Mémoire présenté le : 12 septembre 2023

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Kué Gilles GABA

Titre : Apport des modèles cliométriques composites à la construction des tables de mortalité prospectives des pays émergents et moins avancés

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

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

Membres présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : Actuariatech

Signature :

Directeur de mémoire en entreprise :

Nom :

Signature :

Invité

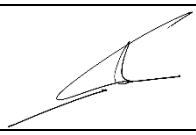
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat



Résumé

Dans ce mémoire, nous avons utilisé la cliométrie pour améliorer les projections actuarielles de mortalité, afin de mieux gérer les risques en assurance vie dans les pays en développement et émergents. Les pays développés ont un historique de taux de mortalité par âge datant de plusieurs générations, tandis que les pays émergents et en développement disposent de données moins complètes.
D'abord, nous avons développé un modèle interne à facteurs PCR-optimal ou PLS pour améliorer la modélisation de la mortalité par âge.
Ensuite, nous avons introduit une stratégie composite pour modéliser les différents âges d'une population, permettant une modélisation indépendante de la mortalité de chaque âge.
Par ailleurs, nous avons développé un modèle mixte cliométrique et composite pour améliorer les prévisions à long terme des mortalités dans les pays en rattrapage transitionnel en utilisant l'histoire quantitative des mortalités des pays transitionnellement plus âgés.
De plus, nous avons développé un ensemble de modèles cliométriques composites à référence externe pour les pays les moins avancés qui souffrent souvent d'un manque de données historiques. Nous avons ainsi créé plusieurs modèles externes cliométriques et composites adaptés de différents modèles classiques pour aider à surmonter le manque de données historiques.
Enfin, les résultats empiriques montrent que ces modèles peuvent aider à améliorer les projections de mortalité dans les pays en développement et émergents, en particulier pour les pays les moins avancés. Cette approche offre donc de nouvelles possibilités pour la modélisation de la mortalité par âge et la gestion des risques en assurance vie.

Mots-clés : cliométrie, projections actuarielles, modèles de mortalité, prévisions à long terme, table de mortalité, PCR optimal, PLS, série temporelle cliométrique, modèle composite, modèle mixte cliométrique, référence externe, pays en développement, pays les moins avancés, modèle externe cliométrique.

Abstract

In this dissertation, we used cliometrics to improve actuarial mortality projections to better manage life insurance risk in developing and emerging countries. Developed countries have generations of historical age-specific mortality rates, while emerging and developing countries have less complete data.
First, we developed an internal PCR-optimal or PLS factor model to improve age-specific mortality modeling.
Second, we introduced a composite strategy to model different ages in a population, allowing independent modeling of mortality at each age.
In addition, we developed a mixed cliometric and composite model to improve long-term predictions of mortality in transitional catch-up countries by using the quantitative mortality history of transitionally older countries.
In addition, we have developed a set of externally referenced composite cliometric models for least developed countries that often suffer from a lack of historical data. In this way, we created several external cliometric and composite models adapted from different classical models to help overcome the lack of historical data.
Finally, empirical results show that these models can help improve mortality projections in developing and emerging countries, especially for the least developed countries. This approach thus offers new possibilities for age-specific mortality modeling and risk management in life insurance.

Keywords: cliometrics, actuarial projections, mortality models, long-term forecasts, life table, optimal RCP, PLS, cliometric time series, composite model, cliometric mixed model, external benchmark, developing countries, least developed countries, external cliometric model.

Remerciements

Mes premiers remerciements s'adressent à Stéphane Loisel, pour sa direction et son encadrement technique et humain. Ses apports méthodologiques et sa rigueur scientifique sans concession sont source d'enrichissement et de progrès.

Anani Olympio a été présent tout le long de mon parcours actuariel. Il a été d'un précieux conseil et source d'inspiration. Je lui témoigne ici ma gratitude.

Je suis particulièrement reconnaissant à Caroline Hillairet et Florence Picard de m'avoir invité à présenter mes travaux lors du webinaire du 22 mai 2023, organisé par le Groupe de travail *Anticiper en univers incertain* de l'Institut des actuaires.
Je remercie les nombreux actuaires ayant participé à ce webinaire, pour leur intérêt, leurs questions et remarques fort utiles.

J'ai également une pensée pour les enseignants et les chercheurs de l'ISFA. Je n'oublie pas les membres de l'administration de l'ISFA, bien évidemment.

Mes remerciements vont également à mes collègues et partenaires de France et du Togo.

Table des matières

Introduction générale

1. Préambule

En fonction de leurs trajectoires et situations historiques (démographiques, sanitaires, éducatives, économiques), les pays du monde rencontrent des problématiques assez différentes concernant la qualité, le détail et le volume des données dont ils disposent.

Les méthodes actuarielles de construction des tables prospectives (règlementaires ou d'expérience) doivent donc permettre une opérabilité technique satisfaisante et prudente en adéquation avec les différents profils de pays et avec leurs contraintes de données.

Nous contribuons à cet objectif, en utilisant la cliométrie (histoire globale, quantitative, multidimensionnelle, de temps long) dont le champ d'application traditionnel est l'investigation quantitative de l'histoire longue, pour améliorer les projections actuarielles de mortalité.

Nous avons choisi ce sujet de mémoire car il nous semble utile d'un triple point de vue.

D'abord, ce sujet nous permet de nous confronter aux limitations et difficultés techniques qui gênent les actuaires dans les pays en développement et donc ralentissent le développement des produits de l'assurance sur la vie dans ces pays. Les difficultés techniques liées aux tables de génération impactent aussi bien les assureurs que les autorités de régulation et de contrôle dans leurs missions respectives.

Ensuite, le développement de notre cabinet d'actuariat conseil sur le marché des pays émergents et en développement, passe par la maîtrise et l'utilisation de techniques actuarielles adaptées à ces environnements spécifiques. Ce mémoire s'inscrit dans le cadre des préparations techniques préalables.

Enfin, sur le plan de la recherche scientifique, le sujet de ce mémoire permet d'approfondir et d'améliorer les méthodes déjà initiées dans le cadre de notre thèse (Gaba, 2021).

In fine, nos travaux visent à améliorer les projections actuarielles de mortalité en utilisant la cliométrie, afin de répondre aux enjeux spécifiques rencontrés par les actuaires dans les pays en développement et émergents. Nous espérons ainsi apporter des contributions significatives à l'amélioration des projections actuarielles de mortalité pour une meilleure gestion des risques en assurance sur la vie.

Dans le reste de cette introduction générale, nous commençons par décrire plus précisément nos problématiques actuarielles et à expliquer l'originalité de nos approches.
Ensuite nous exposons la structure du mémoire.
Enfin, pour faciliter la lecture du mémoire, nous proposons une courte synthèse qui rassemble et rappelle nos contributions et principaux résultats.

2. Problématiques et originalités de notre approche

Afin d'illustrer l'originalité de l'approche proposée dans ce mémoire, nous allons dans un premier temps partir de trois profils de pays (ou de populations) qui ont des problématiques de données d'expériences bien distinctes.
Selon les problématiques de données et des limites détectées sur les modèles classiques, nous allons proposer des approches originales pour tenter d'améliorer ces modèles.

2.1. Profils de pays et problématiques de données

2.1.1. Pays développés

Les pays dits développés correspondent globalement aux pays de l'OCDE (Organisation de coopération et de développement économique).
Les historiques de données démographiques (dont les taux de mortalité par âges) disponibles, sont longs de plusieurs générations (20-30 ans par génération). Il est habituel de disposer de données de plus d'un siècle, soit au moins 4 générations de 25 ans chacune.
Les instituts nationaux de statistiques produisent des données standardisées et de fiabilité « acceptable ».
Les principales bases de données internationales disponibles pour les chercheurs sont par exemple : la base HMD (Human Mortality Database) publiée par l'Université de Californie (Berkeley) et le Max Planck Institute for Demographic Research, la base World Population Prospects (WPP) publiée par l'ONU.

2.1.2. Pays émergents

Les critères du FMI pour qualifier les pays dits émergents sont : PIB par habitant, PIB, l'ouverture aux échanges et la stabilité financière. Ces pays ont un PIB par habitant inférieur à celui des pays développés, mais sont caractérisés par une croissance économique rapide, et un indice de développement humain convergeant rapidement vers ceux des pays développés (Jaffrelot *et al.*, 2008).
Parmi les pays récemment considérés comme émergents, nous pouvons citer : Afrique du Sud, Brésil, Chine, Inde, Indonésie, Malaisie, Mexique, Philippines, Thaïlande, Turquie.

Les historiques de données démographiques (dont les taux de mortalité par âges) disponibles sont variables selon les pays. Il en va de même pour la qualité des données.
Les principales bases de données internationales disponibles pour les chercheurs sont par exemple : la base HLD (Human Life-Table Database) publiée par le Max Planck Institute for Demographic Research, la base World Population Prospects (WPP) publiée par l'ONU.

Concernant la base WPP, il est important de noter qu'en dehors des pays développés, les historiques de mortalité des autres pays sont des reconstitutions issues de modélisations diverses effectuées par la Population Division de l'ONU. Ces données ne sont donc pas issues de statistiques nationales ou d'études de terrain. Par conséquent, il serait légitime d'avoir de fortes réserves quant à l'usage des données de la base WPP pour des usages de projections actuarielles prospectives. Nous avons effectué quelques tests de prévisions avec ces données et avons constaté que seuls les modèles de type Lee-Carter sont systématiquement les meilleurs. Ce résultat semble cohérent mais aussi tautologique, étant donné que

des méthodes adaptées du modèle Lee-Carter sont utilisées pour reconstituer et lisser les historiques de ces bases. Ces premiers tests nécessitent cependant des investigations complémentaires pour confirmation.
Par contre les données de la base HLD sont issues de publications officielles et de recherches de terrain : elles semblent donc plus adaptées pour les études de prospective actuarielle.

2.1.3. Autres pays en voie de développement et pays les moins avancés

Les autres pays en voie de développement et pays les moins avancés constituent le reste des pays du monde en dehors des pays émergents et des pays développés. Même si ces pays sont assez faibles en termes d'indice de développement humain, il n'en demeure pas moins qu'en dehors de quelques pays fortement sinistrés (guerres par exemple), ces pays évoluent démographiquement, éducativement et économiquement sur une pente de développement (Gaba, 2021).

Les historiques de données démographiques et de mortalité par âges de ces pays sont relativement courts, i.e. de l'ordre de quelques années à une vingtaine d'années (Kamega, 2011). La qualité de données s'améliore avec le temps (United Nation Population Division, 2022).
Dans ces pays, seule la mortalité infantile fait l'objet de statistiques officielles suivies et d'études de terrain régulières (United Nation Population Division, 2022).
Par exemple, l'Institut National de la Statistique (INS) de la Côte d'Ivoire réalise régulièrement (une fois par décennie environ) l'Enquête Démographique et de Santé, ainsi que d'autres enquêtes sanitaires de terrain. Nous avons pu ainsi reconstituer les taux de mortalités infanto-juvéniles (0-5 ans) depuis 1965.

Les principales bases de données internationales disponibles pour les chercheurs sont par exemple : la base HLD (souvent parcellaire, mais issue de compilations d'enquêtes), la base WPP (à propos de laquelle nous avions précédemment émis de fortes réserves pour un usage à des fins de prospective actuarielle, en raison des historiques des pays hors OCDE presque entièrement issus de modèles et non d'observations).

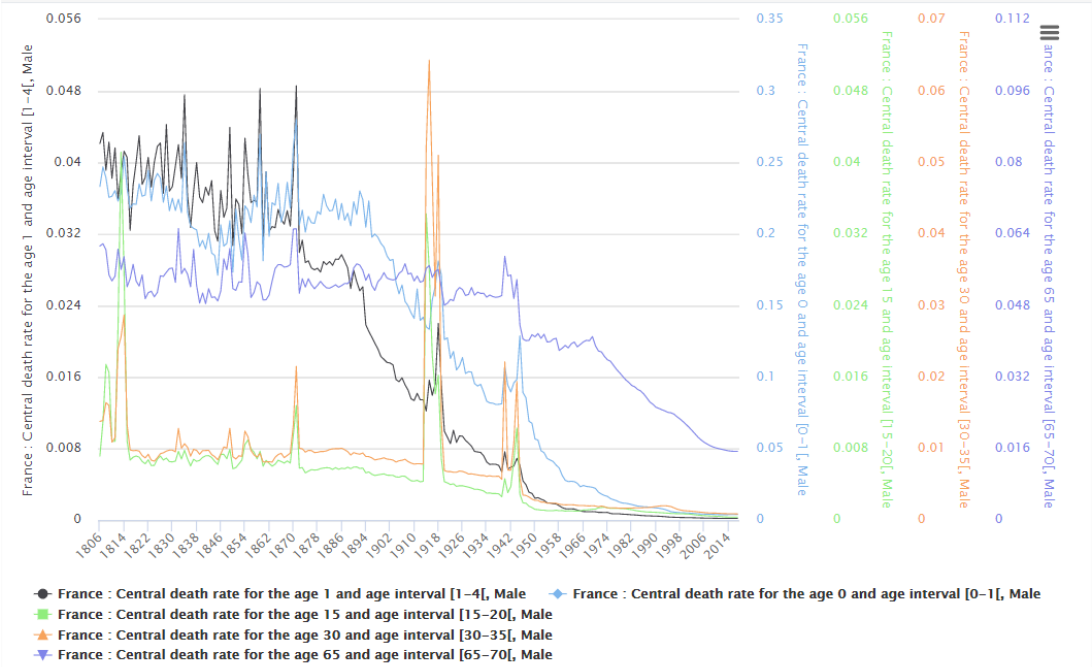
2.2. Problématiques méthodologiques communes à tous les pays

2.2.1. Décalages temporels et différences d'allure entre les courbes des taux de mortalité selon les âges

Nous remarquons empiriquement des **décalages temporels** et des **différences d'allure** dans les courbes des taux de mortalité selon les âges (cf. exemple de la France dans *Figure 1*).
Par conséquent, si une composante temporelle unique reflète la **transition démographique** du point de vue de la mortalité globale, la prise en compte de la **transition sanitaire** qui affecte les âges de manière différenciée, nécessite des composantes temporelles différenciées elles aussi par tranche d'âge. Et cela afin de mieux intégrer ces différences dans l'évolution des taux de mortalité selon les âges.

Une originalité de notre approche consiste à utiliser le constat ci-dessus pour généraliser les modèles de mortalité classiques (notamment les modèles de Lee-Carter et CBD), en sélectionnant de manière optimale (minimisation du *MSE* de validation croisée), une ou plusieurs composantes principales (régression PCR) ou PLS (régression PLS), pour chaque âge modélisé.
Nous appellerons ces extensions, des modèles internes à **facteurs PCR-optimal (dit PCR-O) ou PLS**.

Figure 1 : Décalages temporels et différences d'allure des taux mortalité selon les tranches d'âge.



2.2.2. Limites de la modélisation avec une même classe de modèles de mortalité quel que soit l'âge

Nous désignons par classe de modèles, un ensemble de modèles caractérisés par la même spécification sur les critères suivants :

- Ensemble des variables (ou séries) explicatives initiales
- Procédure de sélection des variables (ou séries) explicatives retenues
- Formulation mathématique
- Loi et méthode d'estimation

Une limitation de nombreux modèles classiques de mortalités provient du fait qu'une même classe de modèles soit appliquée à la modélisation de toutes les courbes des taux de mortalité quel que soit l'âge.

Une des originalités de ce mémoire est de modéliser de façon indépendante la mortalité de chaque âge. Cette indépendance entre les modélisations par âge se traduit par la possibilité d'attribuer différentes classes de modèles à différents âges.

Nous parlons de modèle **composite**, lorsqu'il est possible de mélanger plusieurs classes de modèles potentielles pour modéliser les différents âges d'une population. Un modèle composite est donc une super-classe de modèles, une classe de classes de modèles.

Par exemple cette indépendance de modélisation entre les âges peut se traduire par le choix différencié des facteurs PCR-O ou des facteurs PLS selon les âges modélisés.
Nous proposons alors un modèle multifactoriel **composite** : modèle **composite** à facteurs PCR-O/PLS.

Dans ce cas, la modélisation sera réalisée en deux itérations pour sélectionner dans un premier temps la classe de modèles la mieux adaptée à chaque âge, puis dans un second temps d'estimer les performances **sans biais** des modèles sélectionnés à l'itération précédente.
Les deux itérations se présentent comme suit :

- Itération 1 : Pour chaque âge, choix **a posteriori** de sa meilleure classe de modèle (Best Model Class)
    - Historique : t_0 ; t_1
    - Prévisions : $t_1 + 1$; t_2
- Itération 2 : Pour chaque âge, utilisation **a priori** de sa Best Model Class issue de l'itération 1 pour en estimer la performance sans biais
    - Historique : t_0 ; t_2
    - Prévisions : $t_2 + 1$; t_3

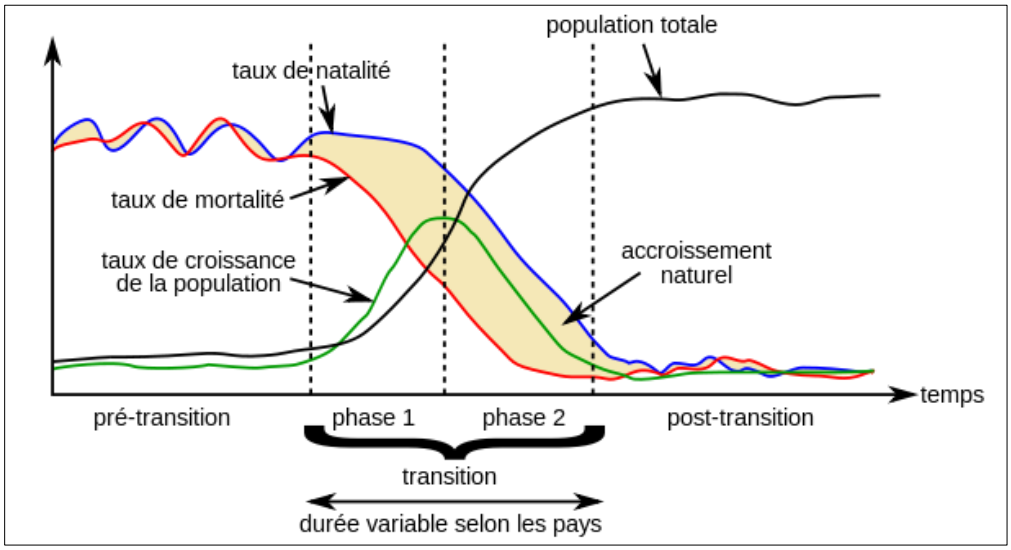
Les temps t_0, t_1, t_2, t_3 étant des bornes de découpages arbitraires en 4 intervalles temporels de l'historique utilisé. Le temps est considéré discret.

2.3. Problématiques spécifiques aux pays émergents et pays en voie de développement

2.3.1. Préambule : universalité de la transition démographique

La transition démographique moderne est le processus historique par lequel une population passe d'un régime démographique relativement stable caractérisé par un taux de mortalité global et un taux de natalité élevés, à un nouveau régime également relativement stable caractérisé par des valeurs nettement plus faibles (division par 3, voire 4) de ces deux taux.

Figure 2 : Schéma général de la transition démographique

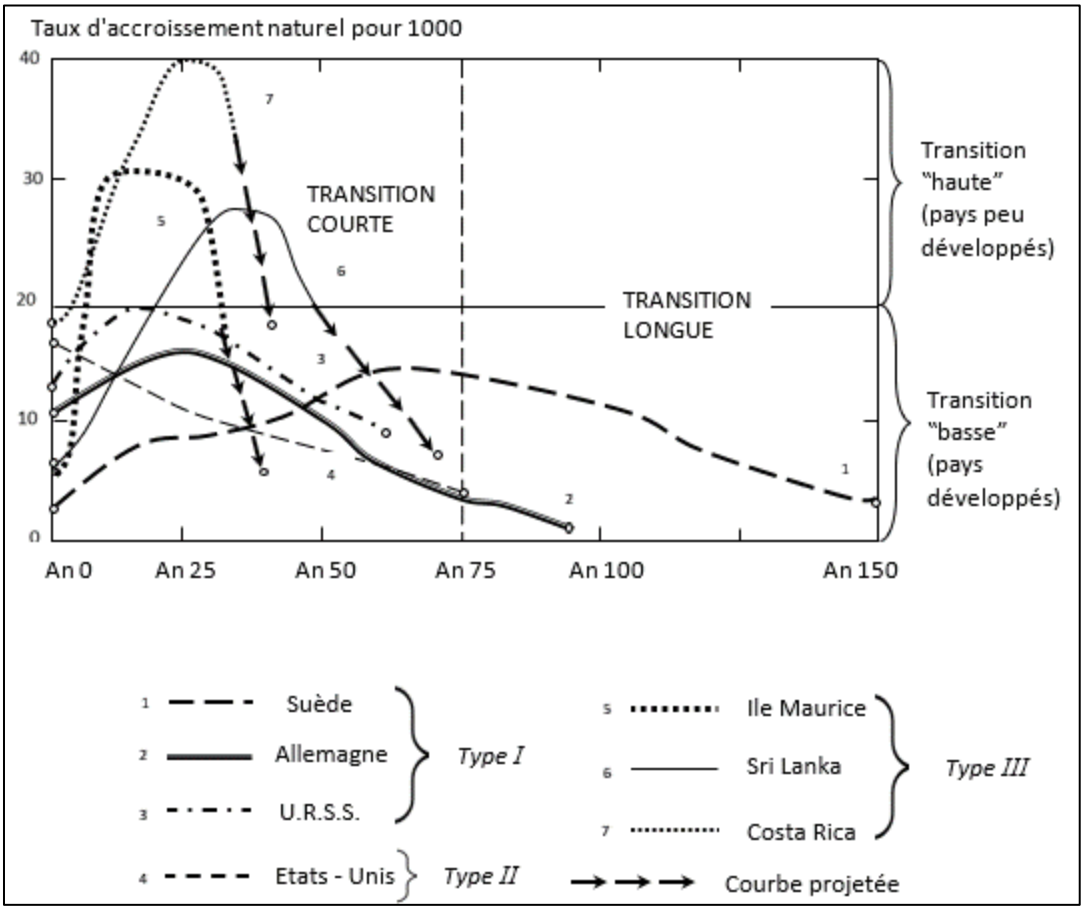


Source : Avdeev A., 2014.

Le constat historique est celui de l'universalité du processus de transition démographique, se déroulant par vagues (ou cohortes) de pays, avec des rythmes des changements transitionnels qui sont assez

variables d'une cohorte de pays à l'autre. En effet, les pays récemment entrés en transition ont généralement des rythmes de changement plus rapides.
Certains pays sont donc plus « *précoces* » dans la transition démographique que d'autres qui sont alors considérés comme étant en situation de « *rattrapage* » relatif : par conséquent, les pays transitionnellement précoces sont transitionnellement « plus âgés », et les pays en rattrapage transitionnel sont transitionnellement « moins âgés ».

Figure 3 : Profils principaux de transition démographique



Source : Chesnais (1986), p. 270

2.3.2. Pays ayant un historique suffisant pour les modèles internes de prospective

Une originalité de notre approche consiste à utiliser l'histoire quantitative des mortalités des pays transitionnellement « plus âgés », pour tenter d'améliorer les prévisions à long terme des mortalités dans les pays transitionnellement « moins âgés » (cf. notion d'âge transitionnel défini à la section 10.1).
En pratique, ces améliorations pourraient s'appliquer aux études prospectives de la mortalité dans les pays émergents ou en développement.

Nous définissons le *temps transitionnel* (cf. section 10.1) comme étant une série temporelle qui est *fortement* concordante avec le temps. Autrement dit, un *temps transitionnel* est une série temporelle qui a une dépendance monotone *forte* avec le temps. Nous notons également que toute transformée

symétrique d'une série temporelle discordante au temps, est concordante au temps, donc peut jouer le rôle de temps transitionnel. (Cf. section 10.1.3)
De même, nous définissons l'âge transitionnel d'un pays, comme étant la différence entre le temps transitionnel et sa valeur minimale (quel que soit le pays) dans la période de transition. Cette valeur minimale est un seuil de passage, quel que soit le pays.

