

Faculté des sciences

Agrégation bayésienne contextuelle de modèles de mortalité :

Une approche par *stacking* hiérarchique
selon l'âge et l'horizon

Auteur·es : Minko Benjamin André Junior

Promoteur·rices : Karim Barigou

Lecteur·rices : Nathalie Lucas

Année académique 2025–2026

Université catholique de Louvain

Institut de statistique, biostatistique et actuariat

Agrégation bayésienne contextuelle de modèles de mortalité

Une approche par *stacking* hiérarchique
selon l'âge et l'horizon

Mémoire présenté en vue de l'obtention du
Master en Sciences actuarielles

par

Minko Benjamin André Junior

Promoteur : Karim Barigou

Lectrice : Nathalie Lucas

Année académique 2025–2026

Résumé

La projection de la mortalité constitue un élément central de la mesure du risque de longévité auquel sont exposés les assureurs-vie, les fonds de pension et les régimes de retraite. Bien que de nombreux modèles stochastiques aient été développés pour représenter l'évolution de la mortalité, leurs performances demeurent hétérogènes selon les âges considérés et les horizons de projection. Cette variabilité rend le choix d'un modèle unique particulièrement délicat et introduit une incertitude structurelle qui s'ajoute à l'incertitude paramétrique.

Afin de limiter cette dépendance à une spécification particulière, plusieurs approches d'agrégation bayésienne ont été proposées, notamment le pseudo-BMA et le stacking fondé sur la validation Leave-Future-Out. Ces méthodes combinent les distributions prédictives de plusieurs modèles en leur attribuant des poids reflétant leurs performances hors échantillon. Toutefois, dans leur formulation classique, ces pondérations sont globales et ne permettent pas de capturer l'hétérogénéité des performances selon l'âge ou l'horizon de projection.

Ce mémoire introduit une extension du stacking bayésien dans laquelle les poids attribués aux modèles deviennent des fonctions de l'âge et de l'horizon. Une structure hiérarchique est imposée aux coefficients de pondération afin de régulariser ces variations et d'éviter qu'elles ne reflètent uniquement le bruit de validation. L'approche est évaluée au moyen d'une étude de simulation, puis appliquée aux données belges de mortalité pour les âges de 50 à 90 ans et les années 1970 à 2024.

Les résultats montrent que le stacking contextuel hiérarchique parvient à identifier les variations pertinentes lorsque celles-ci sont effectivement présentes, tout en préservant une stabilité des poids lorsque l'information disponible est limitée. Sur la période de test 2016 – 2024, cette méthode obtient les meilleures performances globales selon le score logarithmique, le CRPS et l'erreur absolue moyenne, et fournit des intervalles prédictifs correctement calibrés. Les effets sur les probabilités de survie et les rentes temporaires demeurent modérés aux âges intermédiaires, mais deviennent plus marqués aux âges élevés.

Mots-clés : mortalité stochastique, risque de longévité, agrégation de modèles, stacking bayésien, validation Leave-Future-Out, modèle hiérarchique.

Remerciements

Je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

Je remercie tout d'abord mon promoteur, Karim Barigou, pour son accompagnement attentif, ses conseils éclairés et la confiance qu'il m'a accordée tout au long de ce travail. Sa disponibilité et son exigence scientifique ont été déterminantes dans l'aboutissement de cette recherche.

Je tiens également à remercier mes amis, dont le soutien constant, les encouragements et la bienveillance ont rendu ce parcours plus léger et plus riche. Leur présence à mes côtés a été un véritable moteur dans les moments de doute comme dans ceux d'enthousiasme.

Ma reconnaissance va aussi aux membres de mon église, qui m'ont porté par leurs prières, leurs paroles de foi et leur affection. Leur soutien spirituel a été une source de paix et de force tout au long de ce projet.

Je remercie de tout cœur ma maman, Edo Afane Rose Jahel. Son amour inconditionnel, sa patience et ses sacrifices ont rendu possible chacune des étapes de mon parcours. Elle a financé mes études avec une générosité et une détermination qui dépassent les mots, et c'est en grande partie grâce à elle que ce mémoire a pu voir le jour. Sa confiance en moi demeure l'un des plus grands cadeaux de ma vie.

Enfin, je rends grâce à la magnifique et sublime personne du Saint-Esprit, par qui j'ai puisé la force d'avancer et de persévérer. Sa présence, sa lumière et sa direction ont guidé mes pensées, mes décisions et mes efforts.

À toutes et à tous, merci.

Table des matières

Résumé

Remerciements

Introduction	1
1. Modélisation stochastique de la mortalité	3
1.1. Modèles stochastiques de mortalité	3
1.1.1. Modèle de Lee-Carter (LC)	6
1.1.2. Modèle de Renshaw-Haberman (RH)	7
1.1.3. Modèle âge-période-cohorte (APC)	7
1.1.4. Modèle de Cairns-Blake-Dowd (CBD)	7
1.1.5. Modèle M6	8
1.2. Projection des composantes temporelles	8
1.3. Ensemble des modèles candidats	9
2. Agrégation bayésienne et pondérations contextuelles	10
2.1. Fondements de l'estimation et de l'agrégation bayésiennes	11
2.1.1. Théorème de Bayes appliqué à la modélisation de la mortalité . . .	11
2.2. Spécification des lois <i>a priori</i>	12
2.3. Approche par moyennage bayésien de modèles (BMA)	15
2.4. Approche par pseudo-BMA	18
2.5. Approche par <i>stacking</i> bayésien	21
2.5.1. Limites des pondérations fixes et justification d'une extension contex- tuelle	23
2.6. Stacking contextuel et régularisation hiérarchique	23
2.6.1. Construction des observations de validation temporelle	24
2.6.2. Standardisation de l'âge et de l'horizon	26
2.6.3. Transformation <i>softmax</i>	27
2.6.4. Critère prédictif commun aux trois formes de <i>stacking</i>	27
2.6.5. Régularisation hiérarchique	28
2.6.6. Distribution a posteriori généralisée des paramètres de pondération	29
2.6.7. Distribution prédictive finale	30

2.7.	Métriques de performance et règles de score probabilistes	30
2.7.1.	Le score logarithmique (LogS)	31
2.7.2.	Le <i>Continuous Ranked Probability Score</i> (CRPS)	32
2.7.3.	Erreur absolue moyenne	33
2.7.4.	Couverture des intervalles prédictifs	34
2.8.	Construction des quantités actuarielles	34
3.	Évaluation des méthodes et implications actuarielles	36
3.1.	Enseignements de l'étude de simulation	36
3.1.1.	Scénarios simulés	37
3.1.2.	Récupération des surfaces de poids	39
3.1.3.	Faible information et variations contextuelles artificielles	42
3.1.4.	Performance prédictive hors échantillon	43
3.1.5.	Validité numérique et conclusion de la simulation	45
3.2.	Application à la mortalité belge	46
3.2.1.	Données et stratégie d'évaluation temporelle	46
3.2.2.	Hétérogénéité des performances et estimation des poids	47
3.2.3.	Performance prédictive sur la période de test	49
3.2.4.	Variation selon l'âge et l'horizon	51
3.2.5.	Incertitude sur les écarts de performance	52
3.3.	Transmission aux quantités actuarielles	53
3.3.1.	Construction des trajectoires	53
3.3.2.	Probabilités de survie	54
3.3.3.	Rentes temporaires	54
3.4.	Discussion et limites	56
3.5.	Conclusion du chapitre	58
	Conclusion	59
	Bibliographie	
A.	Implémentation numérique et reproductibilité	
A.1.	Estimation des modèles de mortalité	
A.2.	Boucle de validation <i>Leave-Future-Out</i>	
A.3.	Estimation des trois formes de <i>stacking</i>	
A.4.	Prévisions et trajectoires actuarielles	
B.	Étude de simulation et analyses complémentaires	
B.1.	Processus générateur de la simulation	
B.2.	Diagnostics numériques des répétitions	
B.3.	Analyses de sensibilité complémentaires	

Table des figures

3.1.	Récupération des surfaces de poids selon le scénario simulé.	41
3.2.	Surfaces de poids vraies et estimées dans le scénario dépendant de l'âge et de l'horizon.	42
3.3.	Performance prédictive selon le scénario simulé. La première ligne présente le LogS et la seconde le CRPS.	44
3.4.	Log-taux bruts de mortalité en Belgique selon l'âge et l'année, 1970 à 2024.	47
3.5.	Hétérogénéité des performances individuelles selon l'âge et l'horizon.	48
3.6.	Écarts entre les poids hiérarchiques et non régularisés selon l'âge et l'horizon.	49
3.7.	Performance prédictive des méthodes de <i>stacking</i> selon l'âge et l'horizon. . .	51
3.8.	Probabilités de survie après agrégation des forces sur les dix premières années de projection. Les bandes représentent les intervalles centraux à 95 %.	54

Introduction

L'amélioration de la longévité représente un enjeu important pour les assureurs-vie, les fonds de pension et les régimes de retraite. Lorsque les assurés ou les bénéficiaires vivent plus longtemps que prévu, les prestations doivent être versées pendant une durée plus longue. La mesure de ce risque repose donc en grande partie sur la capacité à représenter la mortalité actuelle et à prévoir son évolution future (Barigou et al. 2023).

De nombreux modèles stochastiques ont été proposés pour décrire la mortalité en fonction de l'âge, de l'année calendaire et, dans certains cas, de la cohorte de naissance. Le modèle de Lee-Carter représente l'évolution temporelle au moyen d'un facteur commun (Lee et Carter 1992). Le modèle de Renshaw-Haberman ajoute un effet de cohorte (Renshaw et Haberman 2006). Le modèle de Cairns-Blake-Dowd utilise deux facteurs temporels pour représenter le niveau et la pente de la mortalité aux âges élevés (Cairns, Blake et Dowd 2006), tandis que le modèle M6 complète cette structure par un effet de cohorte (Cairns, Blake, Dowd et al. 2009). Ces modèles, présentés dans la section 1.3, reposent sur des hypothèses différentes et peuvent donc produire des performances hétérogènes selon les âges et les horizons de projection.

Le choix d'un modèle unique introduit ainsi une incertitude structurelle, qui s'ajoute à l'incertitude paramétrique propre à chaque modèle. Pour réduire cette dépendance à une seule spécification, plusieurs approches d'agrégation bayésienne ont été développées. Le moyennage bayésien de modèles (BMA) combine les distributions prédictives en pondérant les modèles par leurs probabilités *a posteriori*. Pour des objectifs de prévision, il peut toutefois être préférable de fonder les pondérations sur des performances réellement observées hors échantillon. Le pseudo-BMA et le *stacking* suivent cette logique (Yao, Vehtari et al. 2018).

Dans le cas de séries temporelles, l'évaluation prédictive doit respecter l'ordre chronologique des observations. Une validation classique peut permettre à des données futures d'intervenir indirectement dans la prévision du passé. Pour cette raison, Barigou et al. (2023) adaptent le pseudo-BMA et le *stacking* à la prévision de la mortalité au moyen d'une validation *Leave-Future-Out*, dans laquelle les modèles sont estimés sur une période passée puis évalués sur des années futures qui n'ont pas été utilisées lors de leur ajustement.

Dans sa forme habituelle, le *stacking* produit un seul vecteur de poids, identique pour tous les âges et tous les horizons. Cette hypothèse peut être restrictive lorsque les performances relatives des modèles varient dans l'espace de prévision. Des pondérations dépendant de l'horizon ont déjà été étudiées pour des prévisions ponctuelles de mortalité (Kessy

et al. 2022), mais une extension permettant de conserver une agrégation de distributions prédictives complètes tout en tenant compte simultanément de l'âge et de l'horizon reste à développer.

La question principale de ce mémoire est donc la suivante : une pondération contextuelle des modèles permet-elle d'améliorer les prévisions de mortalité sans introduire des variations excessives lorsque l'information disponible est limitée ?

Pour répondre à cette question, nous comparons trois méthodes. Le *stacking* global utilise des poids constants. Le *stacking* contextuel non régularisé permet aux poids de varier selon l'âge et l'horizon. Enfin, le *stacking* contextuel hiérarchique utilise la même fonction de pondération, mais introduit des distributions *a priori* sur ses coefficients. Cette régularisation, inspirée du *stacking* hiérarchique de Yao, Pirš et al. (2022), réalise une mise en commun partielle des effets contextuels : lorsque les données contiennent peu d'information, les poids se rapprochent d'une pondération globale ; lorsque les différences entre les modèles sont mieux identifiées, ils peuvent varier davantage. La méthode complète est développée dans la section 2.6.

Les cinq modèles de mortalité sont estimés dans un cadre bayésien binomial négatif, suivant le cadre de Barigou et al. (2023). La loi binomiale négative permet de tenir compte d'une variabilité du nombre de décès supérieure à celle d'une loi de Poisson. Les paramètres et les trajectoires futures sont obtenus à partir de simulations *a posteriori*.

Une étude de simulation est d'abord menée dans un cadre où la structure véritable des poids est connue. Elle permet de vérifier si les méthodes retrouvent les variations réellement présentes et si la régularisation évite d'en créer lorsqu'elles n'existent pas. Les scénarios et les critères d'évaluation sont décrits dans la section 3.1.

Les méthodes sont ensuite appliquées aux données belges de la *Human Mortality Database* (Human Mortality Database 2026). L'analyse porte sur la population totale, pour les âges de 50 à 90 ans et les années 1970 à 2024. Les poids sont estimés à partir de plusieurs exercices historiques de prévision. Les années 2016 à 2024 sont conservées entièrement à l'écart et servent uniquement à l'évaluation finale (section 3.2).

Les distributions prédictives sont comparées au moyen du score logarithmique, du *Continuous Ranked Probability Score* (CRPS), de l'erreur absolue moyenne (MAE) et de la couverture des intervalles prédictifs. Elles sont enfin utilisées pour calculer des probabilités de survie et des valeurs de rentes temporaires, selon la construction présentée dans la section 2.8.

Le premier chapitre présente le cadre probabiliste et les modèles de mortalité retenus. Le deuxième chapitre introduit les méthodes d'agrégation et développe le *stacking* contextuel hiérarchique. Le troisième chapitre rassemble les résultats de l'étude de simulation, l'application à la mortalité belge et les implications actuarielles.

1. Modélisation stochastique de la mortalité

Le risque de longévité correspond au risque que les assurés ou les bénéficiaires vivent plus longtemps que prévu. Il représente un enjeu important pour les assureurs-vie, les fonds de pension et les régimes de retraite, qui doivent verser des prestations pendant une durée potentiellement plus longue qu'anticipé. Pour mesurer ce risque et projeter l'évolution future de la mortalité, la littérature actuarielle a développé de nombreux modèles stochastiques.

Ces modèles décrivent généralement la mortalité à partir de plusieurs dimensions, notamment l'âge, l'année calendaire et, selon les spécifications retenues, la cohorte de naissance. Chaque structure repose sur des hypothèses particulières et peut donc représenter différemment les tendances historiques et leur prolongement dans le futur.

Ce premier chapitre présente le cadre probabiliste et les modèles stochastiques de mortalité retenus dans ce travail. Il expose leurs composantes principales, les contraintes nécessaires à leur identification et la manière dont leurs facteurs temporels sont projetés. L'objectif n'est pas de désigner à ce stade un modèle supérieur aux autres, mais de construire un ensemble de modèles suffisamment diversifié pour étudier ensuite l'incertitude structurelle.

1.1. Modèles stochastiques de mortalité

Les données de mortalité observées à l'échelle d'une population peuvent présenter une variabilité supérieure à celle prévue par une loi de Poisson. Ce phénomène, appelé *surdispersion*, peut notamment s'expliquer par l'hétérogénéité non observée entre les individus d'une même population. Les conditions socio-économiques, l'état de santé, le niveau d'éducation ou encore les comportements individuels peuvent en effet influencer la mortalité et accroître la variabilité du nombre de décès observés.

Afin de tenir compte de cette variabilité supplémentaire, plusieurs travaux proposent de modéliser le nombre de décès au moyen d'une loi binomiale négative plutôt que d'une simple loi de Poisson (Delwarde et al. 2007 ; Wong et al. 2018). Cette approche est également retenue par Barigou et al. (2023) dans leur cadre bayésien de modélisation et d'agrégation des modèles de mortalité.

1. Modélisation stochastique de la mortalité

Pour chaque âge x et chaque année calendaire t , on note $D_{x,t}$ la variable aléatoire représentant le nombre de décès observés et $e_{x,t}$ l'exposition correspondante au risque. Les âges considérés sont notés

$$x = x_1, \dots, x_A,$$

et les années calendaires

$$t = t_1, \dots, t_N,$$

où A et N désignent respectivement le nombre d'âges et le nombre d'années étudiées. On note enfin $\mu_{x,t}$ la force de mortalité à l'âge x durant l'année t .

Une première spécification consiste à supposer que, conditionnellement à la force de mortalité, le nombre de décès suit une loi de Poisson :

$$D_{x,t} \mid \mu_{x,t} \stackrel{\text{ind}}{\sim} \text{Poisson}(\mu_{x,t} e_{x,t}). \quad (1.1)$$

La force de mortalité est ensuite représentée au moyen du prédicteur suivant :

$$\log \mu_{x,t} = \alpha_x + \sum_{i=1}^p \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x} + \log \nu_{x,t}, \quad (1.2)$$

où l'effet aléatoire multiplicatif $\nu_{x,t}$ est défini par

$$\nu_{x,t} \mid \phi \stackrel{\text{ind}}{\sim} \text{Gamma}(\phi, \phi). \quad (1.3)$$

La loi gamma est ici paramétrée par un paramètre de forme et un paramètre de taux tous deux égaux à ϕ . Par conséquent,

$$\mathbb{E}[\nu_{x,t}] = 1 \quad \text{et} \quad \text{Var}(\nu_{x,t}) = \frac{1}{\phi}. \quad (1.4)$$

Le paramètre α_x représente le niveau moyen de mortalité propre à l'âge x . Les indices $\kappa_t^{(i)}$ décrivent les tendances et les variations de la mortalité au cours du temps, tandis que les coefficients $\beta_x^{(i)}$ mesurent la sensibilité de la mortalité à l'âge x à chacun de ces indices de période.

Le terme γ_{t-x} représente l'effet de cohorte associé aux individus nés durant l'année $t - x$. Le coefficient $\beta_x^{(0)}$ permet, lorsqu'il est présent dans le modèle, de faire varier l'intensité de cet effet selon l'âge. Enfin, $\nu_{x,t}$ introduit une variabilité supplémentaire destinée à représenter l'hétérogénéité résiduelle qui n'est pas directement expliquée par les composantes d'âge, de période et de cohorte. Le paramètre ϕ contrôle l'ampleur de cette surdispersion.

Afin d'alléger les notations, on définit le prédicteur linéaire

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^p \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}. \quad (1.5)$$

1. Modélisation stochastique de la mortalité

L'équation (1.2) peut alors s'écrire sous la forme

$$\mu_{x,t} = \exp(\eta_{x,t})\nu_{x,t}. \quad (1.6)$$

Le mélange entre la loi de Poisson définie par l'équation (1.1) et l'effet aléatoire gamma de l'équation (1.3) est équivalent à une loi binomiale négative :

$$D_{x,t} \mid \alpha_x, \beta_x, \kappa_t, \gamma_{t-x}, \phi \stackrel{\text{ind}}{\sim} \text{NegBin}(e_{x,t} \exp(\eta_{x,t}), \phi). \quad (1.7)$$

Dans cette expression, la loi binomiale négative est paramétrée par sa moyenne

$$m_{x,t} = e_{x,t} \exp(\eta_{x,t}) \quad (1.8)$$

et par son paramètre de concentration ϕ .

En appliquant la loi de l'espérance totale, on obtient

$$\begin{aligned} \mathbb{E}[D_{x,t}] &= \mathbb{E}[\mathbb{E}(D_{x,t} \mid \nu_{x,t})] \\ &= \mathbb{E}[e_{x,t} \exp(\eta_{x,t})\nu_{x,t}] \\ &= e_{x,t} \exp(\eta_{x,t})\mathbb{E}[\nu_{x,t}]. \end{aligned} \quad (1.9)$$

Comme $\mathbb{E}[\nu_{x,t}] = 1$, il vient

$$\mathbb{E}[D_{x,t}] = e_{x,t} \exp(\eta_{x,t}). \quad (1.10)$$

La loi de la variance totale donne également

$$\begin{aligned} \text{Var}(D_{x,t}) &= \mathbb{E}[\text{Var}(D_{x,t} \mid \nu_{x,t})] + \text{Var}[\mathbb{E}(D_{x,t} \mid \nu_{x,t})] \\ &= \mathbb{E}[e_{x,t} \exp(\eta_{x,t})\nu_{x,t}] + \text{Var}[e_{x,t} \exp(\eta_{x,t})\nu_{x,t}] \\ &= \mathbb{E}[D_{x,t}] + (e_{x,t} \exp(\eta_{x,t}))^2 \text{Var}(\nu_{x,t}) \\ &= \mathbb{E}[D_{x,t}] + \frac{\mathbb{E}[D_{x,t}]^2}{\phi}. \end{aligned} \quad (1.11)$$

Ainsi,

$$\text{Var}(D_{x,t}) = \mathbb{E}[D_{x,t}] \left(1 + \frac{\mathbb{E}[D_{x,t}]}{\phi} \right) > \mathbb{E}[D_{x,t}]. \quad (1.12)$$

La loi binomiale négative conserve donc la même moyenne que la loi de Poisson correspondante, tout en autorisant une variance plus élevée. Le paramètre ϕ contrôle l'ampleur de la surdispersion : plus ϕ est faible, plus la variance du nombre de décès est importante. À l'inverse, lorsque ϕ tend vers l'infini, la variance se rapproche de la moyenne et le modèle

1. Modélisation stochastique de la mortalité

converge vers un modèle de Poisson standard.

Sous l'hypothèse que les nombres de décès sont conditionnellement indépendants entre les différentes cellules âge-année, la vraisemblance associée aux observations $d_{x,t}$ est donnée par

$$\begin{aligned} \mathcal{L}(\alpha, \beta, \kappa, \gamma, \phi) = \prod_{x,t} & \left[\frac{\Gamma(d_{x,t} + \phi)}{\Gamma(\phi)\Gamma(d_{x,t} + 1)} \right. \\ & \times \left(\frac{e_{x,t} \exp(\eta_{x,t})}{e_{x,t} \exp(\eta_{x,t}) + \phi} \right)^{d_{x,t}} \\ & \left. \times \left(\frac{\phi}{e_{x,t} \exp(\eta_{x,t}) + \phi} \right)^{\phi} \right]. \end{aligned} \quad (1.13)$$

Le prédicteur général défini par l'équation (1.5) permet de représenter plusieurs modèles classiques de mortalité stochastique. Ces modèles se distinguent par les composantes d'âge, de période et de cohorte qu'ils incluent, ainsi que par les contraintes imposées à leurs paramètres. Dans la suite de ce travail, nous retenons les cinq modèles étudiés par [Barigou et al. \(2023\)](#) : le modèle de Lee-Carter, le modèle de Renshaw-Haberman, le modèle âge-période-cohorte, le modèle de Cairns-Blake-Dowd et le modèle M6. Cet ensemble conserve des structures différentes d'âge, de période et de cohorte, tout en permettant une comparaison directe avec leur cadre d'agrégation.

1.1.1. Modèle de Lee-Carter (LC)

Le modèle de Lee-Carter constitue l'un des principaux modèles de référence pour l'analyse et la projection de la mortalité. Proposé par [Lee et Carter \(1992\)](#), il décrit l'évolution des taux de mortalité à partir d'un effet propre à l'âge et d'un facteur temporel commun à l'ensemble des âges.

Dans le cadre probabiliste retenu dans notre travail, son prédicteur s'écrit :

$$\log \mu_{x,t} = \alpha_x + \beta_x \kappa_t^{(1)}. \quad (1.14)$$

Le paramètre α_x représente le niveau moyen de mortalité à l'âge x sur la période d'observation. L'indice temporel $\kappa_t^{(1)}$ décrit l'évolution générale de la mortalité au cours du temps, tandis que le coefficient β_x mesure la sensibilité de chaque âge à cette évolution.

Le principal avantage du modèle LC réside dans sa structure relativement simple et dans l'interprétation directe de ses paramètres. Il suppose toutefois que le profil de sensibilité β_x reste constant au cours du temps. Cette hypothèse peut devenir restrictive lorsque les améliorations de mortalité évoluent différemment selon l'âge. Le modèle ne représente pas non plus explicitement les effets de cohorte, c'est-à-dire les caractéristiques propres aux générations nées durant une même période. Plusieurs extensions ont été proposées afin de répondre à ces limites, notamment le modèle de Renshaw-Haberman.

1.1.2. Modèle de Renshaw-Haberman (RH)

Le modèle de Renshaw-Haberman, proposé par [Renshaw et Haberman \(2006\)](#), prolonge le modèle de Lee-Carter en ajoutant un effet de cohorte. Il permet ainsi de représenter le fait que des individus nés durant une même période peuvent présenter des caractéristiques de mortalité communes tout au long de leur vie.

