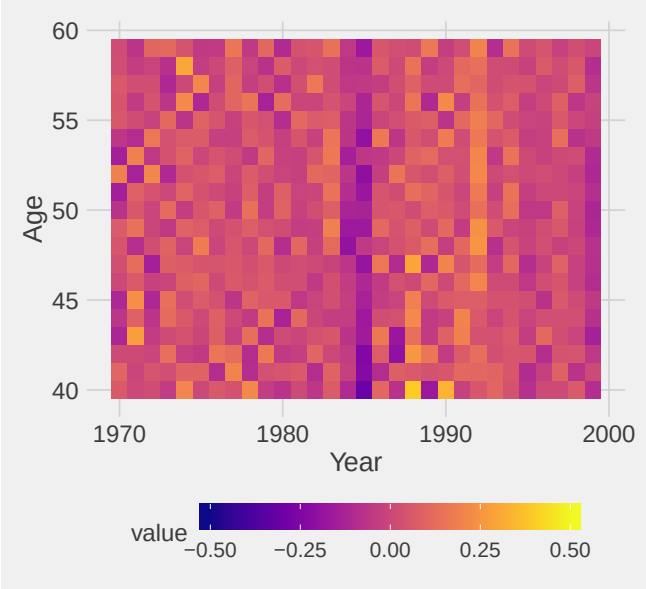


FIGURE 3.24. Taux d'amélioration de mortalité pour la Biélorussie — Données corrigées par réseau de neurones

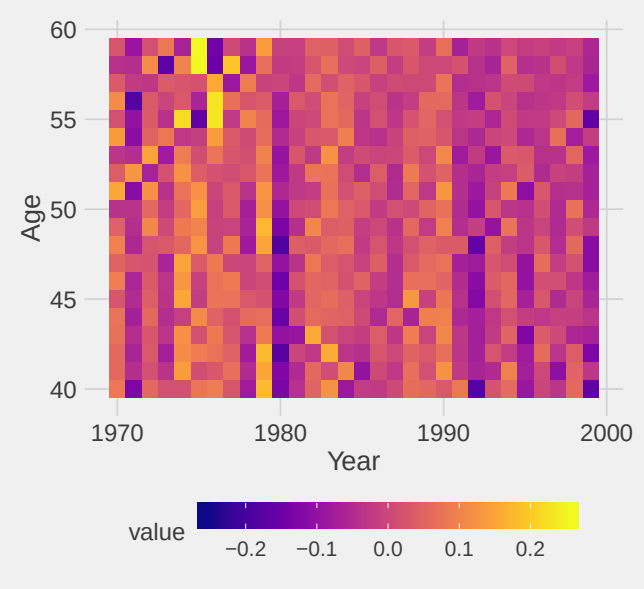


FIGURE 3.25. Taux d'amélioration de mortalité pour la Pologne — Données brutes

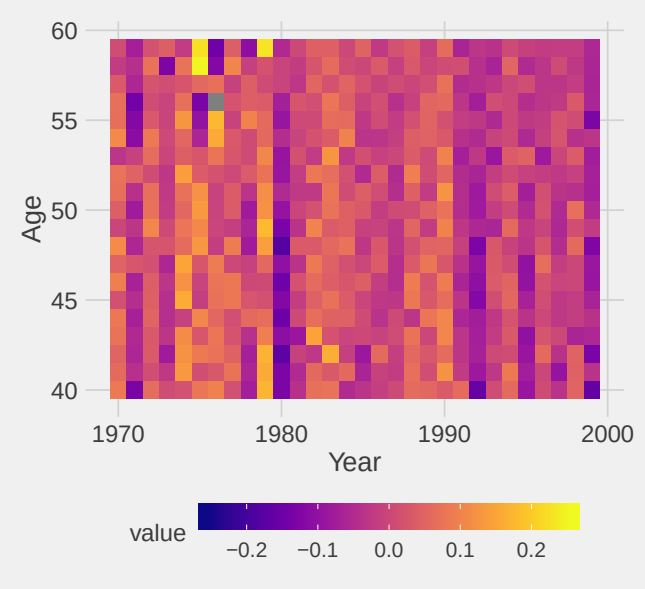


FIGURE 3.26. Taux d'amélioration de mortalité pour la Pologne — Données corrigées par régression linéaire

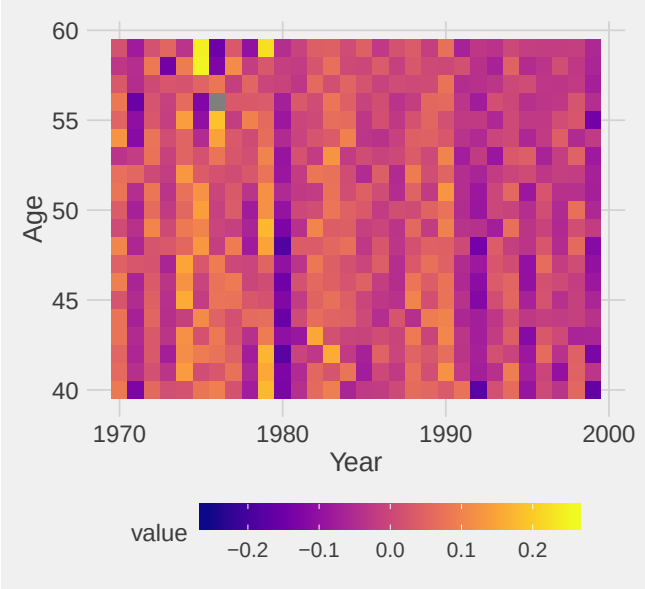


FIGURE 3.27. Taux d'amélioration de mortalité pour la Pologne — Données corrigées par réseau de neurones

3.5.2 Impact de la méthode de correction sur des modèles stochastiques de mortalité

La correction des données de mortalité en général, et la suppression des faux effets cohorte en particulier, a déjà été discutée dans le cadre de la modélisation stochastique de la mortalité par Cairns *et al.* (2016), Boumezoued (2016) et Boumezoued *et al.* (2019). Le but de cette sous-section est de montrer que les conclusions précédentes s'appliquent également à notre méthode de correction, et de discuter des perspectives supplémentaires. Dans la suite, les modèles M1 et M3, ainsi que le modèle de Plat simplifié sont considérés (voir les sections 1.3.1, 1.3.3 et 1.3.5). On rappelle que ces modèles s'écrivent sous les formes suivantes :

$$\eta^{\mathrm{M1}}(a,t) = \beta_a^{(1)} + \beta_a^{(2)} \cdot \kappa_t^{(2)}, \tag{3.15}$$

$$\eta^{\mathrm{M3}}(a,t) = \beta_a^{(1)} + \kappa_t^{(2)} + \gamma_{t-a}^{(3)}, \tag{3.16}$$

$$\eta^{\mathrm{Plat}}(a,t) = \beta_a^{(1)} + \kappa_t^{(2)} + \kappa_t^{(3)}(a-\bar{a}) + \gamma_{t-a}^{(4)}. \tag{3.17}$$

La calibration log-Poisson standard (Brouhns *et al.*, 2002) est considérée en utilisant la librairie R `StMoMo` (Villegas *et al.*, 2015). On considère l'intervalle d'âges $[\![40, 59]\!]$ et la période temporelle $[\![1970, 2000]\!]$. Comme le modèle M3 et le modèle de Plat simplifié comprennent des paramètres de cohorte et en accord avec la pratique standard (voir par exemple Cairns *et al.* (2009) et Haberman et Renshaw (2011)), toutes les cohortes ayant moins de trois observations sont exclues.

Lors de l'évaluation de la qualité de l'ajustement (*goodness-of-fit*) de différents modèles, il est généralement attendu que les modèles ayant plus de paramètres fournissent un meilleur ajustement aux données. Afin d'exclure la possibilité que le meilleur ajustement observé dans un modèle soit le résultat d'une paramétrisation excessive et afin de comparer les performances relatives de plusieurs modèles, il est devenu standard dans la littérature sur la mortalité d'utiliser des critères d'information qui modifient le critère du maximum de vraisemblance en pénalisant les modèles ayant plus de paramètres (Cairns *et al.*, 2009; Villegas *et al.*, 2015). Deux de ces critères sont le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC), une valeur inférieure de l'AIC et du BIC étant préférable. Ils sont définis comme $\mathrm{AIC} = 2\nu - 2\log\mathscr{L}$ et $\mathrm{BIC} = \nu\log\aleph - 2\log\mathscr{L}$, où ν désigne le nombre effectif de paramètres du modèle, $\mathscr{L}$ est la vraisemblance du modèle, et $\aleph$ est le nombre d'observations dans les données.

Les tableaux 3.2 à 3.5 ci-dessous représentent la quantité $-2\log\mathscr{L}$, l'AIC et le BIC des modèles stochastiques de mortalité considérés qui sont calibrés sur les données HMD brutes de l'Allemagne de l'Ouest, la Biélorussie, l'Ukraine et la Pologne, ainsi que sur les données homologues corrigées par notre méthode de correction.

Il apparaît que le modèle M1 est significativement amélioré lorsqu'il est calibré avec des données de mortalité corrigées. Cependant, aucun changement n'est observé pour le modèle M3 et le modèle de Plat simplifié. Cela semble s'expliquer par le fait que ces modèles ont une composante de cohorte qui saisit les irrégularités des tables de mortalité brutes, conduisant à des résidus inchangés entre les tables de mortalité brutes et corrigées. La stabilité en log-vraisemblance, AIC et BIC pour les modèles M3 et Plat simplifié montre également que notre méthode de correction ne perturbe pas les tables de mortalité, à l'exception de la composante de cohorte.

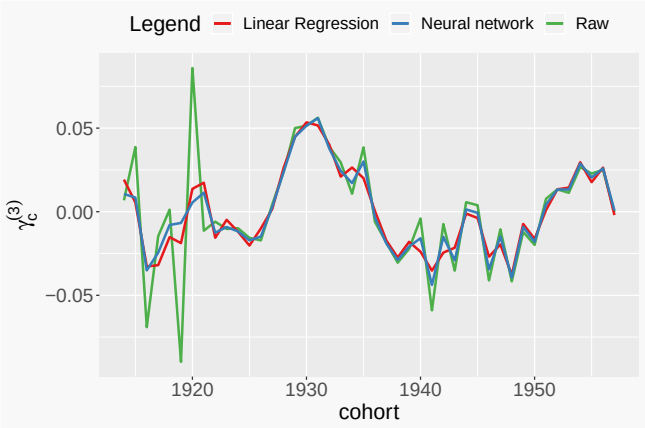


FIGURE 3.28. Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest

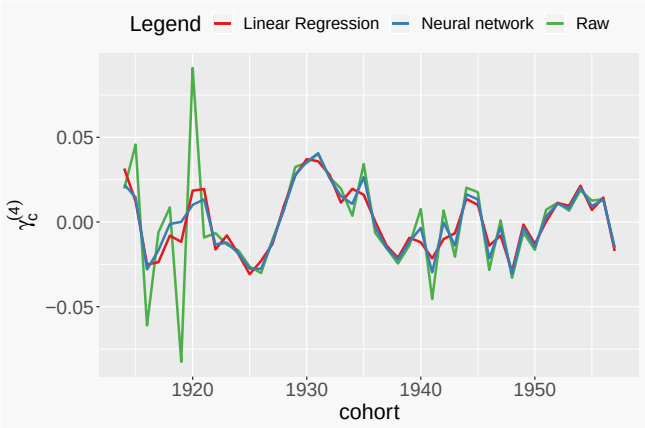


FIGURE 3.29. Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest

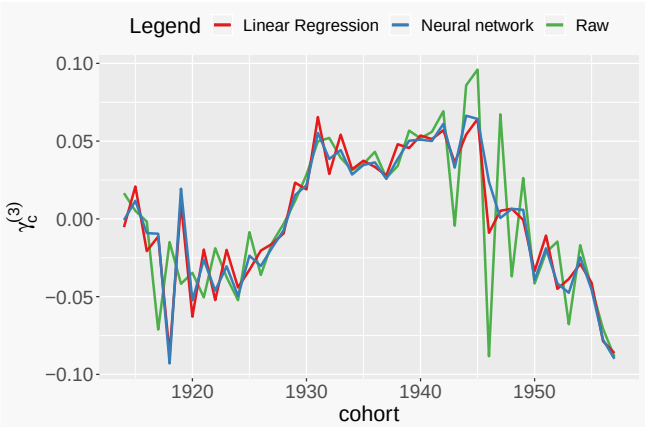


FIGURE 3.30. Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de la Biélorussie

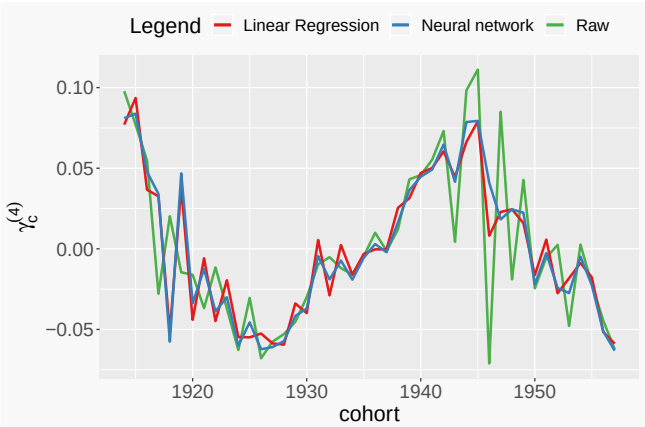


FIGURE 3.31. Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de la Biélorussie

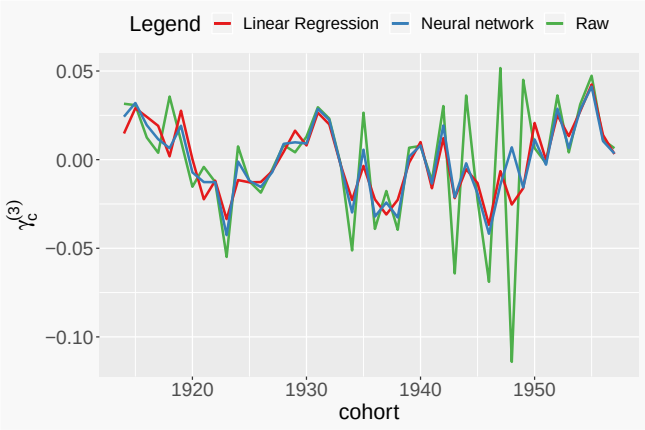


FIGURE 3.32. Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de l'Ukraine

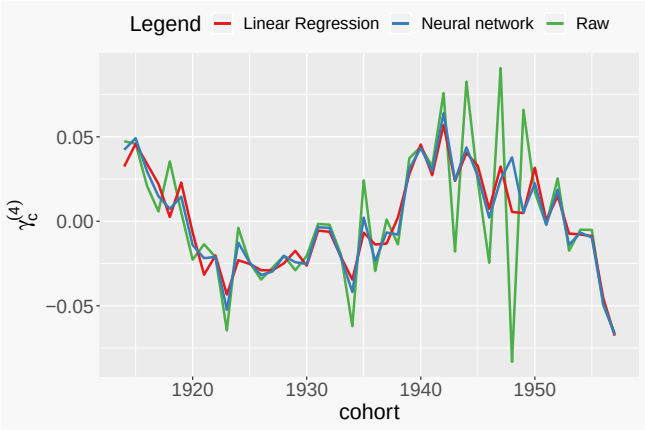


FIGURE 3.33. Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de l'Ukraine

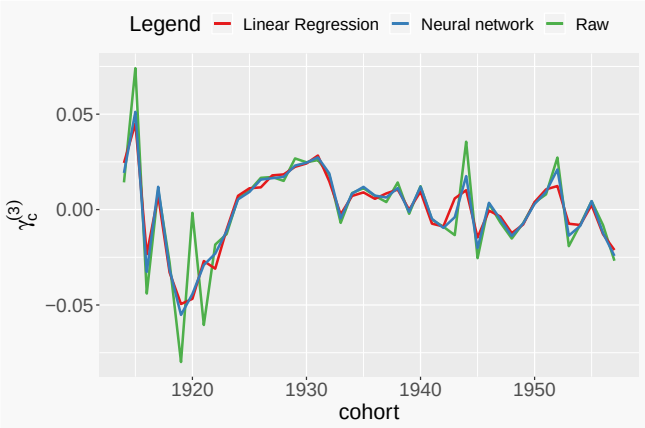


FIGURE 3.34. Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de la Pologne

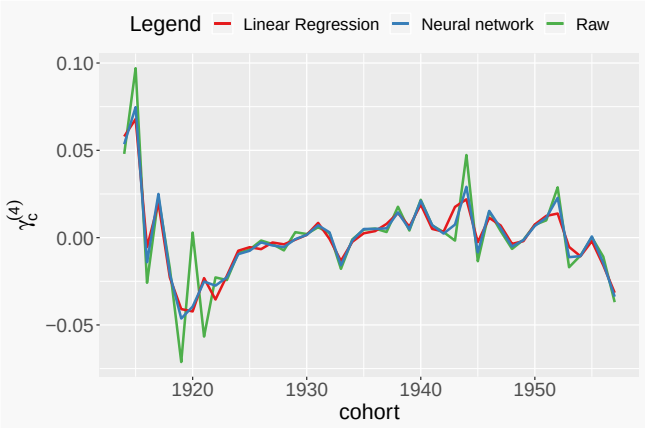


FIGURE 3.35. Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de la Pologne

Plus précisément, cinq conclusions apparaissent à la lumière des résultats obtenus :

- Pour les trois modèles, les log-vraisemblances obtenues sur les données de mortalité corrigées sont proches entre les deux approches de correction par régression linéaire et par réseau de neurones.
- La qualité d'ajustement du modèle M1 est significativement améliorée lorsqu'il est calibré avec des données de mortalité corrigées. Ainsi, notre méthode de correction a le même effet que la méthode de correction de Boumezoued *et al.* (2018) fondée sur des données mensuelles de fertilité. En outre, la log-vraisemblance demeure inchangée pour les modèles stochastiques de mortalité ayant une composante de cohorte (modèles M3 et Plat simplifié), ce qui est également en adéquation avec les résultats obtenus par Boumezoued *et al.* (2018).
- La dynamique des paramètres de cohorte des modèles M3 et Plat simplifié est beaucoup plus régulière lorsque ces modèles sont calibrés sur des données de mortalité corrigées. Par ailleurs, il est à souligner que le classement des modèles en matière de qualité d'ajustement peut même être inversé lorsque des données de mortalité corrigées sont

utilisées dans la calibration. Dans l'exemple des données ukrainiennes, le tableau 3.4 met en évidence que la qualité d'ajustement du modèle M1 devient meilleure que celle du modèle M3 lorsqu'il est calibré sur des données de mortalité corrigées. Ainsi, d'un point de vue statistique, il peut être dans certains cas plus pertinent d'intégrer dans un modèle stochastique de mortalité un paramètre de sensibilité spécifique à l'âge plutôt qu'un paramètre de sensibilité spécifique à la cohorte.

- L'impact de la calibration avec des données de mortalité corrigées sur les projections de mortalité futures a déjà été discuté par Cairns *et al.* (2016) et Boumezoued *et al.* (2018), et il ne s'agit pas de l'objet de ce chapitre. On s'attend à obtenir des résultats similaires avec cette nouvelle méthode de correction qui modifie la dynamique des paramètres gamma pour les modèles de mortalité avec une composante de cohorte. Rappelons que la volatilité du gamma projeté d'un modèle avec composante de cohorte est plus proche de la volatilité historique lors de l'utilisation des données de mortalité corrigées (Boumezoued, 2016). Il s'agit d'un phénomène important car cela montre que des modèles stochastiques de mortalité classiques sont capables de répliquer correctement les risques de longévité et de mortalité.
- Pour aller plus loin, notons que même pour le modèle de Plat simplifié qui présente une qualité d'ajustement particulièrement élevée quand il est calibré sur des données brutes, aucun changement n'est observé en matière de log-vraisemblance lorsqu'il est calibré sur des données de mortalité corrigées, alors que la dynamique du paramètre de cohorte gamma a été considérablement modifiée. À l'inverse, le modèle M1 est instable en matière de log-vraisemblance mais ses paramètres $\beta_a^{(1)}$, $\beta_a^{(2)}$ et $\kappa_t^{(2)}$ demeurent stables (voir l'annexe G pour une représentation graphique des paramètres du modèle M1 calibré sur des données de mortalité brutes et corrigées). Le modèle M3 et le modèle de Plat simplifié sont stables en matière de log-vraisemblance mais leurs paramètres de cohorte gamma sont instables (voir figures 3.28 à 3.35).

	$-2\log\mathscr{L}$			Critère BIC			Critère AIC		
Modèles de mortalité	Brut	LR	NN	Brut	LR	NN	Brut	LR	NN
M1	8252	7556	7585	8695	8000	8029	8390	7694	7723
M3	6515	6515	6514	7104	7104	7104	6699	6699	6698
Plat simplifié	6378	6378	6378	7153	7153	7153	6620	6620	6620

TABLEAU 3.2. Calibration sur les données de l'Allemagne de l'Ouest

	$-2\log\mathscr{L}$			Critère BIC			Critère AIC		
Modèles de mortalité	Brut	LR	NN	Brut	LR	NN	Brut	LR	NN
M1	6324	6128	6113	6768	6571	6557	6462	6266	6251
M3	5767	5767	5767	6357	6357	6357	5951	5951	5951
Plat simplifié	5583	5583	5583	6358	6358	6358	5825	5825	5825

TABLEAU 3.3. Calibration sur les données de la Biélorussie

	$-2\log\mathscr{L}$			Critère BIC			Critère AIC		
Modèles de mortalité	Brut	LR	NN	Brut	LR	NN	Brut	LR	NN
M1	10182	8532	8634	10625	8975	9078	10320	8670	8772
M3	8547	8547	8548	9137	9137	9138	8731	8731	8732
Plat simplifié	6940	6940	6941	7716	7716	7717	7182	7182	7183

TABLEAU 3.4. Calibration sur les données de l'Ukraine

	$-2\log\mathscr{L}$			Critère BIC			Critère AIC		
Modèles de mortalité	Brut	LR	NN	Brut	LR	NN	Brut	LR	NN
M1	7297	7115	7147	7741	7559	7591	7435	7253	7285
M3	6752	6752	6752	7341	7341	7341	6936	6936	6936
Plat simplifié	6567	6567	6567	7343	7343	7343	6809	6809	6809

TABLEAU 3.5. Calibration sur les données de la Pologne

3.6 Bilan de la méthode de correction et ouverture

Ce chapitre a proposé de développer une méthode de correction des estimations de la mortalité en l'absence de données mensuelles de fertilité.