Dans ce contexte de transition démographique par cohortes de pays, avec des rythmes différents de changement, notre démarche consiste à utiliser comme temps transitionnel, la symétrique (par rapport à l'axe horizontal du temps) du taux de mortalité infantile (éventuellement lissé pour accentuer la monotonie et la linéarité).
Les différences des rythmes du processus de transition démographique entre les pays de cohortes distinctes, ne permettent pas de faire des prévisions en utilisant des modèles multivariés classiques de séries temporelles avec variable exogène retardé, car le retard n'est pas stable dans le temps.
Le temps transitionnel permet de faire un appariement temporel entre le pays à modéliser « moins âgé » et le pays transitionnellement « plus âgé ».

La série temporelle *cliométrique de mortalité à l'âge* x du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) est une série créée à partir des données historiques du pays A (transitionnellement « plus âgé » que le pays R), pour jouer un rôle de série exogène dans les modèles multivariés de prévisions de mortalité du pays R.
Cette série cliométrique exogène de mortalité sera construite de sorte à avoir la même allure temporelle que dans les pays « transitionnellement plus âgés », mais adaptée au rythme du pays à modéliser.

La série cliométrique sera ainsi utilisée comme série explicative supplémentaire dans les modèles internes prospectifs de mortalité du pays R : nous avons donc un modèle **mixte** i.e. à la fois interne et externe (via l'utilisation de l'historique du pays A qui joue le rôle de référence externe).
Nous nommons les modèles utilisant la série exogène cliométrique, modèles cliométriques.

En capitalisant sur les apports originaux des sections précédentes, nous proposons le type de modèle suivant : modèle **mixte cliométrique et composite** à facteurs **PCR-optimal/PLS.**

Comme précédemment évoqué, le caractère composite de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles diversifiées (par exemple via le choix du ou des pays A, facteurs PCR-O ou PLS).
Ces modèles seront développés dans la suite de ce mémoire, puis testés empiriquement.

2.3.3. Pays ayant un historique insuffisant pour les modèles internes de prospective

Sont notamment concernés ici les pays les moins avancés (dont l'Afrique sub-saharienne) et les pays en développement hors émergents.
A ce titre nous nous sommes particulièrement intéressés au cas de l'Afrique sub-saharienne francophone rassemblées dans la zone CIMA (Conférence interafricaine des marchés d'assurance).

La CIMA, sise à Libreville (Gabon), est un organisme communautaire du secteur des assurances créé en juillet 1992 par un traité regroupant 14 pays d'Afrique : Bénin, Burkina Faso, Cameroun, Centrafrique, Congo, Côte d'Ivoire, Gabon, Guinée, Guinée équatoriale, Mali, Niger, Sénégal, Tchad et Togo.
La Guinée-Bissau a ratifié le traité en 2002 et est le premier État non-francophone à adhérer à la CIMA.

Ce traité a adopté un code des assurances unique pour les 14 pays, englobant les règles du contrat d'assurance et les règles applicables aux acteurs : assureurs, réassureurs, intermédiaires d'assurance.
Le traité CIMA institue aussi un contrôle régional des assureurs et réassureurs des pays membres, confié à des commissaires-contrôleurs recrutés sur concours commun régional (sans quota de nationalité). Les contrôles nationaux (Directions nationales des assurances) assistent le contrôle régional. L'organe décisionnel (agréments, mesures de redressement, sanctions par exemple), la Commission régionale de contrôle des assurances, comprend 11 membres : 6 représentent les pays par rotation, les autres sont des personnalités qualifiées.

La CIMA et l'ACPR ont signé un accord d'échange d'informations, et les deux institutions coopèrent régulièrement (ACPR, 2023).

Une littérature actuarielle francophone de qualité est disponible sur le contexte et les problématiques de l'assurance vie dans la zone CIMA, dont : Kamega (2011), Kamega et Planchet (2011), Yebouet et al. (2014), Kamega et Pieby (2014).
Ce à quoi s'ajoutent le code des assurances de la zone CIMA (CIMA, 2023) et de nombreuses informations institutionnelles en ligne de la CIMA elle-même ou de partenaires institutionnels comme l'ACPR (ACPR, 2023).

Le taux de pénétration d'une branche d'assurance est le ratio entre le montant des primes émises et le PIB.
Selon les données les plus récentes disponibles (Atlas magazine, 2023), le taux de pénétration de l'assurance vie dans la zone CIMA est de 0.36% en 2020, soit une augmentation de 80% en 10 ans (taux de pénétration de 0.2% en 2011).
En comparaison, le taux de pénétration de l'assurance vie en Afrique est d'environ 2.6%, tandis que la moyenne mondiale est d'environ 7.4% en 2020 (Atlas magazine, 2023).
Il convient de noter que ces chiffres varient considérablement d'un pays à l'autre en Afrique, avec des taux de pénétration de l'assurance vie plus élevés dans certains pays tels que l'Afrique du Sud (13.7%), le Maroc (4.5%) et la Tunisie (2.3%). Les taux sont plus faibles dans d'autres pays, avec 0.8% en Algérie, 0.7% en Egypte et jusqu'à 0.3% au Nigéria.

Cependant, il existe des opportunités de croissance pour les assureurs dans la région, notamment en développant des produits d'assurance adaptés aux besoins et aux capacités financières des populations locales. A ce titre nous pouvons citer le statut d'entreprise de micro-assurance institué par la CIMA, avec des contraintes règlementaires plus légères en matière de capital social par exemple (CIMA, 2023).

Nous nous sommes particulièrement intéressés aux problématiques actuarielles actuelles des professionnels de l'assurance vie dans la zone CIMA.
A ce titre, en plus de la littérature disponible, nous avons complété nos informations par une enquête qualitative de terrain à Lomé (Togo) et à Abidjan (Côte d'Ivoire).
A Lomé (mars 2023), nous avons pu avoir des entretiens approfondis avec le président directeur général d'une société de micro-assurance vie, ainsi qu'avec l'un des fondateurs et actionnaire majeur de la même entreprise en pleine croissance.
A Abidjan (du 22 au 24 mars 2023), nous avons eu de nombreux échanges (lors d'entretiens et d'une conférence que nous avons donnée sur le thème de notre mémoire) avec le directeur technique d'une compagnie d'assurance majeure africaine et aussi avec plusieurs membres d'une brigade de contrôle de la CIMA.

La taille réduite de l'échantillon de professionnels interrogé sur le terrain, ainsi que le devoir de réserve professionnelle qui leur incombe, nous contraint à la discrétion sur leur identité.

Les problématiques actuarielles des professionnels de l'assurance vie interrogés dans la zone CIMA peuvent se résumer aux principaux points suivants :

1. **De nombreux pays de la zone intègrent désormais et progressivement des professionnels du secteur informel dans les couvertures d'assurance santé ou vie**. C'est par exemple le cas au Togo où l'INAM (Institut national d'assurance maladie) qui couvre traditionnellement les fonctionnaires, a récemment étendu son périmètre aux couturières du secteur informel.
   Dans la plupart des pays de la zone, des systèmes de CMU (couverture maladie universelle) sont promulgués et sont en phase de développement, avec le concours financier d'organismes internationaux dont le FMI et la Banque mondiale.
   Le problème qui se pose aux actuaires est celui de la **modification socio-économique de leur portefeuille, ainsi qu'un manque d'historique d'expérience sur les professionnels du secteur informel**.

2. **La zone CIMA ne dispose pas de tables de génération règlementaires pour la tarification des rentes viagères**. Les compagnies d'assurance et mutuelles qui commercialisent des produits de rente viagère en sont réduites depuis des décennies à tarifier leurs polices avec des tables règlementaires du moment.
   Problématique des actuaires de la zone CIMA : face à la baisse soutenue et de long-terme des taux de mortalité en Afrique subsaharienne (transition démographique, transition sanitaire), **un véritable risque de longévité pèse sur les assureurs, du fait de l'allongement continu de l'espérance de vie dans la zone**.
   A cet égard, la prise en compte de l'évolution des taux de mortalité par âge, est devenue un impératif dont tous les acteurs de la filière (autorités comme assureurs) sont éminemment conscients.
   Il faut également noter que les assureurs majeurs de la zone proposent des produits de retraite complémentaire (on peut le confirmer sur leurs sites officiels).

3. **Les tables règlementaires du moment actuelles sont « vieillissantes »** car elles ont été établies à partir de données de 2003-2006 et promulguées en 2012 (Kamega et Pieby, 2014). **Du fait de la baisse constante des taux de mortalité comme conséquence de la transition démographique dans la zone**, le taux d'obsolescence des tables règlementaires du moment est beaucoup plus important en Afrique subsaharienne que dans les pays ayant achevé leur transition démographique (comme les pays européens).
   Pour l'actuaire de la zone CIMA conscient forcément de la réalité des évolutions démographiques « rapides » en cours, il y a une sorte de frustration du fait de la « lenteur » et de la prudence parfois « jugée » excessive des autorités de contrôle.
   S'y ajoutent de manière plus générale, l'impact de ces problématiques sur la nécessité de développement commercial pour les assureurs dans un contexte de hausse démographique et de progression de l'indice de développement humain (notamment dans ses aspects sanitaires et éducatifs, et dans une moindre mesure du point de vue du pouvoir d'achat).

4. **Les tables règlementaires du moment actuelles souffrent également de quelques limites techniques de construction** (Kamega [2011], Kamega et Planchet [2011], Kamega et Pieby [2014]).
   Notamment par rapport à la **représentativité de l'échantillon d'assureurs utilisés, car de**

nombreux pays (parfois importants) ne sont pas inclus dans l'échantillon : Sénégal, Bénin, Gabon et Burkina Faso par exemple.
Cela est d'autant plus épineux qu'il existe des disparités non-négligeables de mortalités et donc d'espérance de vie (mais aussi en termes éducatifs et économiques) entre les pays de la zone CIMA.
Cela rend d'autant plus urgent pour les actuaires de la zone, la mise à jour des tables règlementaires du moment.

5. Enfin, les acteurs interrogés ne semblent pas exprimer la nécessité ou l'urgence d'introduire dans la zone CIMA, une réglementation prudentielle et quantitative de solvabilité similaire à la norme Solvabilité 2 européenne par exemple.
   Les urgences se concentrent sur le développement commercial et technique pour les assureurs et la détection des fraudes aux normes actuelles pour les contrôleurs.

Du point de vue de la construction des tables de génération dans les pays les moins avancés et notamment en Afrique subsaharienne, Kamega (2011) avait proposé une adaptation du modèle de Bongaarts. Toutefois la procédure semble nécessiter un jugement d'expert, ce qui ajoute un risque supplémentaire.

Dans la continuité du modèle mixte original que nous avions introduit ci-dessus pour les pays émergents, nous en proposons une **variante cliométrique à référence externe pour les pays les moins avancés** souffrant comme souvent du manque de profondeur historique pour la mise en œuvre des modèles internes de prospective.

Comme évoqué précédemment à propos de ces pays, seule la mortalité infantile fait l'objet de statistiques officielles suivies et d'études de terrain régulières (United Nation Population Division, 2022).
Nous allons donc profiter de ces données relativement bien renseignées pour réaliser un appariement entre les temps calendaires du pays d'expérience R et ceux du pays de référence A (voir les définitions dans la section précédente), via le temps transitionnel qui leur est commun.
Ces techniques d'appariement seront développées dans la suite de ce mémoire, mais nécessitent d'un point de vue statistique, des degrés de liberté suffisants pour l'estimation de leurs paramètres.

Ainsi l'appariement entre le pays d'expérience R et le pays de référence A (voir les définitions dans la section précédente), permet :

1. de disposer de couples appariés de temps calendaires du pays d'expérience R et ceux du pays de référence A, via le temps transitionnel qui leur est commun

2. de créer plusieurs séries cliométriques pour chaque âge du pays R (par rapport à un ou plusieurs pays A, avec divers paramétrages de sexe ou d'âge). Ces séries cliométriques qui ont par construction des valeurs dans le passé et dans le futur, sont des séries explicatives potentielles dans les modèles externes cliométriques pour le pays d'expérience R

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le premier résultat ci-dessus, nous proposons les 3 classes de modèles externes cliométriques suivants, adaptés des modèles classiques externes :

1. Modèle **externe cliométrique** adapté du modèle externe de BRASS
2. Modèle **externe cliométrique** adapté du modèle externe de COX
3. Modèle **externe cliométrique** adapté du modèle externe de TGH05-TGF05

Les trois modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** mélangeant les adaptations de BRASS/COX/TGH05-TGF05.

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le deuxième résultat ci-dessus, nous proposons les 2 classes de modèles externes cliométriques suivants, adaptés des modèles à facteurs PCR-Optimal et des modèles à facteurs PLS :

1. Modèle **externe cliométrique** à facteurs PCR-Optimal (construits à partir des séries cliométriques)
2. Modèle **externe cliométrique** à facteurs PLS (construits à partir des séries cliométriques)

De même, les deux modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** à facteurs PCR-O/PLS.

Comme précédemment évoqué, le caractère **composite** de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles différentes.

Il nous reste donc à explorer la littérature scientifique et à développer, formuler et évaluer les modèles proposés pour répondre aux problématiques posées.

3. Organisation du mémoire

Ce mémoire est organisé en trois grandes parties.
La première partie est consacrée à la définition et à la présentation des concepts théoriques et actuariels qui nous seront nécessaires dans le reste de nos travaux.
La deuxième partie détaille nos apports méthodologiques pour la construction de nouveaux modèles en réponse à nos problématiques.
La troisième partie est consacrée à la mise en application de certains de nos modèles avec les tests de sensibilités associés à leurs performances.

La première partie, intitulée "Cadre théorique et actuariel", est composée des chapitres 5 et 6. Dans le chapitre 5, nous abordons la transition démographique et la cliométrie, deux concepts théoriques importants pour comprendre les tendances démographiques et les évolutions de la mortalité. Nous présentons les différentes phases de la transition démographique et les outils de la cliométrie utilisés pour étudier les données historiques.

Dans le chapitre 6, nous nous concentrons sur les modèles de mortalité et la construction de tables de mortalité. Nous exposons les principaux modèles démographiques et les techniques de modélisation de la mortalité. Nous examinons également les différentes approches pour construire des tables de mortalité et analyser leur qualité.

La deuxième partie de notre mémoire, "Apports méthodologiques", est composée des chapitres 7 à 11. Dans le chapitre 7, nous étudions la relation entre le modèle Lee-Carter et la régression PCR. Nous présentons le modèle de Lee-Carter et les différentes étapes de son estimation. Nous montrons comment le modèle de Lee-Carter peut être vu comme une régression sur la première composante principale.

Dans le chapitre 8, nous développons le modèle interne PCR optimal. Nous détaillons la technique de la régression sur composantes principales et présentons le modèle Logit-PCR, ainsi que ses améliorations. Dans le chapitre 9, nous explorons le modèle interne PLS. Nous détaillons la technique de la régression PLS et présentons les améliorations du modèle Logit-PCR en utilisant la régression PLS.

Dans le chapitre 10, nous examinons les modèles mixtes cliométriques PCR-Optimal et PLS. Nous introduisons le concept de série temporelle cliométrique et la méthodologie de création d'une série explicative cliométrique. Nous présentons le modèle mixte cliométrique Logit PCR-optimal, le modèle mixte cliométrique Logit-PLS et le modèle mixte cliométrique composite à facteurs PCR-O/PLS. Dans le chapitre 11, nous nous concentrons sur les modèles relationnels (à référence externe) cliométriques. Nous exposons les différentes hypothèses et notations pour les modèles externes à appariement temporel cliométrique et présentons plusieurs modèles externes cliométriques adaptés du modèle externe de BRASS, du modèle externe de COX et du modèle externe de TGH05-TGF05. Nous exposons également les modèles externes cliométriques et composites à facteurs PCR-optimal/PLS.

La troisième partie de notre mémoire, "Résultats empiriques et discussions", est composée des chapitres 12 à 14.
Le chapitre 12 est consacré au choix des modèles classiques internes de référence pour nos travaux empiriques. Nous avons identifié les problèmes d'erreurs de prévision du modèle CBD et les problèmes de

convergence des modèles GAPC avec effet de cohorte. Nous avons finalement choisi les modèles internes classiques de référence pour nos travaux empiriques.

Les chapitres 13 et 14 portent sur la modélisation des mortalités pour les tables prospectives de l'Inde et de l'Equateur respectivement. Dans chaque chapitre, nous présentons les paramétrages associés au pays « R » transitionnellement « moins âgé » (ou pays d'expérience) que nous avons choisi de modéliser. Nous présentons également le meilleur modèle que nous avons obtenu pour chaque pays ainsi que les tests de sensibilité que nous avons effectués pour évaluer la robustesse de notre modèle.

En somme, ce mémoire présente une approche théorique et pratique que nous espérons novatrice pour modéliser la mortalité en utilisant une combinaison de techniques cliométriques et statistiques avancées. Nous avons présenté différents modèles internes, externes, mixtes cliométriques et composites, et nous les avons appliqués à des données empiriques de deux pays. Nous espérons que ce travail contribuera à améliorer la qualité des tables de mortalité et, par conséquent, à mieux comprendre les tendances démographiques et les implications économiques qui en découlent.

4. Contributions et résultats principaux

Dans ce mémoire nous avons tenté d'atteindre les objectifs théoriques et empiriques que nous-nous sommes fixés concernant l'amélioration des prévisions actuarielles de long-terme des taux de mortalité par âge, dans un contexte de transition démographique par cohortes de pays à l'échelle mondiale.

Nous sommes dans un premier temps, partis de trois profils de pays (ou de populations) qui ont des problématiques de données d'expériences bien distinctes.
Selon les problématiques de données et des limites détectées sur les modèles classiques, nous avons développé diverses méthodes pour tenter d'améliorer ces modèles.

4.1. Résultats d'exploration des données d'historique des pays selon leurs profils

Les pays dits développés correspondent globalement aux pays de l'OCDE.
Les historiques des taux de mortalité par âges sont disponibles et longs de plusieurs générations (20-30 ans par génération). Il est habituel de disposer de données de plus d'un siècle, soit au moins 4 générations de 25 ans chacune.

Les critères du FMI pour qualifier les pays dits émergents sont : PIB par habitant, PIB, l'ouverture aux échanges et la stabilité financière.
La disponibilité des historiques des taux de mortalité par âges est variable selon les pays. Il en va de même pour la qualité des données.

Les autres pays en voie de développement et pays les moins avancés constituent le reste des pays du monde en dehors des pays émergents et des pays développés. Même si ces pays sont assez faibles en termes d'indice de développement humain, il n'en demeure pas moins qu'en dehors de quelques pays fortement sinistrés, ces pays évoluent démographiquement, éducativement et économiquement sur une pente de développement.
Les historiques des mortalités par âges de ces pays sont relativement courts, i.e. de l'ordre de quelques années à une vingtaine d'années.
Fait notable dans ces pays, seule la mortalité infantile fait l'objet de statistiques officielles suivies et d'études de terrain régulières.

4.2. Développement du modèle interne à facteurs PCR-optimal ou PLS

Nous remarquons empiriquement des **décalages temporels** et des **différences d'allure** dans les courbes des taux de mortalité selon les âges : par conséquent si une composante temporelle unique reflète la transition démographique du point de vue de la mortalité globale, la prise en compte de la **transition sanitaire** qui affecte les âges de manière différenciée, nécessite des composantes temporelles différenciées elles aussi par tranche d'âge.

Notre approche a consisté à utiliser le constat ci-dessus pour généraliser les modèles de mortalité classiques (notamment les modèles de Lee-Carter et CBD), en sélectionnant de manière optimale, une ou plusieurs composantes principales (régression PCR) ou PLS (régression PLS), pour chaque âge modélisé.
Nous avons appelé ces extensions, des modèles internes à **facteurs PCR-optimal (dit PCR-O) ou PLS**.

4.3. Introduction de la stratégie composite dans la modélisation des taux de mortalité

Une limitation de nombreux modèles classiques de mortalités provient du fait qu'une même classe de modèles soit appliquée à la modélisation de toutes les courbes des taux de mortalité quel que soit l'âge.

Un des résultats de ce mémoire a été de modéliser de façon indépendante la mortalité de chaque âge. Cette indépendance entre les modélisations par âge se traduit par la possibilité d'attribuer différentes classes de modèles à différents âges.

Nous parlons de modèle **composite**, lorsqu'il est possible de mélanger plusieurs classes de modèles potentielles pour modéliser les différents âges d'une population. Un modèle composite est donc une super-classe de modèles, une classe de classes de modèles.
Dans ce cas, la modélisation est réalisée en deux itérations pour sélectionner dans un premier temps la classe de modèles la mieux adaptée à chaque âge, puis dans un second temps d'estimer les performances **sans biais** des modèles sélectionnés à l'itération précédente.

4.4. Développement du modèle mixte cliométrique et composite à facteurs PCR-optimal/PLS

Un autre constat historique est celui de l'universalité du processus de transition démographique, se déroulant par vagues ou cohortes de pays, avec des rythmes des changements transitionnels qui sont assez variables d'une cohorte de pays à l'autre. En effet, les pays récemment entrés en transition ont généralement des rythmes de changement plus rapides.

Certains pays sont donc plus « *précoces* » dans la transition démographique que d'autres qui sont alors considérés comme étant en situation de « *rattrapage* » relatif : par conséquent, les pays transitionnellement précoces sont transitionnellement « plus âgés », et les pays en rattrapage transitionnel sont transitionnellement « moins âgés ».

Notre méthode a consisté à utiliser l'histoire quantitative des mortalités des pays transitionnellement « plus âgés », pour tenter d'améliorer les prévisions à long terme des mortalités dans les pays transitionnellement « moins âgés ».
En pratique, ces améliorations s'appliquent aux études prospectives de la mortalité dans les pays émergents ou en développement.

La série temporelle *cliométrique de mortalité à l'âge* x du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) est une série créée à partir des données historiques du pays A (transitionnellement « plus âgé » que le pays R), pour jouer un rôle de série exogène dans les modèles multivariés de prévisions de mortalité du pays R.
Cette série cliométrique exogène de mortalité sera construite de sorte à avoir la même allure temporelle que dans les pays « transitionnellement plus âgés », mais adaptée au rythme du pays à modéliser.

La série cliométrique est ainsi utilisée comme série explicative supplémentaire dans les modèles internes prospectifs de mortalité du pays R : nous avons donc un modèle **mixte** i.e. à la fois interne et externe (via l'utilisation de l'historique du pays A qui joue le rôle de référence externe).
Nous nommons les modèles utilisant la série exogène cliométrique, modèles cliométriques.

En capitalisant sur les apports originaux des sections précédentes, nous avons développé et formulé le modèle **mixte cliométrique et composite** à facteurs **PCR-optimal/PLS.**

4.5. Développement des modèles à références externes cliométriques et composites

Dans la continuité du modèle mixte original que nous avions introduit ci-dessus pour les pays émergents, nous en proposons une variante cliométrique à référence externe pour les pays les moins avancés souffrant comme souvent du manque de profondeur historique pour la mise en œuvre des modèles internes de prospective.
Comme évoqué précédemment à propos de ces pays, seule la mortalité infantile fait l'objet de statistiques officielles suivies et d'études de terrain régulières. Nous profitons donc de l'existence de ces données renseignées pour réaliser un appariement entre les temps calendaires du pays d'expérience R et ceux du pays de référence A, via le temps transitionnel qui leur est commun.
Ainsi l'appariement entre le pays d'expérience R et le pays de référence A, permet : d'une part de disposer de couples appariés de temps calendaires du pays d'expérience R et ceux du pays de référence A, via le temps transitionnel qui leur est commun, et d'autre part, de créer plusieurs séries cliométriques pour chaque âge du pays R (par rapport à un ou plusieurs pays A, avec divers paramétrages de sexe ou d'âge). Ces séries cliométriques qui ont par construction des valeurs dans le passé et dans le futur, sont des séries explicatives potentielles dans les modèles externes cliométriques pour le pays d'expérience R.