Dans la spécification retenue par [Barigou et al. \(2023\)](#), le modèle s'écrit :

$$\log \mu_{x,t} = \alpha_x + \beta_x \kappa_t^{(1)} + \gamma_{t-x}. \quad (1.15)$$

Le terme γ_{t-x} représente l'effet associé à la cohorte née durant l'année $t - x$. Le modèle conserve donc les composantes d'âge et de période du modèle LC, tout en ajoutant une dimension générationnelle.

Cette extension peut améliorer l'ajustement des données lorsque des effets de cohorte sont effectivement présents. Elle augmente toutefois le nombre de paramètres à estimer et rend l'identification du modèle plus délicate. Des contraintes doivent notamment être imposées afin de distinguer correctement les effets d'âge, de période et de cohorte. Le modèle RH peut également rencontrer des difficultés de convergence numérique, en particulier lorsque l'effet de cohorte est faiblement identifié.

1.1.3. Modèle âge-période-cohorte (APC)

Le modèle âge-période-cohorte, noté APC, distingue trois dimensions principales de la mortalité : l'âge de l'individu, l'année durant laquelle le décès est observé et l'année de naissance de la cohorte à laquelle il appartient.

Son prédicteur est défini par :

$$\log \mu_{x,t} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}. \quad (1.16)$$

Le paramètre α_x représente l'effet propre à l'âge, $\kappa_t^{(1)}$ l'effet commun à l'année calendaire t , et γ_{t-x} l'effet associé à la cohorte de naissance. Contrairement aux modèles LC et RH, l'effet de période n'est pas modulé par un coefficient de sensibilité spécifique à chaque âge. Cette structure est plus simple, mais elle suppose que les variations temporelles de la mortalité affectent tous les âges de manière identique sur l'échelle logarithmique.

1.1.4. Modèle de Cairns-Blake-Dowd (CBD)

Le modèle de Cairns-Blake-Dowd, proposé par [Cairns, Blake et Dowd \(2006\)](#), a été conçu principalement pour représenter la mortalité aux âges élevés. Contrairement au modèle de Lee-Carter, il ne comporte pas de paramètre d'âge libre pour chaque âge. Il représente plutôt le profil de mortalité à l'aide de deux facteurs temporels.

1. Modélisation stochastique de la mortalité

En considérant le modèle binomial négatif et avec le lien logarithmique retenus dans notre travail, la spécification utilisée par Barigou et al. (2023) s'écrit :

$$\log \mu_{x,t} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}, \quad (1.17)$$

où $\bar{x}$ désigne l'âge moyen de l'intervalle d'âges étudié.

Le premier facteur, $\kappa_t^{(1)}$, représente le niveau général de la mortalité durant l'année t . Le second, $\kappa_t^{(2)}$, décrit la pente du profil de mortalité selon l'âge. Lorsque $\kappa_t^{(2)}$ varie, l'évolution de la mortalité peut ainsi être différente aux âges inférieurs et supérieurs à $\bar{x}$.

Remarque 1.1. Dans sa formulation originale, le modèle CBD est généralement défini à partir du logarithme des rapports de chances des probabilités de décès. La formulation présentée ici correspond à son adaptation au prédicteur logarithmique de la force de mortalité utilisé dans le cadre binomial négatif de Barigou et al. (2023).

1.1.5. Modèle M6

Le modèle M6 appartient à la famille de modèles comparés par Cairns, Blake, Dowd et al. (2009). Il prolonge le modèle CBD en lui ajoutant un effet de cohorte. Il permet ainsi de conserver une représentation simple du niveau et de la pente de la mortalité selon l'âge, tout en tenant compte de différences entre les générations.

Son prédicteur est défini par

$$\log \mu_{x,t} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \gamma_{t-x}. \quad (1.18)$$

Le modèle M6 conserve donc les deux facteurs temporels du modèle CBD. Le premier représente le niveau général de la mortalité et le second sa pente selon l'âge. Le terme γ_{t-x} ajoute un effet propre à la cohorte de naissance.

Cette structure complète l'ensemble des modèles candidats. Le modèle RH constitue une extension avec cohorte du modèle LC, tandis que le modèle M6 joue un rôle comparable pour le modèle CBD. Cette symétrie permet d'étudier plus clairement la contribution des effets de cohorte dans deux familles de modèles différentes.

1.2. Projection des composantes temporelles

Les modèles présentés ci-dessus décrivent la mortalité observée à l'aide de différentes combinaisons d'effets d'âge, de période et de cohorte. Pour produire des projections, les facteurs temporels $\kappa_t^{(i)}$ doivent être prolongés au-delà de la période d'observation.

Dans leurs travaux, Barigou et al. (2023) supposent que les facteurs de période suivent une marche aléatoire avec dérive. Pour un modèle comportant q_k facteurs temporels, on

1. Modélisation stochastique de la mortalité

note

$$\boldsymbol{\kappa}_t = \left(\kappa_t^{(1)}, \dots, \kappa_t^{(q_k)} \right)^\top. \quad (1.19)$$

Leur dynamique est définie par

$$\boldsymbol{\kappa}_t = \mathbf{c} + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\varepsilon}_t^\kappa, \quad \boldsymbol{\varepsilon}_t^\kappa \sim \mathcal{N}_{q_k}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (1.20)$$

Le vecteur $\mathbf{c}$ représente la dérive moyenne et la matrice $\boldsymbol{\Sigma}$ décrit la variabilité des innovations. Lorsque le modèle contient un effet de cohorte, celui-ci est projeté séparément au moyen d'une dynamique autorégressive. Les distributions prédictives obtenues intègrent alors l'incertitude sur les paramètres du modèle et l'incertitude liée à l'évolution future de ces composantes.

La qualité des projections peut dépendre de la période historique utilisée pour l'estimation. Une période longue apporte davantage d'observations, mais peut également contenir des régimes de mortalité différents. Cette question sera traitée au moyen d'une validation temporelle, sans utiliser les années futures lors de l'ajustement du modèle.

1.3. Ensemble des modèles candidats

Dans le cadre de notre travail, nous retenons l'ensemble de modèles suivants :

$$\mathcal{M} = \{\text{LC}, \text{RH}, \text{APC}, \text{CBD}, \text{M6}\}. \quad (1.21)$$

Le modèle LC fournit une structure de référence simple fondée sur un effet d'âge et un facteur de période. Le modèle RH ajoute un effet de cohorte au cadre de Lee-Carter. Le modèle APC représente séparément les effets d'âge, de période et de cohorte, sans profil de sensibilité propre à l'âge pour l'effet de période. Le modèle CBD adopte une structure différente, fondée sur le niveau et la pente du profil de mortalité selon l'âge. Le modèle M6 complète cette structure par un effet de cohorte.

Ces cinq modèles ne sont pas choisis parce qu'ils seraient supposés couvrir toutes les structures possibles. Ils constituent un ensemble volontairement limité, dont les composantes reposent sur des hypothèses distinctes. La présence des deux couples LC–RH et CBD–M6 permet en particulier de comparer des modèles avec et sans effet de cohorte dans deux familles différentes.

Le chapitre suivant présente les méthodes utilisées pour combiner leurs distributions prédictives. Il introduit d'abord le moyennage bayésien de modèles, le pseudo-BMA et le *stacking* à poids fixes, avant de développer une pondération contextuelle selon l'âge et l'horizon de projection.

2. Agrégation bayésienne et pondérations contextuelles

Le chapitre précédent a présenté les cinq modèles stochastiques de mortalité que nous retenons dans ce travail. Ils reposent sur des représentations différentes des effets d'âge, de période et de cohorte. Retenir une seule structure revient à produire les projections comme si ce choix était suffisamment adapté, alors que plusieurs modèles peuvent demeurer plausibles.

Cette incertitude est appelée *incertitude structurelle*. Elle ne se confond pas avec l'incertitude sur les paramètres d'un modèle donné. Même lorsque les paramètres sont estimés dans un cadre bayésien, les projections restent conditionnelles à la structure choisie.

Les méthodes de combinaison de modèles, ou *model averaging*, permettent de réduire cette dépendance à une spécification unique. Leur principe consiste à combiner les distributions prédictives de plusieurs modèles candidats en leur attribuant des poids.

Dans le cadre de la prévision de la mortalité, Barigou et al. (2023) étudient le moyennage bayésien de modèles classique, le pseudo-BMA et le *stacking* bayésien. Les deux dernières méthodes déterminent les poids à partir de performances prédictives hors échantillon, évaluées au moyen d'une validation *Leave-Future-Out* qui respecte l'ordre chronologique des observations.

Dans leur forme habituelle, ces méthodes produisent un seul vecteur de poids. Le même poids est donc appliqué à un modèle pour tous les âges et tous les horizons. Cette hypothèse peut être restrictive lorsque les performances relatives des modèles ne sont pas homogènes sur l'ensemble de l'espace de prévision.

L'objectif de ce chapitre est de construire une extension simple et régulière du *stacking*. Les poids seront autorisés à varier selon l'âge et l'horizon de projection. Une structure hiérarchique sera utilisée afin d'éviter une estimation entièrement séparée dans chaque cellule, qui serait trop instable lorsque l'information disponible est limitée.

2.1. Fondements de l'estimation et de l'agrégation bayésiennes

Le cadre bayésien permet de représenter l'incertitude associée aux paramètres des modèles de mortalité. Plutôt que de fournir uniquement une estimation ponctuelle de chaque paramètre, il décrit les valeurs plausibles de celui-ci au moyen d'une distribution de probabilité.

Cette approche diffère notamment de l'estimation par maximisation de la vraisemblance, qui consiste à rechercher les valeurs des paramètres rendant les données observées les plus probables. Dans le cadre bayésien, la vraisemblance est combinée à une distribution *a priori*, qui traduit les informations ou les hypothèses formulées sur les paramètres avant l'observation des données.

Nous nous appuyons sur le cadre bayésien retenu par Barigou et al. (2023) pour estimer les différents modèles de mortalité. L'incertitude relative au choix du modèle sera ensuite prise en compte par les méthodes d'agrégation présentées dans les sections suivantes.

2.1.1. Théorème de Bayes appliqué à la modélisation de la mortalité

Soit

$$\mathcal{M} = \{M_1, \dots, M_K\} \quad (2.1)$$

l'ensemble des K modèles de mortalité candidats. Pour chaque $k \in \{1, \dots, K\}$, le modèle M_k possède un vecteur de paramètres inconnus, noté $\boldsymbol{\theta}^{(k)}$, dont les composantes dépendent de la structure considérée.

Par exemple, un modèle intégrant un effet de cohorte contient le vecteur $\boldsymbol{\gamma}$, tandis que ce paramètre est absent d'un modèle sans composante de cohorte. L'espace de toutes les valeurs possibles des paramètres du modèle M_k est noté Θ_k .

On désigne par $\mathbf{y}_{1:N}$ l'ensemble des données observées entre les années t_1 et t_N . Ces données comprennent les décès $d_{x,t}$ et les expositions $e_{x,t}$ observés pour les différents âges et les différentes années.

Conditionnellement au modèle M_k , le théorème de Bayes permet d'obtenir la distribution *a posteriori* des paramètres :

$$p\left(\boldsymbol{\theta}^{(k)} \mid \mathbf{y}_{1:N}, M_k\right) = \frac{p\left(\mathbf{y}_{1:N} \mid \boldsymbol{\theta}^{(k)}, M_k\right) p\left(\boldsymbol{\theta}^{(k)} \mid M_k\right)}{p\left(\mathbf{y}_{1:N} \mid M_k\right)}. \quad (2.2)$$

Cette relation combine trois éléments.

La vraisemblance

$$p\left(\mathbf{y}_{1:N} \mid \boldsymbol{\theta}^{(k)}, M_k\right)$$

mesure la compatibilité entre les données observées et une valeur donnée des paramètres du modèle M_k . Son expression est donnée par l'équation (1.13).

La distribution *a priori*

$$p(\boldsymbol{\theta}^{(k)} \mid M_k)$$

représente les hypothèses formulées sur les paramètres avant la prise en compte des données. Elle peut notamment être utilisée pour imposer des contraintes, introduire des informations préalables ou régulariser l'estimation.

La vraisemblance marginale

$$p(\mathbf{y}_{1:N} \mid M_k)$$

est obtenue en intégrant le produit de la vraisemblance et de la distribution *a priori* sur l'ensemble de l'espace des paramètres :

$$p(\mathbf{y}_{1:N} \mid M_k) = \int_{\Theta_k} p(\mathbf{y}_{1:N} \mid \boldsymbol{\theta}^{(k)}, M_k) p(\boldsymbol{\theta}^{(k)} \mid M_k) d\boldsymbol{\theta}^{(k)}. \quad (2.3)$$

Pour un modèle M_k fixé, la vraisemblance marginale ne dépend pas de $\boldsymbol{\theta}^{(k)}$. Elle joue donc le rôle de constante de normalisation et garantit que la distribution *a posteriori* est bien une densité de probabilité, c'est-à-dire qu'elle s'intègre à 1. L'équation (2.2) peut alors être écrite sous la forme proportionnelle suivante :

$$p(\boldsymbol{\theta}^{(k)} \mid \mathbf{y}_{1:N}, M_k) \propto p(\mathbf{y}_{1:N} \mid \boldsymbol{\theta}^{(k)}, M_k) p(\boldsymbol{\theta}^{(k)} \mid M_k). \quad (2.4)$$

La distribution *a posteriori* regroupe ainsi les informations apportées par les données et celles contenues dans les lois *a priori*. Elle permet ensuite d'obtenir des estimations ponctuelles des paramètres, par exemple à partir de leur moyenne ou de leur médiane *a posteriori*, ainsi que des intervalles de crédibilité.

Dans les modèles de mortalité considérés, l'intégrale apparaissant dans l'équation (2.3) est généralement difficile à calculer de manière analytique en raison de la dimension élevée de l'espace des paramètres. La distribution *a posteriori* sera donc approchée numériquement. Les principes nécessaires à la lecture des résultats sont présentés dans le corps de notre travail, tandis que le plan d'implémentation et les diagnostics numériques sont regroupés en annexe.

2.2. Spécification des lois *a priori*

Les détails de l'implémentation numérique sont présentés en annexe. La présente section présente les principes généraux des choix probabilistes retenus.

Les paramétrisations et les lois *a priori* exactes peuvent différer selon la structure et l'échelle des modèles. Les spécifications effectivement utilisées dans l'implémentation

numérique sont détaillées dans l'annexe A.

Les modèles de mortalité reprennent autant que possible les lois *a priori* utilisées par Barigou et al. (2023). Elles sont choisies de manière à respecter les contraintes d'identification tout en conservant une information préalable limitée.

Dans la suite, la notation $\mathcal{N}(m, v)$ désigne une loi normale de moyenne m et de variance v . La notation $\mathcal{HN}(0, s)$ désigne une loi demi-normale de paramètre d'échelle s .

1. Effet d'âge de base

Pour chaque âge x , le paramètre α_x reçoit une loi normale diffuse :

$$\alpha_x \sim \mathcal{N}(0, 100), \quad x \in \{x_1, \dots, x_A\}. \quad (2.5)$$

Cette loi correspond à un écart-type égal à 10 sur l'échelle logarithmique de la mortalité.

2. Sensibilité aux facteurs de période

Lorsque le modèle contient un profil de sensibilité par âge, comme dans les modèles LC et RH, le vecteur

$$\boldsymbol{\beta} = (\beta_{x_1}, \dots, \beta_{x_A})^\top \quad (2.6)$$

est soumis aux contraintes

$$\beta_x \geq 0 \quad \text{et} \quad \sum_{x=x_1}^{x_A} \beta_x = 1. \quad (2.7)$$

Une loi de Dirichlet uniforme est alors retenue :

$$\boldsymbol{\beta} \sim \text{Dirichlet}(\mathbf{1}_A), \quad (2.8)$$

où $\mathbf{1}_A$ désigne un vecteur de dimension A dont toutes les composantes sont égales à un.

3. Facteurs temporels de mortalité

Les facteurs temporels sont modélisés par une marche aléatoire avec dérive. Pour un modèle comportant q_k facteurs de période, on note

$$\boldsymbol{\kappa}_t = \left(\kappa_t^{(1)}, \dots, \kappa_t^{(q_k)} \right)^\top. \quad (2.9)$$

Le processus temporel est défini par

$$\boldsymbol{\kappa}_t = \mathbf{c} + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\varepsilon}_t^\kappa, \quad \boldsymbol{\varepsilon}_t^\kappa \sim \mathcal{N}_{q_k}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (2.10)$$

2. Agrégation bayésienne et pondérations contextuelles

Pour assurer l'identifiabilité, les contraintes imposées aux facteurs temporels dépendent de la structure du modèle. Pour les modèles LC, RH et APC, le facteur temporel de la première année est fixé à zéro :

$$\kappa_{t_1} = \mathbf{0}. \quad (2.11)$$

Pour les modèles CBD et M6, qui ne comportent pas d'effet d'âge libre jouant le rôle d'intercept, les états initiaux sont estimés selon la paramétrisation décrite dans l'annexe A.

Les composantes de la dérive reçoivent les lois

$$c_j \sim \mathcal{N}(0, 10), \quad j = 1, \dots, q_k. \quad (2.12)$$

Dans le cas d'un modèle à un seul facteur, la matrice de covariance se réduit à

$$\Sigma = \sigma_\kappa^2, \quad \sigma_\kappa \sim \text{Exp}(0, 1). \quad (2.13)$$

Pour les modèles CBD et M6, qui comportent deux facteurs temporels, la matrice de covariance est définie par

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_\Sigma \sigma_1 \sigma_2 \\ \rho_\Sigma \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}, \quad (2.14)$$

avec

$$\sigma_1, \sigma_2 \sim \text{Exp}(0, 1) \quad \text{et} \quad \rho_\Sigma \sim \mathcal{U}(-1, 1). \quad (2.15)$$

4. Effet de cohorte

Pour les modèles comprenant un effet de cohorte, celui-ci est représenté par un processus autorégressif d'ordre deux :

$$\gamma_c = \psi_1 \gamma_{c-1} + \psi_2 \gamma_{c-2} + \varepsilon_c^\gamma, \quad \varepsilon_c^\gamma \sim \mathcal{N}(0, \sigma_\gamma^2). \quad (2.16)$$

Afin de garantir la stationnarité du processus, les coefficients autorégressifs sont paramétrés à partir de deux autocorrélations partielles, notées ρ_1 et ρ_2 . Celles-ci reçoivent des lois normales tronquées :

$$\rho_1, \rho_2 \sim \mathcal{N}(0, 0.5^2) \quad \text{tronquées à} \quad [-0.99, 0.99]. \quad (2.17)$$

Les coefficients du processus autorégressif sont ensuite obtenus par la transformation

$$\psi_1 = \rho_1(1 - \rho_2), \quad \psi_2 = \rho_2. \quad (2.18)$$

2. Agrégation bayésienne et pondérations contextuelles

Le paramètre d'échelle des innovations reçoit une loi exponentielle de taux égal à 10 :

$$\sigma_\gamma \sim \text{Exp}_{\text{taux}}(10). \quad (2.19)$$

Cette loi possède une moyenne égale à 0,1. Les détails de cette paramétrisation sont présentés dans l'annexe A.

Les contraintes suivantes sont imposées aux cohortes extrêmes :

$$\gamma_{c_1} = 0 \quad \text{et} \quad \gamma_{c_C} = 0. \quad (2.20)$$

Pour le modèle RH, une contrainte supplémentaire de somme nulle est appliquée :

$$\sum_{c=c_1}^{c_C} \gamma_c = 0. \quad (2.21)$$

Dans l'implémentation numérique, les effets de cohorte sont représentés au moyen d'une base orthonormale permettant de les distinguer du niveau constant et de la tendance linéaire. Ce choix permet de mieux séparer l'effet de cohorte des effets d'âge et de période.

5. Paramètre de surdispersion

Dans la paramétrisation retenue, ϕ est le paramètre de concentration de la loi binomiale négative. Plus ϕ est élevé, plus la variance conditionnelle se rapproche de celle d'une loi de Poisson.

En suivant [Barigou et al. \(2023\)](#), la loi *a priori* est placée sur l'inverse de ce paramètre :

$$\frac{1}{\phi} \sim \mathcal{HN}(0, 1). \quad (2.22)$$

Cette spécification porte directement sur l'ampleur de la variabilité supplémentaire, puisque la variance du nombre de décès contient le terme $\mathbb{E}[D_{x,t}]^2/\phi$.

2.3. Approche par moyennage bayésien de modèles (BMA)

La sélection d'un modèle unique conduit à produire les projections comme si la structure retenue était suffisamment adaptée au processus étudié. Cette démarche ignore l'incertitude liée au choix du modèle et peut donner une impression excessive de précision lorsque plusieurs spécifications demeurent plausibles.

Le moyennage bayésien de modèles, ou *Bayesian Model Averaging* (BMA), vise à intégrer cette source d'incertitude en combinant les distributions prédictives de plusieurs modèles candidats. Chaque modèle contribue à la distribution prédictive finale selon sa probabilité *a posteriori*.

Soit

$$\mathcal{M} = \{M_1, \dots, M_K\} \quad (2.23)$$

l'ensemble des K modèles de mortalité candidats, et soit $\mathcal{D}$ l'ensemble des données observées. Pour une quantité future d'intérêt Δ , telle qu'un taux de mortalité, une probabilité de survie ou une espérance de vie, la distribution prédictive obtenue par BMA s'écrit

$$p(\Delta \mid \mathcal{D}) = \sum_{k=1}^K p(\Delta \mid M_k, \mathcal{D}) p(M_k \mid \mathcal{D}). \quad (2.24)$$

Cette expression correspond à un mélange des distributions prédictives produites par les différents modèles.

Pour chaque modèle M_k , le terme

$$p(\Delta \mid M_k, \mathcal{D}) \quad (2.25)$$

désigne la distribution prédictive conditionnelle au modèle. Elle tient compte de l'incertitude sur les paramètres au moyen de leur distribution *a posteriori* :

$$p(\Delta \mid M_k, \mathcal{D}) = \int_{\Theta_k} p(\Delta \mid \boldsymbol{\theta}^{(k)}, M_k, \mathcal{D}) p(\boldsymbol{\theta}^{(k)} \mid M_k, \mathcal{D}) d\boldsymbol{\theta}^{(k)}. \quad (2.26)$$

Le second terme de l'équation (2.24),

$$p(M_k \mid \mathcal{D}), \quad (2.27)$$

est la probabilité *a posteriori* du modèle M_k . Elle constitue le poids attribué au modèle dans la combinaison finale.

En appliquant le théorème de Bayes sur l'espace des modèles, cette probabilité est donnée par :

$$p(M_k \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid M_k) p(M_k)}{\sum_{\ell=1}^K p(\mathcal{D} \mid M_\ell) p(M_\ell)}. \quad (2.28)$$

Le poids attribué au modèle M_k dépend ainsi de deux éléments. La probabilité $p(M_k)$ représente la probabilité *a priori* du modèle, tandis que la quantité

$$p(\mathcal{D} \mid M_k) \quad (2.29)$$

est sa vraisemblance marginale.

En l'absence d'information justifiant de privilégier une structure particulière, une probabilité *a priori* uniforme peut être attribuée aux modèles :

$$p(M_k) = \frac{1}{K}, \quad k = 1, \dots, K. \quad (2.30)$$

Ce choix signifie que les modèles sont considérés comme également plausibles avant l'observation des données. Il ne rend toutefois pas l'analyse indépendante des choix *a priori*, puisque les lois attribuées aux paramètres interviennent également dans le calcul des vraisemblances marginales.

Pour le modèle M_k , la vraisemblance marginale est définie par

$$p(\mathcal{D} \mid M_k) = \int_{\Theta_k} p(\mathcal{D} \mid \boldsymbol{\theta}^{(k)}, M_k) p(\boldsymbol{\theta}^{(k)} \mid M_k) d\boldsymbol{\theta}^{(k)}. \quad (2.31)$$

Cette intégrale moyenne la vraisemblance sur l'ensemble des valeurs possibles des paramètres, pondérées par leur distribution *a priori*. Elle ne mesure donc pas uniquement la qualité d'ajustement obtenue au meilleur jeu de paramètres. Elle tient également compte de la manière dont le modèle répartit sa probabilité sur l'ensemble de son espace paramétrique.

Le BMA prend ainsi en compte deux niveaux d'incertitude. L'incertitude paramétrique est représentée dans chaque distribution prédictive conditionnelle par l'intégration sur $p(\boldsymbol{\theta}^{(k)} \mid M_k, \mathcal{D})$. L'incertitude relative au choix du modèle est représentée par les probabilités $p(M_k \mid \mathcal{D})$.

La mise en œuvre du BMA reste toutefois difficile dans les modèles de mortalité considérés. Les vraisemblances marginales de l'équation (2.31) sont des intégrales de grande dimension qui ne possèdent généralement pas de solution analytique. Leur dimension augmente avec le nombre d'effets d'âge, de période et de cohorte, ainsi qu'avec le nombre de paramètres propres à chaque structure.

Dans leur application, Barigou et al. (2023) estiment les vraisemblances marginales au moyen du *bridge sampling*. Cette méthode utilise les tirages obtenus dans la distribution *a posteriori* afin d'approcher la constante de normalisation de chaque modèle. Une fois ces quantités estimées, les probabilités *a posteriori* sont calculées à partir de l'équation (2.28).