La méthode a été testée sur des pays pour lesquels des données mensuelles de fertilité sont disponibles et les performances de la méthode ont été déterminées en calculant les erreurs MSE (pour *Mean-Squared Errors*) moyennes des prédictions de l'indicateur de correction, l'indicateur de correction observé étant obtenu par comparaison des données des versions 5 et 6 de la *Human Mortality Database*. Le modèle de base de régression linéaire donne des résultats satisfaisants en matière d'erreur de prédiction MSE moyenne, les résultats étant améliorés par l'approche par réseau de neurones.

La méthode a ensuite été appliquée à l'Allemagne de l'Ouest, à l'Ukraine, à la Biélorussie et à la Pologne, pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles pour la majorité des cohortes d'intérêt. Il s'est avéré que la volatilité des taux d'amélioration de la mortalité pour ces pays est considérablement réduite lors de l'utilisation de données de mortalité corrigées produites par la méthode de correction.

L'impact de la méthode de correction a également été étudié sur trois modèles stochastiques de mortalité : le modèle M1, le modèle M3 et le modèle de Plat simplifié. Pour chacun de ces trois modèles, les log-vraisemblances sur les données de mortalité corrigées sont proches entre la régression linéaire et le réseau de neurones. Le modèle M1 est significativement amélioré lorsqu'il est calibré avec des données de mortalité corrigées, tandis que la log-vraisemblance est inchangée pour les modèles stochastiques de mortalité ayant une composante de cohorte. Ainsi, notre méthode de correction a le même effet que les méthodes de correction fondées sur les données mensuelles de fertilité.

La dynamique des paramètres de cohorte des modèles M3 et Plat simplifié est beaucoup plus régulière lorsque ces modèles sont calibrés sur des données de mortalité corrigées. Il est à souligner que le classement des modèles en matière de qualité d'ajustement peut même être inversé dans certains cas lorsque des données de mortalité corrigées sont utilisées dans la calibration. Ainsi, d'un point de vue statistique, il peut être dans certains cas plus pertinent d'intégrer dans un modèle stochastique de mortalité un paramètre de sensibilité spécifique à l'âge plutôt qu'un paramètre de sensibilité spécifique à la cohorte.

Pour aller plus loin, notons que même pour le modèle de Plat simplifié qui présente une qualité d'ajustement particulièrement élevée quand il est calibré sur des données brutes, aucun changement n'est observé en matière de log-vraisemblance lorsqu'il est calibré sur des données de mortalité corrigées, alors que la dynamique du paramètre de cohorte gamma a été considérablement modifiée. À l'inverse, le modèle M1 est instable en matière de log-vraisemblance mais ses paramètres demeurent stables. Le modèle M3 et le modèle de Plat simplifié sont stables en matière de log-vraisemblance mais leurs paramètres de cohorte gamma demeurent instables.

Enfin, il convient de noter que notre méthode de correction peut être améliorée en ajoutant dans l'échantillon d'entraînement d'autres pays pour lesquels on dispose de données mensuelles de fertilité. De plus, des recherches plus approfondies sur l'architecture du réseau de neurones (nombres de couches différents, autres choix d'hyperparamètres), pourraient mener à des résultats plus optimaux que ceux présentés ici. D'autres améliorations possibles de notre méthode comprennent la considération de variables d'entrée supplémentaires ainsi que la considération des données de mortalité des femmes.

CHAPITRE 4

MODÉLISATION PRÉDICTIVE DE LA MORTALITÉ FUTURE

4.1 Introduction

Les prévisions de mortalité sont utilisées dans une grande variété de domaines : pour l'élaboration des politiques de santé, pour orienter la recherche pharmaceutique, la sécurité sociale, la planification des fonds de retraite et l'assurance-vie, pour ne citer que quelques-uns. Au cours du vingtième siècle, il est maintenant bien connu que la mortalité humaine a globalement diminué : dans la plupart des pays industrialisés, la mortalité chez les adultes et les personnes âgées révèle une probabilité de décès annuelle décroissante. (Benjamin et Soliman, 1993; Macdonald *et al.*, 1998) Ceci est bien illustré par la figure 4.1 qui représente le logarithme des estimations HMD de la force de mortalité des hommes de la population française en fonction de l'âge, sur la période temporelle 1950-2016.

Dans ce chapitre, notre approche consiste à discuter le modèle de réseau de neurones multi-populations de Richman et Wüthrich (2019) en comparant sa capacité prédictive avec celles de modèles stochastiques de mortalité ainsi qu'une approche par forêt aléatoire.

Il convient de bien noter que les approches décrites dans ce chapitre ne cherchent pas à incorporer des hypothèses sur les progrès de la science médicale ou des changements environnementaux spécifiques : aucune information autre que l'historique n'est prise en compte. Cela signifie que les modèles décrits sont incapables de prévoir des améliorations soudaines de la mortalité en raison de la découverte de nouveaux traitements médicaux comme des antibiotiques. De même, les futures détériorations causées par les épidémies, l'apparition de nouvelles maladies ou l'aggravation de la pollution ne peuvent pas entrer dans ces modèles. L'actuaire doit garder cela à l'esprit lorsqu'il établit son programme de réassurance.

4.2 Données et définition de l'échantillon de données

4.2.1 Données

La *Human Mortality Database* (HMD) est une initiative conjointe du département de démographie de l'Université de Californie à Berkeley aux États-Unis et de l'Institut Max Planck de recherche démographique de Rostock en Allemagne (Wilmoth et Shkolnikov, 2010). Elle a été créée en 2000 avec comme objectif de regrouper des données détaillées sur la mortalité et la population, et servir de référence pour toute personne s'intéressant à la

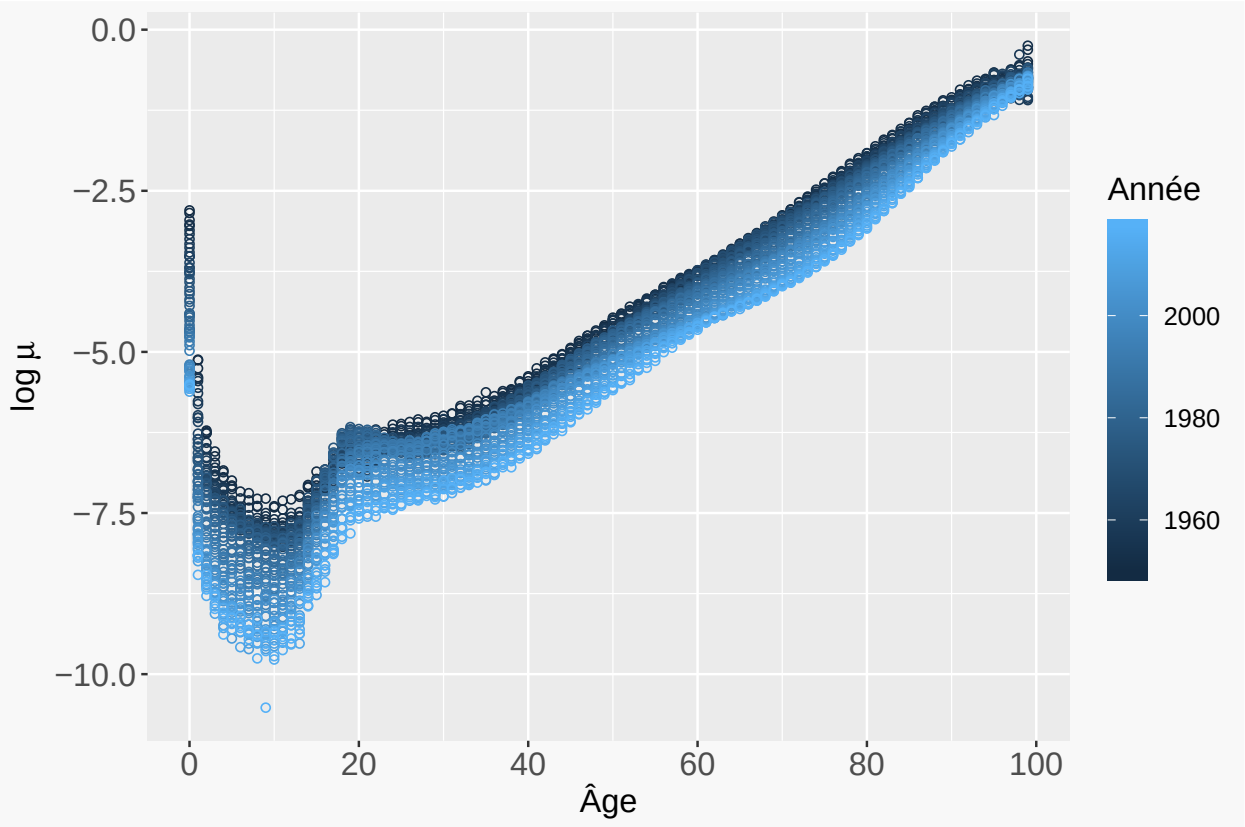


FIGURE 4.1. Logarithme de la force de mortalité des hommes en fonction de l'âge sur la période 1950-2017 — Données HMD de la France

longévité humaine. Elle contient des données sous forme de tables périodiques par année, âge et sexe, et également sous forme de table de cohortes. La profondeur de l'historique des tables périodiques en fonction du pays, est illustrée en figure 4.2. Pour les populations qui subissent des pertes de guerre substantielles, les estimations de la mortalité (en particulier des hommes) basées sur la population civile ne reflètent pas avec précision la période de mortalité en temps de guerre et ne représentent pas la véritable expérience de mortalité pour les générations qui subissent une perte de guerre importante. Dans ce cas, deux séries d'estimations sont disponibles : l'une pour la population civile et l'autre pour la population totale. La mortalité par période est identique dans les deux séries sauf pendant les années de guerre. Des tables de mortalité par cohorte pour la population civile sont fournies, mais ces données sont de mauvaise qualité pour les populations qui ont connu une mortalité de guerre substantielle car elles ne représentent pas la vraie mortalité de cohorte.

Des travaux ont été réalisés pour proposer des procédures d'amélioration de la qualité des données de mortalité, notamment des méthodes de calcul de l'exposition au risque à l'aide de données mensuelles de fertilité (voir le chapitre 3 de ce mémoire qui présente une méthode de correction de données de mortalité applicable en l'absence de données mensuelles de fertilité).

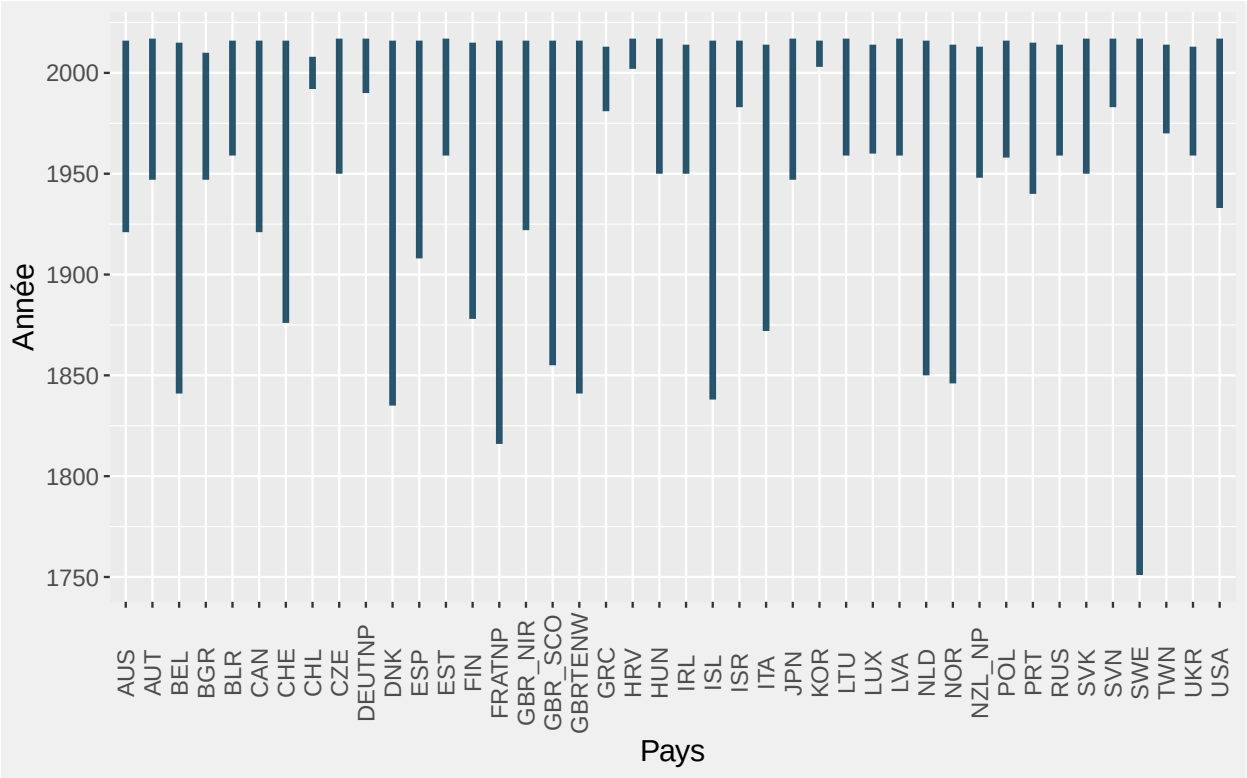


FIGURE 4.2. Disponibilité des données de mortalité de la HMD en fonction du pays

4.2.2 Définition de l'échantillon de données

L'échantillon de données est défini en plusieurs étapes.

Pays retenus L'intervalle d'âges considéré est l'intervalle $[\![0, 99]\!]$. Soient $\mathcal{E} = \{P_1, \ldots, P_N\}, N = 42$, l'ensemble des pays pour lesquels des données de mortalité sont disponibles dans la HMD et (t_i^p) la série des années pour lesquelles les données de mortalité du pays P_p sont disponibles dans la HMD. Parmi les 42 pays présents dans la HMD (voir le tableau C.3 en annexe C pour une liste détaillée), quatre pays ont été exclus :

- Croatie : les données sont disponibles uniquement après l'année 2002,
- Allemagne : les données sont disponibles uniquement après l'année 1990,
- Chili : les données sont disponibles uniquement après l'année 1992,
- Corée du Sud : les données sont disponibles uniquement après l'année 2003.

Ainsi, comme Richman et Wüthrich (2019), 38 des 42 pays disponibles dans la HMD sont inclus dans l'échantillon de données. Comme les observations concernent les genres masculin et féminin, 76 ensembles distincts de taux de mortalité seront prédits.

Remarque Les données HMD considérées pour la Nouvelle-Zélande par Richman et Wüthrich (2019) sont celles de la population néo-zélandaise non-Māorie, de code HMD NZL_NM. Dans la suite, nous avons préféré retenir les données de la population totale, de code NZL_NP.

Traitement des données manquantes de mortalité Richman et Wüthrich (2019) remplacent les valeurs manquantes de taux de mortalité dans les données (qui correspondent

à des valeurs NA ou nulles) par la moyenne des taux de mortalité sur l'ensemble des autres pays considérés pour la même année et le même genre. Cette technique de *Mean imputation* (imputation par la moyenne) consiste donc à remplacer une valeur manquante d'une variable par la moyenne de cette variable sur l'ensemble des données disponibles.

Échantillon de données Aussi, l'échantillon de données (*dataset*) se compose de n observations et s'écrit :

$$\mathcal{D} = \{(\mathbf{x}_i, \log \mu_i) \mid \mu_i \text{ est disponible et } a \in [\![0, 99]\!] \text{ et } t \in [\![1950, 1999]\!]\} \tag{4.1}$$

où $\mathbf{x}$ désigne le vecteur des variables d'entrée (covariables) de dimension D et le logarithme de la force de mortalité (estimée) $\log \mu$ est la variable scalaire de sortie ou cible.

Les entrées du vecteur colonne pour les n observations sont agrégées dans la matrice X de dimension $D \times n$ (*design matrix*), et les cibles sont regroupés dans le vecteur $\mathbf{M}$, de sorte que l'on puisse écrire $\mathcal{D} = (X, \mathbf{M})$.

Dans notre problème de régression, les cibles sont des valeurs réelles continues et positives. Ainsi, l'objectif est d'inférer la relation entre les inputs et les cibles, c'est-à-dire la distribution conditionnelle des cibles en fonction des inputs. En revanche, l'objectif n'est pas de modéliser la distribution des inputs elle-même.

4.2.3 Séparation de l'échantillon de données en un échantillon d'entraînement et un échantillon de test

Deux approches sont traitées par Richman et Wüthrich (2019) et sont nommées respectivement Round 1 et Round 2 dans la suite.

4.2.3.1 Round 1

L'échantillon de données $\mathcal{D}$ comprend les données HMD des pays sélectionnés précédemment pour les années de 1950 à 1999. Il est séparé en un échantillon d'entraînement $\mathcal{D}_{\text{train}}$ qui comprend les données de 1950 à 1990 (soit environ 80% de $\mathcal{D}$) et un échantillon de test $\mathcal{D}_{\text{test}}$ qui comprend les données de 1991 à 1999 (soit environ 20% de $\mathcal{D}$).

4.2.3.2 Round 2

L'échantillon de données $\mathcal{D}$ comprend les données HMD des pays sélectionnés précédemment pour les années de 1950 à 2016. Il est séparé en un échantillon d'entraînement $\mathcal{D}_{\text{train}}$ qui comprend les données de 1950 à 1999 (soit environ 75% de $\mathcal{D}$) et un échantillon de test $\mathcal{D}_{\text{test}}$ qui comprend les données de 2000 à 2016 (soit environ 25% de $\mathcal{D}$).

4.3 Comparaison des approches de prédiction de la mortalité

4.3.1 Modèles stochastiques de mortalité

4.3.1.1 Modèle M1

On rappelle que le modèle M1 (appelé aussi modèle de Lee-Carter) s'écrit sous la forme suivante (voir Section 1.3.1) :

$$\eta(a,t) = \alpha_a + \beta_a^{(1)} \cdot \kappa_t^{(1)}. \tag{4.2}$$

Pour chacun des 72 couples (pays, genre), un modèle M1 est calibré à l'aide de la librairie `StMoMo`. Un extrait simplifié de code R figure ci-dessous :

```
LC <- StMoMo::lc(link = "log")
LC_fit <- StMoMo::fit(LC, Dxt = dxt_mat, Ext = ext_mat,
                      ages = ages, years = years, verbose = FALSE)
```

`Dxt` désigne la matrice des données de décès d'intérêt, `Ext` la matrice des données d'exposition d'intérêt, `ages` le vecteur des âges et `years` le vecteur des années.

La force de mortalité est ensuite projetée sur les années correspondant à l'échantillon de test $\mathcal{D}_{\text{test}}$, en modélisant la série des $\kappa_t^{(1)}$ par un processus ARIMA (pour *Auto-Regressive Integrated Moving Average*) dont les ordres sont choisis par la méthode de Box-Jenkins (Pankratz, 2009). Ceci est réalisé grâce à la fonction `forecast` en définissant l'argument `kt.method = "iarima"` :

```
LC_for <- forecast(LCfit, h = year_max - last_year_train,
                   kt.method = "iarima", kt.order = NULL)
```

En choisissant `kt.order = NULL`, la fonction `auto.arima` de la librairie `forecast` sélectionne le meilleur processus ARIMA pour chaque $\kappa_t^{(1)}$.

4.3.1.2 Modèle M3

On rappelle que le modèle M3 (appelé aussi modèle APC) s'écrit sous la forme suivante (voir Section 1.3.3) :

$$\eta(a,t) = \alpha_a + \kappa_t^{(1)} + \gamma_{t-a}, \tag{4.3}$$

Pour chacun des 72 couples (pays, genre), un modèle M3 est calibré à l'aide de la librairie `StMoMo`. Un extrait simplifié de code R figure ci-dessous :

```
nb_cohort_out <- 3
APC <- StMoMo::apc(link = "log")
wxt <- genWeightMat(ages = ages,  years = years, clip = nb_cohort_out)

APC_fit <- StMoMo::fit(APC, Dxt = dxt_mat, Ext = ext_mat,
                       ages = ages, years = years,
                       wxt = wxt, verbose = FALSE)
```

Comme le modèle M3 comprend un paramètre de cohorte et en accord avec la pratique usuelle (voir *e.g.* Cairns *et al.* (2009) et Haberman et Renshaw (2011)), les cohortes ayant moins de trois observations sont exclues. Ainsi, les trois premières et les trois dernières cohortes sont exclues de la calibration via la matrice de pondération appropriée `wxt`. Cette matrice

contient uniquement des 0 et des 1 et est appliquée à l'aide de la fonction `genWeightMat` de la librairie `StMoMo`.

La force de mortalité est ensuite projetée sur les années correspondant à l'échantillon de test $\mathcal{D}_{\text{test}}$, en modélisant la série des $\kappa_t^{(1)}$ par un processus ARIMA dont les ordres sont choisis par la méthode de Box-Jenkins, la série des γ_{t-a} étant modélisée par un processus ARIMA$(1,1,0)$:

```
APC_for <- forecast(APCfit, h = year_max - last_year_train,
                    gc.order = c(1, 1, 0))
```

4.3.2 Forêt aléatoire

On considère les variables d'entrée suivantes :

- année t,
- âge a,
- cohorte c,
- genre g.

Pour chacun des 36 pays, on calibre une forêt aléatoire à deux populations (Hommes et Femmes) sur l'échantillon d'entraînement $\mathcal{D}_{\text{train}}$ à l'aide de la librairie `randomForest`, en choisissant : `ntrees` $= 300$ (nombre d'arbres). Comme cette procédure a un coût calculatoire élevé, le nombre d'arbres est limité à 300. Ce choix permet de garantir à la fois un pourcentage de variance adéquat expliqué par le modèle et une faible moyenne des résidus au carré.

4.3.3 Réseau de neurones

4.3.3.1 Variables retenues

On considère les variables d'entrée suivantes :

- année t : variable continue
- âge a : variable catégorielle
- genre $g \in \{\text{Homme}, \text{Femme}\}$, variable catégorielle binaire correspondent aux genres Homme et Femme.
- variable pays k : variable catégorielle

Par ailleurs, la variable de sortie retenue est le logarithme de la force de mortalité estimée, notée $\log \mu$ pour simplifier la notation (voir la section 1.1 pour plus de précisions). Ainsi, $\mathbf{x} = (t, a, g, k)$, $y = \log \mu$ et $D = 4$.

4.3.3.2 Couches embedding pour les variables d'entrée catégorielles

Afin de bien gérer chaque variable d'entrée catégorielle a, g, k, on considère l'application embedding (Guo et Berkhahn, 2016) :

$$e_i : x_i \mapsto \mathbf{x}_i. \tag{4.4}$$

Cette application correspond à ajouter une couche de neurones supplémentaire après la couche d'entrée. On définit l'encoding *one-hot* de x_i comme l'application :

$$u_i : x_i \mapsto \delta_{x_i \alpha}, \tag{4.5}$$

où δ est le symbole de Kronecker. Si ν_i est le nombre de valeurs pour la variable catégorielle x_i, alors $\delta_{x_i\alpha}$ est un vecteur de longueur ν_i, où l'élément est différent de 0 pour $\alpha = x_i$. La sortie de la couche supplémentaire de neurones est définie par :

$$\mathbf{x}_i \equiv \sum_{\alpha} w_{\alpha\beta} \cdot \delta_{x_i\alpha} = w_{x_i\beta}, \tag{4.6}$$

où $w_{x_i\beta}$ est le poids qui relie la couche d'encoding *one-hot* à la couche embedding et β est l'indice de la couche embedding. Ainsi, les embeddings correspondent aux poids de cette couche et peuvent être appris de la même façon que les paramètres des autres couches du réseau de neurones.