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le premier résultat ci-dessus, nous avons développé les trois classes de modèles externes cliométriques suivants, adaptés des modèles classiques externes :

- Modèle externe cliométrique adapté du modèle externe de BRASS
- Modèle externe cliométrique adapté du modèle externe de COX
- Modèle externe cliométrique adapté du modèle externe de TGH05-TGF05

Les trois modèles ci-dessus ont été combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** mélangeant les adaptations de BRASS/COX/TGH05-TGF05

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le deuxième résultat ci-dessus, nous avons développé les deux classes de modèles externes cliométriques suivants, adaptés des modèles à facteurs PCR-Optimal et des modèles à facteurs PLS :

- Modèle **externe cliométrique** à facteurs PCR-Optimal (construits à partir des séries cliométriques)
- Modèle **externe cliométrique** à facteurs PLS (construits à partir des séries cliométriques)

De même, les deux modèles ci-dessus ont été combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** à facteurs PCR-O/PLS.

Comme précédemment évoqué, le caractère **composite** de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles différentes.

4.6. Résultats des travaux empiriques

Le modèle **mixte cliométrique et composite** à facteurs **PCR-optimal/PLS** est particulièrement adapté aux pays émergents ayant de l'historique suffisant pour les modèles internes.

Nous avons donc testé ce modèle pour l'Inde et l'Equateur, dans l'objectif de réaliser des prévisions sur un horizon de 25 ans.
Ce modèle a été systématiquement comparé avec les modèles classiques de Lee-Carter, Lee-Carter log-Poisson et CBD log-Poisson.

Pour l'Inde : nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié.
De plus nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Sauf pour les tranches 1-5, 20-25 ans où il y a un très léger écart défavorable au modèle mixte cliométrique composite.

Nous sommes dans le contexte d'un meilleur modèle qui présente un très léger effet de surapprentissage sur le long terme. Le dimensionnement du modèle est donc à la lisière du surapprentissage.
Au vu des deux scénarios alternatifs étudiés *dans les tests de sensibilité*, l'inflation du nombre de choix paramétriques, crée un risque de surapprentissage. Ce phénomène de surdimensionnement dégrade plus fortement les performances par horizon de prévision, que les performances par âge.
L'impact sur les performances par horizon est si dégradé que la hiérarchie des modèles en est vite inversée en défaveur du modèle mixte cliométrique composite.
Un dosage délicat et une série de tests sur le nombre et les modalités paramétriques à explorer, doivent donc être effectués pour tirer le meilleur parti du modèle mixte cliométrique composite.
C'est ce que nous avons effectué en privilégiant les pays A de référence, à transition plus récente en Europe, notamment l'Italie et l'Espagne.

Pour l'Equateur : nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié. Cela est en cohérence forte avec le schéma de la première itération. Ce qui exclut a priori un phénomène de surapprentissage malgré un fort dimensionnement paramétrique. De plus, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Les avantages se sont même largement renforcés.
La hiérarchie de la première itération est conservée et renforcée.
Face à un meilleur modèle performant et riche en options de paramétrages, mais ne présentant aucun signe de surapprentissage, la réduction des choix de paramétrage entraine un risque de réduction des performances. C'est la principale leçon des deux tests de sensibilité de l'Equateur.

Au final en combinant les tests de sensibilité de l'Inde et de l'Equateur, nous concluons que le seuil de surapprentissage est variable selon le pays et ne peut être déterminé qu'après exploration d'un certain nombre de scénarios de dimensionnement paramétrique. Ce seuil de dimensionnement semble plus dépendre de la structure interne des données que de leur volume.

Partie 1 : Cadre théorique et actuariel

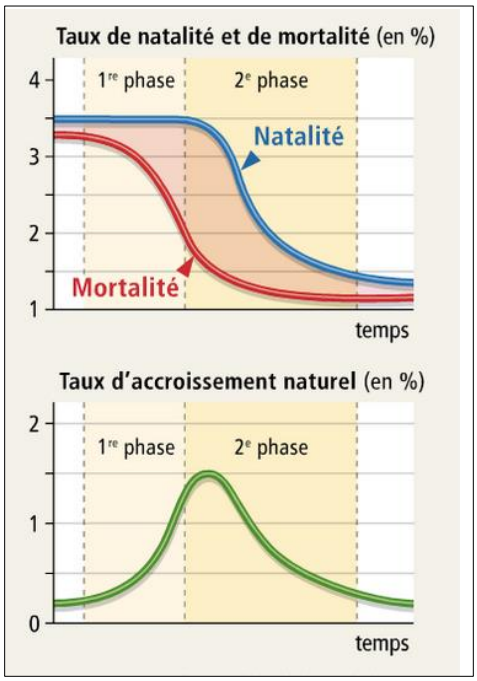
5. Transition démographique et cliométrie

5.1. La transition démographique

5.1.1. Théorie de la transition démographique

Le concept de transition démographique désigne le passage d'un régime démographique relativement stable, où la fécondité et la mortalité sont « élevées » vers un autre régime également relativement stable, où la fécondité et la mortalité sont significativement plus faibles que lors de la phase initiale, en passant d'abord par une chute de la mortalité, suivie quelques décennies plus tard par celle de la fécondité (Mazerolle, 2005).

Figure 4 : Schéma transitionnel, phases et prospective



Source : Durand et al. (2008).

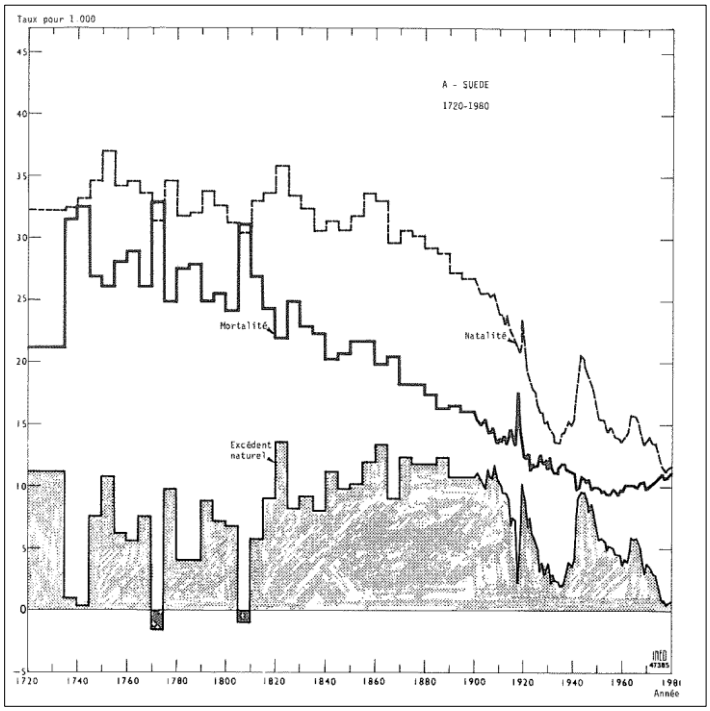
L'expression « transition démographique » a été introduite par le démographe américain Frank Notestein (1945).
Toutefois, Warren Thompson (1929), un autre démographe américain, avait très clairement identifié les deux principales phases de la transition démographique.
Le démographe Français, Adolphe Landry (1934), analysa également le phénomène de changement de régime démographique, dans son ouvrage « La révolution démographique ».

5.1.2. Illustrations historiques

Nous illustrons ci-dessous à travers quelques exemples historiques emblématiques, la généralité des deux phases de cette transition démographique.

Exemples historiques : La Suède et l'Italie

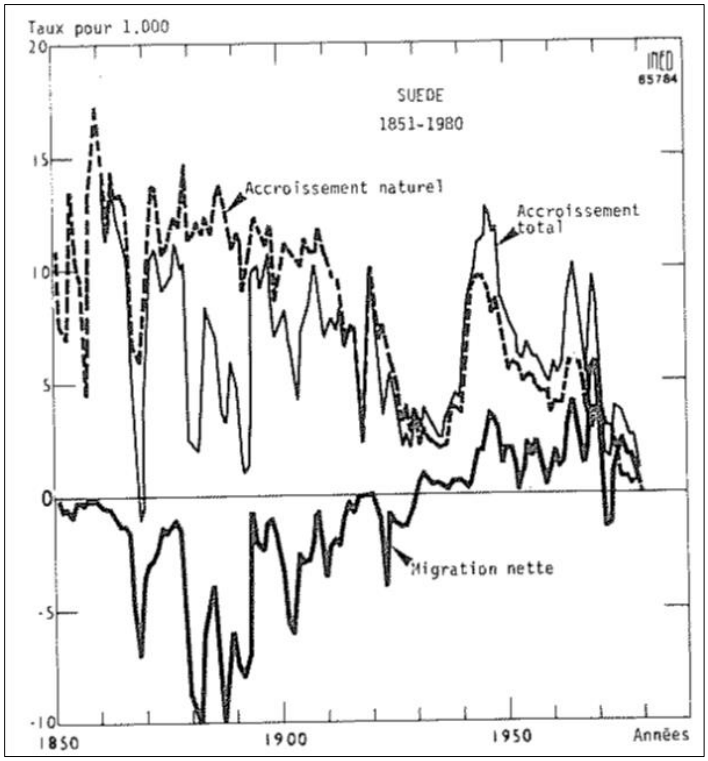
Figure 5 : Mouvement naturel de la population, Suède 1720-1980



Source: Chesnais (1986) «The migratory transition», p. 224.

L'exemple suédois, montre : une phase initiale très fluctuante, avec des niveaux élevés de mortalité et de natalité, puis une amorce de chute de la mortalité dès 1820-1830, suivie quelques décennies plus tard (dès 1850-1860) par une franche chute de la natalité. La phase de stabilisation finale n'est pas encore clairement discernable sur la figure ci-dessus.

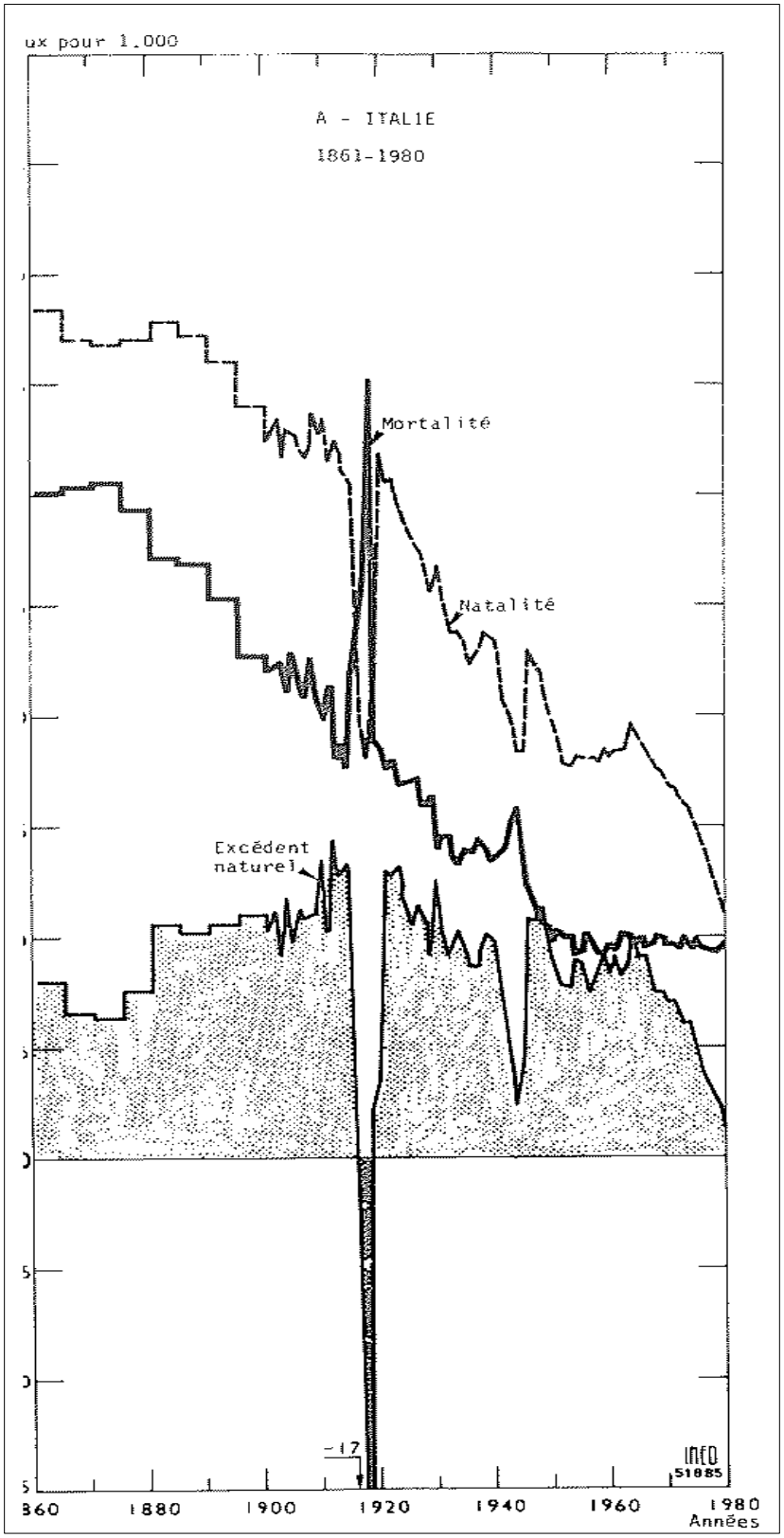
Figure 6 : Migration nette et accroissement naturel, Suède 1851-1980



Source: Chesnais (1986) «The migratory transition», p. 179.

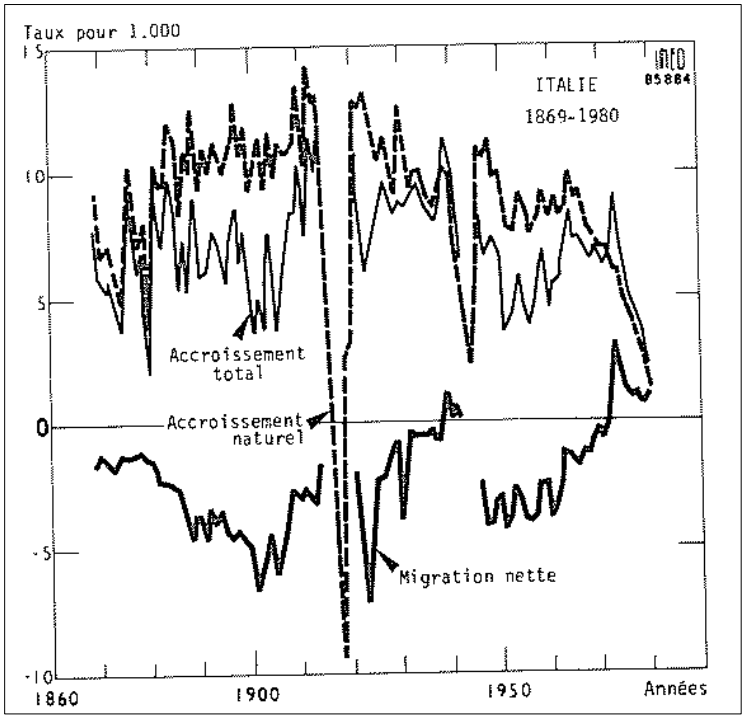
Nous observons que les périodes de plus fort accroissement naturel correspondent aux périodes de solde migratoire négatif, se manifestant par des sorties massives nettes de population.
Nous observons ainsi, un effet miroir entre la courbe de solde migratoire net, et celle de l'accroissement naturel : donc une corrélation forte et négative entre les deux indicateurs.
Cet effet miroir entre solde migratoire et accroissement naturel est concomitant du processus de transition démographique.
On retrouve le même phénomène en Italie (voir ci-après).

Figure 7 : Mouvement naturel de la population, Italie 1861-1980



Source: Chesnais (1986) «The migratory transition», p. 237.

Figure 8 : Migration nette et accroissement naturel, Italie 1869-1980

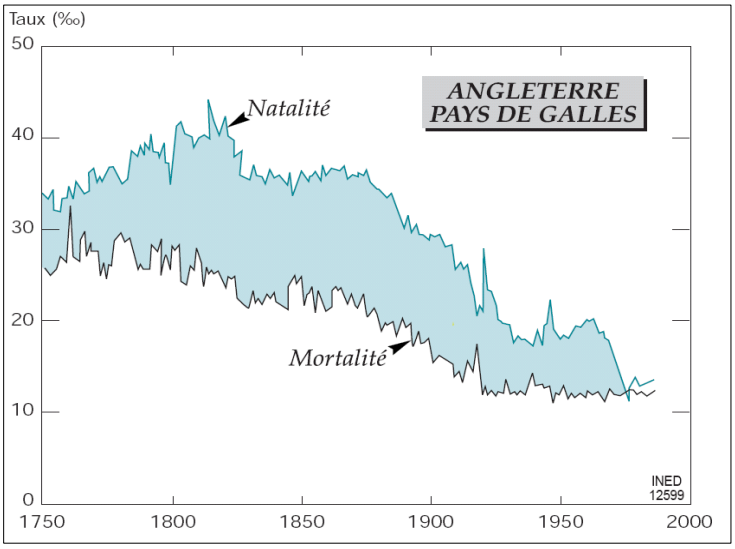


Source: Chesnais (1986) «The migratory transition», p. 179.

Les exemples historiques (cf. graphiques ci-dessous) des vieux pays industriels comme l'Angleterre et l'Allemagne illustrent là encore, avec un décalage temporel par rapport à la Suède, le processus type de la transition démographique : une baisse continue du taux de mortalité, suivie une génération environ plus tard, d'une chute continue du taux de natalité. Après la transition, la mortalité et la natalité se stabilisent à des niveaux significativement plus bas.

Exemple historique : l'Angleterre & Pays de Galles

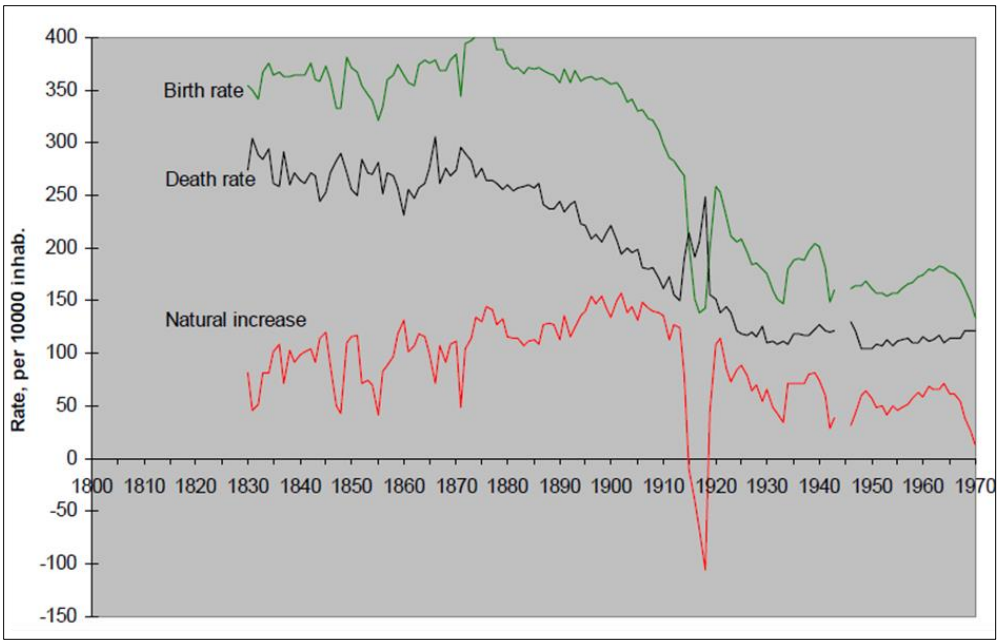
Figure 9 : La transition démographique en Angleterre



Source : INED, Population et Sociétés n° 346, mai 1999.

Exemple historique : l'Allemagne

Figure 10 : La transition démographique en Allemagne

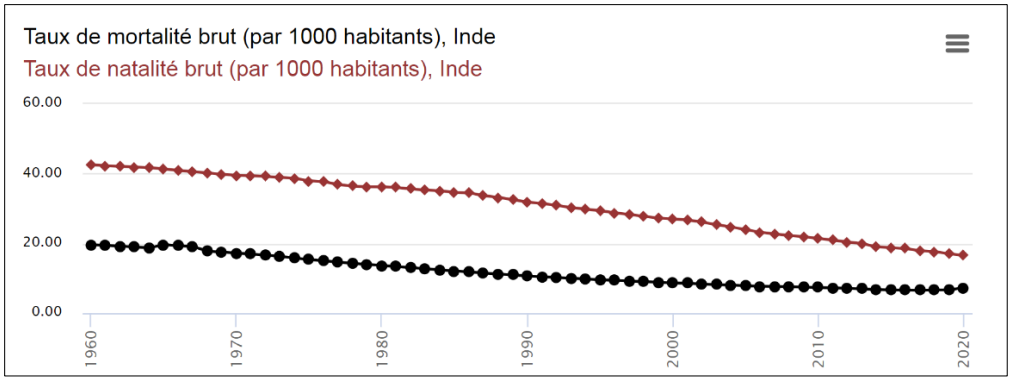


Source : Source : Mitchell, 2007.

Les jeunes nations connaissent également des tendances similaires, mais plus brèves et plus intenses. Les cas de l'Inde et de l'Equateur choisis à dessein, illustrent parfaitement la généralité de la transition démographique.

Exemple historique : l'Inde

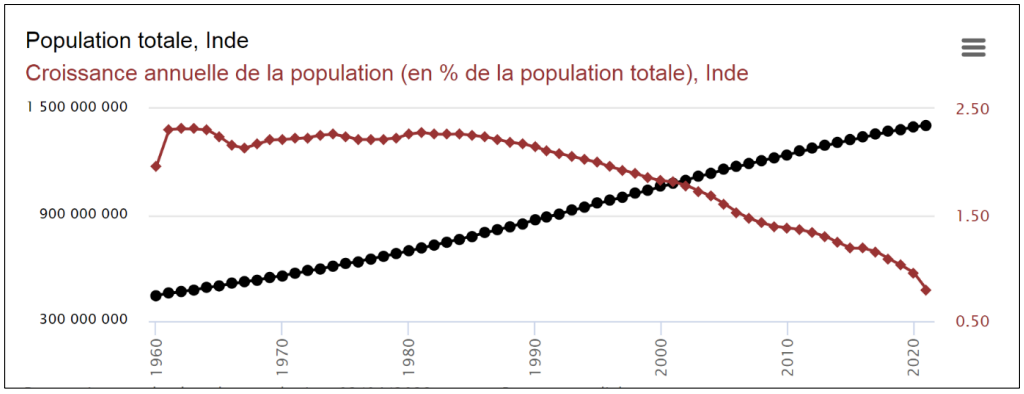
Figure 11 : Mortalité et natalité, Inde



Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

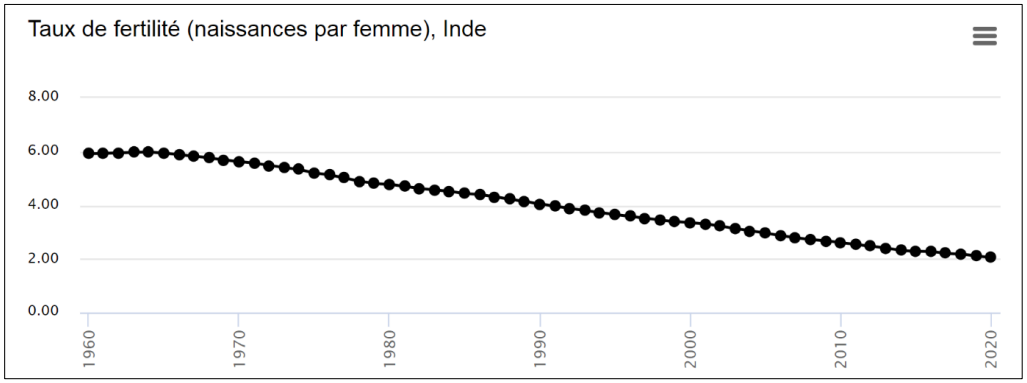
L'Inde est bien dans une transition démographique : baisse continue et écarts importants entre les taux de mortalité et de natalité. La chute de la mortalité et de la natalité a commencé bien avant 1960. L'écart entre les 2 taux étant encore conséquent, la fin de la période transitoire est à prévoir vraisemblablement peu après la décennie 2020.