Au-delà de la difficulté numérique, le BMA présente deux limites importantes pour la prévision. Premièrement, les vraisemblances marginales peuvent être sensibles aux distributions *a priori* attribuées aux paramètres. Deuxièmement, elles évaluent les modèles à partir des données utilisées pour leur estimation et ne mesurent pas directement leur capacité à prévoir de nouvelles observations.

Le BMA est par ailleurs généralement interprété dans un cadre dit *M-closed*, dans lequel le véritable processus générateur des données est supposé appartenir à l'ensemble des modèles candidats. Cette hypothèse est difficile à défendre en pratique lorsque tous les modèles disponibles ne constituent que des approximations du phénomène réel.

Pour la prévision de la mortalité, il peut donc être préférable d'attribuer les poids en fonction de la performance prédictive hors échantillon plutôt qu'à partir des probabilités *a posteriori* des modèles. Le pseudo-BMA, présenté dans la section suivante, adopte cette

logique en construisant ses poids à partir de la densité prédictive évaluée sur une période de validation future.

2.4. Approche par pseudo-BMA

Le BMA classique attribue à chaque modèle un poids égal à sa probabilité *a posteriori*. Le calcul de ces poids nécessite l'estimation des vraisemblances marginales, qui peut être difficile dans les modèles de mortalité de grande dimension. Cette approche présente également une limite prédictive : la vraisemblance marginale évalue principalement la capacité d'un modèle à représenter les données utilisées pour son estimation, et non directement sa capacité à prévoir de nouvelles observations.

Le pseudo-BMA adopte une logique différente. Ses poids ne sont pas des approximations des probabilités *a posteriori* des modèles. Ils sont construits à partir de leurs performances prédictives hors échantillon. L'objectif est ainsi d'attribuer davantage de poids aux modèles qui prédisent le mieux des données qui n'ont pas été utilisées pour leur estimation.

La formulation du pseudo-BMA utilisée ici repose sur les travaux de Yao, Vehtari et al. (2018). Pour la prévision de la mortalité, Barigou et al. (2023) remplacent la validation croisée classique par une validation temporelle de type *Leave-Future-Out*.

Découpage temporel des données

On suppose que les données disponibles sont séparées en deux blocs chronologiquement ordonnés. Le premier bloc,

$$\mathcal{D}_{1:N} = \{d_{x,t}, e_{x,t}\}_{\substack{x=x_1, \dots, x_A \\ t=t_1, \dots, t_N}}, \quad (2.32)$$

constitue la période d'apprentissage. Le second bloc,

$$\mathcal{V}_{N+1:N+M} = \{d_{x,j}, e_{x,j}\}_{\substack{x=x_1, \dots, x_A \\ j=t_{N+1}, \dots, t_{N+M}}}, \quad (2.33)$$

constitue la période de validation future. La séparation chronologique des deux blocs est essentielle. Elle reproduit la situation réelle dans laquelle seules les observations passées sont disponibles au moment de produire une prévision.

Densité prédictive hors échantillon

Pour chaque modèle M_k , la distribution *a posteriori* des paramètres est d'abord estimée à partir du seul échantillon d'apprentissage :

$$p\left(\theta^{(k)} \mid \mathcal{D}_{1:N}, M_k\right). \quad (2.34)$$

2. Agrégation bayésienne et pondérations contextuelles

Pour une cellule d'âge x et une année de validation j , la densité prédictive du nombre de décès s'écrit

$$p(d_{x,j} \mid \mathcal{D}_{1:N}, M_k) = \int_{\Theta_k} p(d_{x,j} \mid \boldsymbol{\theta}^{(k)}, \mathcal{D}_{1:N}, M_k) p(\boldsymbol{\theta}^{(k)} \mid \mathcal{D}_{1:N}, M_k) d\boldsymbol{\theta}^{(k)}. \quad (2.35)$$

Cette distribution tient compte de l'incertitude sur les paramètres ainsi que de l'incertitude associée à la projection des facteurs temporels et, le cas échéant, des effets de cohorte.

L'intégrale de l'équation (2.35) est approchée à partir des tirages *a posteriori*. Si

$$\boldsymbol{\theta}^{(k,1)}, \dots, \boldsymbol{\theta}^{(k,S)} \sim p(\boldsymbol{\theta}^{(k)} \mid \mathcal{D}_{1:N}, M_k), \quad (2.36)$$

alors la densité prédictive peut être estimée par

$$\hat{p}(d_{x,j} \mid \mathcal{D}_{1:N}, M_k) = \frac{1}{S} \sum_{s=1}^S p(d_{x,j} \mid \boldsymbol{\theta}^{(k,s)}, \mathcal{D}_{1:N}, M_k). \quad (2.37)$$

Dans les modèles considérés dans ce travail, chaque terme de cette somme est évalué à partir de la loi binomiale négative et des trajectoires prédictives simulées sous le modèle M_k .

Estimation de la performance prédictive

La performance du modèle M_k est mesurée au moyen de la densité prédictive logarithmique attendue, ou *Expected Log Predictive Density* (ELPD). Dans la validation temporelle retenue, son estimateur correspond à la somme des densités prédictives logarithmiques sur toutes les cellules âge-année du bloc de validation :

$$\widehat{\text{ELPD}}_k^{\text{LFO}} = \sum_{j=t_{N+1}}^{t_{N+M}} \sum_{x=x_1}^{x_A} \log \hat{p}(d_{x,j} \mid \mathcal{D}_{1:N}, M_k). \quad (2.38)$$

Une valeur élevée de $\widehat{\text{ELPD}}_k^{\text{LFO}}$ indique que le modèle attribue, en moyenne, une densité prédictive plus importante aux décès effectivement observés durant la période de validation.

Une validation qui retire isolément une observation n'est pas adaptée à l'objectif de prévision temporelle, car des observations ultérieures peuvent alors contribuer à la prédiction du passé. La validation *Leave-Future-Out* évite cette fuite d'information en évaluant les modèles uniquement sur des années postérieures à la période d'apprentissage (Barigou et al. 2023).

Calcul des poids du pseudo-BMA

Les poids sont obtenus en appliquant une transformation *softmax* aux valeurs estimées de l'ELPD :

$$w_k^{\text{pBMA}} = \frac{\exp\left(\widehat{\text{ELPD}}_k^{\text{LFO}}\right)}{\sum_{\ell=1}^K \exp\left(\widehat{\text{ELPD}}_\ell^{\text{LFO}}\right)}, \quad k = 1, \dots, K. \quad (2.39)$$

Ces poids vérifient

$$w_k^{\text{pBMA}} \geq 0 \quad \text{et} \quad \sum_{k=1}^K w_k^{\text{pBMA}} = 1. \quad (2.40)$$

Pour améliorer la stabilité numérique du calcul, on peut soustraire la plus grande valeur d'ELPD avant d'appliquer l'exponentielle. En posant

$$a = \max_{\ell=1, \dots, K} \widehat{\text{ELPD}}_\ell^{\text{LFO}}, \quad (2.41)$$

on obtient la formulation équivalente

$$w_k^{\text{pBMA}} = \frac{\exp\left(\widehat{\text{ELPD}}_k^{\text{LFO}} - a\right)}{\sum_{\ell=1}^K \exp\left(\widehat{\text{ELPD}}_\ell^{\text{LFO}} - a\right)}. \quad (2.42)$$

La distribution prédictive agrégée est ensuite donnée par

$$p_{\text{pBMA}}(\Delta \mid \mathcal{D}) = \sum_{k=1}^K w_k^{\text{pBMA}} p(\Delta \mid \mathcal{D}, M_k), \quad (2.43)$$

où Δ désigne la quantité future d'intérêt.

Contrairement aux poids du BMA classique, les poids w_k^{pBMA} ne représentent pas la probabilité qu'un modèle soit le véritable processus générateur des données. Ils constituent uniquement des pondérations prédictives fondées sur les performances observées durant la période de validation.

Intérêts et limites

Le principal intérêt du pseudo-BMA est qu'il évalue directement la capacité des modèles à prévoir des observations futures. Il évite également le calcul des vraisemblances marginales et limite la dépendance à un critère fondé uniquement sur l'ajustement aux données historiques.

Cette méthode ne garantit cependant pas l'absence de surajustement. Ses résultats dépendent notamment de la longueur et du positionnement de la période de validation, de l'horizon évalué et de la représentativité des années retenues.

Par ailleurs, l'exponentiation des scores prédictifs peut conduire les poids à se concentrer fortement sur le modèle dont l'ELPD est le plus élevé. Ce phénomène devient particulièrement marqué lorsque le nombre de cellules de validation augmente, car les différences de densité logarithmique s'accumulent.

Le pseudo-BMA évalue enfin chaque modèle séparément avant de transformer son score en poids. Il ne cherche pas directement la combinaison de distributions prédictives qui produit le meilleur score. Il peut donc attribuer presque tout le poids à un seul modèle, même lorsqu'un mélange de plusieurs modèles fournirait une meilleure distribution prédictive.

Cette limite motive l'utilisation du *stacking* bayésien. Contrairement au pseudo-BMA, le *stacking* détermine simultanément les poids en optimisant directement la performance prédictive de la distribution agrégée.

2.5. Approche par *stacking* bayésien

Le pseudo-BMA attribue un score prédictif à chaque modèle séparément, puis transforme ces scores en poids. Cette méthode peut conduire à une forte concentration des pondérations sur le modèle présentant la meilleure valeur estimée de l'ELPD. Elle ne recherche toutefois pas directement la combinaison de distributions prédictives qui produit la meilleure performance hors échantillon.

Le *stacking* adopte une logique différente. Dans sa formulation bayésienne, il combine les distributions prédictives des modèles en choisissant les poids qui maximisent directement une mesure de performance hors échantillon (Yao, Vehtari et al. 2018). Son objectif n'est pas d'estimer la probabilité qu'un modèle soit le véritable processus générateur des données.

Cette approche est particulièrement adaptée à un univers dit $\mathcal{M}$ -open, dans lequel aucun des modèles candidats n'est supposé représenter exactement le processus réel. Dans le cadre de la mortalité, les modèles LC, RH, APC, CBD et M6 reposent sur des représentations différentes des effets d'âge, de période et de cohorte. Il est donc raisonnable de les considérer comme des approximations complémentaires plutôt que comme des descriptions exactes du phénomène étudié.

Soit

$$\mathcal{M} = \{M_1, \dots, M_K\} \quad (2.44)$$

l'ensemble des K modèles candidats. Comme dans la section consacrée au pseudo-BMA, les modèles sont estimés à partir du bloc d'apprentissage $\mathcal{D}_{1:N}$, puis évalués sur le bloc de validation future $\mathcal{V}_{N+1:N+M}$.

2. Agrégation bayésienne et pondérations contextuelles

Pour chaque modèle M_k , chaque âge x et chaque année de validation j , on note

$$\hat{p}_{k,x,j} = \hat{p}(d_{x,j} \mid \mathcal{D}_{1:N}, M_k) \quad (2.45)$$

l'estimation de la densité prédictive hors échantillon du nombre de décès observé. Cette quantité est obtenue à partir des tirages *a posteriori*, conformément à l'équation (2.37).

Le *stacking* cherche un vecteur de poids

$$\mathbf{w} = (w_1, \dots, w_K)^\top \quad (2.46)$$

qui maximise la densité prédictive logarithmique de la distribution mélangée sur l'ensemble du bloc de validation. Les poids optimaux sont définis par

$$\hat{\mathbf{w}}^{\text{stack}} = \arg \max_{\mathbf{w} \in \mathcal{W}} \sum_{j=t_{N+1}}^{t_{N+M}} \sum_{x=x_1}^{x_A} \log \left[\sum_{k=1}^K w_k \hat{p}_{k,x,j} \right]. \quad (2.47)$$

L'espace admissible des pondérations est le simplexe

$$\mathcal{W} = \left\{ \mathbf{w} \in \mathbb{R}^K : w_k \geq 0 \text{ pour tout } k, \quad \sum_{k=1}^K w_k = 1 \right\}. \quad (2.48)$$

Ces contraintes garantissent que la distribution agrégée constitue une combinaison convexe des distributions prédictives individuelles. Les coefficients w_k sont toutefois des poids prédictifs et ne doivent pas être interprétés comme les probabilités que les modèles soient vrais.

Pour une quantité future d'intérêt Δ , la distribution prédictive issue du *stacking* est donnée par

$$p_{\text{stack}}(\Delta \mid \mathcal{D}) = \sum_{k=1}^K \hat{w}_k^{\text{stack}} p(\Delta \mid \mathcal{D}, M_k). \quad (2.49)$$

Contrairement au pseudo-BMA, le *stacking* n'évalue pas chaque modèle indépendamment avant de convertir son score en poids. Les pondérations sont estimées simultanément afin d'optimiser directement la qualité de la distribution mélangée. Plusieurs modèles peuvent ainsi recevoir un poids positif lorsque leurs distributions prédictives apportent des informations complémentaires.

Cette propriété ne garantit cependant pas que plusieurs modèles seront toujours conservés. Lorsque l'un des modèles domine les autres sur la période de validation, la solution peut se situer sur un sommet du simplexe et lui attribuer la totalité, ou la quasi-totalité, du poids. De même, lorsque plusieurs modèles produisent des distributions prédictives très proches, les poids individuels peuvent être faiblement identifiés, même si la distribution agrégée demeure stable. L'utilisation d'une période de validation future réduit le risque d'attribuer des poids élevés à un modèle uniquement parce qu'il s'ajuste bien aux données d'apprentissage.

Elle ne supprime toutefois pas tout risque de surajustement des pondérations. Une période de validation courte, atypique ou insuffisamment représentative peut produire des poids instables. L'évaluation finale des méthodes d'agrégation doit donc être réalisée sur un bloc de test qui n'a servi ni à estimer les modèles individuels ni à calibrer les poids.

2.5.1. Limites des pondérations fixes et justification d'une extension contextuelle

Le *stacking* défini par l'équation (2.47) produit un unique vecteur de poids

$$\hat{\mathbf{w}}^{\text{stack}} = (\hat{w}_1^{\text{stack}}, \dots, \hat{w}_K^{\text{stack}})^\top. \quad (2.50)$$

Ce même vecteur est utilisé pour toutes les classes d'âge et tous les horizons de projection. Il résume donc la performance moyenne des modèles sur l'ensemble du bloc de validation.

Cette agrégation globale peut être restrictive lorsque la performance relative des modèles varie selon le contexte. Un modèle peut fournir des prévisions adaptées à certains âges et moins adaptées à d'autres. De même, une structure efficace à court terme ne conserve pas nécessairement la même performance lorsque l'horizon s'allonge.

La pondération globale peut masquer ces différences, puisque les scores sont additionnés sur toutes les cellules avant l'estimation des poids. Une amélioration limitée à une partie de l'espace âge-horizon peut ainsi être compensée par les autres observations.

Des pondérations dépendant de l'horizon ont déjà été étudiées pour les ensembles de modèles de mortalité, notamment par [Kessy et al. \(2022\)](#). Leur approche porte sur des prévisions ponctuelles. Dans le cadre de notre travail, l'objectif est de conserver l'agrégation de distributions prédictives complètes et le cadre bayésien de [Barigou et al. \(2023\)](#).

Nous remplaçons donc le poids global w_k par une fonction

$$w_k \longrightarrow w_k(x, h), \quad (2.51)$$

où x désigne l'âge et h l'horizon de projection. Cette extension ne suppose pas qu'un modèle particulier doit dominer dans une région donnée. Elle fournit seulement un cadre dans lequel cette variation peut être estimée à partir des performances prédictives historiques.

2.6. Stacking contextuel et régularisation hiérarchique

Le *stacking* hiérarchique généralise la pondération globale en autorisant les poids à dépendre des caractéristiques de la prévision. L'approche proposée par [Yao, Pirš et al. \(2022\)](#) combine deux idées. Premièrement, les poids varient selon les entrées. Deuxièmement, une structure hiérarchique limite les variations qui ne sont pas suffisamment soutenues par les données.

Dans ce travail, les poids dépendent uniquement de l'âge et de l'horizon. Ce choix conserve un modèle interprétable et évite d'introduire un nombre trop important de covariables dans le méta-modèle. L'âge et l'horizon sont directement disponibles au moment de produire une prévision et possèdent une interprétation actuarielle claire.

Trois formes de *stacking* seront comparées à partir des mêmes observations de validation. Dans le *stacking* global, un seul vecteur de poids est utilisé pour tous les âges et tous les horizons. Dans les deux versions contextuelles, les poids dépendent de l'âge et de l'horizon au moyen de la même fonction *softmax*. La différence entre les deux versions contextuelles porte sur la méthode d'estimation des coefficients de cette fonction.

Dans le *stacking* contextuel non régularisé, les coefficients sont obtenus par maximisation directe du critère prédictif. Aucune pénalisation ni aucune distribution *a priori* ne limite leur amplitude.

Dans le *stacking* contextuel hiérarchique, les mêmes coefficients sont estimés dans un cadre bayésien et reçoivent des distributions *a priori* hiérarchiques. Cette structure régularise les effets contextuels et limite les variations excessives des poids lorsque l'information disponible est insuffisante.

La comparaison entre le *stacking* global et le *stacking* contextuel non régularisé permet donc d'évaluer l'intérêt de faire varier les poids selon l'âge et l'horizon. La comparaison entre les deux versions contextuelles permet ensuite d'isoler l'apport de la régularisation hiérarchique.

2.6.1. Construction des observations de validation temporelle

Les poids sont estimés à partir de plusieurs exercices historiques de prévision. Une origine de prévision désigne la dernière année disponible au moment où les modèles sont ajustés et où les prévisions sont produites. On note $\mathcal{O}$ l'ensemble de ces années et $\mathcal{H}$ l'ensemble des horizons considérés.

Pour une origine $r \in \mathcal{O}$, seuls les décès et les expositions observés jusqu'à l'année r sont utilisés pour estimer les modèles de mortalité. L'échantillon d'apprentissage associé à cette origine est donc

$$\mathcal{D}_{1:r} = \{d_{x,t}, e_{x,t} : x \in \{x_1, \dots, x_A\}, t \leq r\}. \quad (2.52)$$

Ainsi, pour un horizon $h \in \mathcal{H}$, la prévision produite à partir de cette origine porte sur l'année cible $r + h$. Par exemple, si $r = 2005$, une prévision à l'horizon $h = 1$ concerne l'année 2006, tandis qu'une prévision à l'horizon $h = 5$ concerne l'année 2010.

Étant donné un âge x , un horizon $h \in \mathcal{H}$ et un modèle M_k , on note

$$\hat{p}_{k,x,r,h} = \hat{p}(d_{x,r+h} \mid \mathcal{D}_{1:r}, M_k) \quad (2.53)$$

la densité prédictive hors échantillon du nombre de décès observé à l'âge x durant l'année

Tableau 2.1. – Exemple d’exercices de prévision à plusieurs origines

Origine r	Horizon h	Année cible $r + h$
2005	1	2006
2005	5	2010
2010	1	2011
2010	5	2015

$r + h$.

Pour chaque origine r , les données comprises dans $\mathcal{D}_{1:r}$ constituent la partie d’apprentissage des modèles de mortalité. Les observations des années $r + h$, avec $h \in \mathcal{H}$, ne sont pas utilisées lors de cet ajustement. Elles servent à évaluer les distributions prédictives produites par les modèles et à estimer les poids de *stacking*. Les différents couples (r, h) forment ainsi la période de calibration des poids. L’utilisation de plusieurs origines permet de répéter le même exercice de prévision à différentes dates historiques. Par exemple, l’horizon $h = 5$ peut être évalué en prévoyant l’année 2005 à partir des données disponibles jusqu’en 2000, puis l’année 2010 à partir des données disponibles jusqu’en 2005. Une période de test plus récente est conservée entièrement à l’écart. Elle n’intervient ni dans l’estimation des modèles individuels, ni dans l’estimation des poids. Elle est utilisée uniquement pour comparer les performances finales des différentes méthodes d’agrégation. L’ensemble des observations utilisées pour estimer les poids est

$$\mathcal{I}_{\text{LFO}} = \{(x, r, h) : x \in \{x_1, \dots, x_A\}, \quad r \in \mathcal{O}, \quad h \in \mathcal{H}, \quad r + h \leq T_{\text{val}}\}, \quad (2.54)$$

où T_{val} est la dernière année autorisée pour la calibration des poids.

Le méta-jeu de données est alors défini par

$$\mathcal{D}_{\text{meta}} = \{d_{x,r+h}, x, h, \widehat{p}_{1,x,r,h}, \dots, \widehat{p}_{K,x,r,h}\}_{(x,r,h) \in \mathcal{I}_{\text{LFO}}}. \quad (2.55)$$

Avec plusieurs origines et plusieurs horizons, une même année cible peut être prévue dans plusieurs exercices. Par exemple, l’année 2010 peut être prévue depuis l’origine 2009 à l’horizon 1, depuis l’origine 2008 à l’horizon 2 ou depuis l’origine 2005 à l’horizon 5.

On note m_t le nombre de couples origine-horizon (r, h) qui conduisent à la même année cible t . Il est défini par

$$m_t = \sum_{r \in \mathcal{O}} \sum_{h \in \mathcal{H}} \mathbf{1}\{r + h = t\} \quad (2.56)$$

Ainsi, si l’année 2010 est atteinte par trois couples origine-horizon, alors $m_{2010} = 3$. Le coefficient

$$\omega_{r,h} = \frac{1}{m_{r+h}} \quad (2.57)$$

attribue alors un poids égal à $1/3$ à chacun de ces trois exercices. La contribution totale

de l'année 2010 au critère reste ainsi égale à un, au lieu d'être multipliée par le nombre de manières de l'atteindre. Cette pondération constitue une correction destinée à éviter la surreprésentation mécanique de certaines années cibles. Elle ne rend pas indépendants les exercices de prévision qui partagent des données d'apprentissage ou une même observation future.

Pour chaque origine, aucune observation postérieure à r n'est utilisée pour ajuster les modèles. Les années réservées au test final restent également exclues de la construction de $\mathcal{D}_{\text{meta}}$.

2.6.2. Standardisation de l'âge et de l'horizon

La standardisation n'est pas nécessaire pour définir la transformation *softmax*, mais elle facilite l'estimation et l'interprétation du méta-modèle. L'âge et l'horizon sont en effet mesurés sur des échelles différentes. Dans l'application envisagée, l'âge peut par exemple varier entre 50 et 90 ans, tandis que l'horizon peut varier entre 1 et 10 ans. Une variation d'une unité n'a donc pas la même signification pour ces deux variables.

Le centrage et la réduction permettent d'exprimer chaque variable en nombre d'écarts-types par rapport à sa moyenne. Les coefficients associés à l'âge et à l'horizon deviennent ainsi plus facilement comparables. Cette transformation améliore également la stabilité numérique de l'estimation et permet d'utiliser des distributions *a priori* de même ordre de grandeur pour les différents effets contextuels.

Ainsi, l'âge standardisé est défini par

$$\tilde{x} = \frac{x - \bar{x}}{s_x}, \quad (2.58)$$

où $\bar{x}$ et s_x sont la moyenne et l'écart-type des âges retenus.

Afin de permettre une variation non linéaire simple selon l'âge, on introduit également

$$q_x = \tilde{x}^2 - \frac{1}{A} \sum_{a=1}^A \tilde{x}_a^2. \quad (2.59)$$

Le terme quadratique est lui-même centré afin que sa moyenne sur l'ensemble des âges soit nulle. Cette opération réduit son recouvrement avec l'intercept et permet à α_k de conserver l'interprétation d'une préférence moyenne pour le modèle M_k .

L'horizon standardisé est défini par

$$\tilde{h} = \frac{h - \bar{h}}{s_h}, \quad (2.60)$$

où les constantes $\bar{h}$ et s_h sont calculées sur les horizons utilisés dans la validation.

2.6.3. Transformation *softmax*

Pour chaque contexte (x, h) , les poids doivent être positifs et leur somme doit être égale à 1. Ils sont donc définis à partir de scores non contraints $f_k(x, h)$ au moyen de la transformation

$$w_k(x, h) = \frac{\exp[f_k(x, h)]}{\sum_{\ell=1}^K \exp[f_\ell(x, h)]}. \quad (2.61)$$

Les fonctions f_k ne sont identifiables qu'à une constante additive commune. Le modèle M_K est donc choisi comme catégorie de référence et

$$f_K(x, h) = 0. \quad (2.62)$$

En nous inspirant de la formulation additive du *stacking* hiérarchique proposée par Yao, Pirš et al. (2022), nous exprimons les scores latents comme des fonctions de caractéristiques contextuelles. Dans le cadre de notre travail, ces caractéristiques sont construites à partir de l'âge et de l'horizon de prévision.

Pour $k = 1, \dots, K - 1$, nous retenons la forme

$$f_k(x, h) = \alpha_k + \beta_{k,x}\tilde{x} + \beta_{k,x^2}q_x + \beta_{k,h}\tilde{h} + \beta_{k,xh}\tilde{x}\tilde{h}. \quad (2.63)$$

Cette équation constitue une adaptation de la formulation générale de Yao, Pirš et al. (2022) au contexte de la prévision de la mortalité. Le choix précis d'un effet linéaire et quadratique de l'âge, d'un effet de l'horizon et d'une interaction âge-horizon n'est pas repris directement de leur application. Il est retenu ici afin de représenter de manière parcimonieuse les variations des performances relatives des modèles selon les deux dimensions étudiées.