Après avoir utilisé des couches embedding pour représenter les variables catégorielles d'entrée, les couches embedding et les variables d'entrée continues sont concaténées. La couche fusionnée est traitée comme une couche d'entrée standard et les couches suivantes peuvent être construites à partir de cette couche. Dans cette méthode, la couche embedding apprend les propriétés intrinsèques de chaque catégorie.

Les dimensions des couches embedding ζ_i sont des hyperparamètres qui doivent être prédéfinis. Les dimensions des entités embedding sont à choisir dans l'intervalle $[1, \nu_i - 1]$ où ν_i est le nombre de valeurs prises par la variable catégorielle x_i.

Les choix suivants ont été effectués :

- âge : $\zeta_a = 5$,
- pays : $\zeta_k = 5$,
- genre : $\zeta_g = 1$.

Remarque : Richman et Wüthrich (2019) choisissent de prendre $\zeta_g = 5$, choix qui n'est pas retenu dans la suite car le nombre de valeurs pris par la variable g est de 2.

4.3.3.3 Normalisation des variables

Les variables continues utilisées n'ont pas des valeurs du même ordre de grandeur. Des écarts importants peuvent nuire aux algorithmes d'apprentissage en les rendant inefficaces. Ainsi, on effectue un changement de variable en remplaçant chacune d'entre elles par leur transformée *Min-Max* $\breve{\xi}_i$ dans $[0, 1]$ (*Min-Max transformation*). Soit ξ une variable d'entrée continue du modèle. La transformation pour la variable ξ de l'exemple i, notée ξ_i, s'écrit :

$$\begin{cases} \mathcal{T} : \xi_i \mapsto \breve{\xi}_i = \dfrac{\xi_i - m}{M - m} \\ \mathcal{T}^{-1} : \breve{\xi}_i \mapsto \xi_i = (M - m) \cdot \breve{\xi}_i + m \end{cases} \tag{4.7}$$

avec

$$\begin{cases} M = \max\limits_{1 \leq i \leq N_{\text{train}}} \xi_i \\ m = \min\limits_{1 \leq i \leq N_{\text{train}}} \xi_i \end{cases} \tag{4.8}$$

où N_{train} désigne le nombre de points de données dans l'échantillon d'entraînement.

Par ailleurs, les variables d'entrée catégorielles ne subissent aucune normalisation. Il est à noter que cette normalisation de variables n'est pas mentionnée par Richman et Wüthrich (2019). Cependant, la fonction sigmoïde utilisée en sortie est à valeurs dans [0,1] et est donc adaptée dans le cas où la variable de sortie est dans [0,1], ce qui est le cas de la variable $\log \mu$ normalisée par la transformation Min-Max, mais pas de la variable $\log \mu$ non normalisée.

4.3.3.4 Initialisation des poids

Les valeurs initiales des poids et des biais du réseau de neurones ont besoin d'être définies afin d'éviter des problèmes relatifs à la vitesse d'apprentissage et la capacité de généralisation. L'initialisation retenue est l'initialisation de Glorot uniforme (Glorot et Bengio, 2010).

4.3.3.5 Architecture du réseau de neurones

Parmi toutes les architectures ayant été testées par Richman et Wüthrich (2019), l'architecture DEEP6 est celle qui donne les meilleurs résultats en matière d'erreur MSE sur l'échantillon de test. Cette architecture a les caractéristiques suivantes :

- cinq couches cachées ayant chacune 128 neurones,
- une fonction d'activation tanh,
- une fonction de sortie sigmoïde,
- une combinaison de couches de régularisation Dropout (paramètre $p = 0.05$) et BatchNorm afin d'éviter le phénomène de sur-apprentissage,
- deux "skip connections" qui relient la couche de sortie à la dernière couche cachée intermédiaire, ainsi qu'à la première couche cachée,
- technique d'apprentissage : descente de gradient version mini-batch avec une taille de mini-batch de 128.

4.3.3.6 Autres choix d'hyperparamètres

Le nombre d'époques maximal choisi est de 50 et le meilleur modèle retenu est celui qui minimise l'erreur de validation. Un échantillon aléatoire de 10% de l'échantillon d'entraînement est utilisé comme échantillon de validation. En d'autres termes, le réseau de neurones est calibré sur 90% des données d'entraînement, et sa performance est évaluée sur 10% de l'échantillon d'entraînement (échantillon de validation). La valeur de 10% au lieu des 5% de Richman et Wüthrich (2019) est retenue afin de diminuer les fluctuations des erreurs de validation pendant la période d'apprentissage. Cela permet aussi de réduire les fluctuations des erreurs de test obtenues.

Pour calibrer le modèle, le package `Keras` de R (Chollet, 2015) est utilisé. On utilise l'optimiseur `Adam` (Kingma et Ba, 2014), avec l'erreur quadratique moyenne (MSE) comme fonction de coût à minimiser et une adaptation du learning rate. Le *learning rate* global est multiplié par $\frac{1}{2}$ si l'erreur de validation ne s'améliore pas sur une durée de 4 époques consécutives. L'apprentissage est arrêté si l'erreur de validation ne s'améliore pas après 5 époques (voir l'annexe D pour la procédure détaillée d'Early Stopping).

4.4 Résultats comparatifs intermédiaires

Les résultats obtenus en matière d'erreur MSE sur la force de mortalité estimée μ sont similaires à ceux obtenus par Richman et Wüthrich (2019) et sont consignés dans le tableau 4.1. Ce tableau montre que pour le Round 1 et le Round 2, les erreurs MSE obtenues avec l'approche par réseau de neurones sont nettement plus faibles que celles obtenues avec le modèle de Lee-Carter. Elles sont également plus faibles que celles obtenues avec le modèle M3 et l'approche par forêt aléatoire mais avec un gain d'erreur beaucoup plus faible. Si l'on

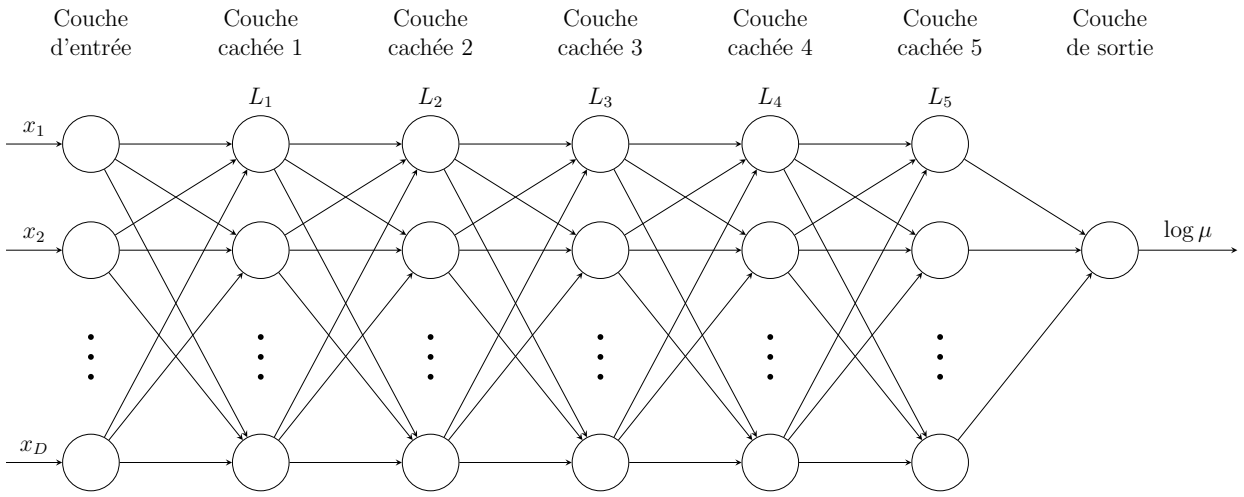


FIGURE 4.3. Architecture du réseau de neurones. Pour des raisons de clarté, toutes les connexions entre neurones ne sont pas représentées ; de plus, les couches de Dropout et de BatchNorm ne sont pas représentées.

exclut le modèle M3 dans la comparaison, la conclusion suivante de Richman et Wüthrich (2019) (qui ne considèrent pas le modèle M3 dans leur approche) semble donc être vérifiée :

"for the purposes of large scale mortality forecasting, deep neural network architectures dramatically outperform traditional single (...) forecasting models".

	Erreurs sur le test (Round 1)	**Erreurs sur le test (Round 2)**
Modèle utilisé	MSE sur μ	MSE sur μ
M3	$5,77 \cdot 10^{-4}$	$4,93 \cdot 10^{-4}$
M1	$9,04 \cdot 10^{-4}$	$5,95 \cdot 10^{-4}$
Réseau de neurones	$5,42 \cdot 10^{-4}$	$3,03 \cdot 10^{-4}$
Forêt aléatoire	$6,26 \cdot 10^{-4}$	$4,63 \cdot 10^{-4}$

TABLEAU 4.1. Comparaison des résultats obtenus par les modèles

4.5 Discussion

4.5.1 Choix de métrique et critères de comparaison

Richman et Wüthrich (2019) choisissent d'utiliser la MSE sur la force de mortalité estimée μ comme métrique d'erreur. Dans cette section, le choix d'une nouvelle métrique sera discuté en commençant par un exemple illustratif très simple.

Supposons que l'échantillon de test soit réduit à deux points A et B, de mortalités réalisées respectives $\mu_A = 0.002$ et $\mu_B = 0.2$. Supposons de plus que les mortalités prédites par le modèle soient respectivement $\tilde{\mu}_A = 0.0018$ et $\tilde{\mu}_B = 0.18$ (soit une erreur relative de 10% pour les deux points A et B). Alors l'erreur MSE sur l'échantillon de test, moyenne des erreurs au carré de A et B, s'écrit :

$$\text{MSE} = \frac{1}{2}(\text{erreur}_A^2 + \text{erreur}_B^2), \tag{4.9}$$

avec $\text{erreur}_A^2 = 4.10^{-8}$ et $\text{erreur}_B^2 = 4.10^{-4} \gg 4.10^{-8} = \text{erreur}_A^2$. Ainsi, les erreurs des points ayant des forces de mortalité élevées en valeur absolue ont un impact prédominant sur l'erreur MSE calculée.

Comme la variable de sortie dans les modèles de mortalité considérés est $\log \mu$ et non μ, il semble plus naturel de considérer comme métrique l'erreur MSE sur $\log \mu$ (que nous appelons LMSE), plutôt que sur μ. En reprenant l'exemple précédent avec cette nouvelle métrique, on obtient cette fois :

$$\text{LMSE} = \frac{1}{2}(\text{log_erreur}_A^2 + \text{log_erreur}_B^2), \tag{4.10}$$

avec $\text{log_erreur}_A^2 = (\log 0.002 - \log 0.0018)^2 \approx 0.0111$ et $\text{log_erreur}_B^2 = \text{log_erreur}_A^2$, cette dernière égalité découlant simplement de la propriété fondamentale du logarithme : $\forall \alpha, \beta > 0, \log(\alpha\beta) = \log \alpha + \log \beta$. Le tableau 4.2 consigne les résultats obtenus dans cet exemple illustratif.

Point	μ	$\tilde{\mu}$	$\log \mu$	$\log \tilde{\mu}$	erreur^2	log_erreur^2
A	0.002	0.0018	-6.21	-6.32	4.10^{-8}	0.0111
B	0.2	0.18	-1.61	-1.71	4.10^{-4}	0.0111

TABLEAU 4.2. Exemple illustratif : comparaison des métriques MSE et LMSE

Le tableau 4.3 reprend le tableau 4.1 en incluant la nouvelle métrique LMSE. Le réseau de neurones ne donne plus les meilleurs résultats en matière d'erreur LMSE pour le Round 1 et le Round 2. C'est la forêt aléatoire qui donne les erreurs LMSE les plus faibles, à la fois pour le Round 1 et le Round 2.

Ainsi, la capacité prédictive de l'approche par réseau de neurones est inférieure à celle de l'approche par forêt aléatoire, à la fois pour le Round 1 et le Round 2, au regard de la LMSE, nouvelle métrique considérée.

Âges 0-99	**Erreurs sur le test (Round 1)**		**Erreurs sur le test (Round 2)**	
Modèle utilisé	MSE	LSME	MSE	LMSE
M3	$5,77 \cdot 10^{-4}$	$1,09 \cdot 10^{-1}$	$4,93 \cdot 10^{-4}$	$1,42 \cdot 10^{-1}$
M1	$9,04 \cdot 10^{-4}$	$8,34 \cdot 10^{-2}$	$5,95 \cdot 10^{-4}$	$1,47 \cdot 10^{-1}$
Réseau de neurones	$5,42 \cdot 10^{-4}$	$7,50 \cdot 10^{-2}$	$3,03 \cdot 10^{-4}$	$1,22 \cdot 10^{-1}$
Forêt aléatoire	$6,26 \cdot 10^{-4}$	$6,35 \cdot 10^{-2}$	$4,63 \cdot 10^{-4}$	$1,06 \cdot 10^{-1}$

TABLEAU 4.3. Comparaison des erreurs en incluant la métrique LMSE

4.5.2 Discussion sur les âges très élevés retenus dans l'échantillon de données

La figure 4.4 représente un nuage de points des estimations HMD des forces de mortalité de l'échantillon de données en fonction de l'âge. Elle met en évidence le fait que pour les âges très élevés (supérieurs à 95 ans), il existe des points de données correspondant à des

forces de mortalité estimées supérieures à 1. Ainsi, on a par exemple pour le point $x_0 = (g = \text{homme}, k = \text{Slovénie}, a = 99, t = 1995)$:

$$\hat{\mu}_{x_0} = \frac{d_{x_0}}{E_{x_0}} = \frac{6}{3.5} \approx 1.714 \tag{4.11}$$

Cette donnée de force de mortalité (réalisée) $\hat{\mu}_{x_0}$ étant une estimation de la véritable force de mortalité μ_{x_0}, il est intéressant de calculer l'intervalle de confiance (IC) à 95% pour μ_{x_0}. On a (Scott, 1981) :

$$\mu_{x_0} \approx \hat{\mu}_{x_0} \pm 1.96\sqrt{\frac{\hat{\mu}_{x_0}}{E_{x_0}}} = \hat{\mu}_{x_0}(1 \pm \frac{1.96}{\sqrt{d_{x_0}}}) \approx 1.714 \pm 1.96 \times 0.700. \tag{4.12}$$

L'IC à 95% est donc approximativement $[0.343, 3.086]$. Il apparaît donc que le coefficient de variation est extrêmement élevé, ce qui implique une incertitude énorme sur μ_{x_0}. Dans la suite de l'analyse comparative, il est donc choisi d'exclure les âges 95-99 dans le calcul des erreurs MSE et LMSE.

4.5.3 Nouvelle comparaison

À présent, les âges 95-99 sont exclus dans le calcul des erreurs MSE et LMSE. Le tableau 4.4 consigne les résultats obtenus et montre qu'en retirant ces âges très élevés du calcul d'erreur, l'approche par réseau de neurones donne l'erreur MSE la plus grande pour le Round 1, l'erreur MSE la plus faible étant obtenue par le modèle de Lee-Carter. Concernant le Round 2, l'approche par réseau de neurones donne l'erreur MSE la plus faible, suivie de l'approche par forêt aléatoire.

Par ailleurs, les résultats obtenus en matière d'erreur LMSE sont quasiment identiques à ceux obtenus précédemment en conservant l'ensemble de la plage d'âges 0-99.

Âges 0-94	**Erreurs sur le test (Round 1)**		**Erreurs sur le test (Round 2)**	
Modèle utilisé	MSE	LMSE	MSE	LMSE
M3	$1,10 \cdot 10^{-4}$	$1,12 \cdot 10^{-1}$	$1,21 \cdot 10^{-4}$	$1,46 \cdot 10^{-1}$
M1	$9,57 \cdot 10^{-5}$	$8,44 \cdot 10^{-2}$	$1,26 \cdot 10^{-4}$	$1,52 \cdot 10^{-1}$
Réseau de neurones	$1,18 \cdot 10^{-4}$	$7,65 \cdot 10^{-2}$	$7,62 \cdot 10^{-5}$	$1,27 \cdot 10^{-1}$
Forêt aléatoire	$1,06 \cdot 10^{-4}$	$6,36 \cdot 10^{-2}$	$9,39 \cdot 10^{-5}$	$1,08 \cdot 10^{-1}$

TABLEAU 4.4. Comparaison finale des résultats obtenus par les modèles

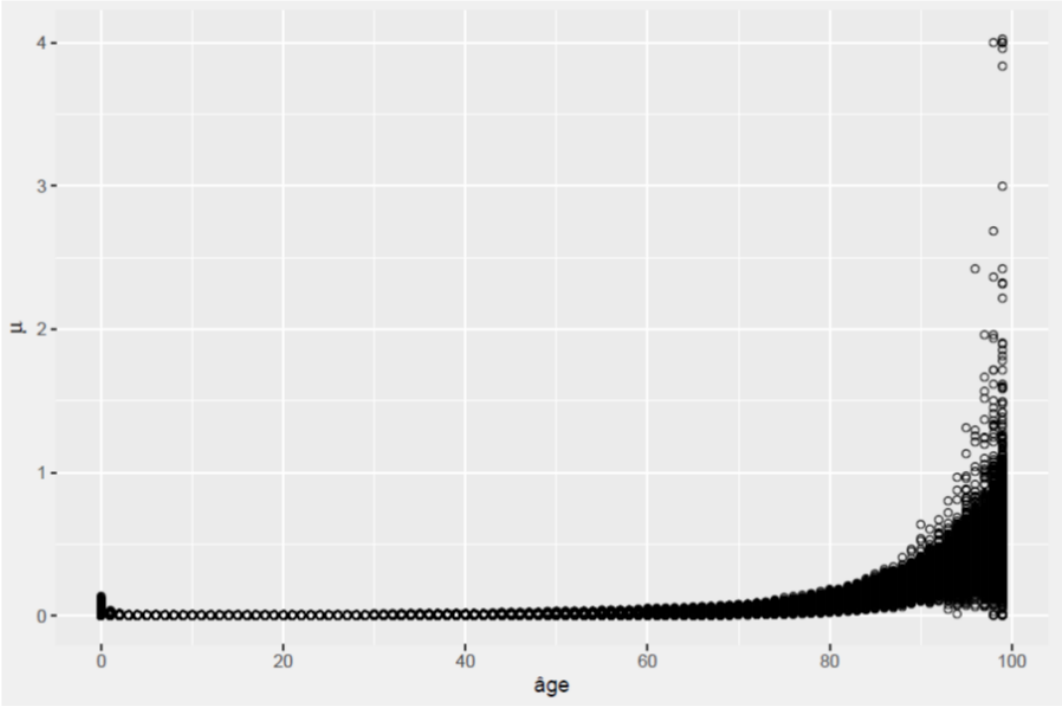


FIGURE 4.4. Nuage de points des données de force de mortalité μ de l'échantillon de données $\mathcal{D}$ en fonction de l'âge

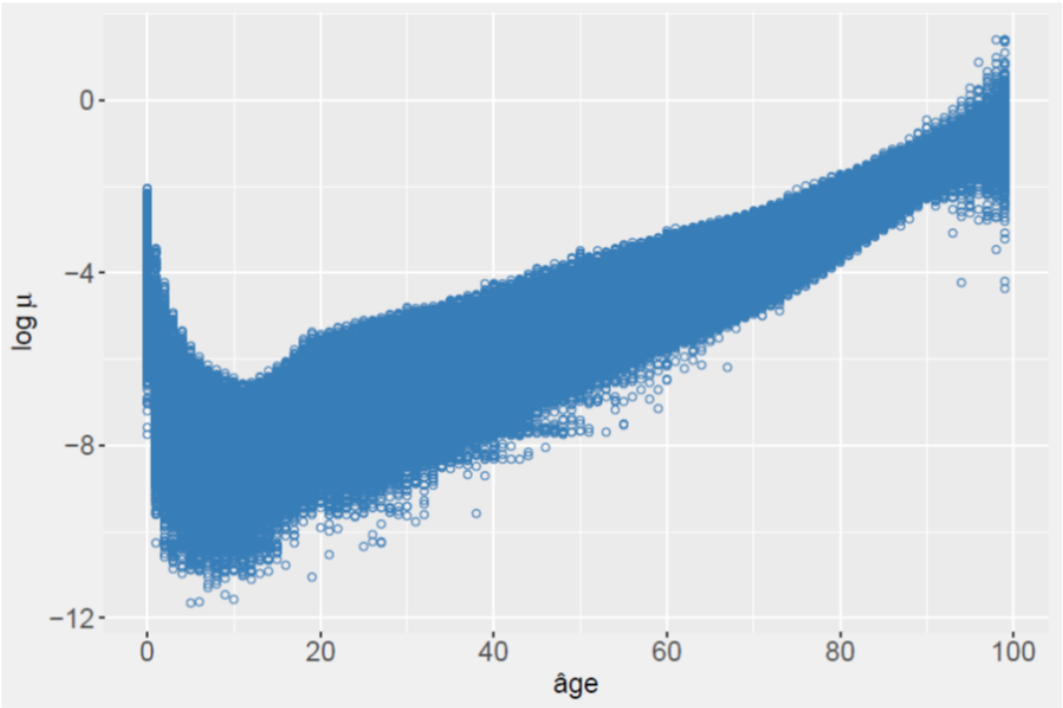


FIGURE 4.5. Nuage de points représentant le logarithme des données de force de mortalité μ de l'échantillon de données $\mathcal{D}$ en fonction de l'âge

4.5.4 Bilan comparatif des approches

Afin d'effectuer une comparaison des approches considérées, en plus du critère de la précision des prédictions (mesurée à partir de l'erreur MSE et LMSE), les critères de qualité d'un modèle de mortalité suivants seront retenus (voir Cairns *et al.* (2009)) :

- Parcimonie : ce critère n'est pas vérifié par l'approche par réseau de neurones qui nécessite d'optimiser un nombre énorme de paramètres (+ de 70 000 paramètres pour l'approche considérée par Richman et Wüthrich (2019)),
- Simplicité à mettre en œuvre : ce n'est pas le cas de l'approche par réseau de neurones, le réglage des nombreux hyperparamètres est en effet un problème complexe, à la fois heuristique et expérimental,
- Possibilité d'application à toute la tranche d'âges : ce n'est pas le cas du modèle M1 (meilleur aux âges faibles) et du modèle M3 (meilleur aux âges élevés),
- Capacité à générer des intervalles de confiance : les intervalles de confiance des prédictions sont facilement obtenus par les modèles M1 et M3, ce qui n'est pas le cas de l'approche par réseau de neurones et par forêt aléatoire (voir annexe E pour une approche envisageable afin d'obtenir des intervalles de prédictions pour un réseau de neurones),
- Interprétabilité du modèle : les modèles M1 et M3 sont facilement interprétables, ce qui n'est pas le cas des approches par réseau de neurones et par forêt aléatoire ("boîtes noires").