Figure 12 : Population, Inde



Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

Figure 13 : Fécondité, Inde



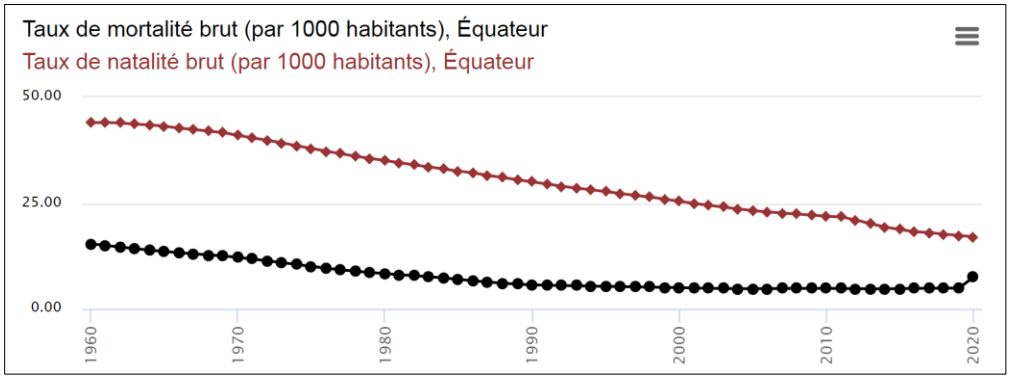
Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

Les données démographiques indiennes, indiquent que la décroissance du solde démographique a commencé entre 1980 et 1990 et que par conséquent, l'Inde est actuellement dans la phase 2 de sa transition démographique.
La fécondité des femmes est en décrue et le seuil des deux enfants par femme est atteint. Ce qui indique une position proche de la fin de la phase 2.

Exemple historique : l'Equateur

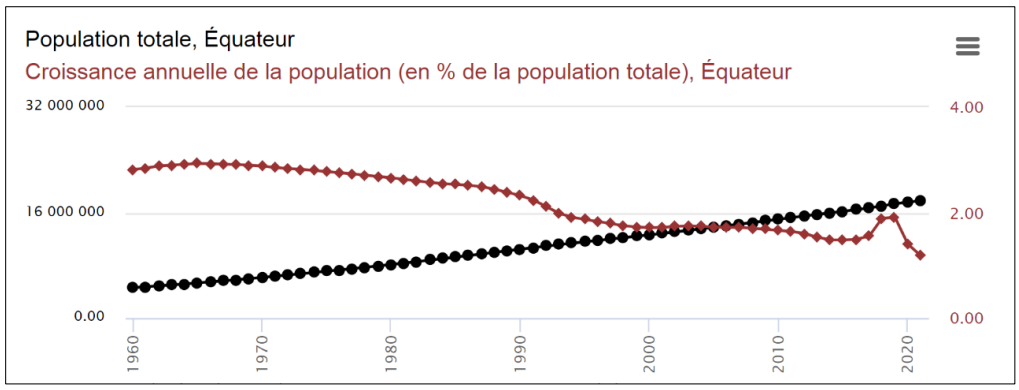
En ce qui concerne l'Equateur, le scénario est sensiblement le même que pour l'Inde, comme l'attestent les trois graphiques ci-dessous. Cependant la décrue du solde démographique a commencé un peu plus tardivement, i.e. entre 1990 et 2000.

Figure 14 : Mortalité et natalité, Equateur



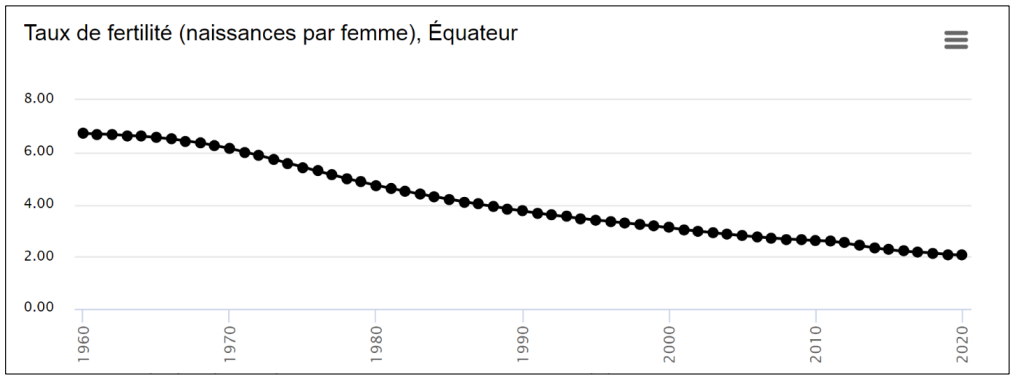
Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

Figure 15 : Population, Equateur



Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

Figure 16 : Fécondité, Equateur



Source : La Banque Mondiale / Perspective Monde, Université de Sherbrooke. 2023

5.1.3. Longévité et vieillissement

Les deux autres conséquences majeures de la transition démographique sont la longévité et le vieillissement, corollaires logiques de la baisse continue de la mortalité.

Phénomène de longévité

Nous faisons un bref exposé du phénomène de longévité inspiré de Mandzij (2011, pp. 9-10).

Depuis les années soixante (et même avant en considérant l'ensemble des périodes de transition démographique), il y a eu une augmentation régulière de l'espérance de vie en Europe et en Amérique du Nord. Ce phénomène s'étend désormais à l'ensemble des pays du globe, sauf périodes de crises sanitaires majeures.

Cette augmentation peut avoir un impact significatif sur les résultats financiers des entreprises offrant des produits d'assurance sur la vie ou des régimes de retraite, car cela implique que les prestations de ces produits doivent être fournies sur une période plus longue que prévu initialement, augmentant ainsi le coût total.

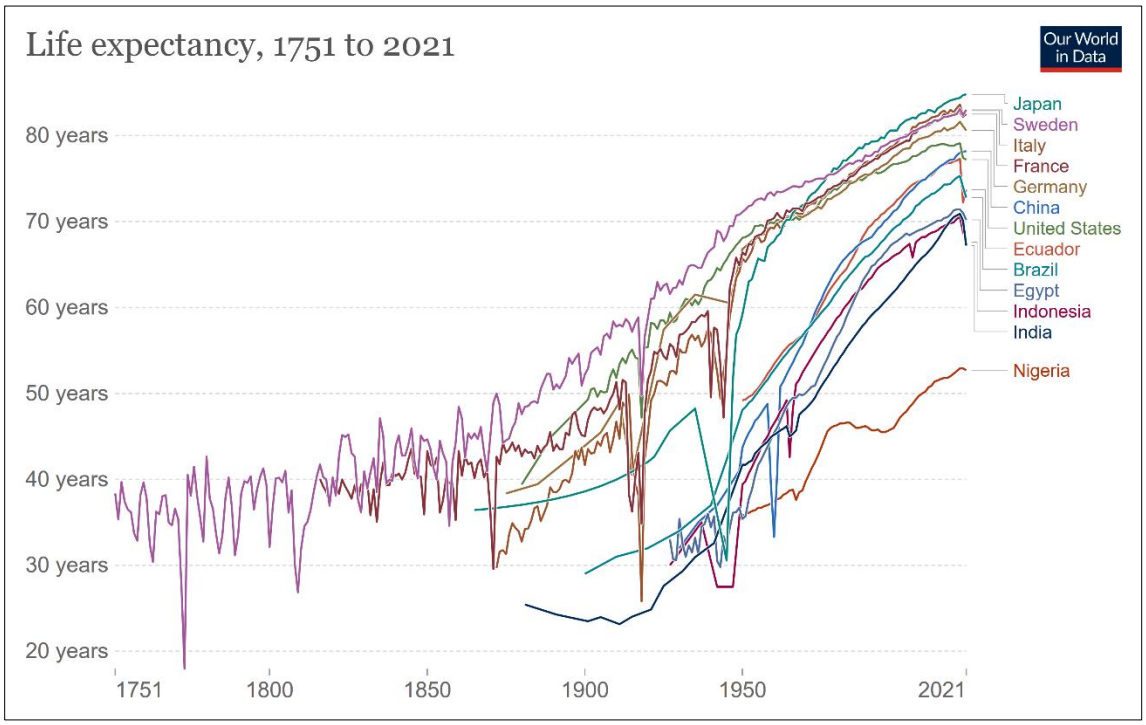
Les régimes de retraites à prestations définies ont été remplacés par des programmes à cotisations définies. Certains pays envisagent de décaler l'âge de départ à la retraite, tandis que d'autres l'ont déjà fait.

Les tables de mortalité prospectives sont utilisées pour gérer le risque de longévité, mais des mises à jour irrégulières peuvent être un problème.

Le phénomène d'antisélection a un impact différent sur les portefeuilles des compagnies d'assurance. Les taux de mortalité et la vitesse d'amélioration de la mortalité varient selon les groupes d'assurés, ce qui rend difficile la gestion du risque de longévité pour les compagnies d'assurance.

La hausse continue et générale de la longévité est illustrée par la figure ci-après.

Figure 17 : Espérance de vie à 10 ans



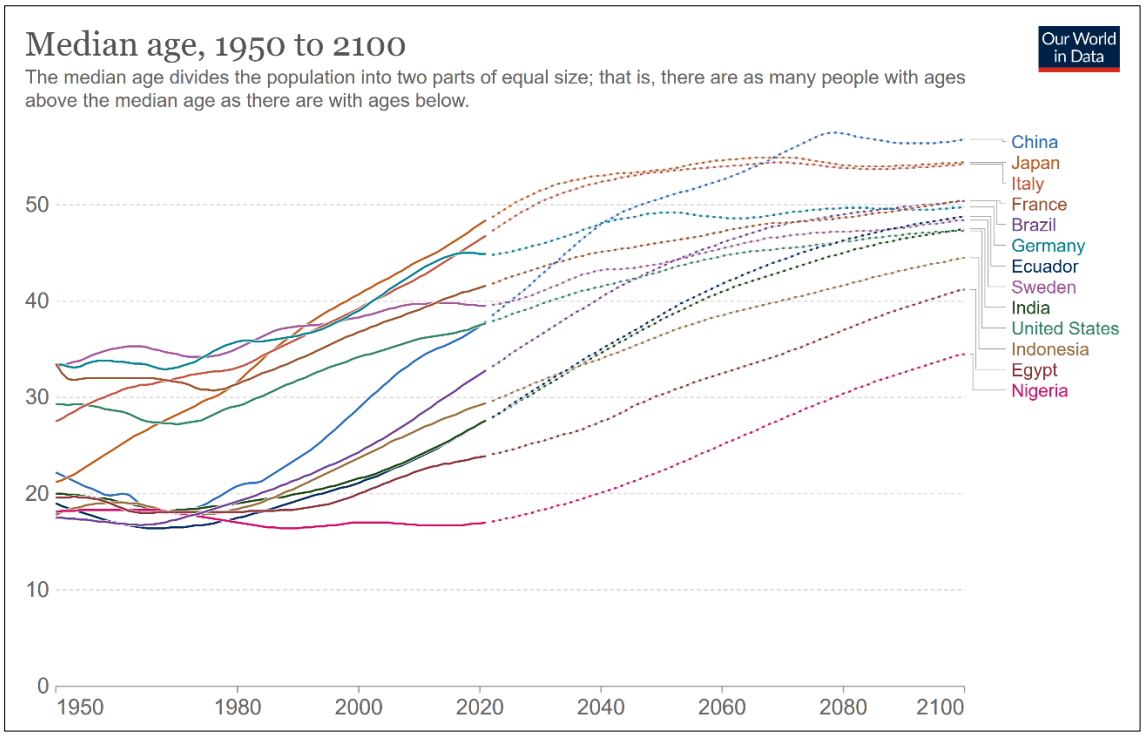
Source: United Nations Population Division and Human Mortality Database/ OurWorldInData.org/life-expectancy/•CC BY-SA. 2023

Vieillissement

La hausse continue de l'âge médian illustre le fait connu du vieillissement des pays ayant terminé leur transition démographique depuis plusieurs générations. L'âge médian divise la population en deux parties

de taille égale : il y a autant de personnes dont l'âge est supérieur à la médiane que d'âges inférieurs à la médiane.
Le graphique ci-dessous illustre le premier temps de rajeunissement, suivi d'une longue phase de vieillissement.

Figure 18 : Âge médian



Source: United Nations Population Division (Median Age) 2022 Revision OurWorldInData.org/life-expectancy/•CC BY-SA.

Note : Les années 1950 à 2020 montrent des estimations historiques. À partir de 2021, les projections de l'ONU (variante moyenne) sont présentées.

5.2. La cliométrie

La cliométrie se définit comme l'application de modèles économiques enrichis d'applications économétriques à l'histoire quantitative et institutionnelle. Elle permet une lecture des évènements historiques dans une perspective de temps long, avec une capacité de prospective à long terme dans les domaines étudiés (Gaba, 2021).

La cliométrie utilise des données quantitatives, applique des modèles économétriques, et adopte souvent une approche multidisciplinaire dans le cadre historique (Daudin, 2007). La cliométrie permet donc une lecture quantitative des évènements historiques dans une perspective de temps long, et de faire de la prospective à long terme et multigénérationnelle dans les domaines étudiés.

En France, cette approche multidisciplinaire, quantitative et de long terme est initiée par le courant de l'École des Annales dont les recherches sont publiées dans la revue « Annales d'histoire économique et sociale », fondée en 1929 par les historiens Lucien Febvre (1878-1956) et Marc Bloch (1886-1944), et dont Fernand Baudrel (1933 à 1985) prend la direction en 1956 après le décès de Lucien Febvre. Emmanuel Le Roy Ladurie chercheur de l'école des Annales publia une série remarquée d'ouvrages sur l'histoire

quantitative du climat (Le Roy Ladurie [1997, 2004, 2006, 2009], Le Roy Ladurie et Séchet [2009]). Le Roy Ladurie une figure tutélaire pour Emmanuel Todd, historien dans la tradition des Annales.
Aux Etats-Unis, le courant de pensée cliométrique naît sous le nom de la *new economic history* et publie ses travaux dans la revue *Explorations in Economic History* dès 1949. La *Cliometrics society* est fondée en 1983, et en 1993, le « Prix Nobel » d'économie est attribué aux cliomètres Douglass North (1920-2015) et Robert Fogel (1926-2013).

Exemples d'indicateurs mesurables sur le long terme (plusieurs générations) :

- Données climatiques
- Indicateurs démographiques (survie, reproduction, migrations, …)
- Données sociales, économiques

L'approche quantitative et multidisciplinaire de l'Histoire permet :

- Une lecture des évènements Historiques par rapport aux tendances de long terme
- De faire de la prospective à long terme (multigénérationnelle) aussi bien dans les domaines démographique, social ou économique

Notre notion de temps transitionnel permet de situer chaque pays dans son processus de transition (économique par exemple), à un temps calendaire donné de son histoire (cf. section 10.1).
Nous définissons le temps transitionnel, comme une série temporelle qui a une dépendance monotone forte avec le temps. Nous mesurons cette dépendance monotone (concordante ou discordante) par le coefficient de corrélation des rangs de Kendall τ entre la série temporelle testée et le temps calendaire. Nous pouvons ainsi tester l'hypothèse qu'une série temporelle puisse être utilisée comme temps transitionnel.
Cette originalité consiste également à développer une notion de temps transitionnel différent du temps calendaire, qui s'écoule au rythme des phases et évènements transitionnels : ce temps transitionnel a pour fonction de comparer la dynamique d'entités (géographiques ou autre) différentes traversées par un même phénomène transitionnel, mais qui se déroule à des périodes historiques (donc calendaires) différentes.

6. Modèles de mortalité et construction de tables de mortalité

6.1. Modèle démographique

Pour décrire, modéliser et analyser la mortalité, il est important de formaliser les concepts de base qui y sont associés. Cela implique de définir certaines notations couramment utilisées en mathématiques actuarielles.
Nous faisons un rappel de ces notations en nous inspirant de Wandji (2015, pp. 18-21).

6.1.1. Durée de vie restante

La durée de vie d'un individu de la population de référence est représentée par la variable aléatoire T.
$\forall x \in \{1,2,\dots\}$, la durée de vie restante d'un individu d'âge x est définie par la variable aléatoire T_x telle que :

$$P[T_x > t] = P[T > t + x | T > x]$$

L'individu vivant à l'âge x décèdera à l'âge $x + T_x$.
Le temps moyen restant à vivre pour un individu d'âge x est donné par :

$$e_x^{\uparrow} = \mathbb{E}[T_x] = \int_0^{\infty} P[T_x > t]dt$$

La probabilité pour qu'un individu vive t années augmente avec l'âge. Ceci s'explique par le fait qu'entre deux âges x et $y(x < y)$:

$$P[x + T_x > t] \leq P[y + T_x > t]$$

Par exemple, un individu de 55 ans a plus de chance de vivre jusqu'à 70 ans qu'un individu de 25 ans.

6.1.2. Fonction de survie et quotient de mortalité

Probabilité de survie

Elle est définie comme étant la probabilité qu'un individu d'âge x vive t années de plus.

$$ {}_tp_x = p_{x,t} = P(x,t) = P[T_x > t] = P[T > t + x | T > x]$$

Probabilité de décès

La probabilité qu'un individu d'âge x décède avant t années est :

$$ {}_tq_x = q_{x,t} = q(x,t) = P[T_x \leq t] = P[T \leq t + x | T > x]$$

Fonction de survie

Soit L_x le nombre de vivants d'âge x d'une cohorte de L_0 individus nés à la même date. Supposons que $L_0 = l_0$, alors le nombre moyen de survivants à l'âge x est donné par :

$$l_x = \mathbb{E}[L_x] = l_0 \, {}_xp_0$$

Le nombre de décès observés parmi les individus d'âge x est donné par : $D_x = L_x - L_{x+1}$.
Quant au nombre moyen de décès à l'âge x, il est défini par :

$$\delta_x = \mathbb{E}[D_x] = l_x q_x$$

La fonction de survie $t \to S(t)$ est définie par :

$$S(t) = P[T > t]$$

Quotient de mortalité

La probabilité de décès d'un individu d'âge x avant l'âge $x + t$ est également définie comme étant le quotient du nombre moyen de décès sur la période $[x,\ x + t]$ rapporté à l'effectif au début de la période.

$$q_{x,t} = 1 - p_{x,t} = 1 - \frac{l_{x+t}}{l_x} = \frac{\delta_{x,t}}{l_x}$$

6.1.3. Exposition au risque

L'exposition au risque sur la période $[x,\ x + t]$ est définie comme étant le temps durant lequel les individus sont exposés au risque de décès. En pratique, il s'agit du nombre d'années vécues pendant la période d'observation (dans ce cas-ci $[\![x, x + t]\!]$) :

$${}_tER_x = ER_{x,t} = ER(x,t) = \int_{\tau=0}^{t} L_{x+\tau} d\tau$$

Nous définissons la moyenne de l'exposition au risque par :

$$\mathbb{E}\big[ER_{x,t}\big] = {}_t\varepsilon_x = \varepsilon_{x,t} = \varepsilon(x,t) = \int_{\tau=0}^{t} l_{x+\tau} d_\tau$$

En introduisant la variable aléatoire $\tau_{x,i}$ définissant le temps durant lequel l'individu i d'âge x est exposé au risque de décès dans l'année d'observation, l'exposition au risque sur l'année observée est égale à :

$$ER_x = \sum_{i=0}^{L_x} \tau_{x,i}$$

Le nombre d'années restants à vivre aux survivants d'une cohorte au-delà de l'âge x est défini par :

$$ER_{x^\bullet} = \int_{\tau=0}^{\infty} L_{x,\tau} d_\tau$$

Nous pouvons également définir le nombre moyen d'années restants à vivre aux survivants d'une cohorte au-delà de l'âge x par :

$$\mathbb{E}[ER_{x\bullet}] = \varepsilon_{x\bullet} = \int_{\tau=0}^{\infty} l_{x+\tau} d_\tau$$

6.1.4. Taux et force de mortalité

Taux de mortalité

La notion d'exposition au risque précédemment introduite permet de définir la notion de taux de mortalité.

Le taux de mortalité est défini comme le rapport du nombre moyen de décès sur le nombre de personnes exposées au risque de décès. Sur une période observée $[x,\ x+t]$, il vaut :

$$_t m_x = m_{x,t} = m(x,t) = \frac{\delta_{x,t}}{\varepsilon_{x,t}}$$

Force de mortalité

Le taux instantané de mortalité ou force de mortalité mesure le risque pour un individu de mourir dans un intervalle de temps proche de zéro (instantanément).

- Le taux instantané de mortalité d'un individu d'âge x est donné par :

$$\mu_{x+1} = \lim_{\delta t \to 0^+} \frac{P[t < T_x \leq t + \delta t | T_x > t]}{\delta t}$$

- Lien entre taux et force de mortalité

$$\lim_{\delta t \to 0} m_{x,\delta t} = \mu_x$$

- Lien entre probabilité de survie et taux instantané de mortalité

$$\mu_{x+t} = \frac{1}{p_{x,t}} \frac{\partial q_{x,t}}{\partial_t} = -\frac{1}{p_{x,t}} \frac{\partial p_{x,t}}{\partial_t} \Longrightarrow p_{x,t} = exp\left(- \int_{\tau=0}^{t} \mu_{x+\tau} d_\tau \right)$$

Hypothèse des taux instantanés de mortalité constants par morceaux

Les données de mortalité sont généralement collectées sur une base annuelle, ce qui signifie que le nombre de décès pour un âge donné est connu pour une année entière. Toutefois, pour tenir compte de la répartition des décès au cours de l'année, il est nécessaire de faire des hypothèses sur leur distribution temporelle.

L'hypothèse des taux instantanés de mortalité constants par morceaux postule :

$$\mu_{x+t} = \mu_x \quad \forall t \in [0,1]$$

La probabilité de décès d'un individu d'âge x à un moment $t \in [0,1]$ dans l'année vaut :

$$q_{x,t} = 1 - (1 - q_x)^t$$

D'où :

$$\mu_{x+t} = -\ln(1 - q_x) = \mu_x$$

6.1.5. Espérance de vie résiduelle

L'espérance de vie résiduelle à l'âge x est la durée moyenne restant à vivre après x ans :

$$e_x^{\uparrow} = \int_0^{\infty} p_{x,t} dt = \frac{1}{l_x} \int_{t \geq 0} l_{x+t} dt$$

Il est à noter que le calcul de l'espérance de vie résiduelle nécessite la formulation d'une hypothèse de répartition des décès dans l'année.

6.1.6. Effets âge, période, cohorte

L'analyse de la mortalité centrale repose sur trois concepts clés captés par un modèle statistique :

- La probabilité de décès augmente avec l'âge pour une date donnée, c'est l'effet âge.
- Pour un âge donné, la probabilité de décéder est moins élevée aujourd'hui qu'il y a de cela quelques années, c'est l'effet période.
- L'effet cohorte est un facteur structurel qui entraîne une mortalité plus forte ou plus faible pour l'ensemble de la vie d'une génération donnée.