Le paramètre α_k représente la préférence moyenne pour le modèle M_k relativement au modèle de référence. Les coefficients $\beta_{k,x}$ et β_{k,x^2} décrivent la variation du poids selon l'âge. Le coefficient $\beta_{k,h}$ décrit la variation selon l'horizon, tandis que $\beta_{k,xh}$ permet à l'effet de l'horizon de différer selon l'âge.

Cette spécification reste volontairement limitée. Elle permet de représenter les principales variations recherchées sans estimer un poids indépendant pour chaque cellule âge-horizon.

2.6.4. Critère prédictif commun aux trois formes de *stacking*

Pour une fonction de poids $w_k(x, h \mid \Psi)$, on définit le critère

$$\ell_{\text{LFO}}(\Psi) = \sum_{(x,r,h) \in \mathcal{I}_{\text{LFO}}} \omega_{r,h} \log \left[\sum_{k=1}^K w_k(x, h \mid \Psi) \hat{p}_{k,x,r,h} \right]. \quad (2.64)$$

Les observations issues des différentes origines sont partiellement dépendantes. L'expres-

2. Agrégation bayésienne et pondérations contextuelles

sion (2.64) est donc utilisée comme un critère prédictif composite. La comparaison finale des méthodes repose sur un bloc de test entièrement distinct, et non sur la valeur de ce critère durant la calibration.

Dans le *stacking* global proposé par Barigou et al. (2023), les poids ne dépendent pas du contexte. Ils sont estimés par

$$\hat{\mathbf{w}}^{\text{global}} = \arg \max_{\mathbf{w} \in \mathcal{W}} \sum_{(x,r,h) \in \mathcal{I}_{\text{LFO}}} \omega_{r,h} \log \left[\sum_{k=1}^K w_k \hat{p}_{k,x,r,h} \right]. \quad (2.65)$$

Dans le *stacking* contextuel non régularisé, les poids suivent les équations (2.61) et (2.63), mais les coefficients sont obtenus uniquement par optimisation :

$$\hat{\Psi}^{\text{nonreg}} = \arg \max_{\Psi} \ell_{\text{LFO}}(\Psi). \quad (2.66)$$

Le terme *non régularisé* ne signifie pas qu'un vecteur de poids entièrement indépendant est estimé pour chaque cellule âge-horizon. Les poids restent liés par le prédicteur commun défini à l'équation (2.63). L'absence de régularisation signifie uniquement que les coefficients de ce prédicteur sont estimés sans pénalisation et sans distribution *a priori*.

Cette méthode sert de point de comparaison. Elle permet d'évaluer l'intérêt de faire varier les poids selon l'âge et l'horizon, mais elle ne contient aucun mécanisme qui limite des coefficients excessifs lorsque les densités prédictives sont peu informatives.

2.6.5. Régularisation hiérarchique

Le modèle hiérarchique conserve exactement la transformation *softmax* et le prédicteur contextuel définis aux équations (2.61) et (2.63). La différence avec la version non régularisée réside uniquement dans l'introduction de distributions *a priori* sur les coefficients et leurs paramètres d'échelle.

On regroupe les effets contextuels dans l'ensemble

$$\mathcal{J} = \{x, x^2, h, xh\}. \quad (2.67)$$

Pour chaque type d'effet $j \in \mathcal{J}$, les coefficients propres aux modèles partagent un paramètre d'échelle :

$$\beta_{k,j} \mid \tau_j \sim \mathcal{N}(0, \tau_j^2), \quad k = 1, \dots, K-1. \quad (2.68)$$

Les intercepts reçoivent la loi

$$\alpha_k \sim \mathcal{N}(0, 1,5^2). \quad (2.69)$$

Les effets principaux sont associés aux lois

$$\tau_x, \tau_h \sim \mathcal{HN}(0, 1), \quad (2.70)$$

tandis que les termes plus flexibles sont davantage régularisés :

$$\tau_{x^2}, \tau_{xh} \sim \mathcal{HN}(0, 0.5). \quad (2.71)$$

Cette structure réalise une mise en commun partielle. Lorsque les données contiennent peu d'information sur une variation donnée, le paramètre d'échelle correspondant est ramené vers zéro et les poids se rapprochent d'une pondération fixe. Lorsque l'information est plus forte, les coefficients peuvent s'écarter davantage de zéro.

Le *stacking* à poids fixes apparaît ainsi comme un cas particulier du modèle contextuel. Si les effets d'âge et d'horizon sont nuls, seuls les intercepts demeurent et les poids ne varient plus selon le contexte.

2.6.6. Distribution a posteriori généralisée des paramètres de pondération

On note Ψ l'ensemble des paramètres et hyperparamètres du méta-modèle. La distribution associée au *stacking* hiérarchique est définie par

$$\tilde{p}(\Psi \mid \mathcal{D}_{\text{meta}}) \propto \exp[\ell_{\text{LFO}}(\Psi)] p(\Psi). \quad (2.72)$$

Cette cible est une distribution *a posteriori* généralisée au sens de l'actualisation bayésienne par une fonction de perte (Bissiri et al. 2016). Le terme $\exp\{\ell_{\text{LFO}}(\Psi)\}$ est construit à partir d'un critère prédictif composite. Il ne constitue pas la vraisemblance d'un modèle génératif du méta-jeu de données. Les intervalles obtenus décrivent donc l'incertitude sous cette cible généralisée et non sous une vraisemblance ordinaire.

La correction de multiplicité $\omega_{r,h}$ évite qu'une année cible représentée par plusieurs couples origine-horizon domine mécaniquement le critère. Elle ne rend pas indépendants les exercices qui partagent une partie de leur échantillon d'apprentissage ou une même observation cible. Aucun paramètre de température supplémentaire n'est introduit. Cette précision modifie l'interprétation de la cible, mais ni le programme **Stan**, ni les poids estimés, ni les résultats numériques.

La construction est réalisée en deux étages. Les modèles de mortalité sont d'abord estimés et évalués hors échantillon. Les densités prédictives ainsi obtenues sont ensuite utilisées pour estimer les fonctions de pondération. Le méta-modèle ne modifie donc pas les paramètres internes des modèles de mortalité.

Les tirages de Ψ permettent d'étudier l'incertitude sur les poids et de construire des

intervalles de crédibilité pour $w_k(x, h)$. Les détails de l'estimation numérique sont présentés en annexe.

2.6.7. Distribution prédictive finale

Après l'estimation des poids, chaque modèle candidat est réestimé sur l'ensemble des données disponibles jusqu'à la dernière année T . Pour le *stacking* hiérarchique, on définit le poids moyen sous la distribution *a posteriori* généralisée

$$\bar{w}_k(x, h) = \mathbb{E}[w_k(x, h) \mid \Psi] \mid \mathcal{D}_{\text{meta}}]. \quad (2.73)$$

Afin d'alléger la rédaction, nous utiliserons ensuite l'expression poids moyen *a posteriori*.

Pour une projection à l'âge x et à l'horizon h , la distribution prédictive finale est

$$p_{\text{HStack}}(d_{x,T+h} \mid \mathcal{D}_{1:T}, \mathcal{D}_{\text{meta}}) = \sum_{k=1}^K \bar{w}_k(x, h) p(d_{x,T+h} \mid \mathcal{D}_{1:T}, M_k). \quad (2.74)$$

Cette écriture est équivalente à l'intégration de la combinaison sur la distribution de Ψ , car le mélange est linéaire par rapport aux poids. Les tirages du méta-modèle restent utiles pour décrire l'incertitude sur les fonctions de pondération, mais la distribution prédictive marginale dépend des poids moyens *a posteriori*.

Pour le *stacking* contextuel non régularisé, la même expression est utilisée en remplaçant $\bar{w}_k(x, h)$ par $w_k(x, h \mid \hat{\Psi}^{\text{nonreg}})$. Pour le *stacking* global, elle est utilisée avec les poids constants $\hat{w}_k^{\text{global}}$.

2.7. Métriques de performance et règles de score probabilistes

Les méthodes présentées dans ce chapitre produisent des distributions prédictives complètes. Leur évaluation ne doit donc pas reposer uniquement sur une prévision ponctuelle, telle que la moyenne ou la médiane *a posteriori*. Une mesure fondée uniquement sur une estimation ponctuelle ne tient pas compte de la dispersion de la distribution prédictive ni de l'incertitude associée aux projections.

Nous utilisons des règles de score probabilistes afin d'évaluer simultanément la position et la dispersion des distributions prédictives. Une règle de score est dite propre lorsque sa valeur moyenne est optimale si la distribution annoncée correspond à la véritable distribution du phénomène. Elle encourage ainsi la production de prévisions probabilistes correctement calibrées.

Les méthodes sont comparées sur une période de test qui n'est utilisée ni pour estimer les modèles de mortalité, ni pour calculer les poids du pseudo-BMA, du *stacking* global ou

du *stacking* contextuel. On note

$$\mathcal{I}_{\text{test}} = \{(x, t) : x \in \{x_1, \dots, x_A\}, \quad t \in \mathcal{T}_{\text{test}}\} \quad (2.75)$$

l'ensemble des cellules âge–année appartenant à cette période. Son nombre d'éléments est noté

$$N_{\text{test}} = |\mathcal{I}_{\text{test}}|. \quad (2.76)$$

Pour chaque cellule, la quantité observée est le taux brut de mortalité

$$y_{x,t} = \frac{d_{x,t}}{e_{x,t}}. \quad (2.77)$$

La variable aléatoire prédite correspondante est

$$Y_{x,t} = \frac{D_{x,t}}{e_{x,t}}. \quad (2.78)$$

Pour une méthode de prévision a , on note

$$F_{x,t}^{(a)}(u) = \Pr_a(Y_{x,t} \leq u) \quad (2.79)$$

sa fonction de répartition prédictive. On note également

$$p_{x,t}^{(a)}(d) = \Pr_a(D_{x,t} = d) \quad (2.80)$$

la probabilité prédictive associée au nombre de décès.

Ces distributions comprennent l'incertitude sur les paramètres, l'incertitude liée à la projection des facteurs temporels et la variabilité du nombre de décès. Pour le *stacking* hiérarchique, l'incertitude sur les poids est présentée séparément au moyen d'intervalles de crédibilité. La distribution prédictive marginale est construite à partir de leurs moyennes *a posteriori*.

Nous retenons deux règles de score complémentaires : le score logarithmique et le *Continuous Ranked Probability Score*. Ces deux scores sont orientés négativement : une valeur plus faible indique une meilleure performance prédictive (Barigou et al. 2023).

2.7.1. Le score logarithmique (LogS)

Le score logarithmique mesure la probabilité attribuée par la distribution prédictive à l'observation effectivement réalisée. Pour une cellule (x, t) et une méthode a , il est défini par

$$\text{LogS}_{x,t}^{(a)} = -\log \left[p_{x,t}^{(a)}(d_{x,t}) \right]. \quad (2.81)$$

Une prévision reçoit donc un score faible lorsqu'elle attribue une probabilité élevée

au nombre de décès observé. À l'inverse, une distribution qui attribue une probabilité très faible à cette observation reçoit une pénalité importante. Le score logarithmique est ainsi particulièrement sensible aux observations situées dans les queues de la distribution prédictive.

Le score moyen sur l'ensemble de la période de test est donné par

$$\text{LogS}^{(a)} = \frac{1}{N_{\text{test}}} \sum_{(x,t) \in \mathcal{I}_{\text{test}}} \text{LogS}_{x,t}^{(a)}. \quad (2.82)$$

La probabilité prédictive est approchée à partir des tirages *a posteriori*. Si $\boldsymbol{\vartheta}^{(a,1)}, \dots, \boldsymbol{\vartheta}^{(a,S)}$ désignent les S tirages utilisés pour construire la prévision, alors

$$\hat{p}_{x,t}^{(a)}(d_{x,t}) = \frac{1}{S} \sum_{s=1}^S p(d_{x,t} \mid \boldsymbol{\vartheta}^{(a,s)}). \quad (2.83)$$

Chaque terme de cette somme est calculé à partir de la loi binomiale négative du modèle considéré. Pour une méthode d'agrégation, la probabilité prédictive correspond au mélange des probabilités produites par les différents modèles. Le calcul est effectué sur l'échelle logarithmique au moyen d'un calcul stable sur l'échelle logarithmique. Les détails numériques sont reportés en annexe.

Le score logarithmique est une règle de score locale. Il dépend uniquement de la probabilité attribuée à la valeur observée. Il ne décrit donc pas directement la forme de la distribution prédictive en dehors de cette observation. Cette limite justifie l'utilisation d'un second critère fondé sur l'ensemble de la fonction de répartition.

2.7.2. Le *Continuous Ranked Probability Score* (CRPS)

Le *Continuous Ranked Probability Score*, ou CRPS, compare l'ensemble de la fonction de répartition prédictive à l'observation réalisée. Pour une cellule (x, t) , il est défini par

$$\text{CRPS}(F_{x,t}^{(a)}, y_{x,t}) = \int_{-\infty}^{+\infty} \left[F_{x,t}^{(a)}(u) - \mathbf{1}\{y_{x,t} \leq u\} \right]^2 du, \quad (2.84)$$

où $\mathbf{1}\{\cdot\}$ désigne la fonction indicatrice.

Cette définition reste valable lorsque la distribution prédictive est discrète, comme dans le cas du taux $D_{x,t}/e_{x,t}$. Contrairement au score logarithmique, le CRPS utilise toute la distribution prédictive. Il évalue à la fois sa proximité avec l'observation et son degré de concentration.

Le score moyen sur la période de test est

$$\text{CRPS}^{(a)} = \frac{1}{N_{\text{test}}} \sum_{(x,t) \in \mathcal{I}_{\text{test}}} \text{CRPS}(F_{x,t}^{(a)}, y_{x,t}). \quad (2.85)$$

2. Agrégation bayésienne et pondérations contextuelles

Le **CRPS** possède la même unité que la variable évaluée. Nous le calculons ici sur les taux de mortalité. Cette formulation facilite la comparaison des prévisions entre les cellules dont les expositions au risque sont différentes.

Le score peut être estimé directement à partir des simulations prédictives. Soient

$$Y_{x,t}^{(a,1)}, \dots, Y_{x,t}^{(a,S)} \sim F_{x,t}^{(a)} \quad (2.86)$$

les tirages prédictifs du taux de mortalité. Une estimation du **CRPS** est donnée par

$$\widehat{\text{CRPS}} \left(F_{x,t}^{(a)}, y_{x,t} \right) = \frac{1}{S} \sum_{s=1}^S \left| Y_{x,t}^{(a,s)} - y_{x,t} \right| - \frac{1}{2S(S-1)} \sum_{\substack{s,s'=1 \\ s \neq s'}}^S \left| Y_{x,t}^{(a,s)} - Y_{x,t}^{(a,s')} \right|. \quad (2.87)$$

Le premier terme mesure la distance entre les simulations prédictives et l'observation. Le second tient compte de la dispersion interne de la distribution prédictive. Lorsque la prévision se réduit à une valeur ponctuelle, le **CRPS** devient égal à l'erreur absolue.

Le score logarithmique et le **CRPS** apportent donc des informations complémentaires. Le premier pénalise fortement les distributions qui attribuent une probabilité très faible à l'observation. Le second fournit une mesure plus globale de la qualité de la distribution prédictive. Les méthodes seront principalement classées selon leurs scores moyens sur la période de test. Des scores par âge et par horizon seront également examinés afin d'identifier les contextes dans lesquels leurs performances diffèrent.

2.7.3. Erreur absolue moyenne

Les règles de score précédentes évaluent les distributions prédictives complètes. Une mesure ponctuelle est également retenue afin de comparer les moyennes prédictives. Pour une méthode a , on note

$$\widehat{\mu}_{x,t}^{(a)} = \mathbb{E}_a [\mu_{x,t} \mid \mathcal{D}] \quad (2.88)$$

la moyenne prédictive de la force de mortalité. L'erreur absolue moyenne sur les nombres de décès est définie par

$$\text{MAE}^{(a)} = \frac{1}{N_{\text{test}}} \sum_{(x,t) \in \mathcal{I}_{\text{test}}} \left| d_{x,t} - e_{x,t} \widehat{\mu}_{x,t}^{(a)} \right|. \quad (2.89)$$

Cette mesure ne remplace pas le score logarithmique ou le **CRPS**, puisqu'elle ne tient pas compte de la dispersion prédictive. Elle fournit un résumé complémentaire de l'erreur sur la prévision centrale.

2.7.4. Couverture des intervalles prédictifs

La calibration des intervalles sera évaluée pour des niveaux nominaux de 80% et 95%. Pour un niveau $1 - \alpha$, on note

$$\left[L_{x,t}^{(a,\alpha)}, U_{x,t}^{(a,\alpha)} \right] \quad (2.90)$$

l'intervalle prédictif associé à la méthode a .

La couverture empirique est

$$\text{Cov}^{(a,\alpha)} = \frac{1}{N_{\text{test}}} \sum_{(x,t) \in \mathcal{I}_{\text{test}}} \mathbf{1} \left\{ y_{x,t} \in \left[L_{x,t}^{(a,\alpha)}, U_{x,t}^{(a,\alpha)} \right] \right\}. \quad (2.91)$$

La largeur moyenne correspondante est

$$\text{Largeur}^{(a,\alpha)} = \frac{1}{N_{\text{test}}} \sum_{(x,t) \in \mathcal{I}_{\text{test}}} \left(U_{x,t}^{(a,\alpha)} - L_{x,t}^{(a,\alpha)} \right). \quad (2.92)$$

Une couverture proche du niveau nominal doit être examinée conjointement avec la largeur des intervalles. Une méthode peut en effet obtenir une couverture élevée en produisant des intervalles très larges.

2.8. Construction des quantités actuarielles

Les scores précédents sont calculés cellule par cellule à partir de mélanges de distributions prédictives. Les probabilités de survie et les valeurs de rente nécessitent toutefois une trajectoire complète de mortalité. Une règle supplémentaire est donc définie pour construire une surface agrégée cohérente.

Pour chaque indice de simulation s , un tirage de trajectoire est prélevé séparément dans chacun des modèles candidats. La force de mortalité agrégée est définie par

$$\mu_{x,T+h}^{(s,\text{agr})} = \sum_{k=1}^K \tilde{w}_k(x, h) \mu_{x,T+h}^{(s,k)}, \quad (2.93)$$

où $\tilde{w}_k(x, h)$ désigne le poids correspondant à la méthode étudiée : poids global, poids contextuel non régularisé ou poids moyen *a posteriori* du modèle hiérarchique.

Cette règle est utilisée uniquement pour les quantités qui dépendent simultanément de plusieurs âges et de plusieurs horizons. Elle est distincte du mélange des densités prédictives utilisé pour le score logarithmique et le CRPS.

Sous l'hypothèse d'une force de mortalité constante au cours de chaque année de projection, la probabilité conditionnelle de survivre durant la j -ième année est

$$p_{x+j-1,T+j}^{(s,\text{agr})} = \exp \left(-\mu_{x+j-1,T+j}^{(s,\text{agr})} \right). \quad (2.94)$$

2. Agrégation bayésienne et pondérations contextuelles

La probabilité qu'un individu âgé de x à la date T survive pendant u années est alors

$${}_u p_{x,T}^{(s,\text{agr})} = \prod_{j=1}^u p_{x+j-1,T+j}^{(s,\text{agr})} = \exp \left(- \sum_{j=1}^u \mu_{x+j-1,T+j}^{(s,\text{agr})} \right). \quad (2.95)$$

Pour un facteur annuel d'actualisation v , la valeur actuelle d'une rente temporaire immédiate, annuelle et payable à terme échu, de durée n est définie par

$$a_{x,T:\overline{n}|}^{(s,\text{agr})} = \sum_{u=1}^n v^u {}_u p_{x,T}^{(s,\text{agr})}. \quad (2.96)$$

Les distributions de ces quantités sont obtenues en répétant le calcul pour l'ensemble des simulations s . Les choix relatifs au taux d'actualisation et à la durée des rentes seront précisés dans l'application empirique.

3. Évaluation des méthodes et implications actuarielles

Ce chapitre évalue les trois formes de *stacking* présentées au chapitre précédent et examine leurs conséquences sur des quantités actuarielles. L’objectif est double. D’une part, il s’agit de vérifier dans quelle mesure les pondérations contextuelles permettent de récupérer les différences de performance entre les modèles selon l’âge et l’horizon. D’autre part, nous souhaitons mesurer l’impact de ces choix de pondération sur des probabilités de survie et des valeurs de rentes temporaires.

L’analyse repose sur deux volets complémentaires. Une étude de simulation est d’abord menée dans un cadre où la structure véritable des poids est connue. Elle permet de contrôler précisément les scénarios étudiés et de vérifier si les méthodes parviennent à retrouver les variations présentes sans en créer artificiellement lorsque l’information est limitée. Les scénarios et les critères d’évaluation sont décrits dans la [section 3.1](#).

Les méthodes sont ensuite appliquées aux données belges de la *Human Mortality Database* ([Human Mortality Database 2026](#)), pour les âges de 50 à 90 ans et les années 1970 à 2024. Les exercices historiques de prévision sont construits selon une validation *Leave-Future-Out*, et les années 2016 à 2024 sont conservées entièrement à l’écart pour l’évaluation finale ([section 3.2](#)). Cette application permet d’examiner le comportement des méthodes dans un contexte actuariel concret et de relier les résultats aux enjeux de risque de longévité.

Enfin, les distributions prédictives obtenues sont utilisées pour calculer des probabilités de survie et des valeurs de rentes temporaires, selon la construction présentée dans la [section 2.8](#). Ces quantités permettent de traduire les différences de performance entre les méthodes en termes d’implications actuarielles, en particulier aux âges élevés et pour les horizons de projection plus longs.

3.1. Enseignements de l’étude de simulation

L’étude de simulation examine le comportement des trois formes de *stacking* dans un cadre où la structure véritable des poids est connue. Elle répond à deux questions principales. La première est de déterminer si une pondération dépendant de l’âge et de l’horizon permet de retrouver une variation réellement présente. La seconde est d’évaluer si la régularisation hiérarchique empêche la méthode contextuelle de créer des variations artificielles lorsque

l'information disponible est limitée.

3.1.1. Scénarios simulés

Chaque répétition utilise les 41 âges entiers de 50 à 90 ans et les dix horizons de 1 à 10, soit 410 cellules. Une base de validation est générée pour estimer les poids, puis une base de test indépendante est produite selon le même mécanisme afin d'évaluer les prévisions hors échantillon. Trente répétitions sont réalisées pour chacun des quatre scénarios, ce qui conduit à 120 expériences indépendantes.

Contrairement à une construction entièrement synthétique, le mécanisme générateur est ici calibré à partir de données de mortalité observées. Les données de la *Human Mortality Database* ([Human Mortality Database 2026](#)) pour la Belgique, population **Total**, sont utilisées sur la période 1970–2015 et pour les âges de 50 à 90 ans. Les modèles LC, RH, APC, CBD et M6 sont ajustés une seule fois sur cette période. Pour chacun d'eux, 1 000 surfaces postérieures de force de mortalité et les 1 000 tirages correspondants du paramètre de dispersion sont conservés. Les observations réelles de 2016 à 2024, utilisées dans l'application empirique, n'interviennent donc pas dans le calibrage de la simulation.

La grille simulée prend pour origine l'année 2015. Les horizons 1 à 10 correspondent ainsi aux années simulées 2016 à 2025. Les expositions reprennent le profil observé en Belgique en 2015 et sont répétées à chaque horizon. Elles varient de 24 287,81 à 167 039,10, pour une médiane égale à 99 197,02. Ce choix conserve un profil d'exposition réaliste selon l'âge sans utiliser d'information provenant des années de test.

Pour chaque base simulée, un indice de tirage postérieur est sélectionné une fois pour chacun des cinq modèles. Ce même tirage fixe, sur l'ensemble de la grille, la surface de force de mortalité et le paramètre de dispersion propre au modèle. Pour chaque cellule d'âge x et d'horizon h , un indice $Z_{x,h}$ identifiant la composante génératrice est ensuite tiré selon les poids vrais :

$$Z_{x,h} \sim \text{Cat} (w_1^*(x, h), \dots, w_5^*(x, h)) . \quad (3.1)$$

Conditionnellement à la composante sélectionnée, le nombre de décès suit une loi binomiale négative selon la paramétrisation NB2 :

$$D_{x,h} \mid Z_{x,h} = k \sim \text{NB2} \left(E_{x,h} \mu_{x,h}^{(k)}, \phi_k \right) , \quad (3.2)$$

où $E_{x,h}$ représente l'exposition, $\mu_{x,h}^{(k)}$ la force de mortalité issue du tirage postérieur du modèle k et ϕ_k le paramètre de dispersion associé au même tirage. Les bases de validation et de test sont générées indépendamment avec des graines distinctes.

Les quatre scénarios conservent les mêmes structures de poids vrais que dans le plan

3. Évaluation des méthodes et implications actuarielles

initial. Afin d'alléger les notations, on définit

$$\tilde{x} = \frac{x - 70}{20}, \quad \tilde{h} = \frac{h - 5.5}{4.5}, \quad q(x) = \tilde{x}^2 - 0.35.$$

Dans les scénarios contextuels, les poids sont obtenus par transformation *softmax*,

$$w_k^*(x, h) = \frac{\exp\{\eta_k(x, h)\}}{\sum_{j=1}^5 \exp\{\eta_j(x, h)\}},$$

avec **M6** comme catégorie de référence.