Un dernier critère retenu est la capacité du modèle à gérer des populations multiples. Les modèles M1 et M3 en sont incapables, et les extensions des modèles de mortalité classiques à plusieurs populations (par exemple le modèle de Li et Lee (2005)), sont difficiles à calibrer (voir El Bekri (2018) qui considère des modèles de mortalité à populations multiples dans un but de complétion de données de mortalité) et ont des qualités de prédictions décevantes (voir Richman et Wüthrich (2019)). Les approches par réseau de neurones et par forêt aléatoire ont elles l'avantage d'être des approches capables de gérer des populations multiples.

L'ensemble des critères de comparaison sont consignés dans le tableau 4.5.

Critère	M1	M3	NN	RF
Précision des prédictions	+	-	+	++
Parcimonie	++	++	-	+
Simplicité à mettre en œuvre	+	+	-	+
Application à toute la tranche d'âges	-	-	+	+
Capacité à générer des intervalles de confiance	+	+	-	-
Capacité à gérer des populations multiples	-	-	+	+
Interprétabilité	+	+	-	-

TABLEAU 4.5. Bilan comparatif des approches. Le tableau met en évidence les critères satisfaits par chacun des modèles

4.6 Application : étude d'un portefeuille de rentes viagères

4.6.1 Objectifs

Dans cette partie, on se propose d'étudier un portefeuille fictif de rentes viagères. On s'intéresse à l'impact de l'utilisation de tables de mortalité alternatives sur le bilan d'un assureur européen, par le calcul de projections de capitaux à verser aux assurés du portefeuille. Il est supposé que la date d'aujourd'hui est le 31 décembre 2009.

Dans une première approche, on calcule les projections sur cinq années des capitaux à verser aux assurés du portefeuille-exemple en faisant varier le modèle de projection de mortalité utilisé ainsi que le pays. Le choix d'une durée de projection de cinq années seulement s'explique par le fait que cela rend possible le calcul d'une erreur de prédiction à partir des données de mortalité HMD disponibles.

Dans une deuxième approche, on calcule les projections des capitaux à verser jusqu'à l'extinction du portefeuille-exemple pour les pays France et Allemagne. Contrairement à l'approche précédente, aucune erreur de prédiction n'est calculable car les données de mortalité HMD sont disponibles pour ces pays jusqu'en 2017 seulement.

Enfin, dans une troisième approche, les résultats obtenus pour la France sont comparés à ceux obtenus dans le cas où les tables réglementaires TGF 05 et TGH 05 sont utilisées dans le calcul de projections.

4.6.2 Analyse descriptive du portefeuille fictif de rentes

4.6.2.1 Variables disponibles

Soient N le nombre total d'assurés du portefeuille fictif et $i \in [\![1, N]\!]$ le numéro d'identification d'un assuré, qui correspond à un unique contrat. On dispose pour chaque assuré i de sa date de naissance b_i, son sexe g_i, le montant de rente annuelle R_i, ainsi que le taux technique τ_i.

Les montants de rentes R_i sont normalisés selon la transformation suivante :

$$R_i \leftarrow \frac{R_i}{\sum_{j=1}^{N} R_j} \tag{4.13}$$

L'âge x_i de l'assuré i est calculé par différence entre la date d'aujourd'hui du 31 décembre 2009 et la date de naissance b_i, arrondi par défaut lorsqu'il s'est écoulé moins de 6 mois depuis la dernière date d'anniversaire, ou par excès dans le cas contraire.

Les assurés ont un âge compris entre 57 ans et 86 ans au 31 décembre 2009, et l'âge maximum des assurés est fixé à l'âge limite de 120 ans.

4.6.2.2 Répartition des assurés par âge

La figure 4.6 représente la répartition du portefeuille-exemple en montants et effectifs par âge à la date du 31 décembre 2009. La proportion de rentiers et des montants sont respectivement maximaux aux âges de 71 ans et 67 ans, et décroissent ensuite progressivement avec l'âge.

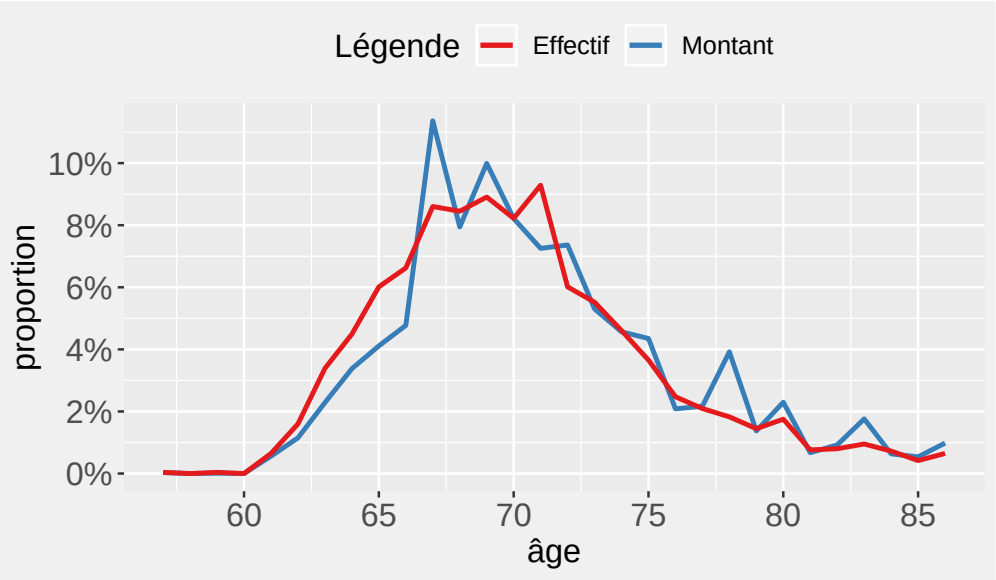


FIGURE 4.6. Répartition du portefeuille de rentes en montants et effectifs par âge au 31 décembre 2009

4.6.2.3 Répartition des hommes et des femmes

Le portefeuille est composé de 19.7% de femmes et de 80.3% d'hommes. La figure 4.7 représente la proportion de femmes par âge en montants et effectifs au 31 décembre 2009. Pour les âges très élevés, les proportions sont instables car il ne reste que peu d'assurés résiduels.

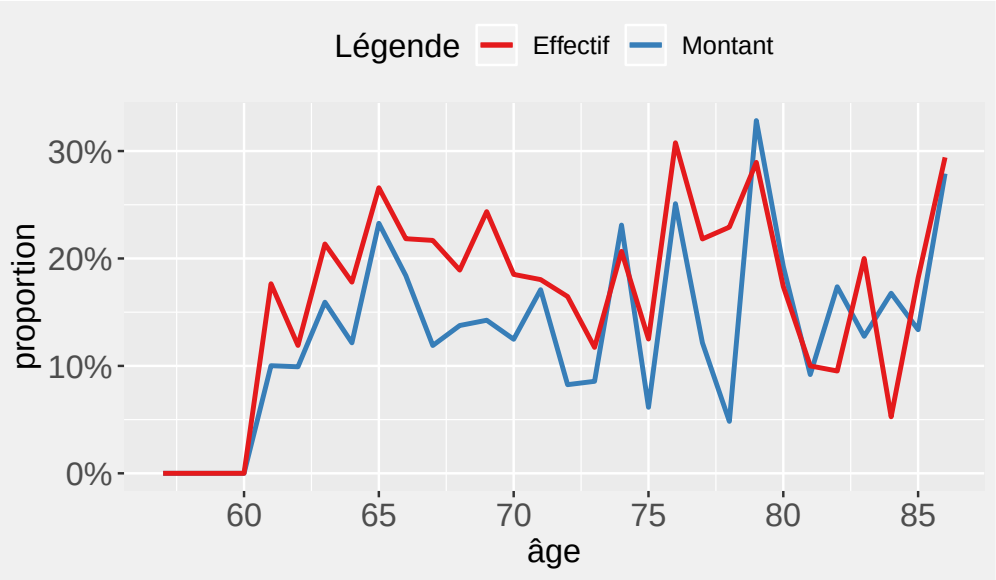


FIGURE 4.7. Proportion de femmes par âge en montants et effectifs au 31 décembre 2009

4.6.2.4 Répartition du montant des rentes annuelles du portefeuille

La figure 4.8 représente la répartition du montant des rentes annuelles du portefeuille au 31 décembre 2009, les unités et échelles des axes étant omises. Cette figure montre qu'une

majorité de rentes est de faible montant. Néanmoins, la distribution a une variance élevée et est donc très étalée vers les montants élevés. En outre, les rentes les plus élevées ont un poids significatif dans le portefeuille, les 5% des rentes les plus élevées représentant plus de 35% des prestations servies.

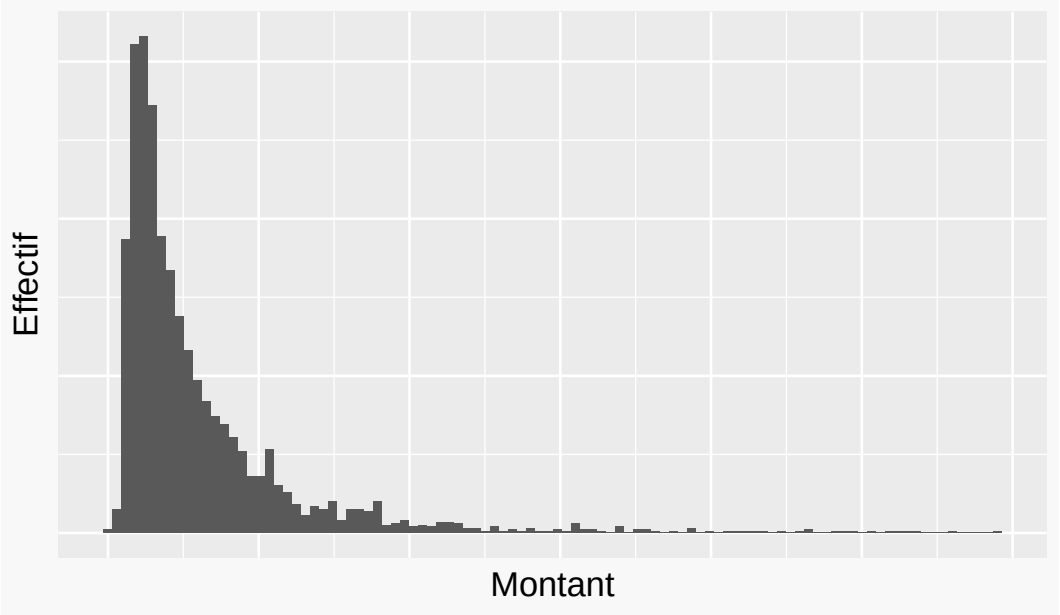


FIGURE 4.8. Répartition du montant des rentes annuelles du portefeuille au 31 décembre 2009

4.6.2.5 Répartition du montant des taux techniques du portefeuille

Le taux technique est un paramètre financier qui permet d'escompter les produits financiers futurs dont les conditions sont fixées par le code des assurances. Ce paramètre, dont la répartition dans le portefeuille-exemple est décrit par la figure 4.9, est compris entre 0% et 4,50%.

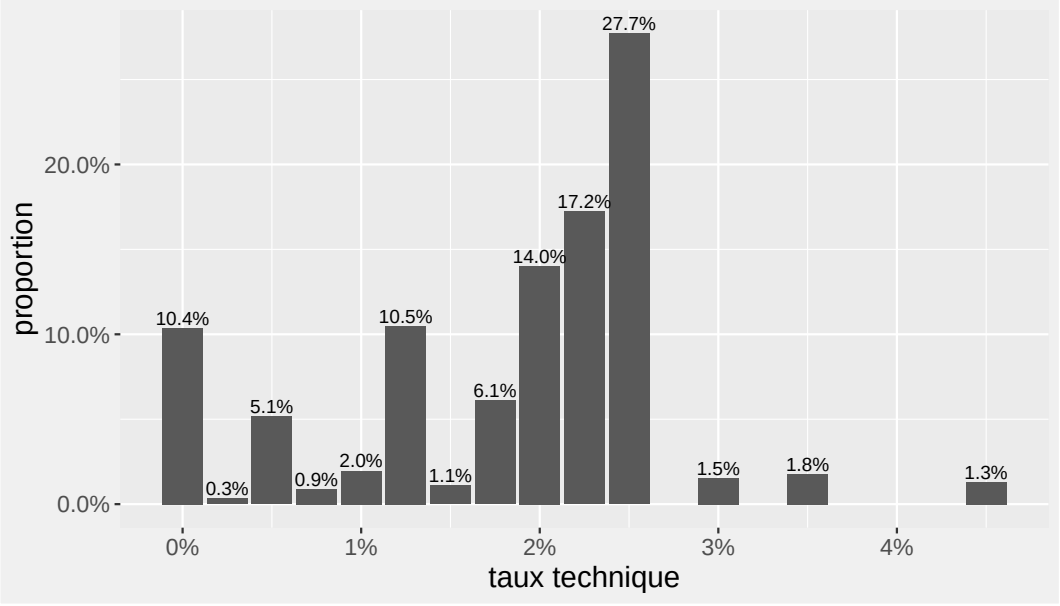


FIGURE 4.9. Répartition du montant des taux techniques du portefeuille

4.6.3 Calculs des projections de mortalité

4.6.3.1 Méthodes de projection de mortalité utilisées

Afin de calculer la somme totale des Valeurs Actuelles Probables (VAP) des rentes versées à tous les assurés du portefeuille-exemple, il convient de choisir un modèle de mortalité permettant de calculer les taux de mortalité futurs.

Il est choisi de considérer les modèles précédemment discutés dans ce chapitre, à savoir les modèles stochastiques de mortalité M1 et M3, ainsi que les approches par réseau de neurones et par forêt aléatoire.

La période temporelle choisie pour la calibration des modèles est la période 1985-2009, l'intervalle d'âges est 50-90 ans, et l'ensemble de dix pays retenus est le suivant : Autriche, République tchèque, Allemagne, Danemark, Finlande, France, Pays-Bas, Norvège, Pologne, Suède.

À partir d'un âge appelé "âge raccord", l'estimation des taux de mortalité pose problème dans la mesure où les individus encore en vie sont de plus en plus rares. Ainsi, les estimateurs utilisés deviennent médiocres et les fluctuations d'échantillonnage sont de plus en plus importantes. Afin de résoudre ce problème, il existe dans la littérature de nombreuses méthodes de fermeture de tables aux grands âges (Quashie et Denuit, 2005). Nous utiliserons pour notre étude la méthode de Denuit et Goderniaux (2005) présentée dans la suite.

4.6.3.2 La méthode de Denuit et Goderniaux

La méthode de Denuit et Goderniaux (2005) se base sur les taux de mortalité $q(x,t)$ et introduit une contrainte de fermeture des tables de mortalité. Elle consiste à ajuster par moindres carrés le modèle log-quadratique suivant :

$$\log \hat{q}(x,t) = a_t + b_t x + c_t x^2 + \varepsilon_x(t) \text{ avec } \varepsilon_x(t) \sim \mathcal{N}(0,\sigma^2) \text{ i.i.d.} \tag{4.14}$$

sur la base des observations relatives aux âges les plus élevés, avec les deux contraintes suivantes :

$$q(x_{\max},t) = 1, \forall t \tag{4.15}$$

$$\frac{\partial}{\partial x} q(x,t)\Big|_{x=x_{\max}} = 0, \forall t \tag{4.16}$$

où $\frac{\partial}{\partial x} q(x,t)$ désigne la dérivée première de $q(x,t)$ par rapport à l'âge x et $x_{\max} = 120$ ans désigne l'âge limite. Les contraintes (4.15) et (4.16) conduisent à l'équation :

$$a_t + b_t x + c_t x^2 = c_t (x_{\max} - x)^2. \tag{4.17}$$

Les contraintes (4.15) et (4.16) imposent à la courbe des taux de mortalité une allure concave aux grands âges et l'existence d'une tangente horizontale au point $x_{\max} = 120$. La contrainte (4.16) a pour conséquence d'empêcher une décroissance éventuelle de $q(x,t)$ aux âges avancés.

Dans la suite, les taux de mortalité $q(x,t)$ aux âges élevés sont calculés à partir de l'équation (4.14), avec un âge raccord de 90 ans et un âge limite $x_{\max}$ de 120 ans.

4.6.4 Le bilan économique d'un assureur

Schématiquement, le bilan économique simplifié d'un assureur se présente de la manière suivante :

Bilan	
A_t	FP_t
	RM_t
	BE_t

FIGURE 4.10. Bilan d'un assureur

où A_t, FP_t, RM_t et BE_t représentent respectivement le montant de placements détenus, les fonds propres, la marge pour risque et le Best Estimate.

Dans la suite, on s'intéresse par nos calculs de Valeurs Actuelles Probables (VAP) à l'impact de l'utilisation de différentes tables de mortalité pour le calcul du Best Estimate.

4.6.5 Première approche

Dans cette première approche, on se propose de calculer les projections sur cinq années des capitaux à verser aux assurés du portefeuille-exemple en faisant varier le modèle de projection de mortalité utilisé ainsi que le pays.

Le choix d'une durée de projection de cinq années seulement s'explique par le fait que cela rend possible le calcul d'une erreur de prédiction à partir des données de mortalité HMD disponibles.

4.6.5.1 Calcul de projection des rentes annuelles versées

La VAP des rentes versées à l'assuré i, d'âge x_i et de sexe g_i, sur les cinq prochaines années, de 2010 à 2014, s'écrit :

$$\text{VAP}_i = R_i \cdot \sum_{s=1}^{T} \frac{1}{(1+\tau_i)^s} \cdot \left(\prod_{j=0}^{s-1} [1 - q(x_i + j, t + j, g_i)] \right), \tag{4.18}$$

où $T = 5$ est le nombre d'années de projection, $t = 2010$ est la première année de projection, et τ_i désigne le taux technique.

En utilisant l'approximation $q(x_i+j, t+j, g_i) = 1-\exp(-\mu(x_i+j, t+j, g_i))$ (voir l'équation (1.6)), on obtient donc l'égalité suivante :

$$\text{VAP}_i = R_i \cdot \sum_{s=1}^{T} \frac{1}{(1+\tau_i)^s} \cdot \left(\prod_{j=0}^{s-1} \exp(-\mu(x_i + j, t + j, g_i)) \right). \tag{4.19}$$

On note que les taux de mortalité futurs $\mu(x_i + j, t + j, g_i)$ dépendent du pays européen considéré : Autriche, République tchèque, Allemagne, Danemark, Finlande, France, Pays-Bas, Norvège, Pologne, ou Suède. Ils sont calculés à partir des modèles stochastiques de mortalité M1 et M3, ainsi que les approches par réseau de neurones et par forêt aléatoire.

La somme totale des VAP des rentes versées à tous les assurés du portefeuille s'écrit donc :

$$\text{VAP}_{\text{tot}} = \sum_{i=1}^{N} \text{VAP}_i. \tag{4.20}$$

Afin de comparer les prédictions des sommes des rentes annuelles versées, qui dépendent à la fois du modèle de mortalité choisi et du pays, on choisit d'utiliser l'erreur relative comme métrique d'erreur :

$$\text{Erreur} = \frac{\sum_{i=1}^{N} \text{VAP}_i^{\text{obs}} - \sum_{i=1}^{N} \text{VAP}_i^{\text{pred}}}{\sum_{i=1}^{N} \text{VAP}_i^{\text{obs}}}. \tag{4.21}$$

4.6.5.2 Comparaison

Le tableau 4.6 consigne l'ensemble des résultats obtenus. Pour quatre pays (Allemagne, Danemark, France, Pologne), on obtient l'erreur la plus faible en valeur absolue avec l'approche par réseau de neurones. Le modèle M1 (République tchèque, Finlande, Pays-Bas) et la forêt aléatoire (Autriche, Norvège, Suède) permettent d'obtenir l'erreur la plus faible dans trois cas chacun. Le modèle M3 ne permet d'obtenir l'erreur la plus faible dans aucun des cas. Néanmoins, si l'on considère la moyenne des sommes de projections sur l'ensemble des dix pays considérés, l'erreur la plus faible est celle qui correspond au modèle M3.

Ainsi, aucun modèle ne se dégage de façon claire dans cet exemple de portefeuille fictif de rentes qui montre que les modèles classiques M1 et M3 restent pertinents pour ce calcul de projection de rentes annuelles versées.

		M1		**M3**		**NN**		**RF**	
Pays	Réalisé	Prédit	Erreur	Prédit	Erreur	Prédit	Erreur	Prédit	Erreur
AUT	4,712	4,704	$1,5 \cdot 10^{-3}$	4,709	$6,8 \cdot 10^{-4}$	4,701	$2,2 \cdot 10^{-3}$	4,709	$4,8 \cdot 10^{-4}$
CZE	4,543	4,543	$8,5 \cdot 10^{-5}$	4,541	$6,3 \cdot 10^{-4}$	4,530	$3,0 \cdot 10^{-3}$	4,543	$1,2 \cdot 10^{-4}$
DEUTW	4,693	4,687	$1,3 \cdot 10^{-3}$	4,690	$7,3 \cdot 10^{-4}$	4,697	$-7,1 \cdot 10^{-4}$	4,687	$1,4 \cdot 10^{-3}$
DNK	4,666	4,634	$6,8 \cdot 10^{-3}$	4,650	$3,4 \cdot 10^{-3}$	4,653	$2,8 \cdot 10^{-3}$	4,650	$3,3 \cdot 10^{-3}$
FIN	4,696	4,698	$-5,1 \cdot 10^{-4}$	4,700	$-9,7 \cdot 10^{-4}$	4,689	$1,4 \cdot 10^{-3}$	4,691	$1,0 \cdot 10^{-3}$
FRATNP	4,766	4,758	$1,7 \cdot 10^{-3}$	4,762	$7,8 \cdot 10^{-4}$	4,763	$6,7 \cdot 10^{-4}$	4,757	$1,9 \cdot 10^{-3}$
NLD	4,711	4,712	$-3,3 \cdot 10^{-4}$	4,709	$3,7 \cdot 10^{-4}$	4,701	$2,1 \cdot 10^{-3}$	4,704	$1,4 \cdot 10^{-3}$
NOR	4,729	4,725	$8,3 \cdot 10^{-4}$	4,735	$-1,3 \cdot 10^{-3}$	4,734	$-9,7 \cdot 10^{-4}$	4,730	$-1,3 \cdot 10^{-4}$
POL	4,518	4,491	$5,9 \cdot 10^{-3}$	4,489	$6,3 \cdot 10^{-3}$	4,511	$1,5 \cdot 10^{-3}$	4,494	$5,4 \cdot 10^{-3}$
SWE	4,745	4,743	$3,6 \cdot 10^{-4}$	4,749	$-8,3 \cdot 10^{-4}$	4,676	$1,5 \cdot 10^{-2}$	4,744	$2,0 \cdot 10^{-4}$
Moyenne	4,678	4,670	$1,8 \cdot 10^{-3}$	4,673	$9,9 \cdot 10^{-4}$	4,665	$2,7 \cdot 10^{-3}$	4,671	$1,5 \cdot 10^{-3}$

TABLEAU 4.6. Résultats des sommes de projections obtenues en fonction du pays et du modèle utilisé. Pour chaque pays, les résultats obtenus par le modèle ayant l'erreur la plus faible en valeur absolue sont représentés en couleur bleue.