6.2. Tables de mortalité

Une table de mortalité est un modèle qui décrit le processus de mortalité d'une cohorte d'individus depuis la naissance jusqu'à leur décès. Elle permet d'étudier le nombre de décès, les probabilités de décès ou de survie, ainsi que l'espérance de vie selon l'âge, le sexe ou la génération. Selon le type de table de mortalité utilisé, il est également possible d'analyser l'évolution de la mortalité dans une génération ou pour un âge donné au fil du temps.
Nous faisons un exposé sur la présentation des tables de mortalité ainsi que les différents types de tables existantes d'un point de vue juridique et technique en France, en nous inspirant de Mandzij (2011, pp. 15-17).

6.2.1. Présentation des tables de mortalité

Une table de mortalité contient habituellement deux informations à l'entrée :

- L'âge, noté x
- Le nombre des survivants à l'âge x, noté l_x

l_0 désigne une génération fictive correspondant le plus souvent à 100 000 naissances à l'origine de la table.
Pour chaque âge x donné, le nombre de décès entre les âges x et $x+1$, noté par d_x, peut être déduit par la relation suivante :

$$d_x = l_x - l_{x+1}$$

A partir des données de la table de mortalité, nous pouvons également déduire les probabilités de survie (p_x) et de décès (q_x) à un an, définies auparavant :

$$p_x = \frac{l_{x+1}}{l_x}$$

$$q_x = \frac{d_x}{l_x}$$

6.2.2. Type des tables de mortalité

Les tables de mortalité sont utilisées pour calculer la probabilité de survie et de décès d'un assuré, ce qui est crucial pour l'activité d'un assureur vie. Les assureurs peuvent utiliser des tables de mortalité règlementaires, mais la réglementation leur permet également d'utiliser leurs propres tables d'expérience dans certaines conditions. Ces tables sont utilisées pour établir les tarifs et les provisions de l'assurance.
D'un point de vue juridique, nous pouvons différencier deux catégories de tables :

- Les tables règlementaires
- Les tables d'expérience

D'un autre côté, du point de vue technique nous distinguons :

- Les tables du moment (statiques, instantanées)
- Les tables prospectives (dynamiques)

Une table de mortalité peut concerner la totalité d'une population ou être segmentée suivant des variables influençant de manière significative le risque de décès. Nous pouvons ainsi disposer de tables spécifiques pour les hommes et les femmes.

Point de vue juridique

Tables réglementaires

Les tables réglementaires françaises sont établies par l'INSEE (Institut National de la Statistique et des Etudes Economiques) et sont basées sur des observations sur la population française sur une période donnée.

- **Tables TH TF 00-02**

Les tables homologuées par l'arrêté du 20/12/2005 sont utilisées pour provisionner les engagements en cas de décès, ainsi que pour les engagements en cas de vie avec une sortie en capital moyennant des décalages d'âges. Les tables TH et TF 00-02 ont été élaborées à partir des données de l'INSEE, basées sur des observations effectuées entre 2000 et 2002. Elles sont applicables aux contrats d'assurance vie souscrits depuis le 1er juillet 1993.

- **Tables TGH05-TGF05**

Ces tables de génération sont obligatoires en vertu de la réglementation, telle que spécifiée dans l'arrêté du 01/08/2006. Elles sont utilisées pour provisionner les engagements de rentes viagères immédiates ou différées (Art. A335-1 C. Ass). Les tables TGH05 et TGF05 ont été élaborées à partir de la population des bénéficiaires de contrats de rentes observée sur la période 1993-2005, ainsi que de la population nationale (INSEE) de 1962 à 2000.

Tables d'expérience

Un assureur peut préférer utiliser des tables de mortalité d'expérience plutôt que les tables officielles. Ces tables sont élaborées par les compagnies d'assurance en se basant sur la mortalité observée dans leur propre portefeuille. Toutefois, ces tables doivent être certifiées et suivies par un actuaire indépendant qualifié. Elles sont utilisables si elles sont plus prudentes que les tables règlementaires correspondantes.

Point de vue technique

Tables du moment

Les tables de mortalité du moment décrivent la mortalité de la population actuelle dans son ensemble, en appliquant la même probabilité de décès à toutes les générations de personnes assurées.

Tables prospectives

Les tables prospectives sont des tables de mortalité qui dépendent de l'âge et du temps (et donc aussi de l'année de naissance). Elles peuvent être présentées par génération (année de naissance) ou par année calendaire.
Contrairement aux tables du moment, les tables prospectives prennent en compte les évolutions potentielles de la mortalité dans le temps, ce qui permet en principe une projection plus précise de la mortalité future.

6.3. Modélisation de la mortalité

Pour construire une table de mortalité d'expérience, il est nécessaire de calculer les taux bruts de mortalité d'expérience et de lisser les données pour réduire leur caractère erratique. Pour projeter les taux de mortalité d'expérience, des modèles relationnels sont utilisés à partir des taux de mortalité de référence. Cependant, pour les âges avancés où l'effectif est plus faible, il est difficile d'obtenir des estimations précises en raison du manque de données. Les méthodes de fermeture de table permettent de pallier cela en extrapolant la mortalité aux grands âges pour obtenir des taux de mortalité pour l'ensemble des âges de la population concernée.
Nous présentons une modélisation de la mortalité inspirée de Mandzij (2011, pp. 18-32), de Wandji (2015, pp. 31-52), de Planchet (2021, pp. 19-39) et de Villegas et al. (2015, p. 12).

6.3.1. Construction des taux bruts de mortalité

Deux estimateurs sont généralement utilisés pour calculer les taux de mortalité bruts d'expérience : il s'agit des Estimateurs de Hoem et de Kaplan-Meier. En effet, leur construction intuitive permet une utilisation facile et pratique.

Estimateur de Hoem

L'objectif est de calculer les taux instantanés de mortalité à partir du nombre de décès et de l'exposition au risque.
Nous observons $l(x,t)$ individus d'âge x sur la période t. La variable aléatoire $DC_i(x,t)$ indique si l'individu i est décédé ou non pendant la période t :

$$DC_i(x,t) = \begin{cases} 1 & si\ l'individu\ i\ est\ décédé\ à\ l'âge\ x \\ 0 & sinon \end{cases}$$

Notons $ER_i(x,t)$ la variable aléatoire indiquant le temps durant lequel chaque individu i a été observé dans l'âge x au cours de l'année t. Les réalisations des variables $DC_i(x,t)$ et $ER_i(x,t)$ seront respectivement notées $dc_i(x,t)$ et $er_i(x,t)$.
Nous procéderons à une estimation du modèle par maximum de vraisemblance. La contribution d'un $ième$ individu d'âge x à la vraisemblance est définie par :

$$\underbrace{exp\big(-er_i(x,t)\mu(x,t)\big)}_{\text{probabilité de survie } er_i(x,t)p_x} \quad \times \quad \underbrace{\mu(x,t)^{dc_i(x,y)}}_{\text{probabilité de décès}}$$

Nous rappelons que le taux instantané de mortalité est supposé constant par morceaux, i.e. :

$$\mu(x+u,t+s) = \mu(x,t)\,, \quad 0 \le s,u \le 1$$

La vraisemblance de modèle s'écrit :

$$\prod_{i=1}^{l(x,t)} exp[-er_i(x,t)\mu(x,t)]\mu(x,t)^{dc_i(x,t)} = exp[-er(x,t)\mu(x,t)] \times \mu(x,t)^{dc(x,t)}$$

Avec :

$er(x,t) = \sum_{i=1}^{l(x,t)} er_i(x,t)$ l'exposition au risque à l'âge x,

$dc(x,t) = \sum_{i=1}^{l(x,t)} dc_i(x,t)$ le nombre de décès à l'âge x.

L'estimateur du taux instantané de mortalité est obtenu en maximisant la vraisemblance :

$$\hat{\mu}(x,t) = \frac{dc(x,t)}{er(x,t)}$$

Nous en déduisons le taux de mortalité annuel :

$$\hat{q}(x,t) = 1 - exp\{-\hat{\mu}(x,t)\}$$

Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur non paramétrique de la fonction de survie S définie par $S_x(u) = P(T_x > u)$, où pour rappel, T_x est la variable aléatoire correspondant à la durée de survie conditionnelle à l'âge x.

Nous observons une police (ou un individu) entre des âges x et $x+1$. La durée de vie de la police n'est pas observée avant la date d'entrée en portefeuille (troncature à gauche). Elle n'est observée que jusqu'à sa date de sortie. Lorsque la sortie de portefeuille (ou population de référence) se fait pour une autre raison que le décès, nous parlons d'une censure à droite.

Soient $x = a_1 < a_2 < \cdots < a_n \le x+1$ des âges où se produisent des évènements (troncature, censure, décès), n_i la population sous risque à l'âge a_i , d_i le nombre de décès, e_i le nombre d'entrées en portefeuille, c_i le nombre de censurés à droite. Nous calculons la population sous risque à l'âge a_i à l'aide d'une formule de récurrence :

$$n_i = n_{i-1} - d_{i-1} - c_{i-1} + e_i$$

Nous obtenons ainsi un estimateur de la loi de survie :

$$\begin{aligned} P(T_x > a_k) &= P(T_x > a_k | T_x > a_{k-1}) \times P(T_x > a_{k-1}) \\ &= \left[\prod_{i=2}^{k} P(T_x > a_i | T_x > a_{i-1})\right] \times 1 \end{aligned}$$

Sachant que :

$$P(T_x > a_i | T_x > a_{i-1}) = 1 - P(T_x \le a_i | T_x > a_{i-1}) = 1 - \frac{d_i}{n_i}$$

Nous obtenons alors un estimateur du taux de mortalité annuel à l'âge x :

$$\hat{q}(x,t) = 1 - P(T_x > a_n) = 1 - \left[\prod_{i=2}^{n} 1 - \frac{d_i}{n_i}\right]$$

6.3.2. Lissage des tables de mortalité

Les taux de mortalité bruts peuvent être instables en raison de fluctuations d'échantillonnage. Pour identifier les caractéristiques de la mortalité propres à une population, il est nécessaire de lisser ces taux.
Les techniques de lissage consistent à remplacer les taux de mortalité bruts par des taux plus réguliers, tout en conservant l'information originale. Il existe différentes méthodes de lissage paramétriques ou non paramétriques. Parmi elles, nous décrirons le modèle paramétrique de Makeham et la méthode non paramétrique de Whittaker-Henderson.
Nous présenterons également un critère de validation des ajustements obtenus.

Modèle de Makeham

Le modèle proposé par Makeham pour modéliser le taux instantané de mortalité $\mu_x$est :

$$\mu_x = \theta_1 + \theta_2\theta_3^x$$

Avec $\theta_1 \geq 0,\ \theta_2 > 0$ et $\theta_3 > 1$.
Le taux instantané de mortalité est composé de deux termes distincts :

- Le premier terme θ_1

Il ne dépend pas de l'âge de l'individu et correspond à la mortalité imprévisible, telle que les décès causés par un accident ou une maladie. Ce type de mortalité peut survenir à tout âge de la vie de l'individu.

- Le deuxième terme $\theta_2\theta_3^x$

Il dépend de l'âge x de l'individu et suit une structure exponentielle qui reflète la tendance à la hausse du risque de mortalité avec l'âge, à mesure que l'individu vieillit.
Le modèle de Makeham permet de prendre en compte à la fois une composante exponentielle et d'autres facteurs, tels que la mortalité imprévisible, dans la modélisation de la mortalité. Cette approche de lissage est particulièrement utile pour analyser les tendances de mortalité sur des périodes plus longues. En effet, en réécrivant le modèle en termes de taux de mortalité annuel, nous obtenons :

$$q_x(\theta_1,\theta_2,\theta_3) = 1 - sg^{\theta_3^x(\theta_3-1)} \Leftrightarrow ln\big(1 - q_x(\theta_1,\theta_2,\theta_3)\big) = -\theta_1 - \frac{\theta_2}{ln\theta_3}\theta_3^x(\theta_3 - 1)$$

Avec $s = exp(-\theta_1)$ et $g = exp\left(\frac{-\theta_2}{ln\theta_3}\right)$.
Les paramètres du modèle de Makeham peuvent être estimés par différentes méthodes (à l'exemple de l'approche par maximum de vraisemblance modifié avec initialisation des paramètres à partir de la méthode de King et Hardy).

Méthode de Whittaker-Henderson

La méthode de Whittaker-Henderson est une méthode de lissage par ajustement statistique. Elle consiste à déterminer statistiquement la meilleure fonction non paramétrique de lissage.
Les taux de mortalité lissés issus du modèle de Whittaker-Henderson minimisent le critère R suivant :

$$\min_{\tilde{q}_x}\left\{R \middle| R = \underbrace{\sum_{x=x_{min}}^{x_{max}} w_x\,(\tilde{q}_x - \dot{q}_x)^2}_{\text{critère de fidélité}} \quad + \quad \underbrace{h\sum_{x=x_{min}}^{x_{max}-z}(\Delta^z\tilde{q}_x)^2}_{\text{critère de régularité}}\right\}$$

où :

- $\dot{q}_x$ le taux de mortalité brut à l'âge x issu des estimations sur la population d'expérience,
- $\tilde{q}_x$ le taux de mortalité lissé à l'âge x,
- w_x la pondération associée à l'âge x,
- Δ^z l'opérateur de différenciation d'ordre z,
- h le poids associé au critère de régularité.

Le critère de minimisation permet de trouver le meilleur taux de mortalité lissé en minimisant l'erreur entre le taux de mortalité brut et le taux de mortalité lissé. Ce critère prend en compte à la fois un critère de fidélité (critère de moindres carrés pondérés) et un critère de régularité (critère de lissage). Les taux de mortalité peuvent être pondérés pour tenir compte de leur importance relative selon différents critères. Si la série des taux de mortalité bruts comporte des données manquantes, nous introduisons la pondération suivante :

$$w_x = \begin{cases} 0 \ \ si \ \dot{q}_x = 0 \\ 1 \ \ sinon \end{cases}$$

Ainsi, les données de mortalité manquantes ne sont pas prises en compte dans la détermination des taux de mortalité lissés.

Il est envisageable d'utiliser un autre critère de pondération pour les taux de mortalité, en tenant compte de la qualité de l'estimation des taux de mortalité bruts. Par exemple, il serait judicieux de surpondérer les données brutes de mortalité des âges ayant un grand nombre d'observations dans la population d'expérience. Dans ce cas, nous utilisons la pondération suivante :

$$w_x = \frac{E_x}{\overline{E}}$$

où E_x représente l'effectif observé à l'âge x et $\overline{E}$ l'effectif moyen sur l'ensemble des âges étudiés. Cette pondération permet de limiter l'importance des taux de mortalité bruts aux grands âges.

Lorsqu'il y a un biais d'échantillonnage, les taux de mortalité bruts peuvent être discontinus. Pour atténuer ces discontinuités, le critère de régularité est utilisé pour pénaliser les ajustements trop irréguliers entre les taux de mortalité lissés et les taux de mortalité bruts.

Selon le degré de différenciation z de la matrice Δ^z, il est possible de pénaliser l'allure de la courbe de manière différente :

- L'utilisation d'une matrice de différenciation du premier ordre permet de rechercher des lisseurs qui engendrent des courbes de taux de mortalité linéaires entre deux âges, en pénalisant les écarts entre deux taux de mortalité bruts successifs.
- En revanche, l'utilisation d'une matrice de différenciation du second ordre permet de pénaliser les écarts du second ordre, ce qui permet d'obtenir des taux de mortalité plus lisses.

Test du Khi-deux des ajustements obtenus

Afin de vérifier la qualité d'ajustement des taux de mortalité, il est nécessaire de réaliser des tests statistiques. L'un de ces tests est le test du Khi-deux qui permet de mesurer la distance entre le taux de mortalité brut et le taux de mortalité lissé.

Ce test consiste à calculer une valeur χ^2_{Emp} telle que :

$$\chi^2_{Emp} = \sum_{x=x_{min}}^{x_{max}} l_x \frac{(\dot{q}_x - \tilde{q}_x)^2}{\tilde{q}_x}$$

Pour un seuil de confiance α compris entre 0 et 1, le test d'adéquation est rejeté lorsque :

$$\chi^2_{Emp} > q_{1-\alpha}\{\chi^2_{th}\}$$

où χ^2_{th} représente la loi du Khi-deux à $(n-1-p)$ degrés de liberté, p étant le nombre de paramètres du modèle, n étant la taille de l'échantillon.
Pour le modèle de Makeham, $p = 3$, et pour le modèle de Whittaker-Henderson, $p = 0$.

6.3.3. Modèles internes (ou intrinsèques)

Ces méthodes proposent un ajustement des taux de mortalité basé sur l'expérience réelle de la mortalité d'une population. Cependant, leur inconvénient est qu'elles nécessitent un historique suffisamment long par rapport à l'horizon de prévision. Elles extrapolent l'information contenue dans les taux de mortalité afin d'effectuer une projection fiable de la mortalité en se basant sur la tendance passée.
Nous distinguons deux catégories au sein de ces modèles internes : les modèles internes classiques et les modèles internes composites à facteurs PCR-O/PLS.

Modèles internes classiques (modèles GAPC)

Les modèles internes classiques peuvent être rassemblés sous une même famille de modèles appelée modèles GAPC *(Generalized Age-Period-Cohort)* ou modèles de mortalité stochastique généralisée âge-période-cohorte.
La présentation de ces modèles requiert certaines notations et hypothèses que nous définissons ci-après.

Hypothèses et notations dans les modèles GAPC

- Variable aléatoire du nombre de décès

Nous rappelons que $D_{x,t}$ désigne la variable aléatoire indiquant le nombre de décès dans une population à l'âge x au cours de l'année civile t. Indiquons également que $d_{x,t}$ est le nombre de décès observés, $E^C_{x,t}$ est l'exposition centrale au risque à l'âge x au cours de l'année t, et $E^0_{x,t}$ est l'exposition initiale correspondante.

- Approximation de l'exposition initiale

En nous référant à Planchet (2021, p. 11), nous pouvons estimer le taux de mortalité central par $\widehat{m}_{x,t}$:

$$\widehat{m}_{x,t} = \frac{d_{x,t}}{(l_{x,t} + l_{x,t+1})/2} = \frac{\mathbb{E}(D_{x,t})}{\mathbb{E}(E^c_{x,t})}$$

Selon le même principe, nous obtenons un taux de mortalité « initial » noté $\widehat{m^0_{x,t}}$ tel que :

$$\widehat{m^0_{x,t}} = \frac{d_{x,t}}{l_{x,t}} = \frac{\mathbb{E}(D_{x,t})}{\mathbb{E}(E^0_{x,t})}$$

Cela permet de déduire :

$$\mathbb{E}(E^0_{x,t}) = l_{x,t} \quad \text{et} \qquad \mathbb{E}(E^c_{x,t}) = \frac{l_{x,t}+l_{x,t+1}}{2}$$

D'où l'estimation :

$$\mathbb{E}(E^0_{x,t}) = l_{x,t} = \frac{l_{x,t} + l_{x,t+1}}{2} + \frac{l_{x,t} - l_{x,t+1}}{2} = \mathbb{E}(E^c_{x,t}) + \frac{1}{2}d_{x,t}$$

Lorsque le contexte est clair, nous pouvons écrire $E_{x,t}$ pour désigner $E_{x,t}^0$ ou $E_{x,t}^c$

- Probabilité de décès à un an

La probabilité de décès à un an pour une personne âgée de x dernier anniversaire et de l'année civile t, notée $q_{x,t}$, peut être estimée comme $\hat{q}_{x,t} = d_{x,t}/E_{x,t}^0$.
La force de mortalité et le taux central de mortalité sont désignés respectivement par $\mu_{x,t}$ et $m_{x,t}$, l'estimation empirique de ce dernier étant $\hat{m}_{x,t} = d_{x,t}/E_{x,t}^c$.
Nous supposons que les décès, $d_{x,t}$, et les expositions centrales $E_{x,t}^c$, ou les expositions initiales $E_{x,t}^0$ sont disponibles sous la forme d'un tableau rectangulaire comprenant les âges (sur les lignes) $x = x_1, x_2, \ldots, x_k$, et les années civiles (sur les colonnes) $t = t_1, t_2, \ldots, t_n$.

- Notion de composante aléatoire

La variable aléatoire du nombre de décès $D_{x,t}$ suit une distribution de Poisson ou une distribution binomiale :

$$D_{x,t} \sim Poisson\left(E_{x,t}^c \mu_{x,t}\right) \qquad \text{ou} \qquad D_{x,t} \sim Binomial\left(E_{x,t}^0, q_{x,t}\right)$$

Avec $\mathbb{E}\left(D_{x,t}/E_{x,t}^c\right) = \mu_{x,t}$ et $\mathbb{E}\left(D_{x,t}/E_{x,t}^0\right) = q_{x,t}$.

- Notion de composante systématique

Dans un modèle GAPC, les effets de l'âge x, de l'année civile t et de l'année de naissance ou de la cohorte $c = t - x$ sont saisis au moyen d'un prédicteur $\eta_{x,t}$ donné par :

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

Le terme α_x est une fonction statique de l'âge saisissant la forme générale de la mortalité par âge.
$N \geq 0$ est un nombre entier indiquant le nombre de termes âge-période décrivant les tendances de la mortalité.
Chaque indice de temps $\kappa_t^{(i)}, i = 1, \ldots, N$, précise la tendance de la mortalité et $\beta_x^{(i)}$module son effet à travers les âges.
Le terme γ_{t-x} rend compte de l'effet de cohorte avec $\beta_x^{(0)}$, et module son effet à travers les âges.
Les termes modulateurs d'âge $\beta_x^{(i)}, i = 0, 1, \ldots, N$, peuvent être soit des fonctions pré-spécifiées de l'âge, c'est-à-dire $\beta_x^{(i)} = f^i(x)$, comme dans les modèles de type CBD, ou des termes non paramétriques sans structure préalable qui doivent être estimés comme dans le modèle de Lee-Carter.
Les indices de période $\kappa_t^{(i)}, i = 1, \ldots, N$, et l'indice de cohorte γ_{t-x} sont des processus stochastiques. C'est la principale caractéristique clé qui permet la projection stochastique des modèles GAPC et donc la génération de prévisions probabilistes des taux de mortalité futurs.

- Lien entre la composante aléatoire et la composante systématique

La fonction de liaison g associant la composante aléatoire et la composante systématique est de sorte que :

$$\mathrm{g}\left(\mathbb{E}\left(\frac{D_{x,t}}{E_{x,t}}\right)\right) = \eta_{x,t}$$

Modèle de Lee-Carter log-Poisson

Brouhns et al. (2002) ont mis en œuvre le modèle de Lee-Carter en supposant une distribution de Poisson du nombre de décès et en utilisant la fonction de liaison logarithmique en ce qui concerne la force de mortalité μ_{xt}. La structure prédictive proposée par Lee et Carter (1992) suppose qu'il existe une fonction d'âge statique, α_x, un terme unique non paramétrique de la période d'âge $(N = 1)$, et aucun effet de cohorte.

Ainsi, le prédicteur est donné par :

$$\eta_{xt} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)}$$

Afin de projeter la mortalité, l'indice temporel $\kappa_t^{(1)}$ est modélisé et prévu à l'aide des processus ARIMA.

Pour assurer l'identifiabilité du modèle, Lee et Carter (1992) suggèrent l'ensemble suivant de contraintes de paramètres

$$\sum_x \beta_x^{(1)} = 1, \qquad \sum_t \kappa_t^{(1)} = 0,$$

Modèle de Renshaw et Haberman (RH)

Renshaw et Haberman (2006) généralisent le modèle de Lee-Carter en incorporant un effet de cohorte pour obtenir le prédicteur :

$$\eta_{xt} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(0)}\gamma_{t-x}$$

Les projections de mortalité pour ce modèle sont calculées à l'aide de séries chronologiques prévisionnelles des estimations $\kappa_t^{(1)}$ et γ_{t-x}, générées à l'aide de processus ARIMA univariés dans l'hypothèse d'indépendance entre la période et les effets de cohorte.