Lorsque les poids dépendent uniquement de l'horizon, les prédicteurs sont

$$\begin{aligned} \eta_{\text{LC}} &= -0.8 - 0.7\tilde{h}, & \eta_{\text{RH}} &= 0.4 + 1.1\tilde{h}, \\ \eta_{\text{APC}} &= 0.3 - 0.4\tilde{h}, & \eta_{\text{CBD}} &= -0.2 + 0.25\tilde{h}, \\ \eta_{\text{M6}} &= 0. \end{aligned}$$

Dans le scénario dépendant de l'âge et de l'horizon,

$$\begin{aligned} \eta_{\text{LC}} &= -0.7 - 0.9\tilde{x} - 0.35\tilde{h}, \\ \eta_{\text{RH}} &= 0.3 + 0.8\tilde{h} + 0.55\tilde{x}\tilde{h}, \\ \eta_{\text{APC}} &= 0.5 - 0.55q(x) - 0.25\tilde{h}, \\ \eta_{\text{CBD}} &= -0.2 + 0.65\tilde{x} - 0.45\tilde{x}\tilde{h}, \\ \eta_{\text{M6}} &= 0. \end{aligned}$$

Le scénario de faible information conserve des poids vrais constants et plus diffus. Les expositions y sont multipliées par 0,10. Les distributions candidates sont également rapprochées en contractant les écarts entre leurs log-forces autour de la moyenne inter-modèles. En notant

$$\bar{\ell}_{x,h} = \frac{1}{5} \sum_{k=1}^5 \log \mu_{x,h}^{(k)},$$

la force utilisée dans ce scénario vérifie

$$\log \mu_{x,h}^{(k,\text{faible})} = \bar{\ell}_{x,h} + 0.25 \left(\log \mu_{x,h}^{(k)} - \bar{\ell}_{x,h} \right).$$

Les tirages de dispersion ne sont pas modifiés.

Les caractéristiques des quatre scénarios et la question associée à chacun d'eux sont résumées dans le [tableau 3.1](#).

Tableau 3.1. – Définition des quatre scénarios de l'étude de simulation HMD-calibrée.

Scénario	Structure des poids vrais	Construction des données	Question évaluée
Poids constants	(0,06; 0,54; 0,24; 0,10; 0,06)	Profil d'exposition HMD 2015 et distributions postérieures non contractées	La méthode hiérarchique conserve-t-elle une pondération stable lorsque la structure vraie est constante ?
Poids selon l'horizon	Transformation <i>softmax</i> de prédicteurs dépendant uniquement de $\tilde{h}$	Profil d'exposition HMD 2015 et distributions postérieures non contractées	La méthode retrouve-t-elle une variation selon l'horizon sans créer d'effet d'âge artificiel ?
Poids selon l'âge et l'horizon	Transformation <i>softmax</i> de prédicteurs comprenant $\tilde{x}$, $q(x)$, $\tilde{h}$ et leur interaction	Profil d'exposition HMD 2015 et distributions postérieures non contractées	La méthode retrouve-t-elle une surface contextuelle dépendant simultanément de l'âge et de l'horizon ?
Faible information	(0,18; 0,24; 0,22; 0,18; 0,18)	Expositions multipliées par 0,10 et écarts de log-force contractés par 0,25	La régularisation empêche-t-elle l'apparition de variations contextuelles artificielles lorsque le signal est faible ?

Le dernier scénario demande une attention particulière. Les poids vrais y sont constants. Il ne cherche donc pas à vérifier si une surface contextuelle complexe peut être retrouvée malgré un signal faible, mais à déterminer si une méthode suffisamment flexible pour dépendre de l'âge et de l'horizon crée artificiellement une structure lorsque les distributions candidates deviennent difficiles à distinguer.

Le *stacking* global et le *stacking* contextuel non régularisé sont estimés par optimisation. Le modèle hiérarchique est estimé intégralement par MCMC sous **Stan**, avec quatre chaînes. Les critères numériques sont fixés avant l'interprétation des résultats et une relance plus longue est prévue lorsqu'un ajustement ne les satisfait pas.

3.1.2. Récupération des surfaces de poids

La qualité de récupération est mesurée par la racine de l'erreur quadratique moyenne entre les poids estimés et les poids vrais. La fréquence des poids inférieurs à 1 % ou supérieurs à 99 % permet d'évaluer le degré d'extrémisation des solutions. Pour le modèle hiérarchique, la couverture des poids vrais par les intervalles crédibles à 95 % et la largeur

3. Évaluation des méthodes et implications actuarielles

moyenne de ces intervalles sont également examinées.

Les résultats du [tableau 3.2](#) reposent sur les répétitions finales disponibles après application des critères numériques. Les scénarios à poids constants et dépendant de l'horizon comportent respectivement 29 et 24 répétitions valides pour la comparaison avec le modèle hiérarchique. Les deux autres scénarios disposent de 30 répétitions hiérarchiques valides. Dans le scénario de faible information, une optimisation du modèle contextuel non régularisé ne converge pas ; ses indicateurs reposent donc sur 29 répétitions.

Tableau 3.2. – Récupération des poids et stabilité des estimations dans les quatre scénarios.

Scénario	Méthode	n	RMSE moyenne	RMSE médiane	Poids extrêmes
Poids constants	Global	29	0,086 4	0,079 7	12,41 %
Poids constants	Contextuel non régularisé	29	0,258 2	0,255 9	50,96 %
Poids constants	Contextuel hiérarchique	29	0,154 5	0,140 9	1,48 %
Poids selon l'horizon	Global	24	0,163 9	0,149 9	1,67 %
Poids selon l'horizon	Contextuel non régularisé	24	0,245 2	0,234 4	35,40 %
Poids selon l'horizon	Contextuel hiérarchique	24	0,161 6	0,145 4	0,87 %
Poids selon l'âge et l'horizon	Global	30	0,170 6	0,162 7	5,33 %
Poids selon l'âge et l'horizon	Contextuel non régularisé	30	0,255 6	0,259 3	39,87 %
Poids selon l'âge et l'horizon	Contextuel hiérarchique	30	0,169 6	0,160 5	1,97 %
Faible information	Global	30	0,166 2	0,169 6	24,67 %
Faible information	Contextuel non régularisé	29	0,397 1	0,398 9	98,97 %
Faible information	Contextuel hiérarchique	30	0,103 8	0,093 4	0,02 %

Pour le modèle contextuel hiérarchique, la couverture des poids vrais par les intervalles crédibles à 95 % vaut 76,18 % dans le scénario à poids constants, 78,66 % lorsque les poids dépendent de l'horizon, 78,33 % dans le scénario dépendant de l'âge et de l'horizon et 98,49 % en faible information. Les largeurs moyennes correspondantes sont 0,278 3, 0,315 4, 0,331 9 et 0,625 3. Les trois premiers scénarios présentent donc une sous-couverture par rapport au niveau nominal de 95 %. Les intervalles deviennent nettement plus larges en faible information et leur couverture se rapproche alors du niveau recherché.

3. Évaluation des méthodes et implications actuarielles

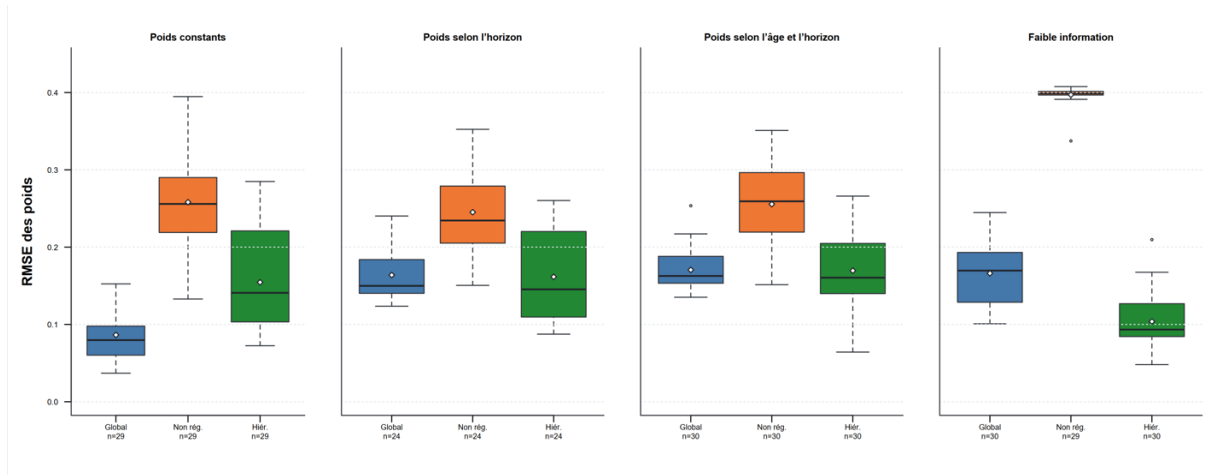


Figure 3.1. – Récupération des surfaces de poids selon le scénario simulé.

Lorsque les poids sont constants, le *stacking* global obtient la RMSE moyenne la plus faible, égale à 0,086 4. Le modèle hiérarchique atteint 0,154 5, mais limite fortement l'apparition de poids extrêmes, dont la fréquence est égale à 1,48 %, contre 12,41 % pour le global. Le modèle non régularisé se comporte moins favorablement, avec une RMSE de 0,258 2 et 50,96 % de poids extrêmes.

Lorsque les poids dépendent uniquement de l'horizon, le modèle hiérarchique obtient une RMSE moyenne de 0,161 6, légèrement inférieure à celle du global, égale à 0,163 9. La différence reste faible, mais le modèle hiérarchique produit moins de poids extrêmes. Le non-régularisé atteint une RMSE de 0,245 2 et présente à nouveau une variabilité plus importante.

Le scénario dépendant de l'âge et de l'horizon conduit à une conclusion proche. La RMSE moyenne vaut 0,170 6 pour le global et 0,169 6 pour le hiérarchique. L'écart entre les deux méthodes est donc limité sur cette mesure agrégée. Le non-régularisé atteint 0,255 6 et produit 39,87 % de poids extrêmes. La régularisation ne conduit pas ici à une amélioration importante de la RMSE par rapport au global, mais elle permet de conserver une surface contextuelle sans reproduire l'instabilité observée dans la solution non régularisée.

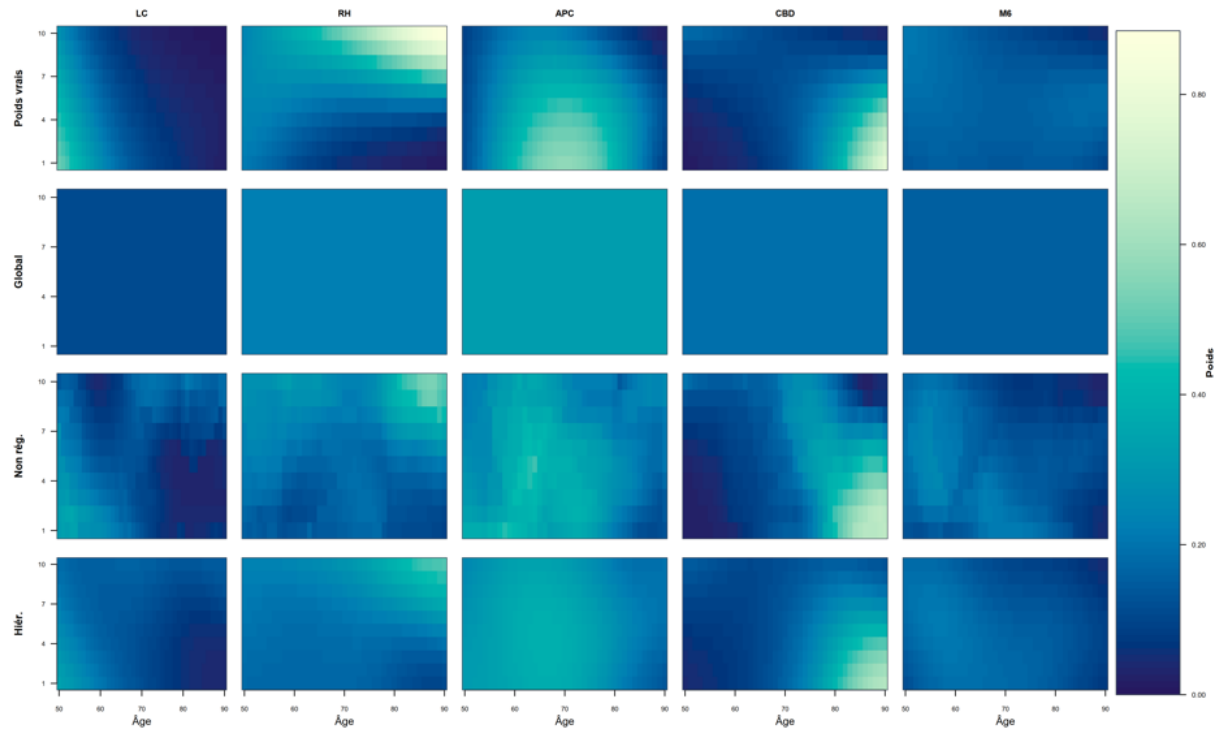


Figure 3.2. – Surfaces de poids vraies et estimées dans le scénario dépendant de l’âge et de l’horizon.

La figure 3.2 compare les surfaces vraies aux surfaces moyennes estimées sur les 30 répétitions valides du scénario dépendant de l’âge et de l’horizon. Les poids globaux restent constants sur l’ensemble de la grille. Les poids non régularisés présentent davantage de variations locales et de zones proches des bornes du simplexe. Les surfaces hiérarchiques conservent une structure plus régulière tout en restant dépendantes de l’âge et de l’horizon.

3.1.3. Faible information et variations contextuelles artificielles

Le scénario de faible information conserve les 410 cellules et les 30 répétitions. Les expositions correspondent à 10 % du profil HMD observé en 2015 et les écarts entre les log-forces des cinq modèles sont multipliés par 0,25. Les paramètres de dispersion restent issus des mêmes distributions postérieures que dans les autres scénarios. Les distributions candidates sont ainsi plus difficiles à distinguer sans modifier les critères de validité numérique.

Les poids vrais restent constants et plus diffus que dans le premier scénario. L’objectif n’est donc pas de vérifier si une surface contextuelle complexe peut être retrouvée malgré un signal faible, mais de s’assurer qu’une méthode autorisant une dépendance à l’âge et à l’horizon ne crée pas artificiellement une structure qui n’existe pas.

Le modèle contextuel non régularisé est particulièrement instable dans ce scénario. Sa RMSE moyenne atteint 0,3971 et 98,97 % de ses poids sont extrêmes. Une des 30

optimisations ne fournit pas de solution convergente et n'est donc pas utilisée pour les indicateurs de cette méthode. Le global reste plus stable, avec une RMSE de 0,166 2. Le modèle hiérarchique obtient la RMSE la plus faible, égale à 0,103 8, et la proportion de poids extrêmes est limitée à 0,02 %.

La couverture des poids vrais atteint 98,49 % pour le modèle hiérarchique et s'accompagne d'une largeur moyenne des intervalles crédibles égale à 0,625 3. Les intervalles sont donc beaucoup plus larges que dans les trois autres scénarios. Ce résultat est cohérent avec la faible information disponible : l'incertitude sur les poids augmente lorsque les distributions candidates deviennent proches, tandis que la régularisation évite que cette incertitude se traduise par des solutions presque dégénérées.

3.1.4. Performance prédictive hors échantillon

La récupération des poids ne suffit pas à évaluer les méthodes. Une surface estimée peut être proche de la vérité sans produire la meilleure prévision. Les trois méthodes sont donc également comparées sur la base de test indépendante. Le LogS présenté correspond à la négative de la log-densité prédictive moyenne. Le LogS et le CRPS doivent tous deux être minimisés.

Pour chaque répétition, les deux scores sont d'abord moyennés sur les 410 cellules de la base de test. Les moyennes du [tableau 3.3](#) utilisent toutes les répétitions finales disponibles pour chaque méthode. Les fréquences de victoire sont calculées uniquement sur l'intersection des répétitions pour lesquelles les trois méthodes sont disponibles et sont attribuées à la méthode dont le score est strictement minimal. Dans le scénario de faible information, la répétition 16 est ainsi conservée pour les moyennes du global et du hiérarchique, mais elle est exclue du calcul des victoires puisque l'optimisation non régularisée n'a pas convergé.

Tableau 3.3. – Performance prédictive hors échantillon dans les quatre scénarios de simulation.

Scénario	Méthode	n	LogS moyen	CRPS moyen	Victoires LogS	Victoires CRPS
Poids constants	Global	29	6,349 4	0,001 586	18/29	19/29
Poids constants	Contextuel non régularisé	29	6,489 5	0,001 687	2/29	4/29
Poids constants	Contextuel hiérarchique	29	6,366 4	0,001 624	9/29	6/29
Poids selon l'horizon	Global	24	6,510 3	0,001 778	10/24	13/24
Poids selon l'horizon	Contextuel non régularisé	24	6,636 2	0,001 895	0/24	2/24
Poids selon l'horizon	Contextuel hiérarchique	24	6,530 8	0,001 823	14/24	9/24
Poids selon l'âge et l'horizon	Global	30	6,496 6	0,001 866	14/30	9/30
Poids selon l'âge et l'horizon	Contextuel non régularisé	30	6,629 6	0,001 975	1/30	3/30
Poids selon l'âge et l'horizon	Contextuel hiérarchique	30	6,512 4	0,001 885	15/30	18/30
Faible information	Global	30	4,146 3	0,001 449	12/29	14/29
Faible information	Contextuel non régularisé	29	4,189 5	0,001 509	0/29	1/29
Faible information	Contextuel hiérarchique	30	4,145 6	0,001 452	17/29	14/29

3. Évaluation des méthodes et implications actuarielles

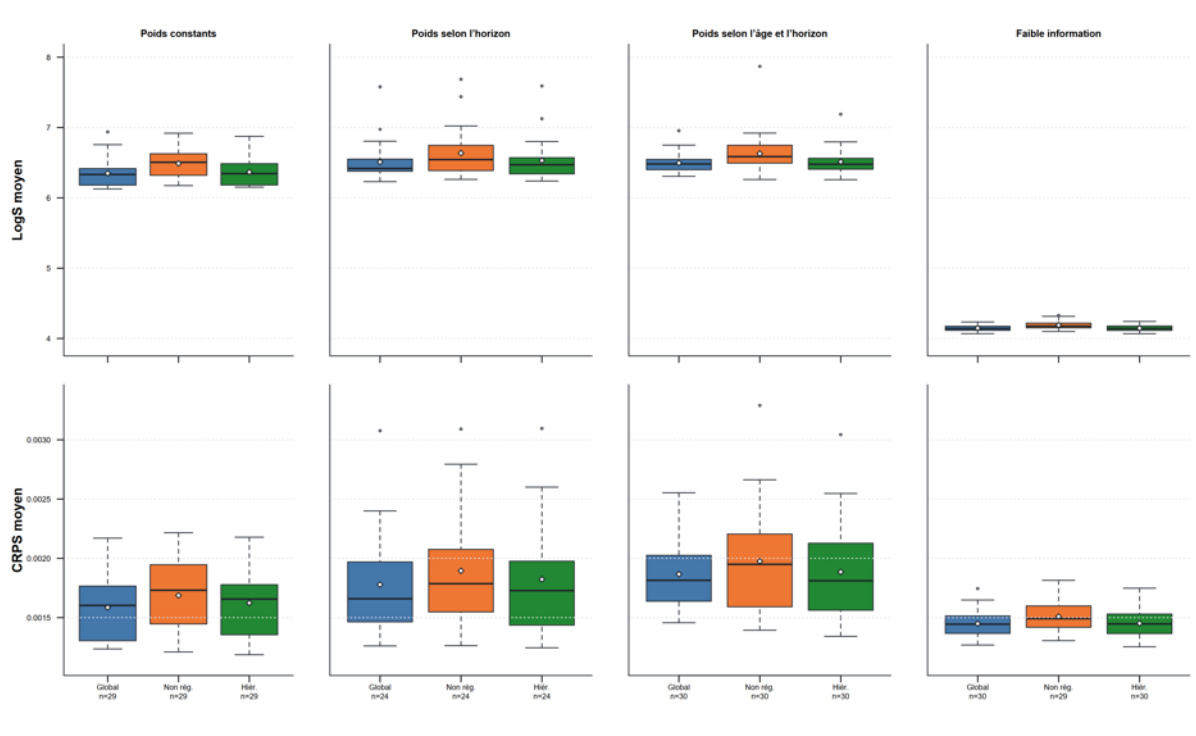


Figure 3.3. – Performance prédictive selon le scénario simulé. La première ligne présente le LogS et la seconde le CRPS.

Dans le scénario constant, le global obtient les meilleurs scores moyens, avec un LogS de 6,349 4 et un CRPS de 0,001 586. Le hiérarchique reste proche, avec des valeurs respectives de 6,366 4 et 0,001 624. Le non-régularisé est moins performant sur les deux critères.

Lorsque les poids dépendent de l'horizon, le global conserve le meilleur LogS moyen et le meilleur CRPS moyen. Le hiérarchique remporte toutefois 14 des 24 répétitions pour le LogS, contre 10 pour le global. Pour le CRPS, le global remporte 13 répétitions et le hiérarchique 9. Les scores moyens et les fréquences de victoire apportent donc des informations différentes : les premiers tiennent compte de l'amplitude des écarts, tandis que les secondes décrivent la fréquence avec laquelle une méthode arrive en première position.

Le même contraste apparaît dans le scénario dépendant de l'âge et de l'horizon. Le global obtient les scores moyens les plus faibles, avec un LogS de 6,496 6 et un CRPS de 0,001 866. Le hiérarchique est très proche, avec 6,512 4 et 0,001 885, mais il remporte 15 répétitions sur 30 pour le LogS et 18 sur 30 pour le CRPS. Le non-régularisé reste en retrait.

Dans le scénario de faible information, les différences entre le global et le hiérarchique deviennent très faibles. Le hiérarchique obtient le meilleur LogS moyen, égal à 4,145 6, contre 4,146 3 pour le global. Pour le CRPS, le global conserve un léger avantage avec 0,001 449, contre 0,001 452. Sur les 29 répétitions comparables entre les trois méthodes, le hiérarchique remporte 17 fois le LogS, tandis que les deux méthodes obtiennent chacune 14 victoires pour le CRPS.

La simulation ne met donc pas en évidence une domination prédictive systématique

du modèle hiérarchique. Le global reste très compétitif et obtient les meilleurs scores moyens dans les trois scénarios les plus informatifs. L'intérêt de la régularisation apparaît surtout dans la stabilité des surfaces de poids et dans le comportement du modèle lorsque l'information devient faible. À l'inverse, le contextuel non régularisé produit des poids beaucoup plus extrêmes et des scores moyens généralement moins favorables.

3.1.5. Validité numérique et conclusion de la simulation

Les 120 répétitions prévues ont été exécutées. Après application des critères numériques et des relances ciblées, 113 ajustements hiérarchiques sont retenus comme valides. Le nombre de répétitions valides est égal à 29 dans le scénario à poids constants, 24 dans le scénario dépendant de l'horizon et 30 dans chacun des deux autres scénarios. Quarante-trois ajustements ont nécessité une relance ciblée : 9 dans le scénario constant, 15 dans le scénario dépendant de l'horizon et 19 dans le scénario dépendant de l'âge et de l'horizon. Sept ajustements restent rejetés après la relance complète prévue par le protocole.

Parmi les 113 répétitions valides, la statistique $\hat{R}$ maximale est égale à 1,0216, l'ESS bulk minimal à 409,29, l'ESS tail minimale à 401,86 et l'E-BFMI minimal à 0,5687. Aucune divergence, aucune atteinte de la profondeur maximale et aucune chaîne bloquée ne sont observées dans les ajustements retenus. Les seuils définis avant l'analyse restent inchangés : $\hat{R} \leq 1,05$, ESS bulk et ESS tail au moins égales à 400 et E-BFMI au moins égale à 0,30.

L'unique échec d'optimisation du *stacking* contextuel non régularisé concerne la répétition 16 du scénario de faible information. Il est distinct des sept échecs Stan. Les données simulées de cette répétition sont valides et les résultats du global et du hiérarchique sont conservés. L'optimisation non régularisée atteint la limite d'itérations avec des coefficients de très grande amplitude, comportement cohérent avec une quasi-séparation lorsque les distributions candidates sont difficiles à distinguer.

Les résultats de la simulation conduisent ainsi à une conclusion plus nuancée que dans un mécanisme entièrement stylisé. Le *stacking* global demeure une référence très compétitive, aussi bien pour la récupération de poids constants que pour les scores prédictifs moyens. Le modèle hiérarchique apporte surtout une stabilisation des surfaces : il réduit fortement les poids extrêmes, obtient une récupération légèrement meilleure dans les deux scénarios contextuels et devient nettement plus précis en faible information. Le non-régularisé reste la méthode la plus instable. La régularisation doit donc être interprétée comme un mécanisme permettant de conserver une flexibilité contextuelle sans laisser cette flexibilité suivre excessivement le bruit de validation.