4.6.6 Deuxième approche

Dans cette deuxième approche, on se propose de calculer les projections des capitaux à verser jusqu'à l'extinction du portefeuille pour les pays France et Allemagne.

Contrairement à l'approche précédente, aucune erreur de prédiction n'est calculable car les données de mortalité HMD sont disponibles pour ces pays jusqu'en 2017 seulement.

4.6.6.1 Calcul de projections

La valeur actuelle probable (VAP) des rentes versées à l'assuré i, d'âge x_i et de sexe g_i, de 2010 jusqu'à l'extinction du portefeuille, s'écrit :

$$\mathrm{VAP}_i = R_i \cdot \sum_{s=1}^{x_{\max}-x_i} \frac{1}{(1+\tau_i)^s} \cdot \left(\prod_{j=0}^{s-1} [1 - q(x_i+j, t+j, g_i)] \right), \tag{4.22}$$

où $x_{\max} = 120$ ans est l'âge limite, $t = 2010$ est la première année de projection, et τ_i désigne le taux technique. Cette équation est similaire à l'équation (4.18) de la première approche en prenant $T = x_{\max} - x_i$.

On note que les taux de mortalité futurs $q(x_i + j, t + j, g_i)$ dépendent du pays considéré : France ou Allemagne. Ils sont calculés à partir de l'approche par réseau de neurones, ainsi que la méthode de Denuit et Goderniaux (2005) pour les âges supérieurs à l'âge raccord de 90 ans.

La somme totale des VAP des rentes versées à tous les assurés du portefeuille s'écrit donc :

$$\mathrm{VAP}_{\mathrm{tot}} = \sum_{i=1}^{N} \mathrm{VAP}_i. \tag{4.23}$$

4.6.6.2 Comparaison France-Allemagne

La figure 4.11 représente la VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 pour les pays France et Allemagne. On rappelle que d'après l'équation de normalisation (4.13), la somme des montants de rentes à l'année 2009 est de 1, i.e. $\sum_{j=1}^{N} R_j = 1$. Par ailleurs, le tableau 4.8 représente la VAP totale des rentes versées aux assurés du portefeuille en fonction de la table de mortalité considérée.

On observe que dans le cas du portefeuille allemand, la VAP des rentes versées aux assurés est toujours inférieure à celle correspondant au portefeuille homologue français. En outre, la VAP totale des rentes versées pour le portefeuille français est supérieure d'environ 6% à celle du portefeuille allemand. Ces observations sont bien cohérentes avec le fait que les taux de mortalité allemands sont supérieurs aux taux de mortalité français. Ainsi, les assurés allemands ont une longévité moindre que celle des assurés français, ce qui implique une VAP moindre dans le cas du portefeuille français.

Table de mortalité	VAP totale
Modèle Allemagne	12,07
Modèle France	12,75

TABLEAU 4.7. VAP totale des rentes versées aux assurés du portefeuille sur la période 2010-2072 en fonction de la table de mortalité considérée

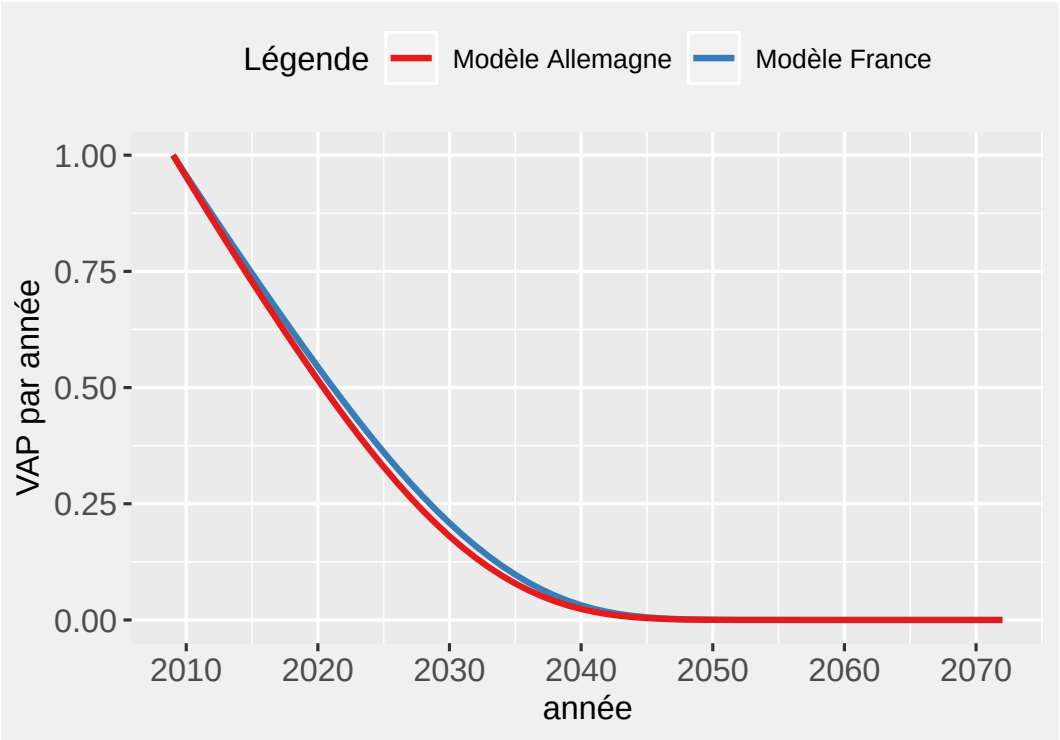


FIGURE 4.11. VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 pour les pays France et Allemagne

4.6.7 Troisième approche

Dans cette troisième approche, on se propose de comparer les résultats obtenus pour la France à ceux obtenus dans le cas où les tables réglementaires TGF 05 et TGH 05 sont utilisées dans le calcul de projection.

4.6.7.1 Les tables réglementaires

Les tables de mortalité utilisées par les assureurs dans leurs calculs actuariels sont imposées par la réglementation des produits d'assurance vie en France. Les autorités du marché assurantiel prévoient à travers ces tables, d'élaborer des hypothèses de mortalité prudentes et pertinentes proches de l'expérience. Dans ce cadre, les tables réglementaires jouent un rôle particulier en assurance vie. Actuellement, les tables utilisées sont :

1. **les tables TF 00-02 et TH 00-02 pour les assurances en cas de décès ou de vie :** La table TF 00-02 est construite à partir d'une population féminine en se basant sur les données de l'Institut national de la statistique et des études économiques (INSEE) sur la période 2000-2002, tandis que la table TH 00-02 est construite à partir d'une population masculine sur la même période. Elles sont homologuées par l'arrêté du 20 décembre 2005 et sont applicables aux contrats de vie ou décès souscrits depuis le 1er juillet 1993. Elles sont utilisées pour la tarification et le provisionnement des garanties décès ou vie, et remplacent les tables TV 88-90 pour les garanties vie et TD 88-90 pour les garanties décès.
2. **les tables de génération TGF 05 et TGH 05 pour les rentes viagères :** Homologuées par l'arrêté du 1er août 2006, ces tables ont été réalisées à partir des données de l'INSEE sur la population générale sur la période 1962-2000 et la mortalité de la population bénéficiaire des contrats de rentes observée de 1993 à 2005. Elles viennent remplacer les Tables Prospectives de Rentes Viagères (TPRV 93) et servent à la tarification et au provisionnement des contrats de rentes viagères depuis le 1er janvier 2007 pour les compagnies d'assurance, et depuis le 1er janvier 2008 pour les institutions de prévoyance et les mutuelles.

On note qu'un arrêt de la Cour de Justice de l'Union Européenne, en date du 1er mars 2011, a rendu impossibles les tarifications en fonction du sexe à compter du 21 décembre 2012, au motif qu'elles constituent des discriminations : "La prise en compte du sexe de l'assuré en tant que facteur de risques dans les contrats d'assurance constitue une discrimination". Seuls les nouveaux contrat souscrits à partir du 21 décembre 2012 bénéficient de cette nouvelle législation.

4.6.7.2 Les tables TGF 05 et TGH 05

Les tables TGF 05 (Table par Génération des Femmes - Version 2005) et TGH 05 (Table par Génération des Hommes - Version 2005) sont des tables prospectives, par génération et par sexe. Elles fournissent les données brutes du nombre de survivants, âge par âge, pour les hommes et pour les femmes, pour toutes les générations comprises entre 1900 et 2005. Les données sont limitées à l'intervalle d'âges 0-120 ans.

Les données utilisées pour construire ces tables sont composées d'observations de 19 portefeuilles de rentiers, 16 en provenance de la FFSA (Fédération Française de l'Assurance) et 3 du CTIP (Centre Technique des Institutions de Prévoyance). Ces observations concernent 700 000 rentes liquidées pour environ 2 millions de lignes au total couvrant la période 1993-2005.

Pour la construction de ces tables de mortalité, il a été retenu des modèles utilisant une référence externe de mortalité, c'est-à-dire en positionnant la mortalité du groupe considéré par rapport à une référence externe afin de pallier au volume limité de données disponibles et donc associées à des échantillons de taille réduite. Pour l'élaboration des tables de mortalité TGH 05 et TGF 05, la référence externe est basée sur une table prospective de l'INSEE construite sur la base des tables du moment de 1962 à 2000 (Planchet, 2005).

Afin de permettre une représentation globale des tables réglementaires, les graphes représentant les données du nombre de survivants L_x ainsi que les taux de mortalité $q(x,t)$ des tables TGF 05 et TGH 05 figurent en annexe I.

4.6.7.3 Calcul de projections à partir des tables réglementaires

La valeur actuelle probable (VAP) des rentes versées à l'assuré i, d'âge x_i et de sexe g_i, de 2010 jusqu'à l'extinction du portefeuille, s'écrit :

$$\text{VAP}_i = R_i \cdot \sum_{k=1}^{x_{\max}-x_i} \frac{1}{(1+\tau_i)^k} \cdot {}_kp_{x_i} = R_i \cdot \sum_{k=1}^{x_{\max}-x_i} \frac{1}{(1+\tau_i)^k} \cdot \frac{L_{x_i+k}}{L_{x_i}}, \tag{4.24}$$

où $x_{\max} = 120$ est l'âge limite et où la quantité ${}_kp_{x_i} = \frac{L_{x_i+k}}{L_{x_i}}$ désigne la probabilité que l'assuré i d'âge x_i est en vie à l'âge $x_i + k$. On note que les grandeurs L_{x_i+k} sont obtenues numériquement à l'aide des tables réglementaires TGF 05 si l'assuré i est une femme et TGH 05 si l'assuré i est un homme. Le vecteur L_i de L_{x_i} pour tous les âges x_i, correspond au nombre de survivants ayant atteint l'âge x_i au sein d'une cohorte de taille L_0 initialement à la naissance, avec $L_0 = 100\ 000$, par convention.

La somme totale des VAP des rentes versées à tous les assurés du portefeuille s'écrit donc :

$$\text{VAP}_{\text{tot}} = \sum_{i=1}^{N} \text{VAP}_i. \tag{4.25}$$

4.6.7.4 Comparaison avec le cas où les tables réglementaires TGF 05 et TGH 05 sont utilisées

La figure 4.12 représente la VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 pour la France dans deux cas : un premier cas où les taux de mortalités futurs sont calculés à partir de l'approche par réseau de neurones ("Modèle France"), et un second cas où les taux de mortalité futurs sont calculés à partir des tables réglementaires TGF 05 et TGH 05 ("Tables TGFH"). On rappelle que d'après l'équation de normalisation (4.13), la somme des montants de rentes à l'année 2009 est de 1, i.e. $\sum_{j=1}^{N} R_j$.

On observe que dans le premier cas, la VAP des rentes versées aux assurés est toujours inférieure à celle correspondant au second cas. En outre, la VAP totale des rentes versées en utilisant le "Modèle France" est inférieure d'environ 16% à celle obtenue en utilisant les tables réglementaires.

Ceci est bien cohérent avec le fait que les tables TGF 05 et TGH 05 sont prudentes pour la calcul de rentes viagères et donc optimistes sur la longévité future.

Table de mortalité	VAP totale
Modèle France	12,75
Tables TGFH	15,22

TABLEAU 4.8. VAP totale des rentes versées aux assurés du portefeuille français sur la période 2010-2072 en fonction de la table de mortalité utilisée

Enfin, la figure 4.13 représente la VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 et des trois tables de mortalité utilisées.

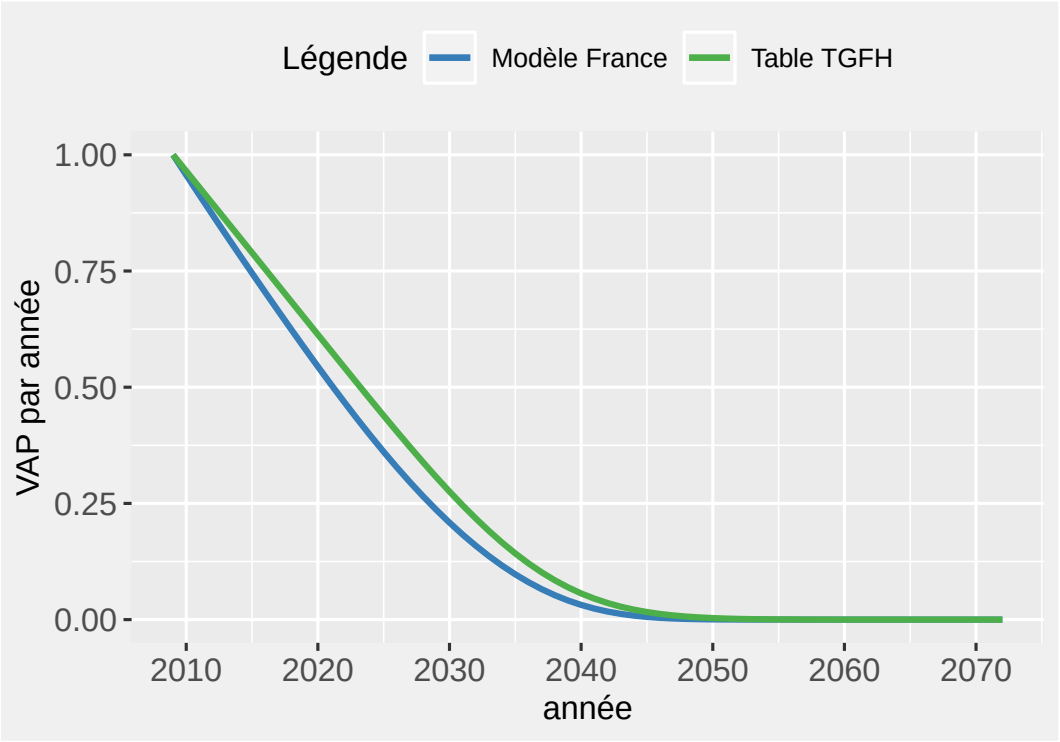


FIGURE 4.12. VAP des rentes versées aux assurés du portefeuille français en fonction des années de projection 2010-2072 et de la table de mortalité utilisée

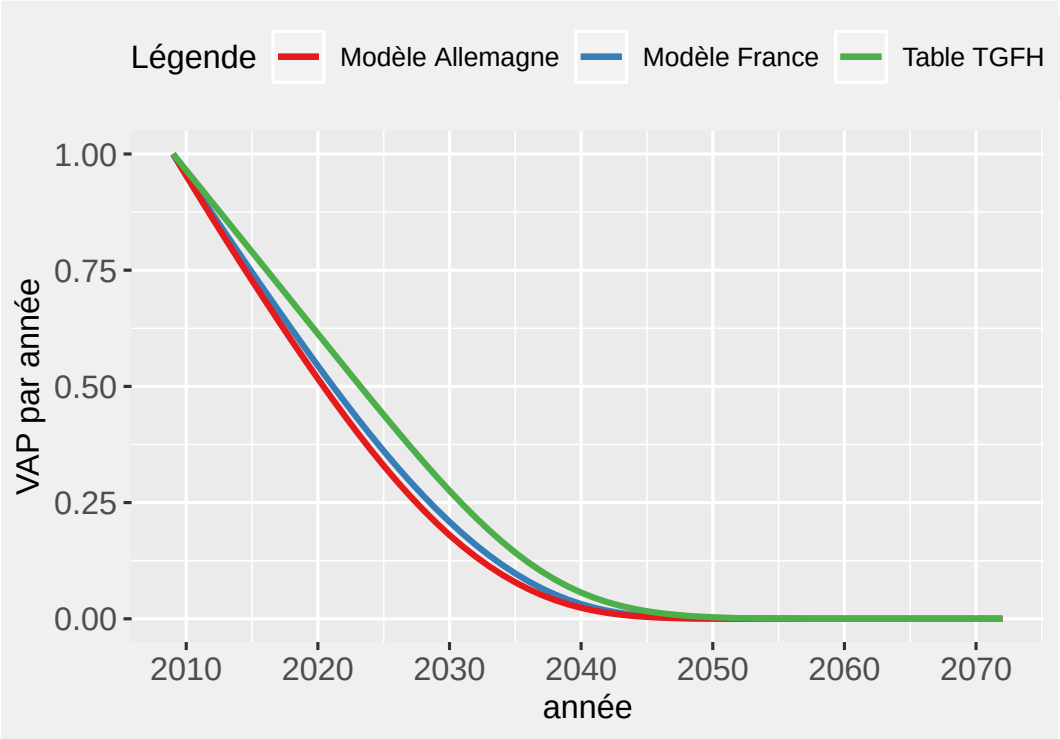


FIGURE 4.13. VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 et des trois tables de mortalité utilisées

4.7 Bilan

Dans ce chapitre, des approches classiques de prédiction de la mortalité ont été comparées avec des approches par apprentissage automatique. La métrique classique de l'erreur MSE sur la force de mortalité n'a pas semblé être satisfaisante, car elle pénalise beaucoup plus les erreurs sur les âges élevés. En raison des fluctuations d'échantillonnage, on obtient des coefficients de variations parfois énormes pour les âges supérieurs à 95 ans. Ceci nous a orienté vers le choix d'une nouvelle métrique d'erreur (LMSE), ainsi que sur le choix de retirer les points correspondant à des âges supérieurs à 95 ans dans l'échantillon de données.

Il est intéressant de faire un lien avec la méthode des moindres carrés qui permet de minimiser l'impact des erreurs expérimentales en "ajoutant de l'information" dans le processus de mesure. Si, comme c'est généralement le cas, on dispose d'une estimation de l'écart-type du bruit qui affecte chaque mesure, on l'utilise pour "peser" la contribution de la mesure : une mesure a d'autant plus de poids que son incertitude est faible. En appliquant ce principe à notre problème, il serait possible de créer une nouvelle métrique d'erreur (appelée erreur MSE pondérée) qui pondèrerait chaque point de façon inversement proportionnelle à sa variance associée.

Tant dans la partie s'intéressant à la prédiction de la mortalité future que dans celle discutant la modélisation d'un portefeuille fictif de rentes dans son ensemble, les résultats obtenus montrent qu'il n'existe pas de modèle qui se dégage de manière claire. Les modèles de mortalité classiques restent pertinents et permettent par exemple d'obtenir facilement des intervalles de confiance, les approches par apprentissage automatique étant notamment pertinents dans un cadre multi-populations (voir le bilan comparatif en section 4.5.4 pour plus de détails sur les avantages et inconvénients de chaque approche).

CONCLUSION

Ce mémoire d'Actuariat a porté sur les dernières avancées en matière de correction d'anomalies dans les tables de mortalité nationales ainsi que la modélisation prédictive de la mortalité future.

Dans une première partie, nous avons présenté et implémenté une approche novatrice par apprentissage automatique afin de produire des tables de mortalité corrigées pour des pays pour lesquels on ne dispose pas de données mensuelles de fertilité.

Cette méthode de correction a d'abord été testée sur des pays pour lesquels des données mensuelles de fertilité sont disponibles et les performances de la méthode ont été déterminées en calculant les erreurs moyennes de prédiction MSE, moyennes des prédictions de l'indicateur de correction, l'indicateur de correction observé étant obtenu par comparaison des données des versions 5 et 6 de la *Human Mortality Database*. Le modèle de base de régression linéaire donne des résultats très satisfaisants en matière d'erreur de prédiction MSE moyenne, les résultats moyens étant améliorés par l'approche par réseau de neurones. La méthode a ensuite été appliquée à des pays pour lesquels les données mensuelles de fertilité ne sont pas disponibles. Il s'est avéré que la volatilité des taux d'amélioration de la mortalité pour ces pays est considérablement réduite lors de l'utilisation de données de mortalité corrigées produites par la méthode de correction.

L'impact de la méthode de correction a également été étudié sur trois modèles stochastiques de mortalité. Le modèle M1 est significativement amélioré lorsqu'il est calibré avec des données de mortalité corrigées, tandis que la log-vraisemblance est inchangée pour les modèles stochastiques de mortalité ayant une composante de cohorte dont la dynamique est beaucoup plus régulière lorsque ces modèles sont calibrés sur des données de mortalité corrigées. Ainsi, cette méthode de correction a le même effet que les méthodes de correction basées sur les données mensuelles de fertilité.

Dans une seconde partie, nous avons présenté, implémenté et comparé des approches de prédiction de la mortalité future : à la fois des approches classiques et des approches par apprentissage automatique. Les approches ont aussi été appliquées dans le cadre de l'étude d'un portefeuille fictif de rentes.

La métrique classique de l'erreur MSE sur la force de mortalité n'a pas semblé être satisfaisante, car elle pénalise beaucoup plus les erreurs sur les âges élevés. En raison des fluctuations d'échantillonnage, des coefficients de variations parfois énormes ont été obtenus pour des âges très élevés. Ceci nous a orienté vers le choix d'une nouvelle métrique d'erreur, ainsi que sur le choix de retirer les points correspondant à des âges dont les fluctuations

d'échantillonnage sont trop importantes.