Afin d'estimer le modèle, Renshaw et Haberman (2006) supposent une distribution de Poisson des décès (composante aléatoire) et utilisent une fonction de liaison logarithmique ciblant la force de mortalité μ_{xt} .

L'identifiabilité du modèle peut être assurée à l'aide de l'ensemble de contraintes de paramètres suivants :

$$\sum_x \beta_x^{(1)} = 1, \sum_t \kappa_t^{(1)} = 0, \sum_x \beta_x^{(0)} = 1, \sum_{c=t_1-x_\kappa}^{t_n-x_1} \gamma_c = 0,$$

Renshaw et Haberman (2006) considèrent également plusieurs sous-structures du prédicteur obtenues en définissant comme constante l'un ou les deux termes modulant l'âge. La sous-structure obtenue en fixant $\beta_x^{(0)} = 1$ est particulièrement intéressante,

$$\eta_{xt} = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \gamma_{t-x}\,,$$

qui a été suggérée par Haberman et Renshaw (2011) comme une structure plus simple qui résout certains problèmes de stabilité du modèle original.

Modèle APC (âge-période-cohorte)

Une autre sous-structure couramment utilisée du modèle de Renshaw et Haberman est ce que l'on appelle modèle âge-période-cohorte (APC), correspondant à $\beta_x^{(1)} = 1, \beta_x^{(0)} = 1$,

$$\eta_{xt} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}\,,$$

qui a une longue tradition dans les domaines de la médecine et de la démographie (voir, par exemple, Clayton et Schifflers (1987), Hobcraft et coll. (1982) et Osmond (1985)), mais qui n'a pas été largement utilisée dans la littérature actuarielle jusqu'à ce qu'elle soit examinée par Currie (2006).
Nous pouvons assurer l'identifiabilité du modèle en imposant l'ensemble des contraintes :

$$\sum_t \kappa_t^{(1)} = 0, \sum_{c=t_1-x_k}^{t_n-x_1} \gamma_c = 0, \sum_{c=t_1-x_k}^{t_n-x_1} c\gamma_c = 0,$$

où les deux dernières contraintes impliquent que l'effet de cohorte fluctue autour de zéro sans tendance linéaire perceptible.

Modèle CBD

Cairns et al. (2006) proposent une structure prédictive avec deux termes âge-période $(N = 2)$ avec des paramètres de modulation d'âge prédéfinis $\beta_x^{(1)} = 1$ et $\beta_x^{(2)} = x - \bar{x}$, aucune fonction d'âge statique et aucun effet de cohorte. Ainsi, le prédicteur du modèle CBD est donné par :

$$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}$$

où $\bar{x}$ est l'âge moyen dans les données.
Cairns et al. (2006) obtiennent des prévisions de mortalité en projetant les effets de période $\kappa_t^{(1)}$ et $\kappa_t^{(2)}$ à l'aide d'une marche aléatoire bivariée avec dérive.
Le modèle CBD n'a pas de problèmes d'identifiabilité et, par conséquent, l'ensemble des contraintes de paramètres est vide. Afin d'estimer le paramètre du modèle CBD, nous pouvons suivre Haberman et Renshaw (2011) et supposer une distribution binomiale des décès en utilisant une fonction de liaison logit ciblant les probabilités de décès à un an $q_{x,t}$.

Modèle M7

Cairns et al. (2009) étendent le modèle original de la CBD en ajoutant un effet de cohorte et un effet d'âge quadratique pour obtenir le prédicteur :

$$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + ((x - \bar{x})^2 - \hat{\sigma}_x^2)\kappa_t^{(3)} + \gamma_{t-x}$$

où $\hat{\sigma}_x^2$ est la valeur moyenne de $(x - \bar{x})^2$.
Pour identifier le modèle, Cairns et al. (2009) imposent l'ensemble des contraintes :

$$\sum_{c=t_1-x_k}^{t_n-x_1} \gamma_c = 0, \sum_{c=t_1-x_k}^{t_n-x_1} c\gamma_c = 0, \sum_{c=t_1-x_k}^{t_n-x_1} c^2\gamma_c = 0,$$

qui garantissent que l'effet de cohorte fluctue autour de zéro et n'a pas de tendance linéaire ou quadratique discernable.
Cairns et al. (2009) considèrent également les structures prédictives plus simples

$$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \gamma_{t-x}$$
$$\eta_{xt} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + (x_c - x)\gamma_{t-x}$$

où x_c est un paramètre constant à estimer. Ces structures sont généralement appelées modèles M6 et M8, respectivement.

Modèle Plat

Plat (2009) combine le modèle de la CBD avec certaines caractéristiques du modèle Lee-Carter pour produire un modèle adapté aux tranches d'âge complètes et qui capture l'effet de cohorte. La structure prédictive proposée suppose qu'il existe une fonction d'âge statique, α_x, trois termes de période d'âge ($N = 3$) avec des paramètres de modulation d'âge prédéfinis $\beta_x^{(1)} = 1$, $\beta_x^{(2)} = \bar{x} - x$, $\beta_x^{(3)} = (\bar{x} - x)^+ = \max(0, \bar{x} - x)$, et un effet de cohorte avec des paramètres de modulation d'âge prédéfinis $\beta_x^{(0)} = 1$. Ainsi, le prédicteur est donné par :

$$\eta_{xt} = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + (\bar{x} - x)^+\kappa_t^{(3)} + \gamma_{t-x}\,.$$

Plat (2009) cible la force de mortalité μ_{xt} avec le lien logarithmique et estime les paramètres du modèle en supposant une distribution de Poisson des décès.
L'ensemble suivant de contraintes de paramètres peut être imposé pour garantir l'identifiabilité :

$$\sum_t \kappa_t^{(1)} = 0, \sum_t \kappa_t^{(2)} = 0, \sum_t \kappa_t^{(3)} = 0 \sum_{c=t_1-x_k}^{t_n-x_1} \gamma_c = 0, \sum_{c=t_1-x_k}^{t_n-x_1} c\gamma_c = 0, \sum_{c=t_1-x_k}^{t_n-x_1} c^2\gamma_c = 0$$

Les trois premières contraintes garantissent que les indices de période sont centrés autour de zéro, tandis que les trois dernières contraintes garantissent que l'effet de cohorte fluctue autour de zéro et n'a pas de tendance linéaire ou quadratique.
Dans les cas où seuls les âges plus âgés sont intéressants, Plat (2009) suggère d'abandonner le terme de la troisième période du prédicteur :

$$\eta_{xt} = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + \gamma_{t-x}\,.$$

Estimation des paramètres des modèles GAPC

Les estimations des paramètres des modèles stochastiques de mortalité de la GAPC peuvent être obtenues en maximisant la log-vraisemblance $\mathcal{L}(d_{x,t}, \hat{d}_{x,t})$ du modèle.

- Dans le cas d'une distribution de Poisson des décès :

$$\mathcal{L}(d_{x,t}, \hat{d}_{x,t}) = \sum_x \sum_t \omega_{x,t}\{d_{x,t}\, log\, \hat{d}_{x,t} - \hat{d}_{x,t} - log\, d_{x,t}\}$$

- Dans le cas d'une distribution binomiale des décès :

$$\mathcal{L}(d_{x,t}, \hat{d}_{x,t}) = \sum_x \sum_t \omega_{x,t} \left\{ d_{x,t}\, log\left(\frac{\hat{d}_{x,t}}{E_{x,t}^0}\right) + \left(E_{x,t}^0 - d_{x,t}\right) log\left(\frac{E_{x,t}^0 - \hat{d}_{x,t}}{E_{x,t}^0}\right) + \binom{E_{x,t}^0}{d_{x,t}} \right\}$$

Dans les deux cas, les $\omega_{x,t}$ sont des poids prenant la valeur 0 si une cellule de données particulière (x, t) est omise ou 1 si la cellule est incluse, et $\hat{d}_{x,t}$ est le nombre attendu de décès prédits par le modèle.

$$\hat{d}_{x,t} = E_{x,t}\, g^{-1}\left(\alpha_x + \sum_{i=1}^{N} \beta_x^{(i)}\, \kappa_t^{(i)} + \beta_x^{(0)}\gamma_{t-x}\right)$$

Modèles internes composites à facteurs PCR-Optimal/PLS

Gaba (2021) a développé les modèles internes à composantes PCR-O ou PLS qui ont permis d'améliorer les prévisions des taux de mortalité des modèles internes classiques dans des cas spécifiques (pays, âges, horizons).
Une originalité de ce mémoire est d'améliorer les modèles internes de Gaba (2021) en leur donnant une flexibilité **composite** qui permet de choisir des classes de modèles éventuellement différentes selon les âges modélisés.
Au sein des modèles internes **composites** à facteurs PCR-optimal/PLS, chaque âge x est modélisé par l'un ou l'autre des deux types de modèles suivants :

- Les modèles internes à facteurs PCR-optimal

ou

- Les modèles internes à facteurs PLS

Pour rappel, dans ce cas, la modélisation est réalisée en deux itérations afin de sélectionner dans un premier temps la classe de modèles la mieux adaptée à chaque âge, puis dans un second temps d'estimer les performances **sans biais** des modèles sélectionnés à l'itération précédente.
Les deux itérations se présentent comme suit :

- Itération 1 : Pour chaque âge, choix **a posteriori** de sa meilleure classe de modèle (Best Model Class)
  - Historique : t_0 ; t_1
  - Prévisions : $t_1 + 1$; t_2
- Itération 2 : Pour chaque âge, utilisation **a priori** de sa Best Model Class issue de l'itération 1 pour en estimer la performance sans biais
  - Historique : t_0 ; t_2
  - Prévisions : $t_2 + 1$; t_3

Les temps $\boldsymbol{t_0, t_1, t_2, t_3}$ étant des bornes de découpages arbitraires en 4 intervalles temporels de l'historique utilisé. Le temps est considéré discret.

6.3.4. Modèles relationnels (ou à référence externe)

Les données de mortalité provenant d'une population d'expérience ne sont pas suffisantes pour construire directement des tables de mortalité prospectives robustes.
Afin de résoudre ce problème, nous relions les taux de mortalité de la population d'expérience aux taux de mortalité d'une population de référence à l'aide d'un modèle relationnel. Il sera ainsi possible de projeter les taux de mortalité d'expérience par l'intermédiaire de la projection préalable des taux de mortalité de référence.

En général, les deux populations peuvent être issues de :

- deux populations distinctes,
- une population et un sous-ensemble de celle-ci,
- une partition de la population considérée,
- une même population à deux époques différentes.

Nous cherchons une fonction f qui vérifie l'une des trois équations suivantes :

$$q_{x,t}^{exp} = f\left(q_{x,t}^{ref}\right)$$
$$\mu_{x,t}^{exp} = f(\mu_{x,t}^{ref})$$
$$\mu_{x,t}^{exp} = f(q_{x,t}^{ref})$$

Avec :
$q_{x,t}^{exp}$ la probabilité de décès d'un individu d'âge x avant le temps t dans la population d'expérience
$q_{x,t}^{ref}$ la probabilité de décès d'un individu d'âge x avant le temps t dans la population de référence
$\mu_{x,t}^{exp}$ le taux instantané de mortalité d'un individu d'âge x au temps t dans la population d'expérience
$\mu_{x,t}^{ref}$ le taux instantané de mortalité d'un individu d'âge x au temps t dans la population de référence
Nous présentons deux catégories de modèles relationnels :

- Les modèles externes à appariement temporel calendaire
- Les modèles externes à appariement temporel cliométrique

Modèles externes à appariement temporel calendaire

Nous désignons par modèles externes à appariement temporel calendaire, les modèles relationnels classiques généralement utilisés. A titre d'exemples, nous présentons ici deux sous-catégories :

- Les modèles logistiques et les modèles log-linéaires
- Les méthodes de positionnement

Modèles logistiques

Modèle logistique de BRASS

Dans le modèle de BRASS, les logits des probabilités de décès d'expérience $q_{x,t}^{exp}$ sont liés aux probabilités de décès de référence $q_{x,t}^{ref}$ à partir d'une relation linéaire de la forme :

$$logit\left(q_{x,t}^{exp}\right) = \theta_1 + \theta_2 \times logit\left(q_{x,t}^{ref}\right), \qquad x = x_{min}, \ldots, x_{max}$$

Les paramètres sont estimés par régression linéaire de l'ensemble des $logit\left(q_{x,t}^{exp}\right)$ sur l'ensemble des $logit\left(q_{x,t}^{ref}\right)$.

Modèle logistique TGH05-TGF05

Dans le modèle logistique TGH05-TGF05, une régression est effectuée en utilisant les logits des probabilités de décès d'expérience et ceux des probabilités de décès de référence. L'équation utilisée pour cette régression est :

$$logit\left(q_{x,t}^{exp}\right) = a_x logit\left(q_{x,t}^{ref}\right) + b_x + \varepsilon_{x,t}$$

Les résidus $\varepsilon_{x,t}$ sont iid et de loi normale centrée.
L'approche utilisée dans le modèle TGH05-TGF05 consiste à utiliser un critère de type moindres carrés pour la régression linéaire dans un cadre de modèle linéaire ordinaire, ce qui rend la mise en œuvre très simple.
Pour construire la table TGH05-TGF05, la régression a été effectuée sur la période 1994-2004, pour laquelle des tables de moment ont été établies. Une régression a été effectuée pour chaque valeur d'âge disponible, de sorte que les paramètres a_x et b_x puissent être spécifiques à chaque âge.
Les résultats ont montré une forte corrélation entre les paramètres a_x et b_x (chacun de ces paramètres étant une fonction de l'âge x), ce qui conduit à contraindre le modèle en posant :

$$a_x = (\alpha + \beta b_x)$$

Le critère des moindres carrés s'écrit :

$$\min_{\alpha,\beta,b_x} \sum_x \sum_t \left[logit\left(q_{x,t}^{exp}\right) - (\alpha + \beta b_x) logit\left(q_{x,t}^{ref}\right) - b_x \right]^2 = \min_{\alpha,\beta,b} P(\alpha, \beta, b)$$

Avec b le vecteur des b_x.

Le programme de minimisation s'écrit de la manière suivante : $\begin{cases} \min\limits_{\alpha,\beta,b} P(\alpha,\beta,b) \\ \nabla P(\alpha,\beta,b_x) \quad = 0, \forall x \\ \nabla^2 P(\alpha,\beta,b_x) \quad > 0, \forall x \end{cases}$

Modèle linéaire généralisé de Poisson

Le modèle linéaire généralisé de Poisson introduit des indicateurs de mortalité qui font varier les taux de décès avec l'âge et l'année.

Le modèle s'écrit :

$$\mu_{x,t}^{exp} = \beta_0 + \beta_1 log(q_{x,t}^{ref}) + \beta_2 x + \beta_3 t + \beta_4 xt + \varepsilon_{x,t}$$

Avec :

$$D_{x,t}^{exp} \sim Poisson(ER_{x,t}^{exp} \times \mu_{x,t}^{exp})$$

Les résidus $\varepsilon_{x,t}$ sont iid et de loi normale centrée.

Les paramètres $\beta_i, i \in [\![0, 4]\!]$ sont estimés par la méthode du maximum de vraisemblance.

Méthodes de positionnement

Modèle proportionnel de COX

Le modèle de COX s'écrit :

$$\mu_{x,t}^{exp} = \theta \times \mu_{x,t}^{ref}, \theta > 0$$

Le modèle à référence externe de COX se propose d'ajuster les taux bruts de la mortalité de la population d'expérience comme un abattement ou une amélioration des taux bruts d'une table de référence.

C'est le modèle relationnel le plus simple car il ne fait intervenir qu'un seul paramètre qui n'est pas fonction de l'âge, ce qui vient en opposition aux paramètres des modèles paramétriques qui le plus souvent dépendent de l'âge ou du temps.

D'une part, cette méthode dispose d'avantages considérables pour l'ajustement des taux bruts, car la relation qu'elle décrit ne dépend que d'un seul paramètre. D'autre part, un certain nombre de chercheurs autour de l'analyse et la projection de données de survie ont émis des réserves sur le modèle de COX car la relation serait trop directe et le caractère proportionnel de la régression présenterait un biais important pour la projection des taux de décès.

Modèle proportionnel SMR (Standardized Mortality Ratio).

Dans cette première approche, les probabilités de décès d'expérience sont ajustées en multipliant les probabilités de décès de référence par un coefficient appelé SMR (Standardized Mortality Ratio).

La SMR est définie comme le rapport du nombre de décès observés sur le nombre de décès prédits dans une population de référence :

$$SMR = \frac{\sum_t \sum_x D_{x,t}^{exp}}{-\sum_t \sum_x ER_{x,t}{}^{ref} \times ln(1 - \widehat{q_{x,t}^{ref}})}$$

Avec :

$D_{x,t}^{exp}$ Le nombre de décès à l'âge x en année t dans la population d'expérience

$ER_{x,t}^{ref}$ L'exposition au risque à l'âge x en année t dans la population de référence

La probabilité de décès d'expérience se calcule par :

$$q_{x,t}^{exp} = SMR \times q_{x,t}^{ref}$$

Positionnement par rapport à une référence externe

Il est également possible de rechercher dans un ensemble de tables prospectives exogènes disponibles la période des tables de référence $[t^{ref}, t^{ref} + h], h > 0$ la plus « proche » de la période $[t^{exp}, t^{exp} + h], h > 0$ issue des données de la table d'expérience. Cela conduit à utiliser comme table d'expérience les tables exogènes **décalées**.

La notion de « la plus proche » suppose l'utilisation d'une distance entre deux tables. Différentes approches pour calculer cette distance sont possibles :

- Le Khi-deux sur les $q_{x,t}$
- La distance déduite des espérances résiduelles ou de leurs intégrales.

Nous ne présentons pas ces modèles.

Modèles externes à appariement temporel cliométrique

Comme indiqué dans les problématiques, ces techniques d'appariement seront développées dans la suite de ce mémoire, mais nécessitent d'un point de vue statistique au moins une vingtaine d'années d'observations pour le pays d'expérience R.
Ainsi l'appariement entre le pays d'expérience R et le pays de référence A (voir les définitions dans la section précédente), permet :

1. de disposer de couples appariés de temps calendaires du pays d'expérience R et ceux du pays de référence A, via le temps transitionnel qui leur est commun

2. de créer plusieurs séries cliométriques pour chaque âge du pays R (par à rapport un ou plusieurs pays A, avec divers paramétrages de sexe ou d'âge). Ces séries cliométriques qui ont par construction des valeurs dans le passé et dans le futur, sont des séries explicatives potentielles dans les modèles externes cliométriques pour le pays d'expérience R

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le premier résultat ci-dessus, nous proposons les 3 classes de modèles externes cliométriques suivants, adaptés des modèles classiques externes :

- Modèle **externe cliométrique** adapté du modèle externe de BRASS
- Modèle **externe cliométrique** adapté du modèle externe de COX
- Modèle **externe cliométrique** adapté du modèle externe de TGH05-TGF05

Les trois modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** mélangeant les adaptations de BRASS/COX/TGH05-TGF05 .

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le deuxième résultat ci-dessus, nous proposons les 2 classes de modèles externes cliométriques suivants, adaptés des modèles à facteurs PCR-Optimal et des modèles à facteurs PLS :

- Modèle **externe cliométrique** à facteurs PCR-Optimal (construits à partir des séries cliométriques)
- Modèle **externe cliométrique** à facteurs PLS (construits à partir des séries cliométriques)

De même, les deux modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** à facteurs PCR-O/PLS .

Comme précédemment évoqué, le caractère composite de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles différentes**.**

6.3.5. Modèles mixtes (à la fois internes et externes)

Les modèles mixtes utilisent simultanément l'historique de la population d'expérience et l'historique de la population de référence (Gaba [2021]).
Un modèle mixte n'est utilisable que si l'historique de la population d'expérience est suffisamment longue par rapport à l'horizon de prévision, et donc compatible avec l'usage d'un modèle interne.

Nos modèles mixtes combinent les prédicteurs des modèles **internes** à facteurs PCR-O ou PLS et ceux des modèles **externes cliométriques** à facteurs PCR-O ou PLS.

Nous proposons dans ce mémoire une amélioration originale du modèle mixte en leur donnant une flexibilité **composite** qui permet de choisir des classes de modèles éventuellement différentes selon les âges modélisés.
Comme nous le verrons dans les résultats empiriques, cela a permis d'améliorer les prévisions pour les âges aux extrémités ; ces âges étant un point faible des modèles mixtes de Gaba (2021).

Nous désignons ces nouveaux modèles mixtes à caractère composite par :
Modèles mixtes cliométriques et composites à facteurs PCR-O/PLS

6.3.6. Fermeture de tables

Les modèles de fermeture sont utilisés pour extrapoler la mortalité aux grands âges, au-delà de l'âge de raccord (âge seuil au-dessus duquel les données de mortalité sont limitées en raison d'un effectif faible). Un modèle de fermeture permet d'extrapoler la mortalité entre l'âge de raccord et l'âge ultime (âge maximal) de la table de mortalité.
Il est fréquent d'introduire un âge dit *âge pivot* qui permet de contrôler l'allure de la courbe de mortalité entre l'âge de raccord et l'âge ultime de la table de mortalité. Le taux de mortalité à cet âge pivot peut être déterminé à partir des résultats des études démographiques.
Différents modèles de fermeture de table peuvent être mis en œuvre. Nous en présentons trois qui sont généralement utilisés.

Modèle de Coale-Kisker

La méthode de Coale et Kisker (Coale et al. [1990]) consiste à extrapoler les taux de mortalité aux grands âges (par exemple, jusqu'à x = 110 ans) en se basant sur les formules :

- Pour $x \geq 65$

Les taux de mortalité aux grands âges x $(x \geq 65)$ sont extrapolés en utilisant la formule :

$$\mu_x = \mu_{65} \times e^{g_x(x-65)}$$

Avec g_x qui est le taux moyen de croissance de la mortalité entre les âges 65 et x.

- Pour $x \geq 80$

Coale et al. (1990) ont empiriquement constaté que les courbes des g_x possèdent en général un pic aux alentours de 80 ans avant de décroître linéairement. Ils ont par conséquent proposé l'équation :

$$g_x = g_{80} + s \times (x - 80), \qquad x \geq 80$$

Les taux de mortalité aux grands âges x $(x \geq 80)$ peuvent ainsi être extrapolés en utilisant la formule :

$$\mu_x = \mu_{x-1} \times e^{g_{80} s (x-80)}, \qquad x \geq 80$$

Avec $s = -\frac{\ln(\mu_{79}+31\times g_{80})}{465}$ et $g_{80} = \frac{\ln\left(\frac{\mu_{80}}{\mu_{65}}\right)}{15}$.

Modèle de Denuit & Goderniaux

Denuit et Goderniaux (2005) proposent un modèle de projection des probabilités de décès aux âges les plus élevés qui impose une contrainte aux grands âges :

$$ln\, q_x = a + bx + cx^2 + \varepsilon_x$$

Avec $\begin{cases} q_{130} = 1 \\ q'_{130} = 0 \end{cases}$

Où q'_x désigne la dérivée de q_x ; les résidus $\varepsilon_x$sont iid et de loi normale centrée.
La première contrainte impose une tangente au point avec $x = 130$, et la seconde contrainte impose une concavité aux grands âges, ce qui empêche une décroissance des probabilités aux âges élevés. Les deux contraintes impliquent :

$$\begin{cases} a = 130^2 c \\ b = -260c \end{cases}$$

D'où :

$$ln\, q_x = c(130^2 - 260x + x^2) + \varepsilon_x$$

Il est nécessaire de déterminer l'âge à partir duquel les taux bruts de mortalité seront remplacés par les valeurs prédites par le modèle.