3.2. Application à la mortalité belge

L'étude de simulation utilise un environnement HMD-calibré dans lequel la structure des poids est connue. L'application aux données belges examine maintenant si les mêmes mécanismes apparaissent lorsque les modèles candidats sont évalués sur des observations réelles et que les poids vrais ne sont pas observables.

3.2.1. Données et stratégie d'évaluation temporelle

Les données proviennent de la *Human Mortality Database* ([Human Mortality Database 2026](#)). Les fichiers utilisés sont les séries annuelles belges de décès et d'exposition au risque par âge simple, *Deaths (period 1x1)* et *Exposure to risk (period 1x1)*. L'analyse porte sur la population **Total**, qui regroupe les femmes et les hommes. Les âges de 50 à 90 ans sont observés de 1970 à 2024, ce qui représente 2 255 cellules âge-année. Aucune valeur manquante, aucune exposition non positive et aucun décès négatif ou fractionnaire ne sont présents dans la base.

Tableau 3.4. – Données et découpage temporel de l'application à la mortalité belge.

Élément	Valeur
Population	Belgique, population Total , femmes et hommes réunis
Âges	De 50 à 90 ans, soit 41 âges
Période disponible	De 1970 à 2024, soit 55 années
Nombre de cellules	2 255 cellules âge et année
Valeurs manquantes	Aucune
Expositions non positives	Aucune
Décès négatifs ou fractionnaires	Aucun
Origines de validation LFO	De 2000 à 2014
Années cibles de validation	De 2001 à 2015
Horizons de validation	De 1 à 10 ans
Ajustement final des modèles	Données de 1970 à 2015
Période de test	De 2016 à 2024
Origine du test	Année 2015, sans réestimation annuelle
Horizons du test	De 1 à 9 ans
Modèles candidats	LC, RH, APC, CBD et M6
Tirages prédictifs	1 000 par méthode et par cellule

La validation temporelle utilise 15 origines. Pour chaque origine, les modèles sont ajustés uniquement sur les années déjà observées, puis évalués sur les années futures disponibles jusqu'en 2015. Les poids sont donc estimés à partir de prévisions réellement hors échantillon.

3. Évaluation des méthodes et implications actuarielles

Les cinq modèles sont ensuite réestimés sur la période 1970–2015. Les observations de 2016 à 2024 ne sont utilisées qu’au moment de l’évaluation finale.

Le test repose sur une origine fixe en 2015 : l’année 2016 correspond à l’horizon 1 et l’année 2024 à l’horizon 9. Cette convention permet d’étudier la dégradation progressive des prévisions à mesure que l’on s’éloigne de l’origine, mais elle implique que l’année cible et l’horizon ne peuvent pas être séparés dans l’interprétation.

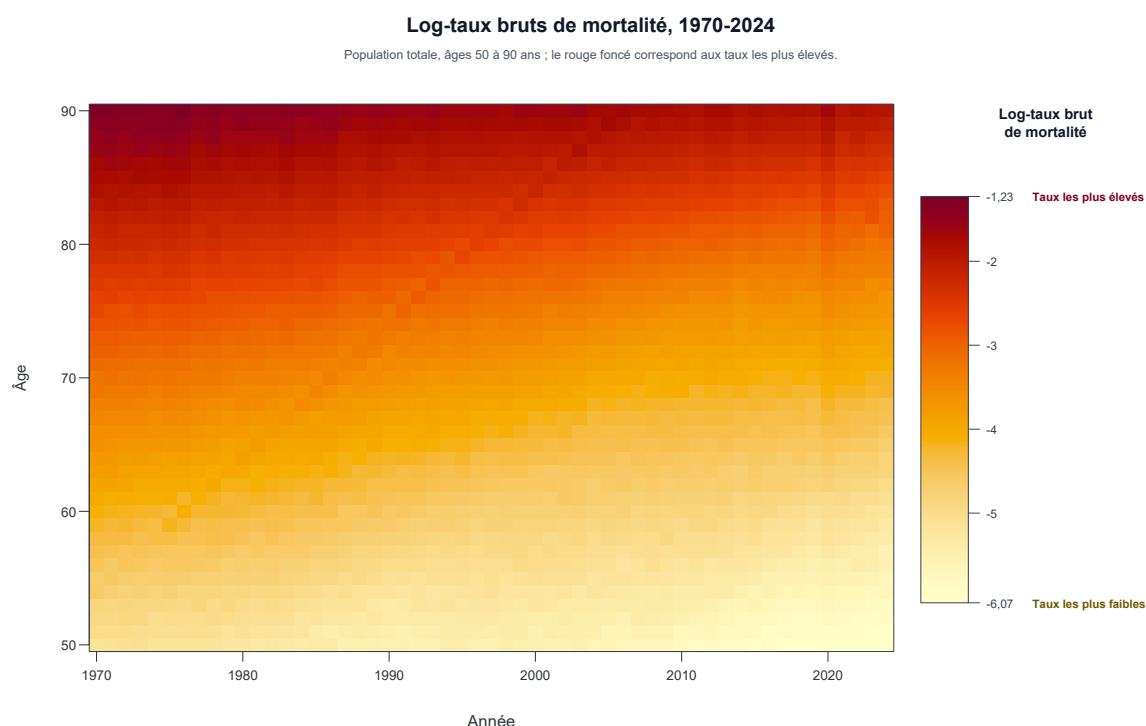


Figure 3.4. – Log-taux bruts de mortalité en Belgique selon l’âge et l’année, 1970 à 2024.

La [figure 3.4](#) met en évidence deux caractéristiques principales. La mortalité augmente fortement avec l’âge, tandis que son niveau diminue généralement au cours du temps. Cette baisse n’est toutefois pas uniforme : des variations annuelles subsistent, notamment autour des années 2020 à 2022. L’effet de cette période sera étudié plus directement au moyen d’une analyse de sensibilité.

3.2.2. Hétérogénéité des performances et estimation des poids

La validation LFO montre que le modèle RH est le plus performant en moyenne. Cette domination n’est toutefois pas uniforme sur l’ensemble de la grille : le meilleur modèle individuel change selon l’âge et l’horizon. Cette hétérogénéité justifie l’étude d’une pondération contextuelle. L’objectif n’est pas seulement d’identifier le meilleur modèle moyen, mais de déterminer si une combinaison peut adapter sa composition sans suivre excessivement le bruit de validation.

3. Évaluation des méthodes et implications actuarielles

Le *stacking* global attribue un poids de 0,887 726 au modèle RH et de 0,112 274 au modèle APC. Les modèles LC, CBD et M6 reçoivent des poids numériquement nuls. La distribution prédictive reste donc très proche d'un mélange dominé par RH.

Les poids contextuels non régularisés présentent des variations beaucoup plus marquées. Ils atteignent fréquemment des valeurs proches de zéro ou de un. Cette solution s'adapte étroitement au méta-échantillon de validation, mais elle peut amplifier des différences locales peu informatives.

Le modèle hiérarchique conserve une dépendance nette à l'âge et à l'horizon. Ses poids postérieurs moyens restent compris entre 0,000 934 et 0,992 072. Il ne se réduit donc pas au vecteur global, mais il atténue les transitions les plus abruptes du non-régularisé.

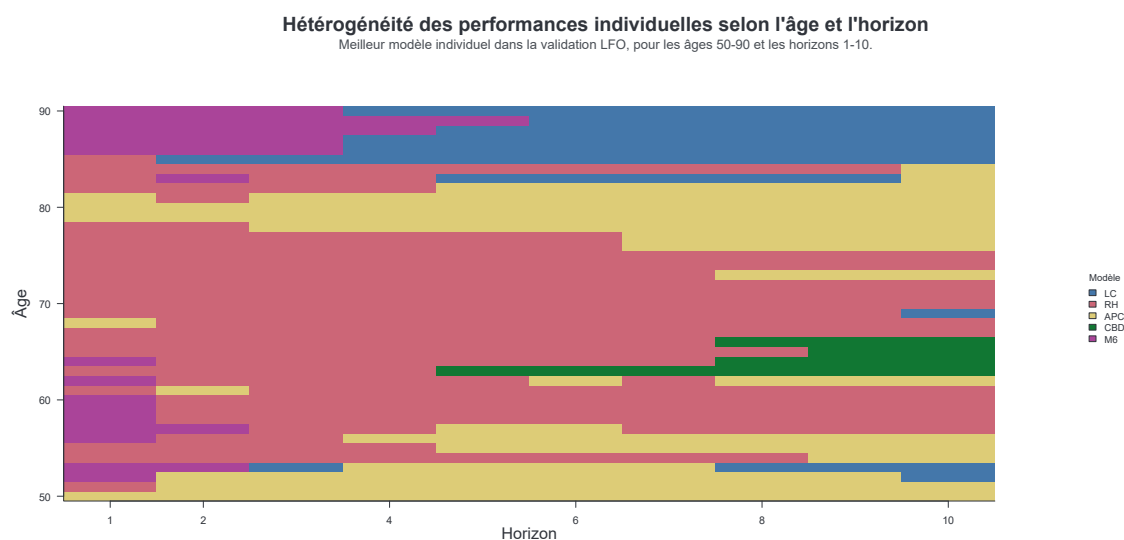


Figure 3.5. – Hétérogénéité des performances individuelles selon l'âge et l'horizon.

3. Évaluation des méthodes et implications actuarielles

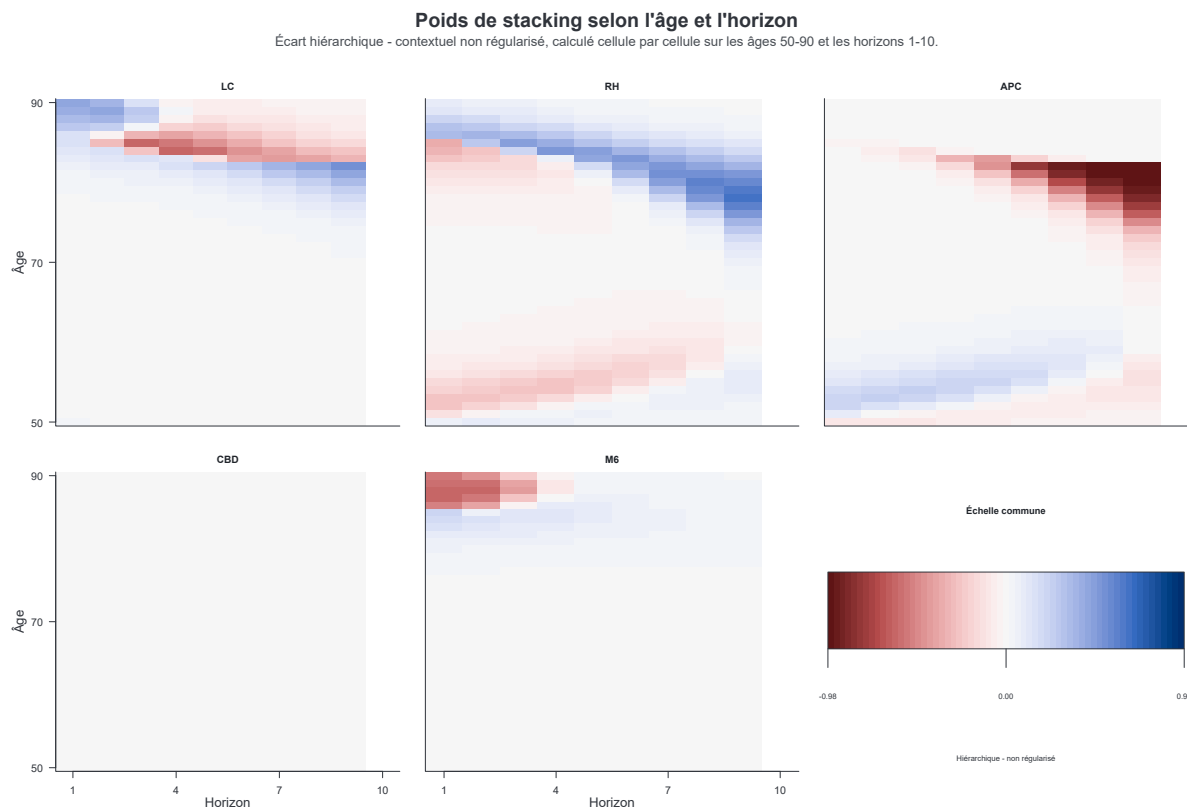


Figure 3.6. – Écarts entre les poids hiérarchiques et non régularisés selon l'âge et l'horizon.

Les figures 3.5 et 3.6 mettent en relation l'hétérogénéité des performances individuelles et l'effet de la régularisation. La figure 3.5 montre que le meilleur modèle individuel change selon l'âge et l'horizon dans la validation LFO. La figure 3.6 présente, pour chaque modèle, la différence entre le poids hiérarchique moyen et le poids contextuel non régularisé. Une valeur positive indique que la régularisation augmente le poids du modèle dans la cellule considérée, tandis qu'une valeur négative indique qu'elle le réduit. Ces écarts montrent que la régularisation ne modifie pas les poids de manière uniforme : elle redistribue leur contribution dans les zones où la solution non régularisée produit les variations les plus marquées.

3.2.3. Performance prédictive sur la période de test

Le LogS présenté correspond à la négative de la log-densité prédictive moyenne. Il doit donc être minimisé, comme le CRPS et la MAE. La MAE est exprimée en nombres de décès. Une couverture est jugée meilleure lorsqu'elle se rapproche de son niveau nominal, et non lorsqu'elle est simplement la plus élevée.

Tableau 3.5. – Performance prédictive globale sur la période de test allant de 2016 à 2024.

Méthode	LogS	CRPS	MAE décès	Couverture 80 %	Couverture 95 %	Largeur 95 %
LC	6,650	0,001 414	142,7	55,0 %	80,2 %	0,006 96
RH	6,394	0,001 408	115,4	71,5 %	87,5 %	0,006 75
APC	6,775	0,002 742	174,1	62,9 %	81,8 %	0,007 20
CBD	6,541	0,002 987	191,1	75,6 %	90,8 %	0,010 04
M6	9,264	0,002 333	256,4	29,8 %	46,6 %	0,010 26
Stacking global	6,355	0,001 469	118,5	74,5 %	89,2 %	0,007 64
Contextuel non régularisé	6,138	0,001 153	101,8	79,1 %	92,4 %	0,007 34
Contextuel hiérarchique	6,106	0,001 098	96,2	81,3 %	94,6 %	0,007 95

Le *stacking* contextuel hiérarchique obtient le meilleur résultat pour les trois critères principaux. Son LogS moyen est égal à 6,106, son CRPS à 0,001 098 et sa MAE à 96,2 décès. Le non-régularisé arrive en deuxième position, tandis que le global reste proche du modèle RH, qui reçoit l’essentiel de son poids.

Par rapport au global, le hiérarchique réduit le LogS de 0,249, le CRPS de 0,000 371 et la MAE de 22,3 décès. La diminution relative atteint environ 25,3 % pour le CRPS et 18,8 % pour la MAE. La variation des poids selon l’âge et l’horizon apporte donc une amélioration qui ne dépend pas d’une seule métrique.

La calibration conduit à la même conclusion. Le hiérarchique atteint une couverture de 81,3 % pour l’intervalle nominal à 80 % et de 94,6 % pour l’intervalle nominal à 95 %. Ces niveaux sont les plus proches des valeurs recherchées. Ses intervalles à 95 % sont légèrement plus larges que ceux du global, mais cette augmentation reste modérée et s’accompagne d’une amélioration nette de la couverture.

3.2.4. Variation selon l'âge et l'horizon

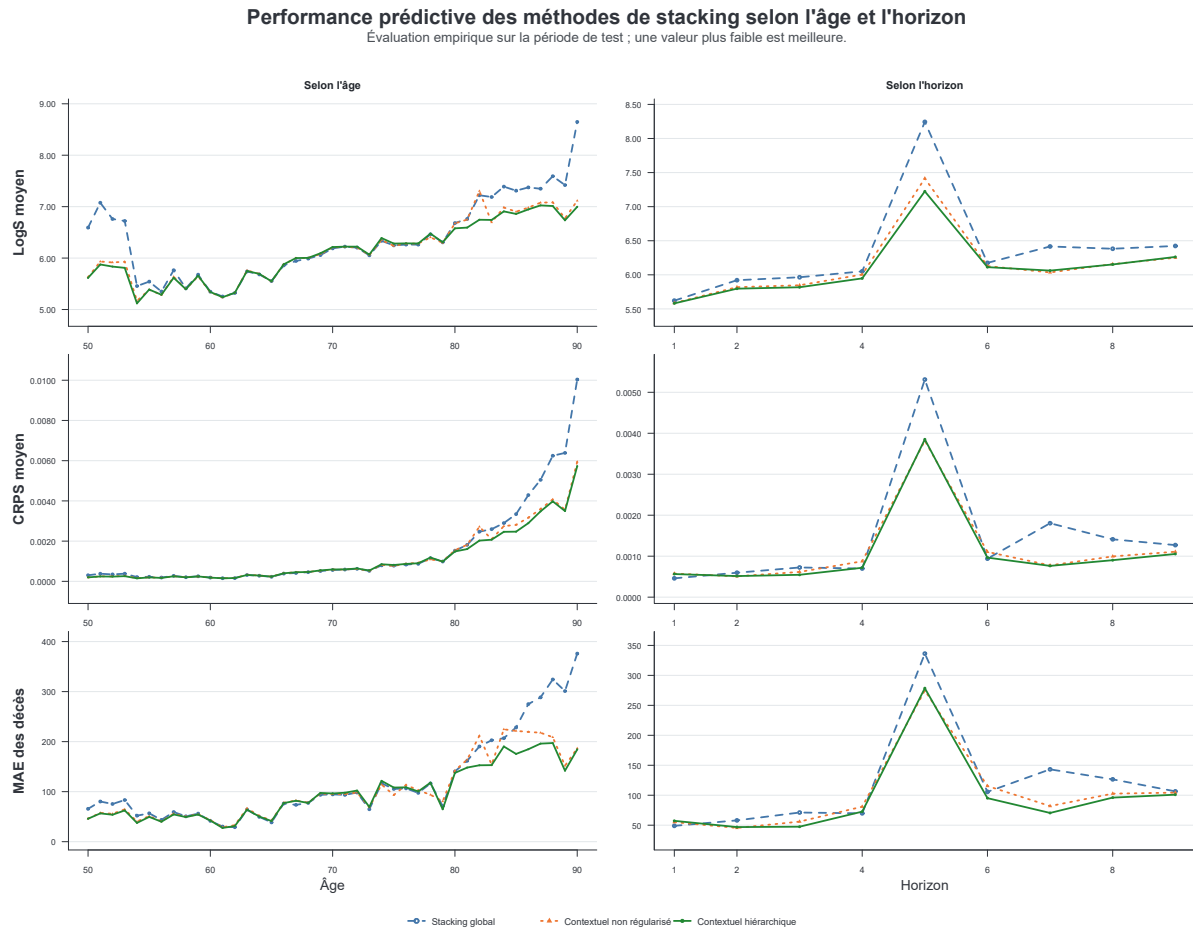


Figure 3.7. – Performance prédictive des méthodes de *stacking* selon l'âge et l'horizon.

La figure 3.7 montre que l'avantage du modèle hiérarchique n'est pas uniforme. Les niveaux du LogS et du CRPS augmentent aux âges élevés, tandis que la MAE dépend à la fois du niveau de mortalité et des expositions. Le classement local peut donc changer même lorsque le hiérarchique obtient le meilleur résultat moyen sur l'ensemble du test.

La variation selon l'horizon doit être interprétée conjointement avec les années cibles. L'origine étant fixée en 2015, les horizons 5 à 7 correspondent aux années 2020 à 2022. La dégradation visible sur ces horizons reflète donc à la fois l'éloignement de l'origine et les perturbations propres à ces années.

L'avantage des méthodes contextuelles ne disparaît pas lorsque les années 2020 à 2022 sont retirées. Sur les années restantes, le hiérarchique obtient un LogS de 5,926, un CRPS de 0,000 716 et une MAE de 70,2 décès. Les valeurs correspondantes sont 5,944, 0,000 780 et 74,1 pour le non-régularisé, puis 6,061, 0,000 860 et 80,2 pour le global. La période pandémique augmente donc le niveau des erreurs, mais elle n'explique pas à elle seule le classement général.

Une nuance apparaît sur les années 2021 à 2024. Le non-régularisé obtient un LogS

légèrement inférieur à celui du hiérarchique, avec 6,143 contre 6,147. Le hiérarchique conserve toutefois un meilleur **CRPS**, égal à 0,000 921 contre 0,000 996, et une meilleure MAE, égale à 90,6 contre 101,1 décès. Sur l'ensemble de la période 2016 à 2024, il reste premier pour les trois critères.

3.2.5. Incertitude sur les écarts de performance

Les différences moyennes sont complétées par un bootstrap réalisé au niveau des années cibles. À chaque répétition, les années de test sont tirées avec remise et toutes les observations d'âge correspondantes sont conservées. Cette procédure respecte la dépendance entre les 41 âges observés au cours d'une même année.

Tableau 3.6. – Différences de performance entre les méthodes et intervalles bootstrap à 95 %. Une différence négative favorise la première méthode.

Comparaison	Métrique	Différence moyenne	Borne 2,5 %	Borne 97,5 %
Contextuel non régularisé moins global	LogS	−0,217 329	−0,390 531	−0,087 464
Contextuel non régularisé moins global	CRPS	−0,000 315	−0,000 721	0,000 005
Contextuel non régularisé moins global	MAE décès	−16,609 407	−34,652 621	−0,089 540
Hiérarchique moins global	LogS	−0,249 027	−0,454 820	−0,109 434
Hiérarchique moins global	CRPS	−0,000 371	−0,000 735	−0,000 077
Hiérarchique moins global	MAE décès	−22,296 569	−39,907 734	−7,176 194
Hiérarchique moins contextuel non régularisé	LogS	−0,031 698	−0,076 548	−0,000 731
Hiérarchique moins contextuel non régularisé	CRPS	−0,000 056	−0,000 095	−0,000 017
Hiérarchique moins contextuel non régularisé	MAE décès	−5,687 163	−10,611 541	−1,138 134

Les intervalles associés à la comparaison entre le hiérarchique et le global restent entièrement négatifs pour les trois métriques. Les différences observées restent favorables au modèle hiérarchique dans les rééchantillonnages considérés, sous réserve du faible nombre d'années disponibles pour cette analyse. Elles restent également favorables par rapport au modèle contextuel non régularisé, bien que les écarts soient plus faibles.

La comparaison entre le non-régularisé et le global est plus nuancée. Les différences moyennes favorisent le non-régularisé, mais l'intervalle du **CRPS** touche légèrement zéro. Le résultat principal ne porte donc pas seulement sur l'intérêt d'une pondération contextuelle : il montre aussi que la régularisation améliore la capacité de généralisation de cette pondération.

Les intervalles obtenus doivent donc être interprétés comme une mesure descriptive de la stabilité des écarts entre années, plutôt que comme des intervalles asymptotiques très précis.

Les analyses complémentaires conduisent au même classement général. Le retrait successif de chaque année cible ne change pas le signe des principales différences. Les variantes supprimant l'effet quadratique de l'âge ou l'interaction, les modifications des lois *a priori*,

l'utilisation de plusieurs graines et le passage de 1 000 à 4 000 tirages prédictifs ne renversent pas les conclusions. Ces résultats détaillés sont reportés en annexe.

3.3. Transmission aux quantités actuarielles

L'amélioration des scores prédictifs ne garantit pas nécessairement une différence importante pour une quantité actuarielle. Les probabilités de survie et les valeurs de rente dépendent d'une succession de forces de mortalité. Les écarts entre méthodes peuvent donc se cumuler, mais ils peuvent également rester faibles lorsqu'ils concernent une durée limitée.

3.3.1. Construction des trajectoires

Les résultats de cette section portent sur la population belge **Total**. Ils décrivent donc des fonctionnelles d'une mortalité agrégée et ne constituent pas une tarification différenciée selon le sexe.

Pour les scores prédictifs cellule par cellule, la distribution agrégée reste un mélange des distributions candidates. La construction d'une trajectoire de survie est différente. Pour chaque tirage s et chaque intervalle j , les forces de mortalité sont combinées par une moyenne pondérée :

$$\mu_{x+j-1, T+j}^{(s, \text{agr})} = \sum_{k=1}^5 w_k(x+j-1, j) \mu_{x+j-1, T+j}^{(s, k)}. \quad (3.3)$$

Aucun modèle n'est sélectionné par un tirage catégoriel au cours de la trajectoire. Pour la version hiérarchique avec propagation, un même tirage postérieur des coefficients est utilisé sur l'ensemble de la surface âge et horizon d'une trajectoire. Cette construction préserve la cohérence entre les poids appliqués aux différentes années.

Les calculs reposent sur 1 000 tirages par méthode, avec une origine de projection fixée en 2015. L'analyse principale est limitée aux dix premières années, qui correspondent à la plage d'horizons couverte par la validation.