En plus du critère de précision des prédictions obtenues par les différentes approches, d'autres critères comparatifs ont été discutés. Il est apparu que les modèles de mortalité classiques restent pertinents au regard de plusieurs critères et permettent par exemple d'obtenir facilement des intervalles de confiance, les approches par apprentissage automatique étant notamment pertinentes dans un cadre à plusieurs populations.

BIBLIOGRAPHIE

Arnold, V. (1957). On functions of three variables. *Proceedings of the USSR Academy of Sciences*, 114:679–681. English translation : Amer. Math. Soc. Transl.

Benjamin, J. et Soliman, A. (1993). Mortality on the move. *Institute of Actuaries, Oxford*, 2:105–130.

Boumezoued, A. (2016). Improving HMD mortality estimates with HFD fertility data. `https://hal.archives-ouvertes.fr/hal-01270565`.

Boumezoued, A., Hoffmann, M. et Jeunesse, P. (2018). A new inference strategy for general population mortality tables. `https://hal.archives-ouvertes.fr/hal-01773665`.

Boumezoued, A., Hoffmann, M. et Jeunesse, P. (2019). Nonparametric adaptive inference of birth and death models in a large population limit. *arXiv preprint arXiv :1903.00673*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R. et Stone, C. (1984). *Classification and regression trees. Monterey, Calif., USA : Wadsworth*. Inc.

Brouhns, N., Denuit, M. et Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance : Mathematics and economics*, 31(3):373–393.

Cairns, A. J., Blake, D. et Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty : theory and calibration. *Journal of Risk and Insurance*, 73(4): 687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. et Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J., Blake, D., Dowd, K. et Kessler, A. R. (2016). Phantoms never die : living with unreliable population data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 179(4):975–1005.

Chollet, F. (2015). Keras : The Python Deep Learning Library.

Chollet, F. et Allaire, J. J. (2018). R interface to Keras. RStudio, Google. `https://cloud.r-project.org/web/packages/keras/index.html`.

Currie, I. D. (2006). Smoothing and forecasting mortality rates with P-splines. *Talk given at the Institute of Actuaries*.

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Dalalyan, A. (2018). Cours d'Apprentissage statistique 2018/2019. *Notes de cours ENSAE*.

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

Deprez, P., Shevchenko, P. V. et Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7(2):337–352.

El Bekri, N. (2018). Mise à jour des tables de mortalité par approche multi-population. Application au calcul de choc et au transfert du risque de longévité. *Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires*.

Gal, Y. et Ghahramani, Z. (2016). Dropout as a Bayesian approximation : Representing model uncertainty in deep learning. *In international conference on machine learning*, pages 1050–1059.

Glorot, X. et Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *In Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Gompertz, B. (1825). XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F.R.S. &c. *Philosophical transactions of the Royal Society of London*, (115):513–583.

Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Guo, C. et Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv :1604.06737*.

Haberman, S. et Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance : Mathematics and Economics*, 48(1):35–55.

Hainaut, D. (2018). A neural network analyzer for mortality forecast. *ASTIN Bulletin : The Journal of the IAA*, 48(2):481–508.

He, K., Zhang, X., Ren, S. et Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

HFD (2018). The Human Fertility Database. Max Planck institute for demographic research (Germany), and Vienna institute of demography (Austria). `https://www.humanfertility.org`.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv :1207.0580*.

HMD (2018). The Human Mortality Database. University of California, Berkeley (USA), and Max Planck institute for demographic research (Germany). `https://www.mortality.org`.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Huang, G., Liu, Z., Van Der Maaten, L. et Weinberger, K. Q. (2017). Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Ioffe, S. et Szegedy, C. (2015). Batch normalization : Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv :1502.03167*.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to Statistical Learning*, volume 112. Springer.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2017). *An Introduction to Statistical Learning : With Applications in R*. New York : Springer.

Keyes, K. M., Utz, R. L., Robinson, W. et Li, G. (2010). What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. *Social science & medicine*, 70(7):1100–1108.

Kingma, D. P. et Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*.

Kolmogorov, A. (1956). The representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Proceedings of the USSR Academy of Sciences*, 108:179–182. English translation : Amer. Math. Soc. Transl.

Lee, R. D. et Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419):659–671.

Levantesi, S. et Pizzorusso, V. (2019). Application of machine learning to mortality modeling and forecasting. *Risks*, 7(1):26.

Lexis, W. H. R. A. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. KJ Trübner.

Li, N. et Lee, R. (2005). Coherent mortality forecasts for a group of populations : An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

Liaw, A. et Wiener, M. (2002). Classification and Regression by `randomForest`. *R News*, 2(3):18–22.

Ludkovski, M., Risk, J. et Zail, H. (2018). Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin*, 48(3):1307–1347.

Macdonald, A. S., Cairns, A. J. G., Gwilt, P. et Miller, K. (1998). An international comparison of recent trends in population mortality. *British Actuarial Journal*, 4(1):3–141.

Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries*, 8(6):301–310.

McCulloch, W. S. et Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Ng, A. (2017). Massive Open Online Course: Deep Learning Specialization. Coursera (2017). `https://www.coursera.org/specializations/deep-learning`.

Pankratz, A. (2009). *Forecasting with univariate Box-Jenkins models : Concepts and cases*, volume 224. John Wiley & Sons.

Planchet, F. (2005). Tables de mortalité d'expérience pour des portefeuilles de rentiers. *Note méthodologique de l'Institut des Actuaires*.

Plat, R. (2009). On stochastic mortality modeling. *Insurance : Mathematics and Economics*, 45(3):393–404.

Prechelt, L. (1998). Early stopping-but when? *In Neural Networks : Tricks of the trade*, pages 55–69. Springer.

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

Renshaw, A. E. et Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance : Mathematics and economics*, 38(3):556–570.

Richards, S. (2008). Detecting year-of-birth mortality patterns with limited data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 171(1):279–298.

Richman, R. (2018). AI in Actuarial Science. *Actuarial Society 2018 Convention, Cape Town, 24–25 October 2018*.

Richman, R. et Wüthrich, M. V. (2019). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, pages 1–21. `https://doi.org/10.1017/S1748499519000071`.

Scott, W. (1981). Some applications of the Poisson distribution in mortality studies. *Transactions of the Faculty of Actuaries*, 38:255–263.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R. (2014). Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Team, R. C. (2018). R : A language and environment for statistical computing. Vienna, Austria : R Foundation for Statistical Computing. `https://www.R-project.org`.

Thome, N. (2019). Fondements théoriques du Deep Learning. *Note de cours du Master MVA*.

Turner, H. et Firth, D. (2007). Generalized nonlinear models in R : An overview of the `gnm` package.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Villegas, A., Kaishev, V. K. et Millossovich, P. (2015). `StMoMo` : An R package for stochastic mortality modelling. *In 7th Australasian Actuarial Education and Research Symposium*.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine*. 2nd ed. Cambridge : Massachusetts Institute of Technology Press.

Wilmoth, J. R., Andreev, K., Jdanov, D. et Glei, D. A. (2007). Methods protocol for the Human Mortality Database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. [Last Revised : May 31, 2007 (Version 5)]*. `http://v5.mortality.org`.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A. et Riffe, T. (2019). Methods protocol for the Human Mortality Database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. [Last Revised : October 5, 2019 (Version 6)]*. `http://mortality.org`.

Wilmoth, J. R. et Shkolnikov, V. M. (2010). Human Mortality Database. University of California.

ANNEXES

A Acronymes et symboles

Acronyme	Signification
ARIMA	Auto-Regressive Integrated Moving Average
CART	Classification And Regression Trees
GAPC	Generalized Age-Period-Cohort
HFD	Human Fertility Database
HMD	Human Mortality Database
IC	Intervalle de confiance
i.i.d.	indépendantes et identiquement distribuées
Modèle APC	Modèle Age-Period-Cohort
Modèle CBD	Modèle de Cairns-Blake-Dowd
MSE	Mean Squared Error
NN	Neural Network
RF	Random Forest
RSS	Residual Sum of Squares

Tableau A.1. Acronymes utilisés

Symbole	Signification
$\mathbb{E}$	Espérance
$\mathbb{P}$	Probabilité
$\mathbb{V}$	Variance
$\mathcal{B}(n,p)$	Loi binomiale de paramètres n et p
$\mathcal{N}(\mu,\sigma^2)$	Loi Normale de moyenne μ et de variance σ^2
$\mathcal{P}(\lambda)$	Loi de Poisson de paramètre λ
$\xrightarrow[n\to+\infty]{\mathcal{L}}$	Convergence en loi
$\xrightarrow[n\to+\infty]{p.s.}$	Convergence presque sûre

Tableau A.2. Symboles utilisés

B Extrait de code R de l'architecture du réseau de neurones

```
Year <- layer_input(shape = c(1))
Age <- layer_input(shape = c(1))
Country <- layer_input(shape = c(1))
Gender <- layer_input(shape = c(1))
Age_embed <- Age %>%
  layer_embedding(input_dim = nb_age,
                  output_dim = 5, input_length = 1) %>%
  keras :: layer_flatten()
Gender_embed <- Gender %>%
  layer_embedding(input_dim = 2,
                  output_dim = 1, input_length = 1) %>%
  keras :: layer_flatten()
Country_embed = Country %>%
  layer_embedding(input_dim = nb_country, output_dim = 5,
                  input_length = 1) %>%
  keras :: layer_flatten()

features <- layer_concatenate(list(Year, Age_embed,
                                   Gender_embed, Country_embed))

middle <- features %>%
  layer_dense(units = nb_units_1, activation = "tanh") %>%
  layer_batch_normalization() %>%
  layer_dropout(dropout_rate) %>%
  layer_dense(units = nb_units_2, activation = "tanh") %>%
  layer_batch_normalization() %>%
  layer_dropout(dropout_rate) %>%
  layer_dense(units = nb_units_3, activation = "tanh") %>%
  layer_batch_normalization() %>%
  layer_dropout(dropout_rate) %>%
  layer_dense(units = nb_units_4, activation = "tanh") %>%
  layer_batch_normalization() %>%
  layer_dropout(dropout_rate)

output <- layer_concatenate(list(features, middle)) %>%
  layer_dense(units = nb_units_5, activation = "tanh") %>%
  layer_batch_normalization() %>%
  layer_dropout(dropout_rate) %>%
  layer_dense (units = 1, activation = "sigmoid", name = "main_output")

NEURAL_NET <- keras_model(inputs = c(Year, Age, Country, Gender),
                          outputs = c(output))
```

C Liste des pays disponibles dans la HMD

	Pays	Code HMD	$t_{\min}$	$t_{\max}$
1	Espagne	ESP	1908	2016
2	Italie	ITA	1872	2014
3	France	FRATNP	1816	2017
4	Autriche	AUT	1947	2017
5	Angleterre et pays de Galles	GBRTENW	1841	2016
6	Danemark	DNK	1835	2016
7	Suède	SWE	1751	2017
8	Finlande	FIN	1878	2015
9	Norvège	NOR	1846	2014
10	Islande	ISL	1838	2016
11	Pays-Bas	NLD	1850	2016
12	Belgique	BEL	1841	2015
13	Portugal	PRT	1940	2015
14	Suisse	CHE	1876	2016
15	Luxembourg	LUX	1960	2014
16	Irlande du Nord	GBR_NIR	1922	2016
17	Écosse	GBR_SCO	1855	2016
18	Irlande	IRL	1950	2014
19	Grèce	GRC	1981	2013
20	République tchèque	CZE	1950	2017
21	Hongrie	HUN	1950	2017
22	Slovaquie	SVK	1950	2017
23	Bulgarie	BGR	1947	2010
24	Biélorussie	BLR	1959	2016
25	Estonie	EST	1959	2017
26	Lituanie	LTU	1959	2017
27	Lettonie	LVA	1959	2017
28	Pologne	POL	1958	2016
29	Russie	RUS	1959	2014
30	Slovénie	SVN	1983	2017
31	Ukraine	UKR	1959	2013
32	Israël	ISR	1983	2016
33	Australie	AUS	1921	2016
34	USA	USA	1933	2017
35	Nouvelle-Zélande	NZL_NP	1948	2013
36	Canada	CAN	1921	2016
37	Japon	JPN	1933	2017
38	Taïwan	TWN	1970	2014
39	Croatie	HRV	2002	2017
40	Allemagne	DEUTNP	1990	2017
41	Chili	CHL	1992	2008
42	Corée du Sud	KOR	2003	2016

Tableau C.3. Liste des pays disponibles dans la HMD

D Early Stopping

L'*Early Stopping* (arrêt prématuré) est une technique qui permet d'anticiper le phénomène de sur-apprentissage en arrêtant prématurément la présentation des exemples à un réseau de neurones.

Dans la plupart des documents traitant des réseaux de neurones, on peut trouver des courbes similaires à celle présentée en figure D.1. Pour une telle courbe, il est aisé de sélectionner l'époque correspondant au meilleur modèle. En effet, on arrête l'entraînement dès que l'erreur de validation à une époque donnée est plus grande que l'erreur de validation de l'époque précédente.

Néanmoins, en pratique, l'erreur de validation lors d'une période d'entraînement d'un réseau de neurones n'évolue pas de manière aussi idéalisée que ce que montre la figure D.1, mais ressemble davantage à ce que montre la figure D.2 (Prechelt, 1998). L'erreur de validation peut encore décroître après avoir commencé à croître. De plus, dans un contexte réaliste, nous ne connaissons jamais l'erreur de généralisation exacte, mais nous estimons celle-ci à l'aide de l'erreur de validation. Il n'y a donc pas de règle évidente pour décider quand le minimum de l'erreur de généralisation est obtenu.

FIGURE D.1. Courbes idéalisées d'entraînement et de validation. Axe vertical : erreurs, axe horizontal : époque

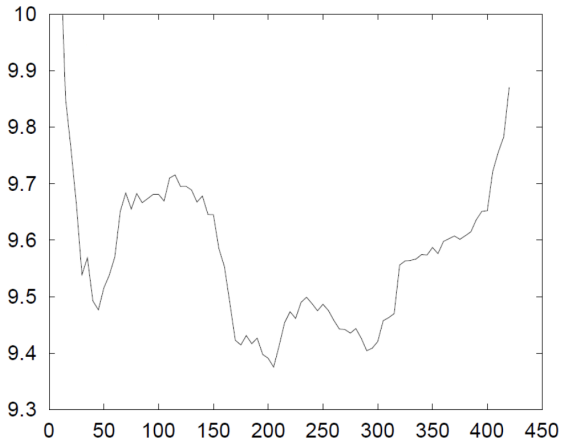


FIGURE D.2. Courbe réelle de validation. Axe vertical : erreur de validation, axe horizontal : époque

Étant donné qu'il n'existe pas de résultats connus à l'heure actuelle sur les propriétés des critères d'arrêt, son choix est fait de façon expérimentale. Soit E la métrique d'erreur de l'algorithme d'apprentissage, par exemple l'erreur quadratique moyenne (MSE). Soient respectivement $E_{\text{train}}(t)$ et $E_{\text{val}}(t)$ les erreurs d'entraînement et de validation mesurées après l'époque t. Alors on considère que l'erreur optimale de validation E_{opt} est dans notre cas :

$$E_{\text{opt}} = \min_{t \leq 100} E_{\text{val}}(t) \tag{4.26}$$

L'époque retenue est donc :

$$T = \operatorname*{argmin}_{t \leq 100} E_{\text{val}}(t) \tag{4.27}$$

Par ailleurs, on adopte une stratégie d'Early Stopping avec plateau en utilisant les fonctions `callback_reduce_lr_on_plateau` et `callback_model_checkpoint` de `Keras`.

Les paramètres suivants de la fonction `callback_reduce_lr_on_plateau` de `Keras` ont été utilisées :

- `monitor = val_loss` : dans notre problème, la métrique d'erreur choisie est l'erreur de validation.
- `min_lr` $= \eta_{\min} = 5.10^{-5}$: borne inférieure du taux d'apprentissage.
- `patience` $= 5$: nombre d'époques sans amélioration de la métrique d'erreur après lequel la période d'entraînement est stoppée.
- `factor` $= 0.8$: facteur φ par lequel le taux d'apprentissage est réduit, $\eta_{new} = \eta_{\text{old}} \cdot \varphi$. Ainsi, le taux d'apprentissage s'écrit sous la forme $\eta_t = \eta_{\text{init}} \cdot \varphi^k$ avec :
  - $\eta_{\text{init}} = 10^{-3}$ le taux d'apprentissage initial,
  - k le nombre de fois que le taux d'apprentissage a été réduit d'un facteur φ,
  - η_t le taux d'apprentissage à l'époque t,
  - $\eta_t \geq \eta_{\min}$.
- `cooldown` $= 5$: nombre d'époques à attendre avant de réduire le taux d'apprentissage d'un facteur `factor`.
- `verbose` $= 1$: permet d'afficher l'évolution de la métrique d'erreur en fonction de l'époque.

La fonction `callback_model_checkpoint` est également utilisée afin de conserver le meilleur modèle parmi tous les modèles considérés au regard des critères choisis lors de la période d'entraînement. On utilise les paramètres suivants :

- `filepath` : lieu du dossier où le meilleur modèle sera stocké, au format `weights.epoch:02d-val_loss:.2f.hdf5`. Grâce à ce format, le meilleur modèle est enregistré directement avec le numéro de l'époque optimale ainsi que le minimum de la métrique considérée (`val_loss`) dans le nom du fichier.
- `monitor = val_loss` : dans notre problème, la métrique d'erreur choisie est l'erreur de validation.
- `verbose` $= 0$: permet de masquer l'évolution de la métrique d'erreur en fonction de l'époque.
- `save_best_only = TRUE` : seul le meilleur modèle correspondant à l'époque retenue T est conservé et enregistré.

Ainsi, l'équation (4.26) devient :

$$E_{\text{opt}} = \min_{t \in \mathcal{C}} E_{\text{val}}(t), \tag{4.28}$$

où $\mathcal{C}$ désigne l'ensemble des époques qui vérifient les critères de la fonction `callback_reduce_lr_on_plateau` mentionnés précédemment.

E Monte Carlo Dropout

L'un des inconvénients des modèles d'apprentissage automatique est la grande difficulté d'obtenir des intervalles de confiance. Dans le cas des réseaux de neurones, il est possible de considérer l'approche par méthode de Monte Carlo Dropout (MCD) (Gal et Ghahramani, 2016).

Le Dropout classique est une technique de régularisation pour éviter le phénomène de sur-apprentissage. Des couches de Dropout sont ajoutées uniquement pendant la période d'entraînement, mais pas pendant la période de test.

Contrairement au Dropout classique, le principe du MCD est d'utiliser le Dropout au moment de l'inférence. L'idée-clef est de laisser les couches de Dropout pendant la période d'entraînement et la période de test.

Le cadre théorique utilise une couche de Dropout avant chaque couche en tant qu'approximation d'inférence bayésienne. Le taux de Dropout p est un hyperparamètre qui doit être ajusté. Un faible taux p élimine tout intérêt de la méthode de MCD. Un taux p important peut conduire à une divergence ou peut nécessiter plus d'itérations pour atteindre une convergence. Ainsi, un taux moyen tel que $p \in [0.1, 0.2]$ est raisonnable.

Au moment de la phase de test, on répète N fois le processus (quelques centaines de fois dans l'approche de Gal et Ghahramani (2016)), c'est-à-dire que l'on donne en entrée au réseau de neurones N fois les mêmes inputs mais avec des Dropouts aléatoires. On considère ensuite la moyenne des N prédictions obtenues et l'on peut générer un intervalle de prédictions. L'idée est donc de générer des prévisions aléatoires et les interpréter comme des échantillons d'une distribution probabiliste, les auteurs parlent d'approximation d'inférence bayésienne. Ainsi, une variance élevée indique une incertitude élevée du réseau de neurones et inversement. Une intervention humaine peut être envisagée afin de résoudre les problèmes d'incertitude élevée.