Modèle TGH05-TGF05 de fermeture de table

Le modèle TGH05-TGF05 a été utilisé pour la construction des tables réglementaires du même nom. Les probabilités de décès pour les âges au-delà de l'âge de raccord sont prolongées à partir d'une forme quadratique :

$$logit\,(q_x) = ln\left(\frac{q_x}{1-q_x}\right) = \theta_1 + \theta_2 x + \theta_3 x^2$$

L'âge de raccord utilisé pour la construction des tables TGH05-TGF05 est de $x_r = 95$ ans. L'âge pivot est de 110 ans et la probabilité de décès associée à cet âge est de 0,5.
Les trois paramètres du modèle doivent respecter les contraintes ci-dessous :

- L'ajustement à l'âge de raccord x_r est continu et dérivable :

$$\text{logit}\left(q_{x_r}\right) = \theta_1 + \theta_2 x_r + \theta_3 x_r^2$$
$$logit\left(q_{x_r}\right) - logit\left(q_{x_r-1}\right) = \theta_2\left(x_r - (x_r - 1)\right) + \theta_3(x_r^2 - (x_r - 1)^2)$$

- Le taux de décès à l'âge pivot x_p est de 0,5 :

$$logit\,(0{,}5) = \theta_1 + \theta_2 x_p + \theta_3 x_p^2$$

Quant à l'âge pivot x_p, il peut être réécrit sous la forme :

$$x_p = at + b$$

où a et b sont des paramètres déterminés à l'aide d'un critère des moindres carrés sur les âges et les années.

Partie 2 : Apports méthodologiques

7. Relation entre le modèle Lee-Carter et la régression PCR

Le modèle de Lee et Carter (1992) est un des modèles classiques de référence dont nous allons comparer les performances avec celles des modèles que nous proposons.
Afin d'en proposer une généralisation, nous étudions et formalisons ici les relations entre le modèle classique de Lee-Carter et la régression sur composantes principales *(principal component regression, PCR)*.

7.1. Processus stochastique et série temporelle

Dans les exposés mathématiques de l'ensemble du présent document, nous faisons l'hypothèse d'un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.

Définitions : Processus stochastique et série temporelle (cf. Boutaharet, Royer-Carenzi [2019], p.13)
Un *processus stochastique* est une série de variables aléatoires indexées sur le temps, notées $(\mathrm{S_t})_{\mathrm{t}\in\mathcal{T}}$.
A chaque instant $\mathrm{t} \in \mathcal{T}$, $\mathrm{S_t}$ est une variable aléatoire définie sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$.
Chaque observation sur la durée $\mathcal{T}$, notée $\mathrm{s_t} = \mathrm{S_t}(\omega)$, est une réalisation du processus $(\mathrm{S_t})_{\mathrm{t}}$.

Même si le phénomène sous-jacent évolue en continu sur un intervalle temporel ($\mathcal{T} = [a, b], 0 \leq a \leq b$), les données ne sont concrètement récoltées que ponctuellement, aux instants $t_1, \dots, t_n \in \mathcal{T}$, avec $0 \leq t_1 < \cdots < t_n$.
Les valeurs observées aux instants $t_1, \dots, t_n$ forment la *série temporelle* (ou *série chronologique*) et sont notées $s_{t_1}, \dots, s_{t_n}$
Chaque observation s_{t_k}est une réalisation de la variable aléatoire S_{t_k}. On notera S_n le vecteur aléatoire $(S_{t_1}, \dots, S_{t_n})$ dont est issue la série des observations $s_{t_1}, \dots, s_{t_n}$. Notons que les variables aléatoires $S_{t_1}, \dots, S_{t_n}$ ne sont pas supposées a priori *i.i.d*.

Nous supposerons que la durée $t_{k+1} - t_k$ est constante, et que la série ne présente pas de données manquantes ; concrètement nous avons imputé des valeurs aux éventuelles données manquantes par interpolation linéaire.

7.2. Présentation du modèle de Lee-Carter

Le modèle de Lee et Carter (1992) s'intéresse à $m_{x,t}$ qui représente le taux de mortalité central entre les âges x et $x + 1$ mesuré au cours de l'année t. Le modèle de Lee-Carter, peut s'écrire comme suit (Boyer et al. [2015], p. 536):

$$ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}$$

où :

- a_x correspond à la moyenne du logarithme des taux de mortalité par âge calculée sur l'ensemble de la période considérée,
- k_t représente la tendance générale du niveau de mortalité au cours du temps,

- b_x est la sensibilité au facteur commun k_t spécifique à chaque âge $\left(\frac{d\,ln(m_{x,t})}{dt} = b_x \frac{dk_t}{dt}\right)$,
- $\varepsilon_{x,t}$ est le résidu du modèle, supposé *i.i.d.*, d'espérance nulle et de variance σ_ε^2

7.3. Estimation des paramètres du modèle

L'estimation des paramètres s'effectue par le critère des moindres carrés ordinaires et par la décomposition en valeurs singulières.

- Estimation de a_x par un critère des moindres carrés ordinaires (Planchet, Lelieur [2007]; Safitri *et al* [2018]) :

$$\left(\hat{a}_x, \hat{b}_x, \hat{k}_t\right) = \underset{a_x, b_x, k_t}{\arg\, min} \sum_{x,t} \left(ln\left(m_{x,t}\right) - a_x - b_x k_t\right)^2$$

Sous contrainte que $\sum_{x=x_m}^{x_M} b_x^2 = 1$ ou $\sum_{x=x_m}^{x_M} b_x = 1$ et $\sum_{t=t_q}^{t_Q} k_t = 0$,

avec :

$x_m = min(x)$

$x_M = max(x)$

$t_q = min(t)$

$t_Q = max\ (t)$

Il convient donc de résoudre le problème d'optimisation :

$$\frac{\partial \sum_{x,t}\left(ln\left(m_{x,t}\right) - a_x - b_x k_t\right)^2}{\partial a_x} = 2(Q - q + 1)a_x - 2\sum_{t=t_q}^{t_Q}\left(ln\left(m_{x,t}\right) - b_x k_t\right) = 0$$

$$= 2(Q - q + 1)a_x - 2\sum_{t=t_q}^{t_Q} ln\left(m_{x,t}\right) + 2b_x \sum_{t=t_q}^{t_Q} k_t = 0$$

Or $\sum_{t=t_q}^{t_Q} k_t = 0 \Rightarrow (Q - q + 1)a_x = \sum_{t=t_q}^{t_Q} ln\left(m_{x,t}\right)$

$$\hat{a}_x = \frac{1}{Q - q + 1}\sum_{t=t_q}^{t_Q} ln\left(m_{x,t}\right)$$

- Estimation de b_x et k_t

L'estimation de b_x et k_t se fait via le premier terme de la décomposition en valeurs singulières de la matrice $Z = ln\left(m_{x,t}\right) - \hat{a}_x$

Pour obtenir une meilleure visualisation de la composante temporelle k_t, nous modifions la présentation classique en mettant les informations temporelles en lignes et celle liées à l'âge en colonnes :

$$m_{x,t} = \begin{bmatrix} m_{x_m,t_q} & \cdots & m_{x_M,t_q} \\ \vdots & \ddots & \vdots \\ m_{x_m,t_Q} & \cdots & m_{x_M,t_Q} \end{bmatrix}$$

$$Z = \begin{bmatrix} z_{x_m,t_q} = ln\left(m_{x_m,t_q}\right) - \hat{a}_{x_m} & \cdots & z_{x_M,t_q} = ln\left(m_{x_M,t_q}\right) - \hat{a}_{x_M} \\ \vdots & \ddots & \vdots \\ z_{x_m,t_Q} = ln\left(m_{x_m,t_Q}\right) - \hat{a}_{x_m} & \cdots & z_{x_M,t_Q} = ln\left(m_{x_M,t_Q}\right) - \hat{a}_{x_M} \end{bmatrix}$$

Z est d'ordre $(Q - q + 1, M - m + 1)$.

L'approximation de Z donnant $Z \approx \widehat{k_t}\hat{b}_x^T$ (ou encore $\hat{Z} = \widehat{k_t}\hat{b}_x^T$) découle de la formule du modèle de Lee-Carter :
$Z = ln(m_{x,t}) - \hat{a}_x = \widehat{k_t}\hat{b}_x^T + \hat{\varepsilon}_{x,t}$,
avec :
$\hat{b}_x^T = \left(\hat{b}_{x_m}, \dots, \hat{b}_{x_M}\right)$ et $\widehat{k_t}^T = \left(\hat{k}_{t_q}, \dots, \hat{k}_{t_Q}\right)$.

Soit u_i le i$^{\text{ème}}$ vecteur propre normé de ZZ^T de dimension $(Q - q + 1) \times (Q - q + 1)$ correspondant à la valeur propre λ_i on a :

$ZZ^T u_i = \lambda_i u_i$, avec $u_i^T u_i = 1$

En multipliant les deux membres de la relation précédente par Z^T, on a :

$Z^T Z Z^T u_i = Z^T \lambda_i u_i$

$(Z^T Z) Z^T u_i = Z^T \lambda_i u_i$

$(Z^T Z) Z^T u_i = \lambda_i Z^T u_i \Rightarrow$ $Z^T u_i$ est un vecteur propre de $Z^T Z$ associé à la même valeur propre λ_i, mais n'est pas normé car :
$(Z^T u_i)^T (Z^T u_i) = {u_i}^T Z Z^T u_i = {u_i}^T \lambda_i u_i = \lambda_i$, car ${u_i}^T u_i = 1$
Ou encore : $\|Z^T u_i\| = \sqrt{\lambda_i}$

Soit v_i le i$^{\text{ème}}$ vecteur propre normé ($v_i^T v_i = 1$) de $Z^T Z$ d'ordre $(M - m + 1) \times (M - m + 1)$

Donc : $v_i = \frac{1}{\sqrt{\lambda_i}} Z^T u_i$ (formule de transfert)

On a :
$(Z^T Z) v_i = \lambda_i v_i$

En multipliant à gauche par Z, on a :
$(ZZ^T) Z v_i = \lambda_i Z v_i$

$Z v_i$ est un vecteur propre non-normé de ZZ^T associé à la valeur propre λ_i (comme u_i sauf que u_i est normé) car :

$(Z v_i)^T (Z v_i) = {v_i}^T (Z^T Z v_i) = {v_i}^T \lambda_i v_i = \lambda_i$, car ${v_i}^T v_i = 1$
ou encore : $\|Z v_i\| = \sqrt{\lambda_i}$
Donc $u_i = \frac{1}{\sqrt{\lambda_i}} Z v_i$ (2$^{\text{ième}}$ formule de transfert)

La décomposition en valeurs singulières (SVD) de Z est donnée par :

$$Z = U\Sigma V^T$$

Avec

$$U = [u_1 \dots u_{Q-q+1}]$$

$$V^T = \begin{pmatrix} v_1^T \\ \dots \\ v_{M-m+1}^T \end{pmatrix}$$

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

D est d'ordre $(Q-q+1, M-m+1)$ et est la matrice des valeurs singulières non-nulles de Z

Donc : $Z = \sum_{i\geq 1} \sqrt{\lambda_i}\, u_i v_i^T$

Puisque la première valeur propre λ_1 est bien supérieure aux autres, la meilleure approximation de Z est alors donnée par :
$Z \approx \sqrt{\lambda_1} u_1 v_1^T$

En confrontant la relation $Z \approx \widehat{k_t}\widehat{b_x^T}$ et $Z \approx \sqrt{\lambda_1} u_1 v_1^T$, on a : $\widehat{k_t}\widehat{b_x^T} = \sqrt{\lambda_1} u_1 v_1^T$.
En tenant compte des ordres de $\widehat{k_t}$ et $\widehat{b_x}$ et de la contrainte ${b_x}^T b_x = 1$, nous obtenons par identification :
$\widehat{b_x} = v_1$ et $\hat{k}_t = \sqrt{\lambda_1} u_1$

Si nous avions utilisé la contrainte alternative $\sum_{x=x_m}^{x_M} b_x = 1$, nous aurions obtenu le résultat ci-après :

En multipliant et divisant le second membre de la relation $\widehat{k_t}\widehat{b_x^T} = \sqrt{\lambda_1} u_1 v_1^T$ par $\sum_j v_{1j}$, on a :

$$\widehat{k_t}\widehat{b_x^T} = \sqrt{\lambda_1} u_1 v_1^T \times \frac{\sum_j v_{1j}}{\sum_j v_{1j}}$$

$$\widehat{k_t}\widehat{b_x^T} = \frac{v_1^T}{\sum_j v_{1j}} \times \sqrt{\lambda_1} \sum_j v_{1j} u_1$$

De ce qui précède, on a par identification :

$$\widehat{b_x} = \frac{v_1}{\sum_j v_{1j}} \text{ et } \hat{k}_t = \sqrt{\lambda_1} \sum_j v_{1j} u_1$$

avec $\sum_j v_{1j} \neq 0$.

7.4. Modèle Lee-Carter, comme régression sur la 1ère composante principale

7.4.1. Analyse en composantes principales (ACP)

L'Analyse en Composantes Principales (ACP) permet de transformer des variables possiblement corrélées entre elles en de nouvelles variables décorrélées les unes des autres nommées composantes principales ou axes principaux. L'ACP est une analyse factorielle, en ce sens qu'elle produit des facteurs (ou axes principaux) qui sont des combinaisons linéaires des variables initiales, hiérarchisées et indépendantes les unes des autres. Nous en faisons ici un exposé inspiré de Johnston (1988, pp. 628-637).

Bien qu'il existe plusieurs façons de l'aborder, l'ACP peut être considérée comme une méthode qui permet de projeter des observations depuis l'espace à p dimensions des p variables initiales vers un espace à k dimensions (k < p) tel qu'un maximum d'information soit conservée. L'information est mesurée au travers de la variance.
Supposons que nous disposions de données de p variables sur n individus.
Nous travaillerons sur la matrice Z représentant les données des n observations des p variables **centrées** ; soit :

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{bmatrix} = [Z_1 \dots Z_p]$$

$(Z^T Z)$ est alors symétrique et semi-définie positive.
Nous allons considérer la transformation de Z en une nouvelle matrice Z^* dont les colonnes sont les composantes principales de Z. Ces composantes principales sont des combinaisons linéaires des p variables initiales, qui seront non corrélées deux à deux. La première composante principale aura la plus grande variance possible, la seconde étant, parmi les composantes principales qui ne sont pas corrélées avec la première, celle qui a la plus grande variance possible, et ainsi de suite.
Trouver la première de ces composantes principales z_1^* revient à trouver le vecteur v_1 tel que la projection des observations de la matrice Z sur v_1 ait une variance maximale. En adoptant la notation matricielle, z_1^* qui est la projection de l'échantillon de Z sur v_1 s'écrit : $z_1^* = Zv_1$.
Soit $z_1^* = \begin{bmatrix} z_{11}^* \\ \vdots \\ z_{n1}^* \end{bmatrix}$ avec $z_{i1}^* = v_{11}z_{i1} + v_{12}z_{i2} + \cdots + v_{1p}z_{ip}$ pour $i = 1, \dots, n$; et $v_1 = [v_{11} \dots v_{1p}]^T$ un vecteur à p éléments.

$z_1^* = Zv_1$
$= [Z_1 \dots Z_p]v_1$
$= \sum_{i=1}^{p} v_{1i}Z_i$

D'où $\mathbb{E}(z_1^*) = \sum_{i=1}^{p} v_{1i}\mathbb{E}(Z_i) = 0$ car $\mathbb{E}(Z_i) = 0$.

Nous cherchons à maximiser la variance de z_1^* : $Var(z_1^*) = \mathbb{E}(\, {z_1^*}^T z_1^*)$.
Ce qui revient à maximiser : ${z_1^*}^T z_1^* = (Zv_1)^T Zv_1 = {v_1}^T (Z^T Z)v_1$.
Nous devons imposer une contrainte au vecteur v_1, car sinon ${v_1}^T (Z^T Z)v_1$ peut-être aussi grand que l'on veut. Nous allons donc normer le vecteur v_1, ce qui revient à poser : ${v_1}^T v_1 = 1$.
Le problème ici est donc de maximiser ${v_1}^T (Z^T Z)v_1$ sous la contrainte que v_1 soit unitaire, ${v_1}^T v_1 = 1$.
Pour le faire, utilisons la méthode d'optimisation de Lagrange :

Soit la fonction Lagrangienne définie par : $\mathcal{L} = {v_1}^T(Z^TZ)v_1 + \lambda_1(1 - {v_1}^T v_1)$.
Avec $\lambda_1 \in \mathbb{R}$ multiplicateur de Lagrange, nous avons :

$$\begin{cases} \frac{\partial \mathrm{L}}{\partial v_1} = 0 \\ \frac{\partial \mathrm{L}}{\partial \lambda_1} = 0 \end{cases} \Longrightarrow \begin{cases} 2(Z^TZ)v_1 - 2\lambda_1 v_1 = 0 \\ {v_1}^T v_1 = 1 \end{cases}$$

$$\Longrightarrow \begin{cases} (Z^TZ)v_1 - \lambda_1 v_1 = 0 \\ {v_1}^T v_1 = 1 \end{cases}$$

$$\Longrightarrow \begin{cases} (Z^TZ)v_1 = \lambda_1 v_1 \\ {v_1}^T v_1 = 1 \end{cases}$$

Tout ceci montre que v_1 est le vecteur propre de (Z^TZ) associée à la valeur propre λ_1.
La variance de z_1^* est donc :

$Var(z_1^*) = \mathbb{E}(\ {z_1^*}^T z_1^*)$ avec
${z_1^*}^T z_1^* = {v_1}^T(Z^TZ)v_1$
$= {v_1}^T((Z^TZ)v_1)$
${z_1^*}^T z_1^* = {v_1}^T(\lambda_1 v_1)$
$= \lambda_1({v_1}^T v_1)$
$= \lambda_1$

D'où $(z_1^*) = {z_1^*}^T z_1^* = \lambda_1$.
Donc la solution optimale pour v_1 est donnée par le vecteur propre associé à la plus grande valeur propre de (Z^TZ).
Dans le cas où il n'y a pas de colinéarité stricte, la matrice sera définie positive et aura donc des valeurs propres strictement positives. $z_1^* = Zv_1$ est donc la première composante principale de Z.

Pour la seconde composante principale z_2^* :
Soit $z_2^* = Zv_2$ avec v_2 un vecteur à p éléments. Nous voulons choisir v_2 de façon à maximiser la variance de z_2^* sous contrainte que ${v_2}^T v_2 = 1$ et que z_1^* et z_2^* soient non-corrélés.
z_1^*et z_2^* sont non-corrélés $\Leftrightarrow \mathrm{Cov}(z_1^*, z_2^*) = 0$.
Nous avons aussi $\mathbb{E}(z_2^*) = 0$.
Calcul de $\mathrm{Cov}(z_1^*, z_2^*)$:

$$\begin{aligned} \mathrm{Cov}(z_1^*, z_2^*) &= \mathrm{Cov}(Zv_1\ , Zv_2\ ) \\ &= \mathbb{E}[(Zv_1)^T(Zv_2)] \\ &= \mathbb{E}[{v_1}^T(Z^TZ)v_2] \\ &= \mathbb{E}[((Z^TZ)v_1)^T v_2] \\ &= \mathbb{E}[(\lambda_1 v_1)^T v_2] \\ &= \lambda_1 {v_1}^T v_2 \end{aligned}$$

Ainsi $\mathrm{Cov}(z_1^*, z_2^*) = 0 \Rightarrow \lambda_1 {v_1}^T v_2 = 0$.
Comme $\lambda_1 > 0 \Rightarrow {v_1}^T v_2 = 0$.
Ce qui implique que v_1 et v_2 sont orthogonaux. Alors pour trouver v_2, nous devons résoudre le problème d'optimisation suivant : $\max({v_2}^T(Z^TZ)v_2)$ sous contraintes que ${v_2}^T v_2 = 1$ et ${v_1}^T v_2 = 0$.
Soit la fonction de Lagrange donnée par : $\mathcal{L} = {v_2}^T(Z^TZ)v_2\ + \lambda_2(1 - {v_2}^T v_2) + \gamma {v_1}^T v_2$ avec $\lambda_2, \gamma \in \mathbb{R}$.

Nous avons :

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial v_2} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda_2} = 0 \\ \frac{\partial \mathcal{L}}{\partial \gamma} = 0 \end{cases} \Rightarrow \begin{cases} 2(Z^T Z)v_2 - 2\lambda_2 v_2 + \gamma v_1 = 0 \\ 1 - {v_2}^T v_2 = 0 \\ {v_1}^T v_2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} (Z^T Z)v_2 + \frac{\gamma}{2} v_1 = \lambda_2 v_2 \\ {v_2}^T v_2 = 1 \\ {v_1}^T v_2 = 0 \end{cases}$$

En multipliant chaque membre de l'équation $(Z^T Z)v_2 + \frac{\gamma}{2} v_1 = \lambda_2 v_2$ par ${v_1}^T$, nous obtenons :

${v_1}^T (Z^T Z)v_2 + \frac{\gamma}{2} {v_1}^T v_1 = \lambda_2 {v_1}^T v_2$

$((Z^T Z)v_1)^T v_2 + \frac{\gamma}{2} {v_1}^T v_1 = \lambda_2 {v_1}^T v_2$

$\Rightarrow \lambda_1 {v_1}^T v_2 + \frac{\gamma}{2} = \lambda_2 {v_1}^T v_2$

$\Rightarrow \gamma = 2(\lambda_2 - \lambda_1){v_1}^T v_2$

$\Rightarrow \gamma = 0$ car ${v_1}^T v_2 = 0$.

Et donc la relation $(Z^T Z)v_2 + \frac{\gamma}{2} v_1 = \lambda_2 v_2$ devient $(Z^T Z)v_2 = \lambda_2 v_2$.

D'où v_2 est le vecteur propre de $(Z^T Z)$ associé à la deuxième plus grande valeur propre λ_2.

Et nous avons $Var(z_2^{*T} z_2^*) = {v_2}^T((Z^T Z)v_2) = \lambda_2$.

Pour les composantes principales restantes :

En utilisant l'induction, supposons que les $v_1, v_2, \dots, v_{i-1}$ sont des vecteurs propres unitaires de $(Z^T Z)$ associés à leur valeur propre $\lambda_1, \lambda_2, \dots, \lambda_{i-1}$ respectivement classées par ordre décroissant, et soit v_i le vecteur unitaire de la i[ième] composante principale $z_i^* = Z v_i$.

Nous obtenons v_i en résolvant le problème suivant : $\max({v_i}^T (Z^T Z)v_i)$ sous contraintes que ${v_i}^T v_i = 1$ et que les vecteurs $z_1^*, z_2^*, \dots, z_{i-1}^*$ et z_i^* soient non-corrélés deux à deux.

$(Z^T Z)v_j = \lambda_j v_j$ et ${v_i}^T (Z^T Z)v_j = \lambda_j v_i^T v_j = 0$, $\forall\, j = 1, \dots, i-1$.

Comme $\lambda_j > 0$, on a donc $v_i^T v_j = 0, \forall\, j = 1, \dots, i-1$.

Nous construisons le Lagrangien : $\mathcal{L} = {v_i}^T (Z^T Z)v_i + \lambda_i(1 - {v_i}^T v_i) + \sum_{j=1}^{i-1} \gamma_j {v_i}^T v_j$ avec $\lambda_i, \gamma_j \in \mathbb{R}$.

$\frac{\partial \mathcal{L}}{\partial v_i} = 0 \Rightarrow (Z^T Z)v_i + \sum_{j=1}^{i-1} \frac{\gamma_j}{2} v_j = \lambda_i v_i$

$\frac{\partial \mathcal{L}}{\partial \lambda_i} = 0 \Rightarrow {v_i}^T v_i = 1$

$\frac{\partial \mathcal{L}}{\partial \gamma_j} = 0 \Rightarrow v_i^T v_j = 0, \forall\, j = 1, \dots, i-1$

Nous obtenons ${v_j}^T (Z^T Z)v_i + \sum_{j=1}^{i-1} \frac{\gamma_j}{2} v_j^T v_j = \lambda_i {v_j}^T v_i$ par multiplication de la relation $(Z^T Z)v_i + \sum_{j=1}^{i-1} \frac{\gamma_j}{2} v_j = \lambda_i v_i$ par v_j^T

$\Rightarrow \lambda_j v_j^T v_i + (i-1)\frac{\gamma_j}{2} = \lambda_i v_j^T v_i$ (Car $(Z^T Z)v_j = \lambda_j v_j \Rightarrow v_j^T (Z^T Z) = \lambda_j v_j^T$)

$\Rightarrow \gamma_j = \frac{2}{(i-1)} (\lambda_i - \lambda_j) v_j^T v_i$

Or $v_j^T v_i = 0 \Rightarrow \gamma_j = 0, \forall\, j = 1, \dots, i-1$.