3.3.2. Probabilités de survie

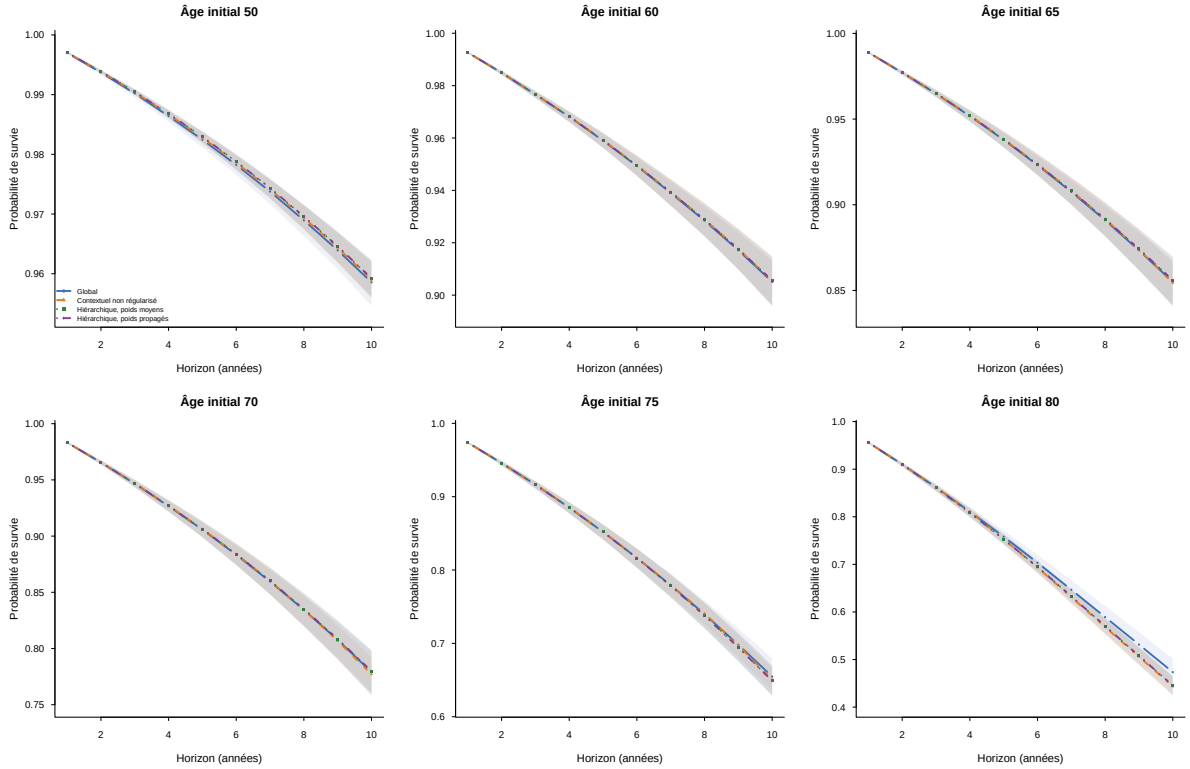


Figure 3.8. – Probabilités de survie après agrégation des forces sur les dix premières années de projection. Les bandes représentent les intervalles centraux à 95 %.

Les probabilités de survie restent très proches entre les quatre constructions aux âges initiaux 50, 60 et 65 ans. Les différences s’accumulent progressivement lorsque l’âge initial augmente. À 75 ans, elles commencent à devenir visibles aux horizons les plus longs. À 80 ans, la courbe du *stacking* global se situe au-dessus de celles des méthodes contextuelles.

Les bandes d’incertitude restent cependant largement superposées. Une différence entre les courbes moyennes ne doit donc pas être interprétée comme une séparation nette des distributions. La version hiérarchique à poids propagés présente une moyenne presque identique à celle obtenue avec les poids postérieurs moyens. La propagation de l’incertitude sur les poids agit principalement sur la dispersion.

3.3.3. Rentes temporaires

La quantité étudiée est une rente temporaire immédiate, annuelle et payable à terme échu. Pour un âge initial x , une durée n et un taux d’intérêt annuel i , sa valeur pour le tirage s est

$$a_{x:\overline{n}|}^{(s)} = \sum_{t=1}^n (1+i)^{-t} {}_t p_x^{(s)}. \quad (3.4)$$

3. Évaluation des méthodes et implications actuarielles

L'analyse principale utilise une durée de dix ans et un taux annuel de 2 %. Le premier paiement est conditionné à la survie pendant une année.

Tableau 3.7. – Rentes temporaires immédiates de dix ans au taux annuel de 2 %. Chaque cellule indique la moyenne suivie de l'intervalle central à 95 %.

Âge	Stacking global	Contextuel non régularisé	Hiérarchique, poids moyens	Hiérarchique, poids propagés
50	8,803 1 [8,788 6 ; 8,817 2]	8,805 7 [8,794 5 ; 8,816 7]	8,807 0 [8,796 2 ; 8,817 3]	8,807 0 [8,795 8 ; 8,817 9]
60	8,565 9 [8,535 3 ; 8,595 4]	8,568 0 [8,534 8 ; 8,601 0]	8,567 9 [8,535 1 ; 8,600 6]	8,567 9 [8,534 6 ; 8,601 2]
65	8,348 8 [8,301 5 ; 8,392 8]	8,351 4 [8,300 0 ; 8,399 8]	8,352 8 [8,300 5 ; 8,402 7]	8,352 8 [8,299 9 ; 8,404 1]
70	8,020 1 [7,945 9 ; 8,088 6]	8,020 2 [7,951 1 ; 8,088 7]	8,024 4 [7,946 5 ; 8,099 8]	8,024 3 [7,946 7 ; 8,098 7]
75	7,472 9 [7,368 4 ; 7,574 9]	7,471 1 [7,388 3 ; 7,548 4]	7,462 9 [7,369 8 ; 7,557 7]	7,462 7 [7,368 8 ; 7,558 9]
80	6,583 6 [6,452 7 ; 6,710 2]	6,492 4 [6,407 5 ; 6,579 2]	6,495 7 [6,412 5 ; 6,583 0]	6,495 4 [6,406 2 ; 6,588 7]

Note : les calculs reposent sur une origine de projection fixée en 2015 et sur 1 000 tirages par méthode.

La rente est annuelle, immédiate et payable à terme échu.

Aux âges 50 à 70 ans, les valeurs moyennes restent presque identiques. La différence entre le hiérarchique à poids moyens et le global varie entre 0,002 00 et 0,004 26. Ces écarts représentent une très faible proportion de la valeur de la rente.

À 75 ans, la différence moyenne devient négative et atteint $-0,009 94$. Elle est négative dans 75,3 % des tirages appariés, mais son intervalle central à 95 % contient zéro. À 80 ans, l'écart devient plus marqué. La différence moyenne entre le hiérarchique à poids moyens et le global est égale à $-0,087 87$ et elle est négative dans 97,3 % des tirages.

Afin de préciser la quantité $P(\Delta < 0)$ utilisée dans le tableau suivant, on note, pour un âge x et un tirage apparié s ,

$$\Delta_x^{(s)} = a_x^{(s, \text{hiér. moyen})} - a_x^{(s, \text{global})}.$$

La quantité $P(\Delta < 0)$ correspond ici à la proportion empirique des tirages appariés pour lesquels cette différence est négative :

$$\hat{P}(\Delta_x < 0) = \frac{1}{S} \sum_{s=1}^S \mathbf{1} \{ \Delta_x^{(s)} < 0 \}.$$

Avec $S = 1\,000$, une valeur de 75,3 % signifie donc que la rente obtenue avec le *stacking* hiérarchique à poids moyens est inférieure à celle du *stacking* global dans 753 des 1 000 tirages appariés. Cette proportion décrit le signe des différences simulées ; elle doit être interprétée conjointement avec leur amplitude et avec l'intervalle central à 95 %.

Tableau 3.8. – Différences appariées principales pour les rentes temporaires de dix ans au taux de 2 %.

Âge	Hiér. moyen moins global	IC 95 %	$P(\Delta < 0)$	Hiér. propagé moins hiér. moyen	IC 95 %
50	0,003 92	[−0,006 00; 0,013 76]	21,1 %	−0,000 01	[−0,003 98; 0,004 36]
60	0,002 00	[−0,002 01; 0,006 14]	17,9 %	0,000 03	[−0,001 98; 0,001 26]
65	0,003 92	[−0,003 14; 0,011 44]	14,4 %	−0,000 01	[−0,002 12; 0,001 61]
70	0,004 26	[−0,004 96; 0,013 66]	20,2 %	−0,000 09	[−0,006 45; 0,005 65]
75	−0,009 94	[−0,036 67; 0,016 81]	75,3 %	−0,000 24	[−0,017 60; 0,015 41]
80	−0,087 87	[−0,174 12; 0,001 13]	97,3 %	−0,000 35	[−0,037 25; 0,032 80]

La comparaison avec le contextuel non régularisé montre que les deux méthodes contextuelles conduisent à des valeurs proches. À 80 ans, la différence moyenne entre le hiérarchique à poids moyens et le non-régularisé est égale à 0,003 33, avec un intervalle à 95 % de [−0,022 45; 0,029 53]. L'écart observé à cet âge concerne donc principalement la comparaison entre les poids contextuels et le vecteur global.

La propagation de l'incertitude sur les poids modifie très peu les moyennes. À 80 ans, elle diminue la rente moyenne de 0,000 35. Son effet apparaît surtout dans la dispersion : l'écart-type passe de 0,044 16 à 0,047 57 et la largeur de l'intervalle central à 95 % augmente de 0,170 55 à 0,182 51.

Ces résultats montrent que l'amélioration des scores prédictifs ne se traduit pas automatiquement par une différence actuarielle importante. Sur une durée de dix ans, les écarts restent faibles aux âges modérés. Ils deviennent plus visibles aux âges élevés, lorsque la mortalité est plus forte et que les différences entre modèles s'accumulent davantage dans les probabilités de survie.

3.4. Discussion et limites

La simulation et l'application réelle répondent à deux questions complémentaires. La simulation permet d'évaluer les méthodes lorsque la structure des poids est connue, tandis que l'application belge examine leur capacité de généralisation lorsque ces poids ne sont pas observables.

Les résultats de simulation mettent d'abord en évidence le rôle stabilisateur de la régularisation. Le modèle contextuel non régularisé présente la RMSE la plus élevée et la plus forte fréquence de poids extrêmes dans chacun des quatre scénarios. Le modèle hiérarchique réduit fortement cette instabilité. Son avantage sur le global pour la récupération des poids reste toutefois modeste dans les scénarios dépendant de l'horizon ou simultanément de l'âge et de l'horizon. Il devient beaucoup plus net lorsque l'information est faible. Dans ce dernier cas, la RMSE moyenne du hiérarchique est égale à 0,103 8, contre 0,166 2 pour le global et 0,397 1 pour le non-régularisé.

L'interprétation des performances prédictives doit être plus nuancée. Le *stacking* global

3. Évaluation des méthodes et implications actuarielles

obtient les meilleurs LogS et CRPS moyens dans les trois scénarios les plus informatifs. Le hiérarchique reste proche de ces valeurs et remporte davantage de répétitions pour certains critères, notamment dans les scénarios dépendant de l'âge et de l'horizon ou de faible information. La simulation ne montre donc pas qu'une pondération contextuelle hiérarchique domine systématiquement une pondération globale. Elle montre plutôt que la régularisation permet de conserver une dépendance au contexte tout en évitant l'instabilité beaucoup plus marquée du modèle non régularisé.

L'application belge apporte un résultat complémentaire. Sur les années 2016 à 2024, le hiérarchique obtient les meilleurs scores moyens parmi les méthodes de *stacking*. Le non-régularisé améliore lui aussi les résultats par rapport au global, mais reste moins performant que le hiérarchique. L'amélioration observée sur les données réelles ne doit donc pas être présentée comme une conséquence automatique de la structure contextuelle : elle dépend de l'information effectivement contenue dans la validation et de la capacité de la régularisation à en extraire les variations qui se généralisent.

Le nouveau mécanisme générateur améliore la proximité avec une application de mortalité réelle, puisqu'il utilise des surfaces postérieures issues de cinq modèles ajustés aux données belges de 1970 à 2015, des expositions observées en 2015 et des paramètres de dispersion tirés de leurs distributions postérieures. Il reste néanmoins semi-synthétique. Les cinq modèles ne sont pas réestimés à chaque répétition, les expositions de 2015 sont répétées aux dix horizons et la composante génératrice est sélectionnée cellule par cellule selon des poids imposés. La simulation vise donc toujours à contrôler la structure des poids plutôt qu'à reproduire l'ensemble de la dynamique démographique belge.

Une seconde limite concerne l'incertitude sur les poids. Dans les trois scénarios les plus informatifs, la couverture des intervalles crédibles à 95 % du modèle hiérarchique se situe entre 76,18 % et 78,66 %. Cette sous-couverture indique que ces intervalles sont trop étroits au regard de la variabilité répétée du mécanisme générateur. En faible information, leur couverture atteint 98,49 %, mais au prix d'une largeur moyenne nettement plus importante. Les conclusions sur les poids moyens sont donc plus solides que celles portant sur la calibration exacte de leur incertitude.

Les effectifs effectivement utilisés varient également selon les scénarios. Sur les 120 répétitions exécutées, 113 satisfont les critères numériques après relance. Les comparaisons hiérarchiques reposent ainsi sur 29 répétitions dans le scénario constant, 24 dans le scénario dépendant de l'horizon et 30 dans les deux autres scénarios. Une optimisation du contextuel non régularisé échoue en faible information. Ces échecs sont conservés dans le bilan numérique et ne sont pas remplacés par des estimations ne satisfaisant pas les critères définis avant l'analyse.

L'application empirique porte sur la population totale. Les femmes et les hommes peuvent présenter des niveaux et des évolutions de mortalité différents. Les poids obtenus ici ne peuvent donc pas être interprétés comme ceux qui seraient nécessairement estimés

dans une analyse séparée par sexe.

Le test final comprend neuf années et utilise une origine fixe en 2015. L'année cible et l'horizon sont alors confondus. Une évaluation roulante, dans laquelle les modèles seraient réestimés chaque année, répondrait à une question différente. L'ensemble de modèles candidats reste également limité aux cinq structures présentées dans le premier chapitre. Une combinaison ne peut pas corriger une erreur commune à tous les modèles.

Enfin, l'illustration actuarielle est limitée à une rente temporaire de dix ans. Cette durée permet de rester dans la plage des horizons validés, mais elle ne représente pas l'ensemble des engagements de long terme. Les valeurs calculées sur la population totale doivent être interprétées comme une illustration de la transmission des prévisions, et non comme une base de tarification individuelle.

3.5. Conclusion du chapitre

L'étude de simulation HMD-calibrée montre que la régularisation hiérarchique agit d'abord sur la stabilité des poids. Le *stacking* global reste la méthode la plus précise lorsque les poids vrais sont constants et obtient les meilleurs scores prédictifs moyens dans les trois scénarios les plus informatifs. Le modèle hiérarchique reste proche de ces performances, réduit fortement la fréquence des poids extrêmes et obtient une récupération légèrement meilleure lorsque les poids dépendent du contexte. Son avantage devient beaucoup plus net en faible information. À l'inverse, le modèle non régularisé reste instable et peut conduire à une quasi-séparation lorsque les distributions candidates sont difficiles à distinguer.

Sur les données belges, le *stacking* contextuel hiérarchique obtient le meilleur LogS, le meilleur CRPS et la meilleure MAE sur la période de test allant de 2016 à 2024. Ses couvertures sont également les plus proches des niveaux nominaux. Les analyses par âge, par horizon et par période indiquent que ce résultat ne dépend pas uniquement des années 2020 à 2022.

Les conséquences actuarielles restent plus modérées. Les probabilités de survie et les valeurs de rente sont très proches aux âges intermédiaires pour une durée de dix ans. Les écarts deviennent plus visibles aux âges élevés, principalement entre le *stacking* global et les méthodes contextuelles. La propagation de l'incertitude sur les poids modifie peu les moyennes, mais augmente légèrement la dispersion.

Ces résultats soutiennent une interprétation prudente de la pondération contextuelle. Son intérêt ne réside pas dans une domination systématique du *stacking* global, mais dans la possibilité d'adapter les poids lorsque la performance relative des modèles varie, tout en limitant les solutions extrêmes lorsque l'information est insuffisante. La structure hiérarchique fournit ce compromis entre adaptation et stabilité.

Conclusion

Ce travail avait pour objectif de déterminer si une pondération des modèles de mortalité dépendant de l'âge et de l'horizon pouvait améliorer les prévisions, sans introduire de variations excessives lorsque l'information disponible est limitée. Cette question s'inscrit dans un contexte où aucun modèle stochastique ne domine nécessairement pour tous les âges et tous les horizons, de sorte que le choix d'une spécification unique laisse subsister une incertitude structurelle importante.

Cinq modèles de mortalité ont été considérés : Lee-Carter, Renshaw-Haberman, âge-période-cohorte, Cairns-Blake-Dowd et M6. Ils ont été estimés dans un cadre bayésien binomial négatif, permettant de représenter la surdispersion des nombres de décès ainsi que l'incertitude associée aux paramètres et aux trajectoires futures. Leurs distributions prédictives ont ensuite été combinées au moyen d'une validation *Leave-Future-Out*, afin que les poids reposent sur des performances réellement observées hors échantillon.

Trois méthodes d'agrégation ont été comparées. Le *stacking* global utilise un seul vecteur de poids pour l'ensemble des âges et des horizons. Le *stacking* contextuel non régularisé autorise ces poids à varier selon l'âge, l'horizon et leur interaction. Enfin, le *stacking* contextuel hiérarchique conserve cette flexibilité, mais introduit une régularisation destinée à limiter les variations qui ne sont pas suffisamment soutenues par les données.

L'étude de simulation a été calibrée à partir des données belges de mortalité observées entre 1970 et 2015. Les cinq modèles ont été ajustés une fois sur cette période et leurs distributions postérieures ont servi de composantes au mécanisme générateur. Les résultats montrent que le *stacking* global reste très compétitif. Il récupère le mieux les poids lorsqu'ils sont constants et obtient les meilleurs scores prédictifs moyens dans les trois scénarios les plus informatifs. Le modèle hiérarchique reste proche de ces performances, réduit fortement la fréquence des poids extrêmes et obtient une récupération légèrement meilleure lorsque les poids dépendent de l'horizon ou simultanément de l'âge et de l'horizon. Son avantage devient surtout visible en faible information, où il reste stable alors que le modèle non régularisé produit presque uniquement des poids extrêmes et rencontre un cas de non-convergence.

Cette simulation conduit également à nuancer l'interprétation de l'incertitude sur les poids. Dans les trois scénarios les plus informatifs, la couverture des intervalles crédibles hiérarchiques à 95 % reste inférieure au niveau nominal, tandis qu'elle se rapproche de 95 % lorsque l'information est faible au prix d'intervalles plus larges. La régularisation

3. Évaluation des méthodes et implications actuarielles

améliore donc principalement la stabilité des estimations ponctuelles, sans garantir une calibration parfaite de l'incertitude sur les poids.

L'application aux données belges confirme l'intérêt de cette régularisation dans un contexte où les poids vrais ne sont pas observables. L'analyse porte sur la population totale, pour les âges de 50 à 90 ans et les années 1970 à 2024. Les poids sont estimés à partir d'exercices historiques de prévision dont les années cibles s'arrêtent en 2015, tandis que la période 2016–2024 est conservée entièrement à l'écart pour l'évaluation finale.

Sur cette période de test, le *stacking* contextuel hiérarchique obtient les meilleures performances globales parmi les méthodes comparées. Son score logarithmique moyen est égal à 6,106, son CRPS moyen à 0,001 098 et son erreur absolue moyenne à 96,2 décès. Par rapport au *stacking* global, il réduit le CRPS d'environ 25,3 % et l'erreur absolue moyenne d'environ 18,8 %. Ses couvertures empiriques, égales à 81,3 % pour les intervalles nominaux à 80 % et à 94,6 % pour les intervalles nominaux à 95 %, sont également les plus proches des niveaux recherchés.

Les analyses par âge et par horizon montrent toutefois que cet avantage n'est pas uniforme. Les performances relatives des modèles varient dans l'espace de prévision, ce qui justifie l'étude de poids contextuels. La période de test reposant sur une origine fixe en 2015, l'effet propre de l'horizon ne peut cependant pas être complètement séparé de celui de l'année cible. Les horizons 5 à 7 correspondent notamment aux années 2020 à 2022. Le retrait de ces années ne modifie toutefois pas le classement général des méthodes, ce qui indique que l'amélioration observée ne repose pas uniquement sur cette période particulière.

Les conséquences sur les quantités actuarielles sont plus modérées que les différences mesurées par les scores prédictifs. Pour une projection de dix ans, les probabilités de survie et les valeurs de rente restent très proches aux âges intermédiaires. Les écarts deviennent plus visibles aux âges élevés, où les différences entre les forces de mortalité se cumulent davantage. La propagation de l'incertitude sur les poids modifie peu les valeurs moyennes, mais augmente légèrement la dispersion des quantités actuarielles.

Ces résultats conduisent à répondre positivement, mais avec nuance, à la question principale du mémoire. Une pondération dépendant de l'âge et de l'horizon peut améliorer les prévisions lorsque les performances relatives des modèles varient effectivement selon le contexte, comme l'indique l'application belge. La simulation montre toutefois que cette amélioration n'est pas automatique et que le *stacking* global peut rester difficile à dépasser en moyenne. L'intérêt de la structure hiérarchique réside surtout dans sa capacité à autoriser des variations contextuelles tout en limitant les solutions extrêmes et l'adaptation excessive au bruit de validation.

Plusieurs limites doivent être rappelées. La simulation reste semi-synthétique malgré son calibrage sur les données HMD : les cinq modèles sont ajustés une seule fois, les expositions de 2015 sont répétées aux différents horizons et la structure des poids vrais est

3. Évaluation des méthodes et implications actuarielles

imposée. La couverture des intervalles sur les poids reste également inférieure au niveau nominal dans trois scénarios. L'application porte sur la population belge totale et ne permet pas de conclure que les mêmes pondérations seraient obtenues pour les femmes et les hommes séparément. L'ensemble des candidats est limité à cinq modèles classiques et une combinaison ne peut pas corriger une erreur commune à toutes les spécifications. La période de test ne comprend que neuf années et repose sur une origine de prévision fixe. Enfin, l'étude actuarielle est limitée à une rente temporaire de dix ans, afin de rester dans la plage des horizons effectivement validés.

Des prolongements pourraient examiner séparément les populations féminine et masculine, élargir l'ensemble des modèles candidats ou utiliser une évaluation roulante permettant de mieux distinguer l'effet de l'horizon de celui de l'année cible. Une extension de la simulation pourrait également réestimer les modèles à chaque répétition ou faire évoluer les expositions dans le temps. L'étude de quantités actuarielles de plus longue durée constituerait enfin une extension naturelle, à condition de disposer d'une validation adaptée à ces horizons.

Dans l'ensemble, ce travail montre que l'intérêt du *stacking* contextuel ne réside pas dans une flexibilité maximale. Il repose sur un compromis entre adaptation et stabilité. La structure hiérarchique permet de conserver une dépendance à l'âge et à l'horizon lorsque celle-ci est utile, tout en réduisant les comportements extrêmes lorsque l'information est insuffisante. Elle constitue ainsi une extension pertinente du *stacking* global, sans remettre en cause le fait que ce dernier demeure une référence solide pour l'agrégation probabiliste de modèles de mortalité.

Bibliographie

- Barigou, K., P.-O. Goffard, S. Loisel et Y. Salhi (2023). « Bayesian model averaging for mortality forecasting using leave-future-out validation ». In : *International Journal of Forecasting* 39.2, p. 674-690.
- Bissiri, P. G., C. C. Holmes et S. G. Walker (2016). « A General Framework for Updating Belief Distributions ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 78.5, p. 1103-1130.
- Cairns, A. J. G., D. Blake et K. Dowd (2006). « A Two-Factor Model for Stochastic Mortality with Applications to Longevity Risk Management ». In : *North American Actuarial Journal* 10.4, p. 123-143.
- Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong et I. Balevich (2009). « A Quantitative Comparison of Stochastic Mortality Models Using Data from England and Wales and the United States ». In : *North American Actuarial Journal* 13.1, p. 1-35. url : <https://doi.org/10.1080/10920277.2009.10597538>.
- Delwarde, A., M. Denuit et C. Partrat (2007). « Negative Binomial Version of the Lee–Carter Model for Mortality Forecasting ». In : *Applied Stochastic Models in Business and Industry* 23.5, p. 385-401.
- Human Mortality Database (2026). *Belgium : Deaths (Period 1x1) and Exposure to Risk (Period 1x1)*. Fichiers modifiés le 21 octobre 2025 et téléchargés le 13 juillet 2026. Max Planck Institute for Demographic Research, University of California, Berkeley, et French Institute for Demographic Studies. url : <https://www.mortality.org/> (visité le 13/07/2026).
- Kessy, S. R., M. Sherris, A. M. Villegas et J. Ziveyi (2022). « Mortality forecasting using stacked regression ensembles ». In : *Scandinavian Actuarial Journal* 2022.7, p. 591-626.
- Lee, R. D. et L. R. Carter (1992). « Modeling and Forecasting U.S. Mortality ». In : *Journal of the American Statistical Association* 87.419, p. 659-671.
- Minko, B. A. J. (2026). *Code et résultats reproductibles du mémoire : agrégation bayésienne contextuelle de modèles de mortalité*. Version 1.0.0. url : <https://github.com/bminko/stacking-bayesien-mortalite> (visité le 30/07/2026).
- Renshaw, A. E. et S. Haberman (2006). « A cohort-based extension to the Lee–Carter model for mortality reduction factors ». In : *Insurance : Mathematics and Economics* 38.3, p. 556-570.