F Indicateurs de correction prédits pour chaque échantillon de test, moyennés par âge au sein de chaque cohorte

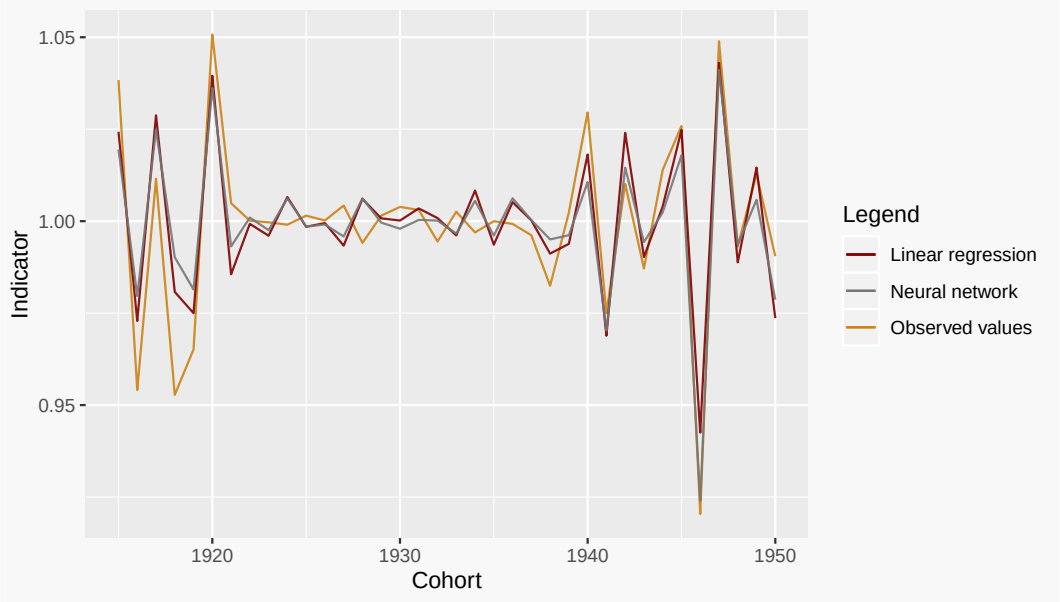


FIGURE F.3. Autriche : Indicateurs de correction prédits

FIGURE F.4. France : Indicateurs de correction prédits

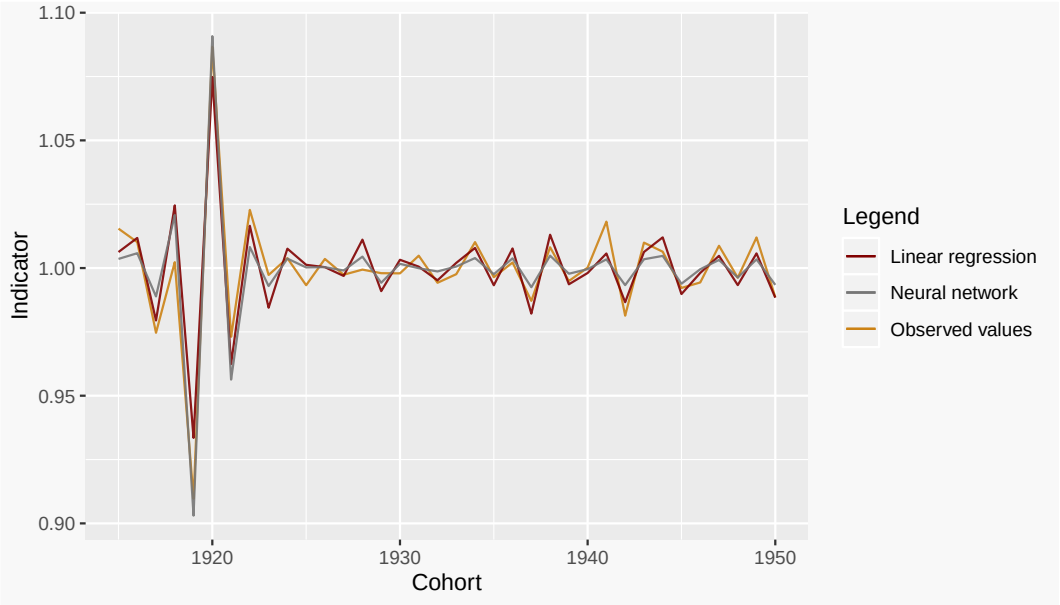


FIGURE F.5. Italie : Indicateurs de correction prédits

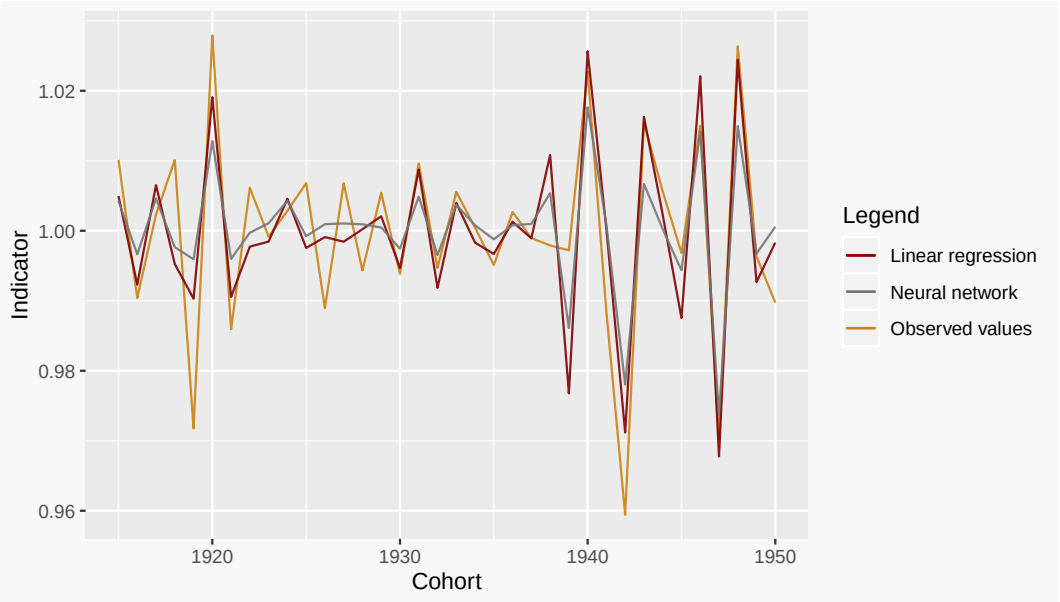


FIGURE F.6. Espagne : Indicateurs de correction prédits

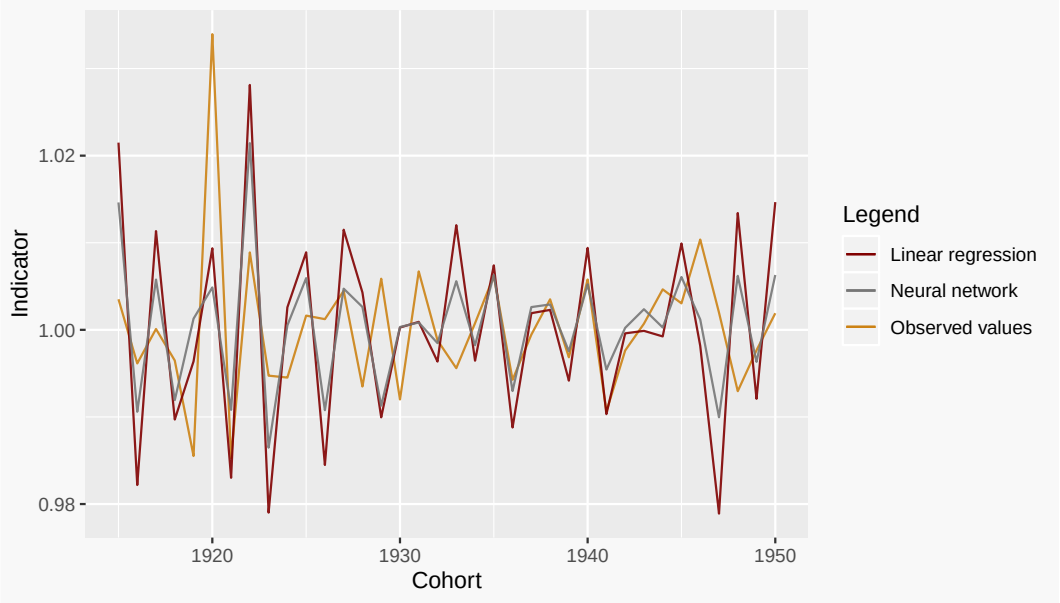


FIGURE F.7. Danemark : Indicateurs de correction prédits

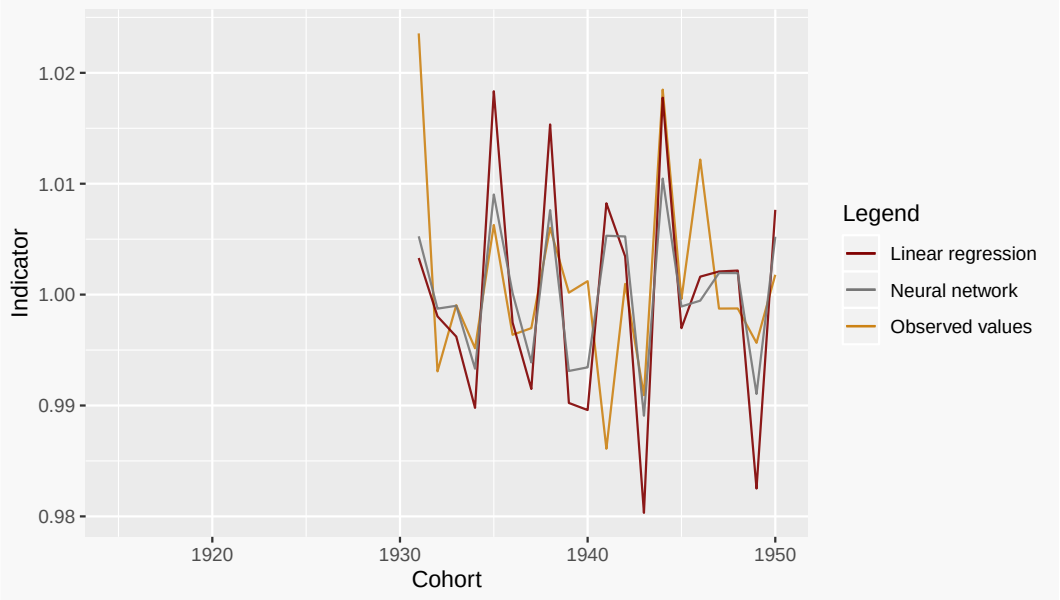


FIGURE F.8. Norvège : Indicateurs de correction prédits

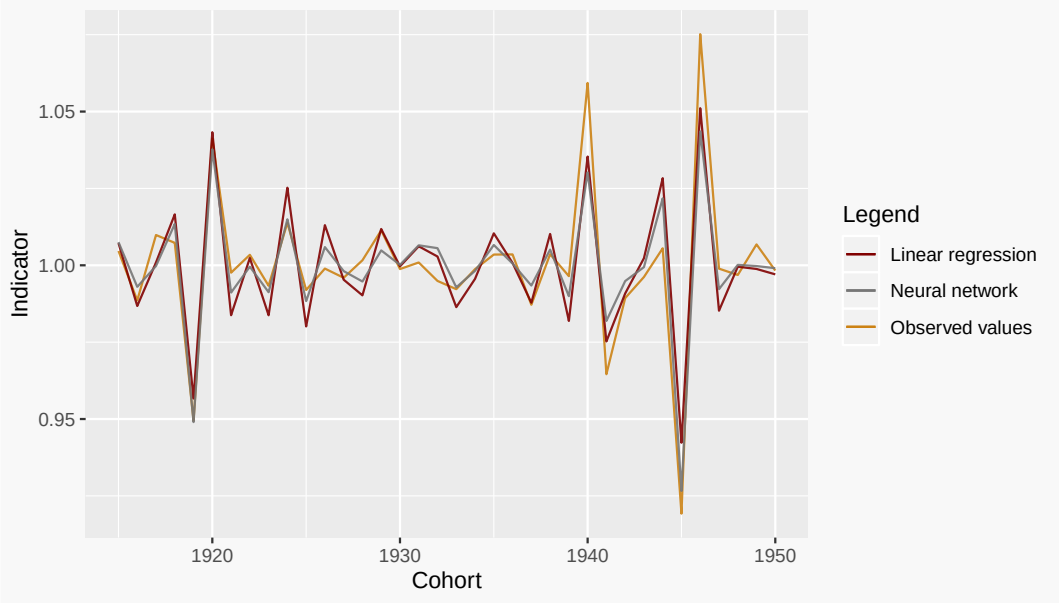


FIGURE F.9. Finlande : Indicateurs de correction prédits

FIGURE F.10. Suède : Indicateurs de correction prédits

G Paramètres du modèle M1, calibré sur des données de mortalité brutes et corrigées

G.1 Calibration sur les données de l'Allemagne de l'Ouest

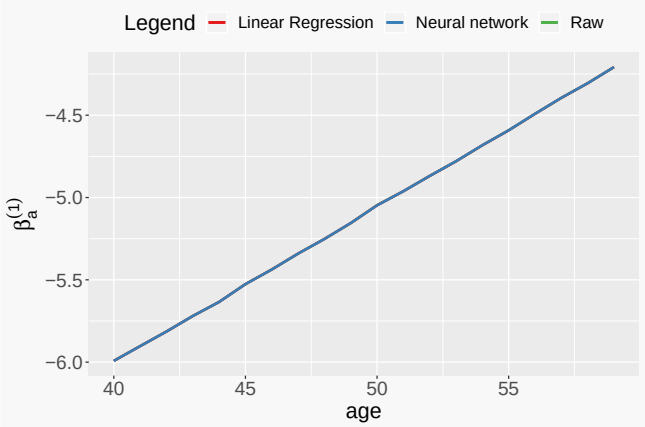


FIGURE G.11. Paramètre $\beta_a^{(1)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest

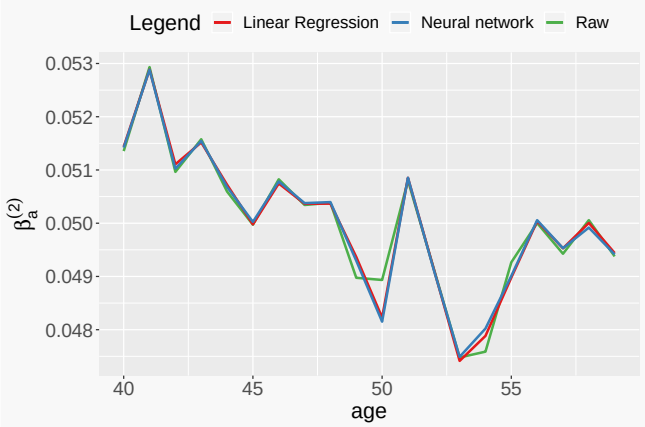


FIGURE G.12. Paramètre $\beta_a^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest

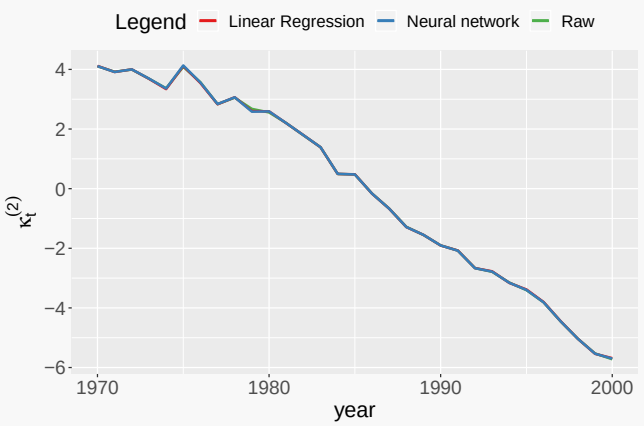


FIGURE G.13. Paramètre $\kappa_t^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest

G.2 Calibration sur les données de la Biélorussie

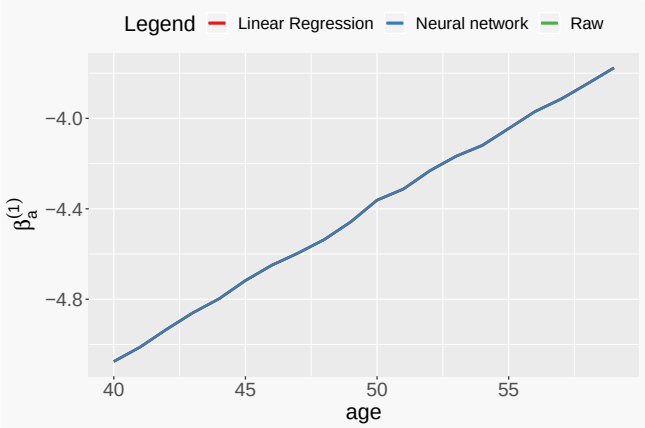


FIGURE G.14. Paramètre $\beta_a^{(1)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Biélorussie

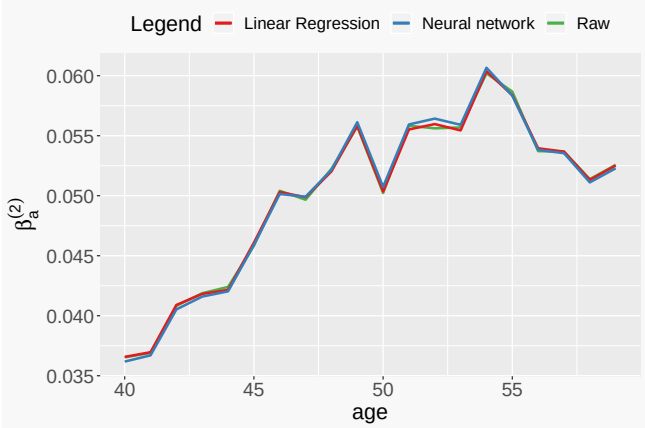


FIGURE G.15. Paramètre $\beta_a^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Biélorussie

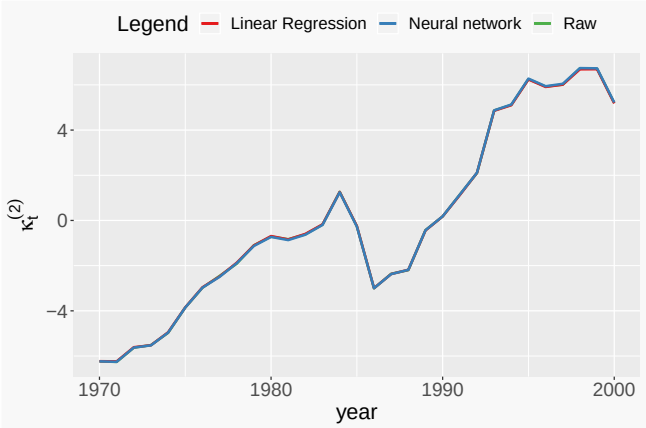


FIGURE G.16. Paramètre $\kappa_t^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Biélorussie

G.3 Calibration sur les données de l'Ukraine

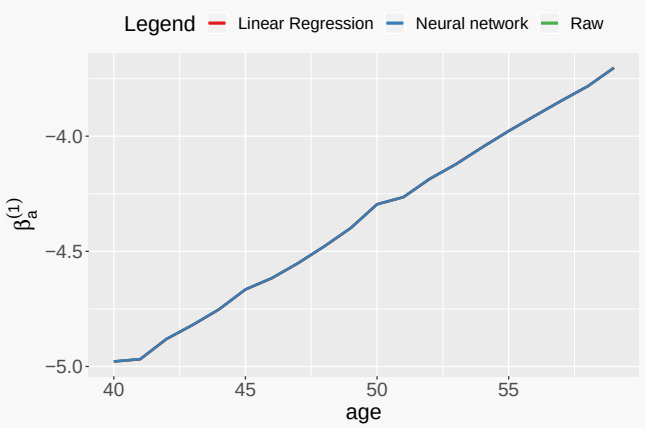


FIGURE G.17. Paramètre $\beta_a^{(1)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Ukraine

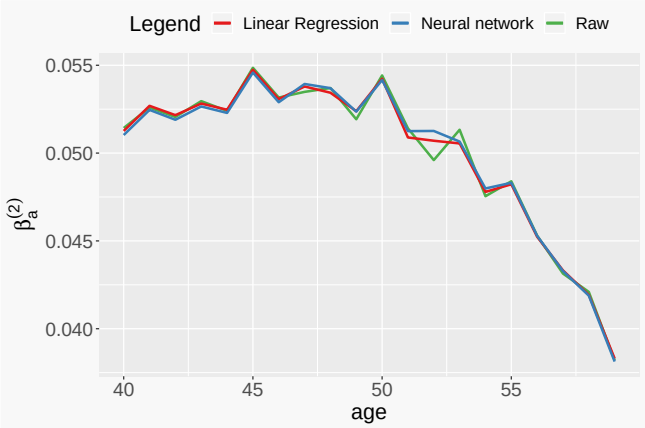


FIGURE G.18. Paramètre $\beta_a^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Ukraine

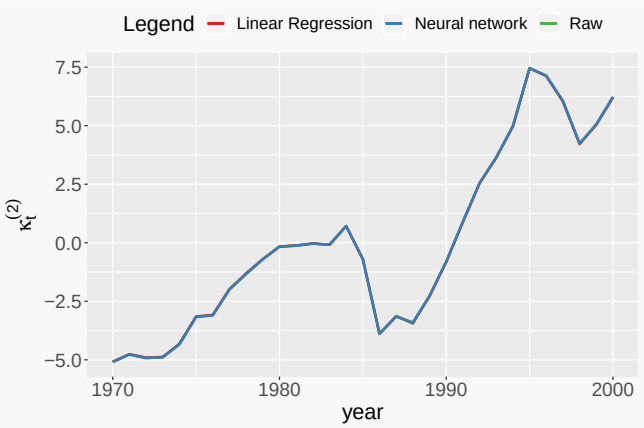


FIGURE G.19. Paramètre $\kappa_t^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de l'Ukraine

G.4 Calibration sur les données de la Pologne

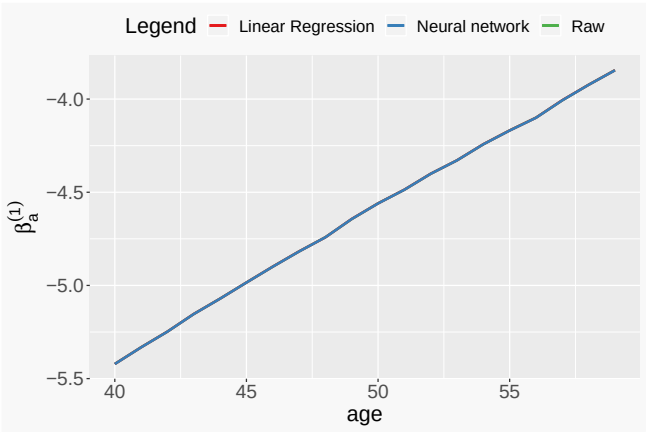


FIGURE G.20. Paramètre $\beta_a^{(1)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Pologne

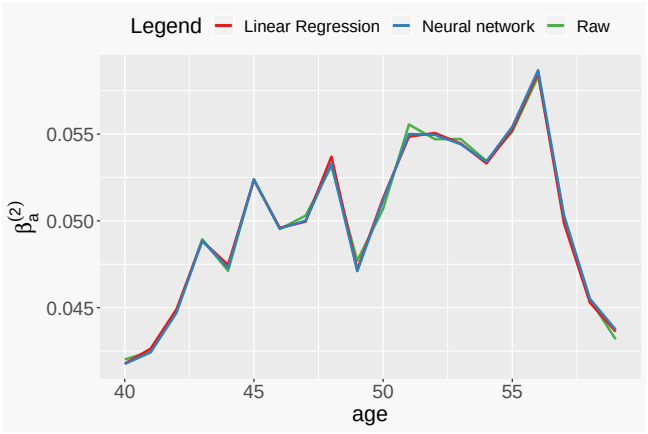


FIGURE G.21. Paramètre $\beta_a^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Pologne

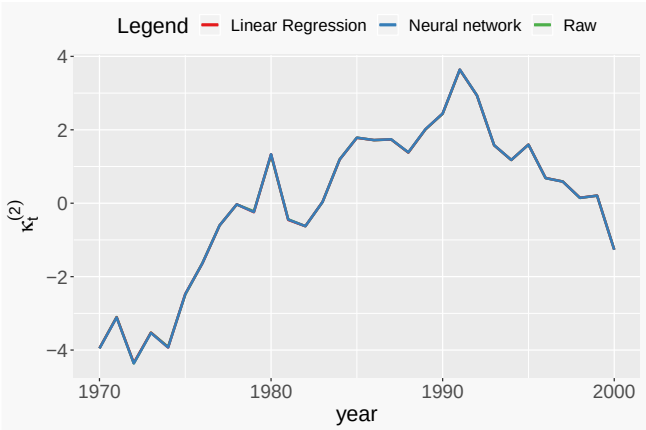


FIGURE G.22. Paramètre $\kappa_t^{(2)}$ du modèle M1 — Données de mortalité brutes et corrigées de la Pologne

H Discussion : Détails du calcul des estimations de mortalité corrigées dans la version 6 de la HMD

Dans la méthode de correction des données de mortalité HMD avec données de fertilité (Wilmoth *et al.*, 2007, 2019), les expositions au risque sont estimées de la manière suivante :

$$\begin{cases} E^{(5)}(a,t) = \frac{1}{2}(P(a,t+1)+P(a,t)) + \frac{1}{6}(D^L(a,t) - D^U(a,t)) \\ E^{(6)}(a,t) = (P(a,t+1)+P(a,t))(1-m_{t-a}) + \frac{D^L(a,t)-D^U(a,t)}{2(1-m_{t-a})}((1-m_{t-a})^2 - \sigma^2_{t-a}) \end{cases}, \tag{4.29}$$

avec

$$\begin{cases} m_{t-a} = \mathbb{E}(B(t-a)) \\ \sigma^2_{t-a} = \mathbb{V}(B(t-a)) \end{cases} \tag{4.30}$$

où $B(t-a)$ designe la variable aléatoire dans $[0,1]$ qui représente la distribution des anniversaires au sein de la cohorte annuelle $t-a$ pour un pays fixé et un genre fixé. Ainsi, la relation suivante est vérifiée :

$$I(a,t) = \frac{(P(a,t+1)+P(a,t))(1-m_{t-a}) + \frac{D^L(a,t)-D^U(a,t)}{2(1-m_{t-a})}((1-m_{t-a})^2 - \sigma^2_{t-a})}{\frac{1}{2}(P(a,t+1)+P(a,t)) + \frac{1}{6}(D^L(a,t) - D^U(a,t))}. \tag{4.31}$$

En notant :

$$\begin{cases} P_\alpha(a,t) = P(a,t+1) + P(a,t) \\ D_\alpha(a,t) = D^L(a,t) - D^U(a,t) \\ \beta^2_{t-a} = \dfrac{1}{2}((1-m_{t-a})^2 - \sigma^2_{t-a}) \end{cases} \tag{4.32}$$

on obtient :

$$I(a,t) = \frac{P_\alpha(a,t)(1-m_{t-a}) + \frac{D_\alpha(a,t)}{(1-m_{t-a})}\beta^2_{t-a}}{E^{(5)}(a,t)}, \tag{4.33}$$

i.e.

$$I(a,t) = A(a,t)(1-m_{t-a}) + B(a,t)\frac{\beta^2_{t-a}}{1-m_{t-a}}, \tag{4.34}$$

avec

$$\begin{cases} A(a,t) = \dfrac{P_\alpha(a,t)}{\frac{1}{2}P_\alpha(a,t) + \frac{1}{6}D_\alpha(a,t)} = \dfrac{P_\alpha(a,t)}{E^{(5)}(a,t)} \\ B(a,t) = \dfrac{D_\alpha(a,t)}{\frac{1}{2}P_\alpha(a,t) + \frac{1}{6}D_\alpha(a,t)} = \dfrac{D_\alpha(a,t)}{E^{(5)}(a,t)} \end{cases} \tag{4.35}$$

Dans cette formule, les variables m_{t-a} et β^2_{t-a} dépendent uniquement de la cohorte $t-a$, mais les variables $P_\alpha(a,t)$ et $D_\alpha(a,t)$ dépendent à la fois de a et de t. Ainsi, il existe une variabilité en âge de la variable I au sein d'une cohorte fixée (voir les figures 3.5 à 3.8).