- Wilmoth, J. R., K. Andreev, D. Jdanov, D. A. Gleijer et T. Riffe (5 août 2025). *Methods Protocol for the Human Mortality Database*. Version 6. url : <https://www.mortality.org/File/GetDocument/Public/Docs/MethodsProtocolV6.pdf> (visité le 30/07/2026).
- Wong, J. S., J. J. Forster et P. W. Smith (2018). « Bayesian Mortality Forecasting with Overdispersion ». In : *Insurance : Mathematics and Economics* 83, p. 206-221.
- Yao, Y., G. Pirš, A. Vehtari et A. Gelman (2022). « Bayesian Hierarchical Stacking : Some Models Are (Somewhere) Useful ». In : *Bayesian Analysis* 17.4, p. 1043-1071.
- Yao, Y., A. Vehtari, D. Simpson et A. Gelman (2018). « Using Stacking to Average Bayesian Predictive Distributions ». In : *Bayesian Analysis* 13.3, p. 917-1003.

A. Implémentation numérique et reproductibilité

Cette annexe décrit les éléments nécessaires pour comprendre et reproduire l'application empirique. L'ensemble du code source, des sorties détaillées et des diagnostics individuels est disponible dans le dépôt public `github` associé à ce travail ([Minko 2026](#)).

A.1. Estimation des modèles de mortalité

L'application porte sur la Belgique et sur la population `Total`, qui réunit les femmes et les hommes. Les décès et les expositions annuels par âge simple proviennent de la *Human Mortality Database* ([Human Mortality Database 2026](#)). Les âges retenus vont de 50 à 90 ans. Les données disponibles couvrent les années 1970 à 2024, mais les ajustements finaux des modèles utilisent uniquement la période 1970 à 2015. Les années 2016 à 2024 restent ainsi hors échantillon.

Les fichiers indiquent le protocole HMD, version 6. La construction générale des taux et des expositions est documentée dans le protocole officiel ([Wilmoth et al. 2025](#)).

Les cinq modèles LC, RH, APC, CBD et M6 sont estimés sous une loi binomiale négative. Pour une cellule (x, t) , l'exposition $E_{x,t}$ intervient comme terme de décalage :

$$D_{x,t} \sim \text{NB}(E_{x,t} \exp\{\eta_{x,t}\}, \phi).$$

Cette écriture distingue l'intensité de mortalité du volume de population exposé au risque et autorise une surdispersion par rapport à la loi de Poisson.

Les modèles sont programmés dans `Stan` et appelés depuis `R` 4.3.1. L'exécution finale a utilisé l'interface `RStan`; le compilateur embarqué est `stanc` 2.32.2. Les packages `posterior` et `bridgesampling` complètent les outils d'estimation et de diagnostic. Chaque ajustement final repose sur quatre chaînes, 1 000 itérations de chauffe et 1 000 itérations conservées par chaîne, soit 4 000 tirages conservés. La configuration utilise `adapt_delta=0,95` et `max_treedepth=12`. Une métrique diagonale est employée pour LC et CBD; une métrique dense est utilisée pour RH, APC et M6.

Les contraintes d'identification sont imposées directement dans les programmes `Stan`. Pour LC et RH, le premier facteur de période est fixé à zéro et le profil d'âge appartient au

simplexe. Pour APC, le premier facteur de période est également fixé à zéro. Dans RH, APC et M6, les deux cohortes extrêmes sont fixées à zéro et les cohortes internes sont exprimées dans une base orthonormale, orthogonale au niveau constant et à la tendance linéaire. Pour CBD et M6, les deux états initiaux sont estimés, car ces modèles ne contiennent pas d'intercept d'âge séparé.

Les facteurs de période univariés de LC, RH et APC sont projetés par marche aléatoire avec dérive. Les deux facteurs de CBD et M6 suivent une marche aléatoire bivariée avec innovations corrélées. Les effets de cohorte de RH, APC et M6 sont projetés par un processus autorégressif d'ordre deux. Sa paramétrisation par autocorrélations partielles garantit la stationnarité. Mille tirages prédictifs sont produits par méthode et par cellule.

Tableau A.1. – Diagnostics des cinq ajustements finaux sur la période 1970 à 2015.

Modèle	$\hat{R}_{\max}$	ESS bulk _{min}	ESS tail _{min}	Div.	Prof.	E-BFMI _{min}	Statut
LC	1,008 1	643	1 918	0	0	0,944	Valide
RH	1,003 3	2 876	2 085	0	0	0,927	Valide
APC	1,003 7	2 840	1 422	0	0	0,970	Valide
CBD	1,007 6	501	717	0	0	0,390	Valide
M6	1,003 9	3 719	2 151	0	0	0,954	Valide

Un ajustement est retenu si $\hat{R} < 1,05$, si les ESS bulk et tail sont supérieures à 400, si aucune divergence ni atteinte de la profondeur maximale n'est observée et si aucune chaîne n'est bloquée.

A.2. Boucle de validation *Leave-Future-Out*

La validation LFO utilise les origines 2000 à 2014 et les horizons 1 à 10, dans la limite des années cibles disponibles jusqu'en 2015. Pour chaque origine r , les cinq modèles sont ajustés uniquement avec les observations disponibles jusqu'à r . Ils produisent ensuite les distributions prédictives des années $r + h$. La log-densité du décès effectivement observé est enregistrée pour chaque âge, chaque horizon admissible et chacun des cinq modèles.

La boucle peut être résumée par les étapes suivantes :

1. ajuster les cinq modèles avec les observations disponibles jusqu'à l'origine r ;
2. simuler les prévisions aux horizons admissibles ;
3. calculer la densité prédictive du décès observé ;
4. enregistrer l'âge, l'origine, l'horizon, l'année cible et les cinq log-densités ;
5. réunir les résultats dans le méta-jeu utilisé pour estimer les poids.

Tableau A.2. – Exemples de couples origine–horizon dans la validation LFO.

Origine r	Horizon h	Année cible $r + h$	Âges évalués
2000	1	2001	50–90
2000	10	2010	50–90
2005	10	2015	50–90
2010	5	2015	50–90
2014	1	2015	50–90

La procédure comprend 15 origines et 5 modèles, soit 75 ajustements LFO. Le méta-jeu final contient 4 305 cellules de validation hors modèle et 21 525 log-densités. Les cinq colonnes de log-densité sont entièrement finies et ne contiennent aucune valeur manquante.

Une même année cible peut être obtenue depuis plusieurs origines. Pour éviter qu’elle soit mécaniquement surreprésentée, chaque exercice reçoit le poids

$$\omega_{r,h} = \frac{1}{m_{r+h}},$$

où m_{r+h} est le nombre d’exercices conduisant à la même année cible. La somme des poids est ainsi égale à un pour une année cible et un âge donnés. Cette correction ne rend toutefois pas indépendants les exercices qui partagent une partie des données d’apprentissage ou une même observation cible.

Tableau A.3. – Résumé des diagnostics des 75 ajustements LFO.

Modèle	Ajust.	Valides	$\hat{R}_{\max}$	ESS bulk _{min}	ESS tail _{min}	Div./Prof.
LC	15	15	1,012 4	416	746	0/0
RH	15	15	1,004 4	2 528	1 500	0/0
APC	15	15	1,004 8	1 992	1 055	0/0
CBD	15	15	1,015 0	417	701	0/0
M6	15	15	1,007 6	2 171	1 375	0/0

A.3. Estimation des trois formes de *stacking*

Les trois méthodes utilisent le même méta-jeu et le même critère LFO pondéré. Les log-densités sont combinées au moyen de `log_sum_exp`, ce qui évite une perte de précision numérique.

Stacking global

Le *stacking* global estime un seul vecteur de cinq poids, soumis à la contrainte du simplexe. Les paramètres libres sont transformés par *softmax*, puis le critère LFO pondéré est maximisé par BFGS. Douze points de départ sont utilisés. Parmi les solutions de code de convergence zéro et d'objectif fini, le programme retient celle qui maximise le critère. L'optimisation finale a convergé après 15 évaluations principales de la fonction.

Stacking contextuel non régularisé

Le prédicteur contextuel est identique à celui du chapitre 2 :

$$\eta_k(x, h) = \alpha_k + \beta_{k,x}\tilde{x} + \beta_{k,x^2}q_x + \beta_{k,h}\tilde{h} + \beta_{k,xh}\tilde{x}\tilde{h},$$

avec

$$q_x = \tilde{x}^2 - \overline{\tilde{x}^2}.$$

L'âge et l'horizon sont standardisés à partir des valeurs uniques du méta-jeu, en utilisant leur moyenne et leur écart-type empirique. Le modèle **M6** est la catégorie de référence : son prédicteur est fixé à zéro. Les poids sont obtenus par transformation *softmax*. Aucun terme de pénalisation et aucune loi *a priori* ne sont ajoutés. L'optimisation BFGS utilise douze initialisations et conserve la meilleure solution convergée. L'ajustement final a nécessité 904 évaluations principales de la fonction.

Stacking contextuel hiérarchique

Le modèle hiérarchique utilise le même prédicteur et la même catégorie de référence. Pour les quatre modèles non référents,

$$\alpha_k \sim \mathcal{N}(0, 1,5^2), \quad \beta_{k,j} = \tau_j z_{k,j}, \quad z_{k,j} \sim \mathcal{N}(0, 1).$$

Les échelles associées à l'âge et à l'horizon suivent une demi-loi normale d'échelle un. Celles du terme quadratique centré et de l'interaction suivent une demi-loi normale d'échelle 0,5.

La cible estimée sous **Stan** est la distribution *a posteriori* généralisée définie à l'équation (2.72). Le terme $\exp\{\ell_{\text{LFO}}(\Psi)\}$ est une utilité prédictive composite et non la vraisemblance d'un modèle génératif du méta-jeu. La correction $\omega_{r,h}$ est appliquée directement à chaque contribution. Aucun paramètre de température supplémentaire n'est utilisé.

L'estimation hiérarchique utilise quatre chaînes, 1000 itérations de chauffe et 1000 itérations conservées par chaîne. La métrique est dense, le seuil d'acceptation est réglé à 0,98 et la profondeur maximale à 12. Son $\hat{R}$ maximal vaut 1,0031, l'ESS bulk minimale 1 422 et l'ESS tail minimale 1 033. Aucune divergence ni atteinte de la profondeur maximale n'est observée. L'E-BFMI minimal par chaîne vaut 0,808.

Tableau A.4. – Comparaison des trois formes de *stacking*.

Méthode	Structure des poids	Régularisation	Estimation	Incertitude sur les poids
Global	Un vecteur constant	Aucune	BFGS, 12 dé-parts	Non
Contextuel non régularisé	Fonction de l'âge et de l'horizon	Aucune	BFGS, 12 dé-parts	Non
Contextuel hiérarchique	Même fonction contextuelle	Mise en commun par les τ_j	NUTS sous Stan	Oui, sous la cible généralisée

A.4. Prévisions et trajectoires actuarielles

Pour l'évaluation cellule par cellule, les poids sont d'abord évalués au couple (x, h) . Chaque tirage sélectionne ensuite un modèle candidat selon une loi catégorielle dont les probabilités sont ces poids, puis prend un tirage de la distribution prédictive du modèle sélectionné. Mille tirages sont utilisés par cellule. Cette procédure produit bien le mélange probabiliste utilisé pour le LogS, le CRPS et les couvertures. Les quantiles du mélange sont calculés à partir des tirages agrégés ; ils ne sont jamais obtenus en moyennant les quantiles des cinq modèles.

La construction actuarielle est différente, car elle doit conserver une trajectoire de forces de mortalité cohérente. Pour chaque indice de simulation s et chaque durée j ,

$$\mu_{x+j-1, T+j}^{(s, \text{agr})} = \sum_{k=1}^5 w_k(x+j-1, j) \mu_{x+j-1, T+j}^{(s, k)}.$$

L'âge $x+j-1$ est l'âge atteint pendant la j -ième année de projection, $T+j$ l'année calendaire et j l'horizon depuis l'origine $T = 2015$. Les forces des cinq modèles sont agrégées à chaque couple âge–horizon avec le même indice de tirage s . Aucun modèle unique n'est sélectionné pour toute la trajectoire et aucune sélection catégorielle indépendante n'est effectuée pour les quantités actuarielles.

La probabilité de survie sur j années est calculée par

$${}_j p_x^{(s)} = \exp \left\{ - \sum_{\ell=1}^j \mu_{x+\ell-1, T+\ell}^{(s, \text{agr})} \right\}.$$

La rente temporaire immédiate, annuelle et payable à terme échu, de durée n est ensuite

$$a_{x:\overline{n}|}^{(s)} = \sum_{j=1}^n (1+i)^{-j} {}_j p_x^{(s)}.$$

L'analyse principale retient $n = 10$ et $i = 0,02$. Elle utilise les âges initiaux 50, 60, 65, 70, 75 et 80 ans. La dernière force de la trajectoire principale atteint donc au plus l'âge 89. La restriction à dix ans maintient les calculs confirmatoires dans la plage d'horizons couverte par la validation LFO. Les contrôles automatiques vérifient notamment la somme des poids, leur appartenance à $[0, 1]$, la décroissance des probabilités de survie et la borne supérieure donnée par la rente certaine.

B. Étude de simulation et analyses complémentaires

Cette annexe complète la présentation du chapitre 3. Elle précise le processus générateur de la nouvelle étude de simulation calibrée sur les données belges de la *Human Mortality Database* et résume les diagnostics numériques des 120 répétitions exécutées. Les analyses de sensibilité de l'application empirique sont ensuite rappelées séparément.

B.1. Processus générateur de la simulation

La grille comprend les âges $x = 50, \dots, 90$ et les horizons $h = 1, \dots, 10$, soit $41 \times 10 = 410$ cellules. La population de référence est la Belgique, population **Total**. Les cinq modèles LC, RH, APC, CBD et M6 sont ajustés une seule fois sur les données observées entre 1970 et 2015, pour les âges de 50 à 90 ans. Pour chacun d'eux, 1 000 surfaces postérieures de force de mortalité et les tirages postérieurs correspondants du paramètre de dispersion sont conservés.

L'origine de projection est fixée en 2015. Les horizons $h = 1, \dots, 10$ correspondent donc aux années simulées 2016 à 2025. Les décès effectivement observés entre 2016 et 2024 dans l'application empirique ne sont pas utilisés pour calibrer le mécanisme générateur. L'étude de simulation reste ainsi séparée de la période de test réelle.

Pour chaque base simulée et pour chaque modèle k , une ligne postérieure est tirée une seule fois. Elle détermine sur l'ensemble de la grille la surface de force de mortalité

$$\mu_k(x, h)$$

et le paramètre de dispersion associé ϕ_k . Les expositions $E_{x,h}$ reprennent le profil observé dans la HMD en 2015 et ce même profil d'âge est répété aux dix horizons. Elles varient de 24 287,81 à 167 039,10, pour une médiane égale à 99 197,02.

Dans chaque cellule (x, h) , un indice $Z_{x,h}$ identifiant la composante génératrice est tiré selon les poids vrais :

$$Z_{x,h} \sim \text{Cat}(w_1^*(x, h), \dots, w_5^*(x, h)).$$

Conditionnellement à $Z_{x,h} = k$, le nombre de décès suit une loi binomiale négative de

paramétrisation NB2 :

$$D_{x,h} \mid Z_{x,h} = k \sim \text{NB2}(E_{x,h}\mu_k(x, h), \phi_k).$$

Les bases de validation et de test sont générées indépendamment avec des graines distinctes. La première sert exclusivement à estimer les poids de *stacking*, tandis que la seconde est utilisée pour l'évaluation prédictive hors échantillon.

Les tirages postérieurs de dispersion diffèrent selon les modèles. Leur distribution est résumée dans le tableau B.1.

Tableau B.1. – Résumé des tirages postérieurs du paramètre de dispersion utilisés dans la simulation.

Modèle	Moyenne	Quantile 2,5 %	Médiane	Quantile 97,5 %
LC	1 117,3	993,5	1 114,3	1 253,1
RH	8 879,6	6 383,7	8 658,1	12 689,9
APC	1 417,7	1 259,0	1 414,4	1 591,6
CBD	199,1	184,8	198,9	213,6
M6	6 883,5	5 110,1	6 772,4	9 439,0

Dans le scénario à poids constants,

$$\mathbf{w}^*(x, h) = (0,06, 0,54, 0,24, 0,10, 0,06).$$

Le scénario de faible information utilise le vecteur constant et plus diffus

$$\mathbf{w}^*(x, h) = (0,18, 0,24, 0,22, 0,18, 0,18).$$

Pour les deux scénarios contextuels, les poids sont obtenus par transformation *softmax*. On note

$$\tilde{x} = \frac{x - 70}{20}, \quad \tilde{h} = \frac{h - 5,5}{4,5},$$

et M6 constitue la catégorie de référence. Lorsque les poids dépendent uniquement de l'horizon,

$$\begin{aligned} \eta_{\text{LC}} &= -0,8 - 0,7\tilde{h}, & \eta_{\text{RH}} &= 0,4 + 1,1\tilde{h}, \\ \eta_{\text{APC}} &= 0,3 - 0,4\tilde{h}, & \eta_{\text{CBD}} &= -0,2 + 0,25\tilde{h}, & \eta_{\text{M6}} &= 0. \end{aligned}$$

Dans le scénario dépendant de l'âge et de l'horizon, $q(x)$ désigne le carré centré de l'âge

standardisé et les scores sont

$$\begin{aligned}\eta_{LC} &= -0,7 - 0,9\tilde{x} - 0,35\tilde{h}, \\ \eta_{RH} &= 0,3 + 0,8\tilde{h} + 0,55\tilde{x}\tilde{h}, \\ \eta_{APC} &= 0,5 - 0,55q(x) - 0,25\tilde{h}, \\ \eta_{CBD} &= -0,2 + 0,65\tilde{x} - 0,45\tilde{x}\tilde{h}, \\ \eta_{M6} &= 0.\end{aligned}$$

Dans les deux scénarios,

$$w_k^*(x, h) = \frac{\exp\{\eta_k(x, h)\}}{\sum_{j=1}^5 \exp\{\eta_j(x, h)\}}.$$

Le scénario de faible information combine deux modifications supplémentaires. Les expositions sont multipliées par 0,10 et les différences entre les distributions candidates sont réduites. Pour chaque cellule, si

$$\bar{\ell}(x, h) = \frac{1}{5} \sum_{k=1}^5 \log \mu_k(x, h),$$

la force utilisée dans ce scénario est définie par

$$\log \mu_k^{\text{FI}}(x, h) = \bar{\ell}(x, h) + 0,25 [\log \mu_k(x, h) - \bar{\ell}(x, h)].$$

Les paramètres de dispersion ϕ_k ne sont pas modifiés. Ce mécanisme rend les candidats plus difficiles à distinguer tout en conservant des niveaux de mortalité issus des ajustements HMD.

Tableau B.2. – Définition des quatre scénarios de simulation HMD-calibrée.

Scénario	Structure des poids vrais	Exposition et séparation des candidats	Question évaluée
Poids constants	Vecteur fixe (0,06, 0,54, 0,24, 0,10, 0,06)	Profil HMD 2015 ; distributions prédictives non contractées	La méthode hiérarchique évite-t-elle d'introduire une variation contextuelle inutile ?
Horizon	<i>Softmax</i> dépendant de $\tilde{h}$	Profil HMD 2015 ; distributions prédictives non contractées	La méthode retrouve-t-elle une variation selon l'horizon sans créer d'effet d'âge artificiel ?
Âge-horizon	<i>Softmax</i> dépendant de $\tilde{x}, \tilde{h}, q(x)$ et de leur interaction	Profil HMD 2015 ; distributions prédictives non contractées	La méthode retrouve-t-elle une surface contextuelle dépendant de l'âge et de l'horizon ?
Faible information	Vecteur fixe (0,18, 0,24, 0,22, 0,18, 0,18)	Expositions multipliées par 0,10 ; écarts des log-forces multipliés par 0,25	La régularisation empêche-t-elle l'apparition de variations artificielles lorsque l'information est faible ?

B.2. Diagnostics numériques des répétitions

L'étude principale comprend 30 répétitions indépendantes dans chacun des quatre scénarios, soit 120 répétitions exécutées. Le *stacking* global et le contextuel non régularisé sont estimés par optimisation BFGS avec deux points de départ, tandis que le modèle hiérarchique est estimé intégralement sous **Stan**.

Les simulations ont été exécutées sous Windows 11 avec **R** 4.5.3, **RStan** 2.32.7, **posterior** 1.7.0 et **Stan** 2.32.2. Cet environnement est distinct de l'exécutable **R** 4.3.1 utilisé pour le pipeline empirique décrit dans l'annexe A.

La configuration légère utilise quatre chaînes, 500 itérations de chauffe et 500 itérations conservées par chaîne, avec `adapt_delta`=0,98, une métrique dense et `max_treedepth`=12. Une répétition est rejetée si $\hat{R} > 1,05$, si une ESS bulk ou tail est inférieure à 400, si une divergence, une atteinte de la profondeur maximale ou une chaîne bloquée est détectée, ou si l'E-BFMI est inférieur à 0,30. Dans ce cas, une relance ciblée utilise 1000 itérations de chauffe, 1000 itérations conservées et `adapt_delta`=0,99. Le critère de validité n'est pas modifié après la relance.

Tableau B.3. – Diagnostics finaux des répétitions, par scénario.

Scénario	Exéc.	Val.	Rel.	Éch.	$\hat{R}_{\max}$	ESS b. _{min}	ESS t. _{min}	E-BFMI _{min}	Temps (h)
Poids constants	30	29	9	1	1,015	409,3	456,1	0,569	5,74
Horizon	30	24	15	6	1,016	428,8	401,9	0,627	5,97
Âge–horizon	30	30	19	0	1,022	449,8	435,4	0,660	6,20
Faible information	30	30	0	0	1,011	674,4	455,8	0,720	2,01

Sur les 120 répétitions exécutées, 113 satisfont l’ensemble des critères numériques après application du protocole de relance. Les sept échecs définitifs correspondent à un ajustement du scénario à poids constants et à six ajustements du scénario dépendant de l’horizon. Ils sont exclus des résultats nécessitant un ajustement hiérarchique valide.

Parmi les 113 répétitions retenues, $\hat{R}_{\max} = 1,0216$, l’ESS bulk minimale vaut 409,29, l’ESS tail minimale 401,86 et l’E-BFMI minimal 0,5687. Aucune divergence, aucune atteinte de la profondeur maximale et aucune chaîne bloquée ne sont observées. Le temps Stan cumulé, relances comprises, est de 19,92 heures.

Une difficulté distincte apparaît pour le *stacking* contextuel non régularisé dans la répétition 16 du scénario de faible information. L’optimisation atteint la limite d’itérations avec des coefficients de très grande amplitude. Ce comportement est cohérent avec une quasi-séparation dans un contexte où les distributions candidates sont volontairement proches. Cette non-convergence n’affecte ni le *stacking* global ni l’ajustement hiérarchique de la même répétition, qui restent conservés. Les statistiques du non-régularisé sont donc calculées sur 29 répétitions dans ce scénario.

B.3. Analyses de sensibilité complémentaires

Les analyses complémentaires du chapitre 3 utilisent uniquement des sorties déjà calculées. Le tableau B.4 en résume les conclusions.

Tableau B.4. – Résumé des analyses de sensibilité effectivement exécutées.

Analyse	Modification testée	Résultat principal	Classement modifié
Tirages prédictifs	1 000 contre 4 000, avec plusieurs graines	Les variations Monte-Carlo diminuent avec 4 000 tirages et ne changent pas l'avantage général du hiérarchique.	Non
Lois <i>a priori</i>	Échelles divisées par deux ou doublées par repondération d'importance	Les scores et les moyennes restent proches ; les ESS d'importance plus faibles invitent à une interprétation prudente.	Non
Terme quadratique	Retrait de q_x dans les versions contextuelles	Le LogS du hiérarchique se dégrade davantage, mais le CRPS et la MAE restent proches de la spécification principale.	Partiel selon le score
Interaction	Retrait de $\tilde{x}\tilde{h}$	Les performances restent voisines de la spécification complète.	Non
Années pandémiques	Retrait des années 2020 à 2022	Le niveau des erreurs baisse, mais le hiérarchique reste premier pour le LogS, le CRPS et la MAE.	Non
Année cible	Retrait successif d'une année de test	Le signe des différences principales reste inchangé.	Non
Incertitude sur les poids	Poids postérieurs propagés au lieu des poids moyens	Les moyennes de rente sont presque inchangées ; la dispersion augmente légèrement.	Non
Taux d'intérêt	Taux annuels de 1, 2 et 3 %	Le niveau des rentes varie comme attendu sans modifier les conclusions sur les écarts entre pondérations.	Non