I Tables réglementaires TGF 05 et TGH 05

Afin de permettre une représentation globale des tables réglementaires, cette annexe comprend les représentations graphiques des données du nombre de survivants L_x, ainsi que les taux de mortalité q_x des tables TGF 05 et TGH 05.

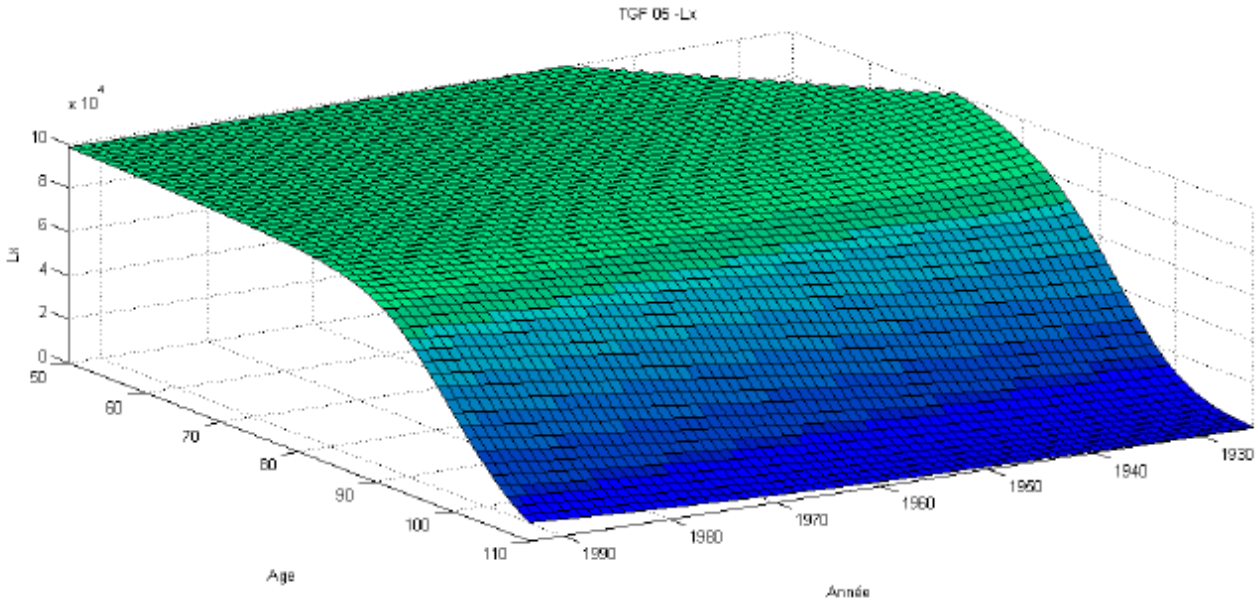


FIGURE I.23. Table TGF 05 : L_x

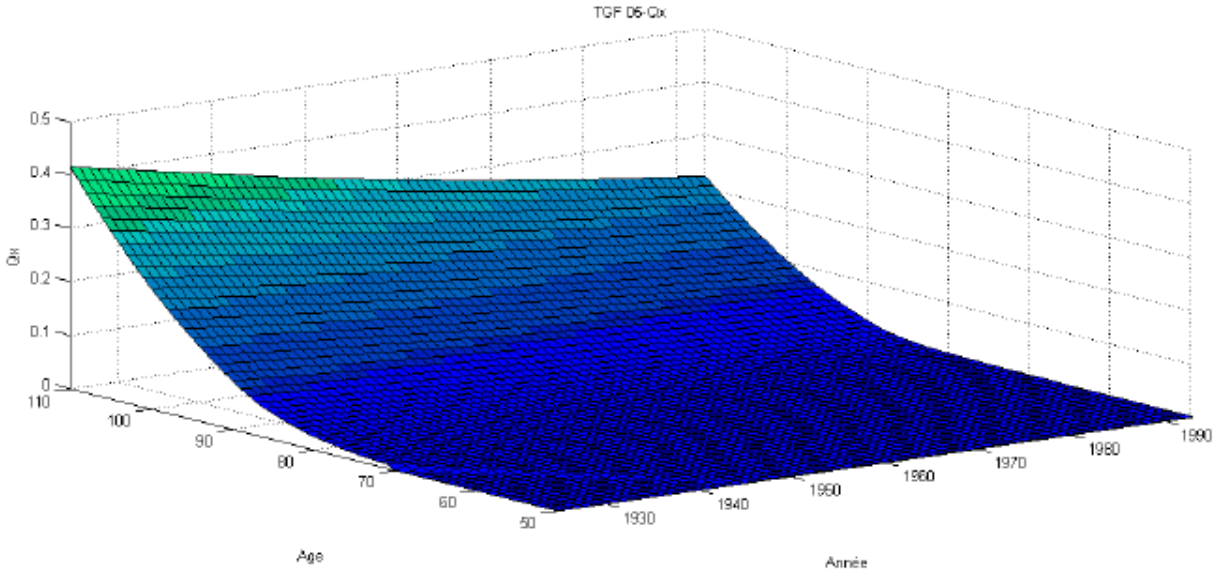


FIGURE I.24. Table TGF 05 : q_x

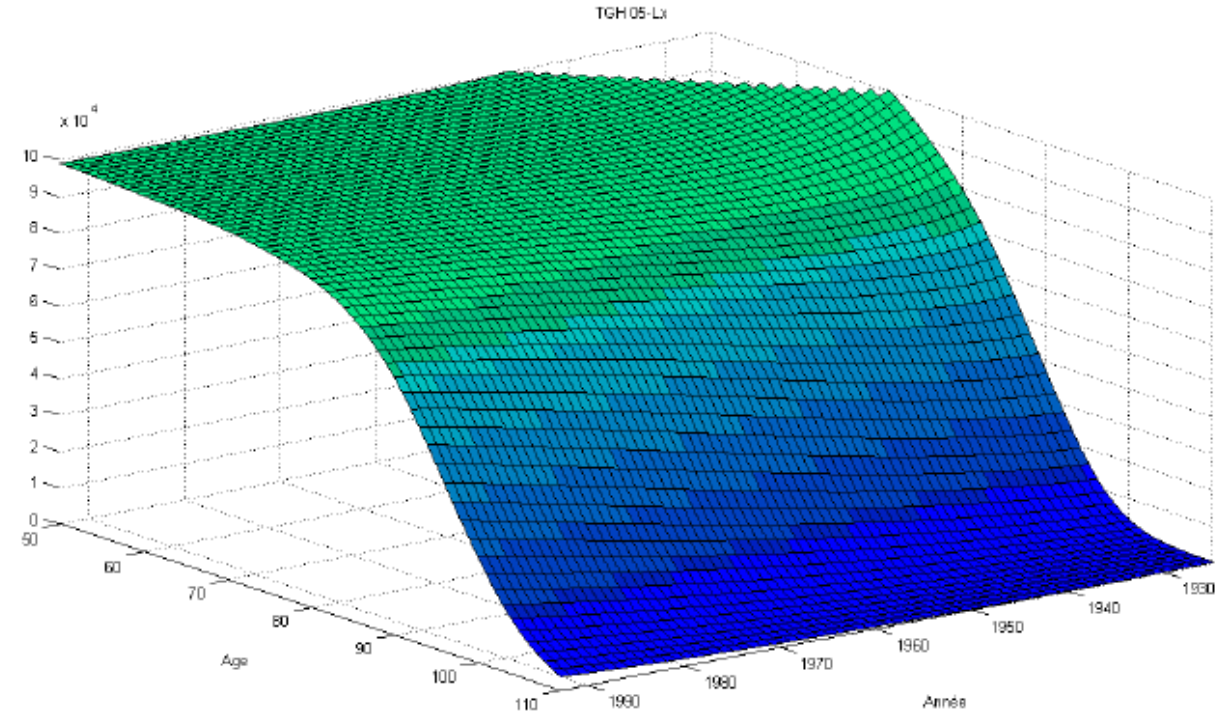


FIGURE I.25. Table TGH 05 : L_x

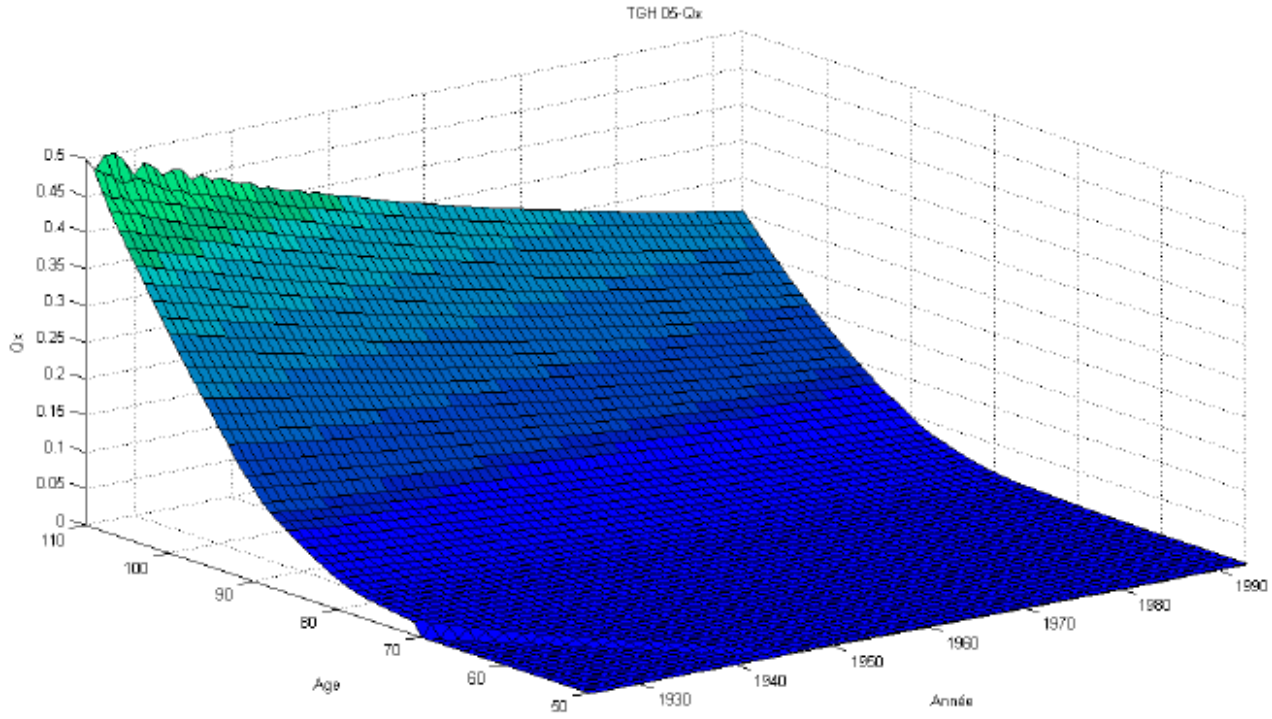


FIGURE I.26. Table TGH 05 : q_x

TABLE DES FIGURES

1 Nuage de points des données de force de mortalité μ de l'échantillon de données en fonction de l'âge ix
2 Scatterplot representing the estimated force of mortality μ as a function of the age in the dataset xiii

1.1 Diagramme de Lexis 4
1.2 Force de mortalité de la population totale française (a) et son logarithme (b) en fonction de l'âge en 2017 — données provenant de la *Human Mortality Database* 6
1.3 Force de mortalité (a) et son logarithme (b) en fonction de l'âge et du sexe en 2017 — données de la population française provenant de la *Human Mortality Database* 7
1.4 Logarithme de la force de mortalité des femmes (a) et des hommes (b) en fonction de l'âge sur la période 1900-2017 — données de la population française provenant de la *Human Mortality Database* 7

2.1 Représentation graphique de la fonction de Heaviside 22
2.2 Représentation graphique de la fonction ReLU 22
2.3 Représentation graphique de la fonction sigmoïde 22
2.4 Représentation graphique de la fonction tanh 22
2.5 Taux d'apprentissage et descente de gradient 24
2.6 Régularisation par Dropout 25
2.7 Représentation d'un raccourci 26
2.8 Illustration de l'approche présentée dans le cas de deux prédicteurs X_1, X_2 et cinq régions $R_1, \ldots, R_5$ 28

3.1 Taux d'amélioration de la mortalité des hommes — données brutes de l'Allemagne de l'Ouest 32
3.2 Nombre de naissances par mois observées en France de 1861 à 2017 32
3.3 Disponibilité des données mensuelles de fertilité dans la HMD et la HFD . . 33
3.4 Diagramme de Lexis 34
3.5 Valeurs de I pour la cohorte 1915 36
3.6 Valeurs de I pour la cohorte 1919 36
3.7 Valeurs de I pour la cohorte 1940 36
3.8 Valeurs de I pour la cohorte 1941 36

3.9 Exemple illustratif simple . . . 39
3.10 Architecture du réseau de neurones . . . 40
3.11 Boîtes à moustaches des erreurs MSE obtenues sur les échantillons de tests . 43
3.12 Indicateurs de correction prédits — Allemagne de l'Ouest . . . 44
3.13 Indicateurs de correction prédits — Ukraine . . . 44
3.14 Indicateurs de correction prédits — Biélorussie . . . 44
3.15 Indicateurs de correction prédits — Pologne . . . 44
3.16 Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données brutes . . . 45
3.17 Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données corrigées par régression linéaire . . . 45
3.18 Taux d'amélioration de mortalité pour l'Allemagne de l'Ouest — Données corrigées par réseau de neurones . . . 45
3.19 Taux d'amélioration de mortalité pour l'Ukraine — Données brutes . . . 46
3.20 Taux d'amélioration de mortalité pour l'Ukraine — Données corrigées par régression linéaire . . . 46
3.21 Taux d'amélioration de mortalité pour l'Ukraine — Données corrigées par réseau de neurones . . . 46
3.22 Taux d'amélioration de mortalité pour la Biélorussie — Données brutes . . . 47
3.23 Taux d'amélioration de mortalité pour la Biélorussie — Données corrigées par régression linéaire . . . 47
3.24 Taux d'amélioration de mortalité pour la Biélorussie — Données corrigées par réseau de neurones . . . 47
3.25 Taux d'amélioration de mortalité pour la Pologne — Données brutes . . . 48
3.26 Taux d'amélioration de mortalité pour la Pologne — Données corrigées par régression linéaire . . . 48
3.27 Taux d'amélioration de mortalité pour la Pologne — Données corrigées par réseau de neurones . . . 48
3.28 Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest . . . 50
3.29 Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de l'Allemagne de l'Ouest . . . 50
3.30 Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de la Biélorussie . . . 50
3.31 Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de la Biélorussie . . . 50
3.32 Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de l'Ukraine . . . 51
3.33 Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de l'Ukraine . . . 51
3.34 Paramètre $\gamma_c^{(3)}$ du modèle M3 — Données de mortalité brutes et corrigées de la Pologne . . . 51
3.35 Paramètre $\gamma_c^{(4)}$ du modèle de Plat simplifié — Données de mortalité brutes et corrigées de la Pologne . . . 51

4.1 Logarithme de la force de mortalité des hommes en fonction de l'âge sur la période 1950-2017 — Données HMD de la France . . . 56
4.2 Disponibilité des données de mortalité de la HMD en fonction du pays . . . 57

4.3 Architecture du réseau de neurones . . . 63
4.4 Nuage de points des données de force de mortalité μ de l'échantillon de données $\mathcal{D}$ en fonction de l'âge . . . 66
4.5 Nuage de points représentant le logarithme des données de force de mortalité μ de l'échantillon de données $\mathcal{D}$ en fonction de l'âge . . . 66
4.6 Répartition du portefeuille de rentes en montants et effectifs par âge au 31 décembre 2009 . . . 69
4.7 Proportion de femmes par âge en montants et effectifs au 31 décembre 2009 . 69
4.8 Répartition du montant des rentes annuelles du portefeuille au 31 décembre 2009 . . . 70
4.9 Répartition du montant des taux techniques du portefeuille . . . 70
4.10 Bilan d'un assureur . . . 72
4.11 VAP des rentes versées aux assurés du portefeuille pour les pays France et Allemagne . . . 75
4.12 VAP des rentes versées aux assurés du portefeuille français en fonction des années de projection 2010-2072 et de la table de mortalité utilisée . . . 79
4.13 VAP des rentes versées aux assurés du portefeuille en fonction des années de projection 2010-2072 et des trois tables de mortalité utilisées . . . 79
D.1 Courbes idéalisées d'entraînement et de validation . . . 91
D.2 Courbe réelle de validation . . . 91
F.3 Autriche : Indicateurs de correction prédits . . . 94
F.4 France : Indicateurs de correction prédits . . . 94
F.5 Italie : Indicateurs de correction prédits . . . 95
F.6 Espagne : Indicateurs de correction prédits . . . 95
F.7 Danemark : Indicateurs de correction prédits . . . 96
F.8 Norvège : Indicateurs de correction prédits . . . 96
F.9 Finlande : Indicateurs de correction prédits . . . 97
F.10 Suède : Indicateurs de correction prédits . . . 97
G.11 Paramètre $\beta_a^{(1)}$ du modèle M1 — Allemagne de l'Ouest . . . 98
G.12 Paramètre $\beta_a^{(2)}$ du modèle M1 — Allemagne de l'Ouest . . . 98
G.13 Paramètre $\kappa_t^{(2)}$ du modèle M1 — Allemagne de l'Ouest . . . 98
G.14 Paramètre $\beta_a^{(1)}$ du modèle M1 — Biélorussie . . . 99
G.15 Paramètre $\beta_a^{(2)}$ du modèle M1 — Biélorussie . . . 99
G.16 Paramètre $\kappa_t^{(2)}$ du modèle M1 — Biélorussie . . . 99
G.17 Paramètre $\beta_a^{(1)}$ du modèle M1 — Ukraine . . . 100
G.18 Paramètre $\beta_a^{(2)}$ du modèle M1 — Ukraine . . . 100
G.19 Paramètre $\kappa_t^{(2)}$ du modèle M1 — Ukraine . . . 100
G.20 Paramètre $\beta_a^{(1)}$ du modèle M1 — Pologne . . . 101
G.21 Paramètre $\beta_a^{(2)}$ du modèle M1 — Pologne . . . 101
G.22 Paramètre $\kappa_t^{(2)}$ du modèle M1 — Pologne . . . 101
I.23 Table TGF 05 : L_x . . . 103
I.24 Table TGF 05 : q_x . . . 103
I.25 Table TGH 05 : L_x . . . 104
I.26 Table TGH 05 : q_x . . . 104

LISTE DES TABLEAUX

1 Erreurs MSE obtenues sur les échantillons de test . . . vii
2 Bilan comparatif des approches . . . ix
3 Test set MSEs of the baseline linear regression and the neural network . . . xi
4 Comparative assessment of approaches . . . xiii

3.1 Erreurs MSE obtenues sur les échantillons de test . . . 42
3.2 Calibration sur les données de l'Allemagne de l'Ouest . . . 53
3.3 Calibration sur les données de la Biélorussie . . . 53
3.4 Calibration sur les données de l'Ukraine . . . 53
3.5 Calibration sur les données de la Pologne . . . 53

4.1 Comparaison des résultats obtenus par les modèles . . . 63
4.2 Exemple illustratif : comparaison des métriques MSE et LMSE . . . 64
4.3 Comparaison des erreurs en incluant la métrique LMSE . . . 64
4.4 Comparaison finale des résultats obtenus par les modèles . . . 65
4.5 Bilan comparatif des approches . . . 67
4.6 Résultats des sommes de projections obtenues en fonction du pays et du modèle utilisé . . . 73
4.7 VAP totale des rentes versées aux assurés du portefeuille sur la période 2010-2072 en fonction de la table de mortalité considérée . . . 75
4.8 VAP totale des rentes versées aux assurés du portefeuille français sur la période 2010-2072 en fonction de la table de mortalité utilisée . . . 78
A.1 Acronymes utilisés . . . 88
A.2 Symboles utilisés . . . 88
C.3 Liste des pays disponibles dans la HMD . . . 90