$(Z^T Z)v_i + \sum_{j=1}^{i-1} \frac{\gamma_j}{2} v_j = \lambda_i v_i$ devient $(Z^T Z)v_i = \lambda_i v_i$. Alors v_i est le vecteur propre de $(Z^T Z)$ associé avec la i[ième] plus grande valeur propre λ_i, orthogonal aux vecteurs $v_1, v_2, \dots, v_{i-1}$.
Dès lors que les valeurs propres de $(Z^T Z)$ sont distinctes, chaque vecteur propre v_i est unique (signé) d'où le résultat : $z_i^* = Zv_i$ est la i[ième] composante principale de Z.

Nous formons ainsi une matrice orthogonale avec les p vecteurs propres $v_1, v_2, \dots, v_p$ de $(Z^T Z)$ associés à ces p valeurs propres. Soit $V = [v_1, v_2, \dots, v_p]$
Les p composantes principales de Z sont alors données par la matrice Z^* de taille $n \times p$:

$$Z^* = ZV$$

En plus, $Z^{*T} Z^* = V^T (Z^T Z) V = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$

Ce qui montre bien que les composantes principales sont non-corrélées deux à deux et que leurs variances sont données par : $z_i^{*T} z_i^* = \lambda_i, \forall\, i = 1, \dots, p$.

7.4.2. Modèle Lee-Carter, comme régression sur la 1ère composante principale

Dans le modèle de Lee-Carter, nous avons :

- Une matrice Z avec des colonnes centrées,
- v_1 le premier vecteur propre normé de $Z^T Z$ associé à la plus grande valeur propre λ_1.

D'après l'exposé sur l'ACP (cf. section 7.4.1), si z_1^* est la première composante principale de Z, alors $z_1^* = Zv_1$.
Nous avions établi à la section 7.3 que sous la contrainte ${b_x}^T b_x = 1$, nous obtenons $\hat{k}_t = \sqrt{\lambda_1} u_1$.

Or d'après la 2ème formule de transfert établi à la section 7.3 : $Zv_1 = \sqrt{\lambda_1} u_1$

Donc :
$\hat{k}_t = \sqrt{\lambda_1} u_1 = Zv_1 = z_1^*$

Ainsi $\hat{k}_t$ est égal à la 1ère composante principale de Z (z_1^*) et donc par conséquent le modèle de Lee-Carter est une régression (avec un terme constant) sur z_1^* :
$ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t} = a_x + b_x z_1^* + \varepsilon_{x,t}$

Si nous avions utilisé la contrainte alternative $\sum_{x=x_m}^{x_M} b_x = 1$, nous aurions obtenu le résultat ci-après :
$\hat{k}_t = \sqrt{\lambda_1} \sum_j v_{1j} u_1 = Zv_1 \sum_j v_{1j} = z_1^* \sum_j v_{1j}$

Donc dans ce cas alternatif, $\hat{k}_t$ est égal à la 1ère composante principale de Z (notée z_1^*) à un facteur multiplicatif ($\sum_j v_{1j}$) près et donc par conséquent le modèle de Lee-Carter reste une régression (avec un terme constant) sur z_1^* :
$ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t} = a_x + b_x (z_1^* \sum_j v_{1j}) + \varepsilon_{x,t}$

8. Modèle interne PCR optimal

8.1. Régression sur composantes principales

La régression sur composantes principales (*Principal Component Regression, PCR*) a été introduite principalement pour gérer les problèmes de multicolinéarité entre variables explicatives. Nous en faisons ici un exposé car il nous sera utile pour proposer une généralisation à d'autres composantes principales du modèle de Lee-Carter.

8.1.1. Définition

La régression sur composantes principales (*Principal Component Regression, PCR*) est une analyse en régression sur les composantes d'une analyse en composantes principales ACP (Cornillon et al., 2011)
Soit :

$$Y = x_1\beta_1 + \cdots + x_p\beta_p + \varepsilon,$$

où

- $\beta_1, \dots, \beta_p$ sont des paramètres à estimer
- Y : variable expliquée, centrée réduite
- $x_1, \dots, x_p$ sont des variables explicatives, centrées réduites

Ce modèle peut être réécrit sous la forme :

$$Y = X\beta + \varepsilon \quad \text{avec} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{et} \quad X = [x_1, \dots, x_p].$$

La PCR est utilisée le plus souvent lorsque la matrice X n'est pas de plein rang ou lorsque $p > n$ avec n qui est le nombre d'observations.

8.1.2. Principe

La PCR introduit des composantes issues d'une ACP à la place des variables explicatives (X) et trouve un modèle équivalent :
$Y = X^*\beta^* + \varepsilon$
X^* étant la matrice des composantes principales.

La méthode se déroule en trois (3) phases :

- Analyse en composantes principales sur X

La matrice (X^TX) est une matrice symétrique donc diagonalisable, et elle peut s'écrire sous la forme

$X^TX = U \wedge U^T$ où U = Matrice des vecteurs propres normalisés de (X^TX) c'est-à-dire $(U^TU = UU^T = I)$.
$\wedge = \text{diag}(\lambda_1, \lambda_2 \dots, \lambda_p)$ est une matrice diagonale des valeurs propres $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p)$

Dans l'ACP, U est la matrice des axes principaux normés à l'unité et $\{\lambda_j\}_{j:1,\dots,p}$ sont les valeurs propres de X^TX.

En écrivant : $X = XUU^T$ *(*car $UU^T = I$*).*
Alors, l'équation $Y = X\beta + \varepsilon$ devient : $Y = XUU^T\beta + \varepsilon$

$Y = X^*\beta^* + \varepsilon$, où $\beta^* = U^T\beta$ et $X^* = XU$

Les colonnes de X^* sont les p composantes principales : $x_j^* = Xu_j$.
Le modèle $Y = X^*\beta^* + \varepsilon$ est appelé modèle de régression sur les composantes principales.
$X^{*^T}X^* = U^TX^TXU = U^TU \wedge U^TU = \wedge \Longrightarrow$ les colonnes de X^* sont orthogonales entre elles et de variance λ_j.

Régression linéaire de la variable indépendante sur les composantes
$x_j^* = Xu_j$ étant la j[ième] composante principale.

- Régression sur la 1[ère] composante

Soit $x_1^* = Xu_1$ où u_1 est un vecteur propre unitaire engendré par la plus grande valeur propre λ_1
$Y = x_1^*\beta_{(1)}^* + \varepsilon$ est le modèle de régression sur la 1[ère] composante principale x_1^*

Utilisons la méthode des moindres carrés pour estimer $\beta_{(1)}^*$
Soit à minimiser $S(\beta_{(1)}^*)$ avec
$S(\beta_{(1)}^*) = \|Y - x_1^*\beta_{(1)}^*\|^2$

$$= Y^TY - Y^Tx_1^*\beta_{(1)}^* - \beta_{(1)}^{T^*}x_1^{*^T}Y + \beta_{(1)}^{T^*}x_1^{*^T}x_1^*\beta_{(1)}^*$$

$$\frac{\partial S(\beta_{(1)}^*)}{\partial \beta_{(1)}^*} = 0 \Leftrightarrow -Y^Tx_1^* - x_1^{*^T}Y + 2x_1^{*^T}x_1^*\beta_{(1)}^* = 0$$
$$\Leftrightarrow -2x_1^{*^T}Y + 2x_1^{*^T}x_1^*\beta_{(1)}^* = 0$$

$$\hat{\beta}_{(1)}^* = \left(x_1^{*^T}x_1^*\right)^{-1}x_1^{*^T}Y$$

$\frac{\partial^2 S(\beta_{(1)}^*)}{\partial^2 \beta_{(1)}^*} = 2x_1^{*^T}x_1^* > 0$ car $x_1^{*^T}x_1^* = \|x_1^*\|^2$ donc $\hat{\beta}_{(1)}^*$ est bien un minimum strict.

- Régression sur k composantes ($k < p$).

$$Y = x_1^*\beta_1^* + \cdots + x_k^*\beta_k^* + \varepsilon$$

$\hat{\beta}^*(k) = \left(X_{[k,1]}^{*^T}X_{[1,k]}^*\right)^{(-1)}X_{[k,1]}^{*^T}Y$, de façon analogue à la régression sur une composante

Etudions la dépendance des paramètres estimés :

Soient $\hat{\beta}_i^*(k) = \left(x_i^{*^T}x_i^*\right)^{-1}x_i^{*^T}Y$ et $\hat{\beta}_j^*(k) = \left(x_j^{*^T}x_j^*\right)^{-1}x_j^{*^T}Y$

$$Cov(\hat{\beta}_i^*(k), \hat{\beta}_j^*(k)) = Cov\left[\left(x_i^{*^T}x_i^*\right)^{-1}x_i^{*^T}Y, \left(x_j^{*^T}x_j^*\right)^{-1}x_j^{*^T}Y\right]$$

$$= \mathbb{E}\left[\left(x_i^{*T} x_i^*\right)^{-1} x_i^{*T} Y Y^T x_j^* \left(x_j^{*T} x_j^*\right)^{-1}\right]$$
$$= \left(x_i^{*T} x_i^*\right)^{-1} x_i^{*T} \mathbb{E}(Y^T Y) x_j^* \left(x_j^{*T} x_j^*\right)^{-1}, \text{ car } Y^T Y = YY^T$$

$Cov\left(\hat{\beta}_i^*(k), \hat{\beta}_j^*(k)\right) = \left(x_i^{*T} x_i^*\right)^{-1} x_i^{*T} V(Y) x_j^* \left(x_j^{*T} x_j^*\right)^{-1}$, *où* $V(Y) = \mathbb{E}(Y^T Y) = Variance\ de\ Y$

Si $i \neq j \Rightarrow x_i^{*T} x_j^* = 0$ car x_i^* et x_j^* sont orthogonaux d'où $Cov\left(\hat{\beta}_i^*(k), \hat{\beta}_j^*(k)\right) = 0$.
Si $i = j$,

$$Cov\left(\hat{\beta}_i^*(k), \hat{\beta}_j^*(k)\right) = \left(x_i^{*T} x_i^*\right)^{-1} x_i^{*T} x_i^* V(Y) \left(x_i^{*T} x_i^*\right)^{-1}$$
$$= V(Y) \left(x_i^{*T} x_i^*\right)^{-1} \left(x_i^{*T} x_i^*\right) \left(x_i^{*T} x_i^*\right)^{-1}$$
$$= V(Y) \left(x_i^{*T} x_i^*\right)^{-1}, \textit{ car } \left(x_i^{*T} x_i^*\right) \left(x_i^{*T} x_i^*\right)^{-1} = 1$$
$$= \sigma^2 \left(x_i^{*T} x_i^*\right)^{-1}$$

$Cov\left(\hat{\beta}_i^*(k), \hat{\beta}_j^*(k)\right) = \sigma^2 \frac{1}{\lambda_i}$, car $x_i^{*T} x_i^* = \lambda_i$

D'où :

$$cov\left(\hat{\beta}_{i(k)}^*, \hat{\beta}_{j(k)}^*\right) = \sigma^2 \begin{cases} \frac{1}{\lambda_i} & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

- Calcul des paramètres de la régression en fonction des variables d'origine

Pour tout k fixé, le paramètre $\hat{\beta}_{(k)}^*$ estimé en PCR est :

$$\hat{\beta}_{PCR}^*(k) = \left(X_{[k,1]}^{*T} X_{[1,k]}^*\right)^{(-1)} X_{[k,1]}^{*T} Y^T$$

En utilisant la relation $\beta^* = U^T \beta$, on peut passer de $\hat{\beta}_{PCR}^*(k)$ à $\hat{\beta}_{PCR}(k)$:
Soit

$$\beta^* = U^T \beta$$

En multipliant chaque membre de $\beta^* = U^T \beta$ par U on a :

$$U\beta^* = UU^T \beta$$

$$U\beta^* = \beta \text{ car } (UU^T = I)$$

On a alors $\hat{\beta}_{PCR}(k) = U_{[1:k]} \hat{\beta}_{PCR}^*(k)$

8.2. Modèle Logit-PCR

Nous avions démontré (cf. section 7.4) que le modèle de Lee-Carter [$ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}$] est une régression sur la 1ère composante principale (à un facteur multiplicatif près) de la matrice $Z = ln(m_{x,t}) - \hat{a}_x$.

Le modèle CBD est basé sur les $\text{logit}\, q_x\,(t)$, aussi bien pour la variable à modéliser, que pour l'estimation des variables explicatives.
En combinant les approches Lee-Carter (et sa relation avec la régression PCR) et CBD, nous pouvons formuler un modèle de régression PCR emboité avec les k premières composantes principales et basé sur les $\text{logit}\, q_x\,(t)$:

$$\text{logit}\, q_x\,(t) = a_x + b_{x,1}x_1^*(t) + b_{x,2}x_2^*(t) + b_{x,3}x_3^*(t) + \cdots + b_{x,k}x_k^*(t) + \varepsilon_{x,t}$$

où :

- $q_x(t)$ est la probabilité de décès à l'âge x, au temps t
- a_x : moyenne des $\text{logit}\, q_x\,(t)$ pour l'âge x calculée sur l'ensemble de la période considérée
- x_i^* est la i ème composante principale de la matrice $\text{X} = \text{logit}\, q_x\,(t) - \hat{a}_x$
- $p = M - m + 1$
- $k\ (k \leq p)$: Nombre de premières composantes principales sélectionnées (par méthode *Foreward*).
- $\varepsilon_{x,t}$ est le résidu du modèle, supposé *i.i.d.*, d'espérance nulle et de variance σ_ε^2

Nous désignons ce modèle par : ***Logit-PCR*** .

8.3. Amélioration du modèle Logit-PCR

Boyer et al [2015] semblent associer le pouvoir explicatif (par rapport à la variable à expliquer) d'une composante principale, au pourcentage de variance qu'elle explique (du côté des variables explicatives).
Cette méthode de sélection des composantes principales à conserver dans le modèle, n'est pas optimale du point de vue explicatif ou prédictif, car la variable à expliquer n'intervient pas dans le choix de ces composantes.
Les améliorations envisagées ci-dessus se justifient d'autant plus que nous remarquons empiriquement des décalages temporels, des différences d'allure dans la dynamique des taux de mortalité selon les tranches d'âge.
Cela questionne l'utilisation d'une composante temporelle unique, quel que soit l'âge modélisé.

Si une composante temporelle unique reflète la transition démographique du point de vue de la mortalité globale, la prise en compte de la transition épidémiologique et sanitaire qui affecte les âges de manière différenciée (en débutant par les plus jeunes âges) nécessite des composantes temporelles différenciées par tranche d'âge.

L'objectif ici est d'améliorer le modèle Logit-PCR en utilisant une (ou "plusieurs" selon le nombre d'observations i.e. années d'historique disponibles) composante principale optimale, pour chaque tranche d'âge (ou âge) modélisée. On appellera ce modèle *Logit-PCR-**optimal*** .

Nous ne souhaitons pas conserver toutes les composantes : c'est une réduction de dimension qui se justifie souvent par le grand nombre de variables explicatives initiales (et donc de composantes principales calculées), leurs colinéarités potentielles, et aussi par le nombre limité d'observations utilisées pour l'estimation (apprentissage du modèle).

Il est donc nécessaire de sélectionner les k_x meilleures composantes principales à conserver dans le modèle de chaque âge x.
Nous pouvons proposer quelques méthodes classiques de sélection de variables explicatives :

a. Critères de sélection du modèle : R^2 ajusté, AIC, BIC ou MSEP. Nous utiliserons préférentiellement MSEP.
b. Procédure de sélection du modèle : *stepwise (*utilisé avec R^2 ajusté, AIC, BIC), ou *recherche exhaustive*. Nous utiliserons préférentiellement la recherche exhaustive.

Le nouveau modèle est désormais formulé séparément pour chaque âge :

$$\text{logit}\, q_x\,(t) = a_x + \beta^*_{c_1} x^*_{c_1} + \beta^*_{c_2} x^*_{c_2} + \cdots + \beta^*_{c_{k_x}} x^*_{c_{k_x}} + \varepsilon_{x,t}$$

où :

$q_x(t)$ est la probabilité de décès à l'âge x, au temps t
a_x : moyenne des $\text{logit}\, q_x\,(t)$ pour l'âge x calculée sur l'ensemble de la période considérée
$x^*_{c_i}$ est une composante principale de la matrice $\mathrm{X} = \text{logit}\, q_x\,(t) - \hat{a}_x$ (il ne s'agit pas de la $i^{\text{ème}}$ composante principale)
$p = M - m + 1$
k_x $(k_x \leq p)$: Nombre de premières composantes principales sélectionnées.
$\varepsilon_{x,t}$ est le résidu du modèle, supposé *i.i.d.*, d'espérance nulle et de variance σ^2_ε

Nous désignons ce modèle par : ***Logit-PCR-optimal***.

Tableau 1 : Comparaison théorique des modèles Logit-PCR et Logit-PCR-Optimal

Etapes	**Modèle Logit-PCR**	**Modèle Logit-PCR-Optimal**
Critère du choix du modèle	MAPE, MSE obtenus par validation croisée	MAPE, MSE obtenus par validation croisée
Procédure de sélection du modèle	Méthode emboîtée (*foreward selon la variance expliquée par composante*)	Recherche exhaustive ou *stepwise*
Modèles à estimer	$\text{logit}\, q_x\,(t) = a_x + b_{x,1} x^*_1(t) + \cdots + b_{x,k} x^*_k(t) + \varepsilon_{x,t}$	$\text{logit}\, q_x\,(t) = a_x + \beta^*_{c_1} x^*_{c_1} + \beta^*_{c_2} x^*_{c_2} + \cdots + \beta^*_{c_{k_x}} x^*_{c_{k_x}} + \varepsilon_{x,t}\,, x = x_m, \ldots, x_M$

Tests de validation du modèle
Idéalement, chaque modèle, avant son utilisation, devrait être validé par les tests suivants :

a. Test d'homoscédasticité des erreurs (test de Harrison-McCabe)
Hypothèse (H_0) :$var(\varepsilon_t) = \sigma^2$ (homoscédasticité) contre (H_1): $var(\varepsilon_t) \neq \sigma^2$(hétéroscédasticité)

b. Test d'autocorrélation des erreurs (test de Breusch-Godfrey)
Hypothèse (H_0):$Cov(\varepsilon_t, \varepsilon_s) = 0$ (non-autocorrélation) contre (H_1): $Cov(\varepsilon_t, \varepsilon_s) \neq 0$ (autocorrélation) $\forall\, t \neq s$

c. Test de normalité des erreurs (test de Jarque-Bera)
Hypothèse (H_0) :ε_t suivent la loi normale contre (H_1): ε_t ne suivent pas la loi normale $\forall\, t$
L'hypothèse de normalité n'est pas indispensable pour valider le modèle.

Si les résidus sont auto-corrélés, il sera nécessaire de les modéliser par un processus ARMA de sorte à obtenir un bruit blanc.

9. Modèle interne PLS

9.1. Régression PLS

Comme la régression sur composantes principales (*Principal Component Regression, PCR*), la régression PLS (*Partial Least Square*) a été introduite principalement pour gérer les problèmes de multicolinéarité entre variables explicatives, à la nuance que ses composantes (appelées composantes PLS) sont calculées en tenant compte de leurs corrélations élevées avec la variable à expliquer. Nous en faisons ici un exposé car il nous sera utile pour proposer une généralisation du modèle de Lee-Carter basée sur les composantes PLS.

Soient les nouvelles variables explicatives $l^{(1)}, l^{(2)}, \dots, l^{(k)}$, combinaisons linéaires des variables de départ, qui sont orthogonales entre elles et classées par ordre décroissant : $l^{(j)} = Xc_j$.
Nous constatons que les composantes principales X_j^* de la matrice X déterminées dans le cadre de l'ACP et du PCR obéissent également à ces critères.

Nous déroulons un exemple de calcul des $l^{(j)}$ lorsque Y est univarié.
Dans le contexte Y univarié, la régression PLS est appelée PLS1 et elle se définit itérativement. Pour cela, une procédure itérative va être utilisée (Cornillon et al. [2011]).

Etape 1 : La matrice X est notée $X^{(1)}$ et Y la variable à expliquer est notée $Y^{(1)}$.
La 1ère composante (ou variable latente) $l^{(1)} = X^{(1)}w_1 \in \mathbb{R}^n$ est choisie telle que :

$l^{(1)} = \underset{l=X^{(1)}w_1, w_1 \in \mathbb{R}^P, \|w_1\|^2=1}{\operatorname{argmax}} \operatorname{cov}\left(l, Y^{(1)}\right)$, où w_1 est un vecteur de poids dans $\mathbb{R}^P$
C'est-à-dire maximiser $\operatorname{cov}\left(l, Y^{(1)}\right)$ sous contrainte $\|w_1\|^2 = 1$.
Pour le faire, utilisons la fonction Lagrangienne qui est définie par :

$$L = \left(Y^{(1)}\right)^T X^{(1)} w_1 + \lambda(1 - \|w_1\|^2)$$

On a alors :
$\frac{\partial L}{\partial w_1} = \left(Y^{(1)}\right)^T X^{(1)} - 2\lambda w_1 = 0$ et $\frac{\partial L}{\partial \lambda} = (1 - \|w_1\|^2) = 0$

L'équation $\left(Y^{(1)}\right)^T X^{(1)} - 2\lambda w_1 = 0$ montre que w_1 est colinéaire au vecteur $\left(Y^{(1)}\right)^T X^{(1)}$ et $\|w_1\|^2 = 1$ montre qu'il est normé. Si l'on veut un maximum, il suffit de prendre le vecteur

$$w_1 = \frac{\left(Y^{(1)}\right)^T X^{(1)}}{\left\|\left(Y^{(1)}\right)^T X^{(1)}\right\|}, \text{ avec } \lambda = \frac{1}{2}.$$

Ensuite, nous effectuons la régression simple de $Y^{(1)}$ sur $l^{(1)}$, $Y^{(1)} = r_1 l^{(1)} + \hat{\varepsilon}_1$ où $r_1 \in \mathbb{R}$ est le coefficient de la régression estimée par moindres carrés et $\hat{\varepsilon}_1 = P_{l^{(1)\perp}} Y^{(1)}$ est le vecteur des résidus de la régression simple sans constante, où $P_{l^{(1)\perp}}$ est l'opérateur de projection orthogonale sur la droite engendrée par $l^{(1)}$ et donné par $P_{l^{(1)\perp}} = (I - P_{l^{(1)}})$ et $P_{l^{(1)}} = l^{(1)}({l^{(1)}}^T l^{(1)})^{-1} {l^{(1)}}^T$

Etape 2 : On note $Y^{(2)} = \hat{\varepsilon}_1 = P_{l^{(1)}\perp} Y^{(1)}$, le résidu de $Y^{(1)}$ sur $l^{(1)}$ ou soit la partie encore non expliquée de Y.
Soit $X^{(2)} = P_{l^{(1)}\perp} X^{(1)}$ la partie de $X^{(1)}$ n'ayant pas encore servi à expliquer, ou soit le vecteur des résidus des régressions des variables de X sur $l^{(1)}$.
La seconde composante PLS $l^{(2)} = X^{(2)} w_2 \in \mathbb{R}^n$ est choisie telle que :

$$l^{(2)} = \underset{l = X^{(2)} w_2, w_2 \in \mathbb{R}^P, \|w_2\|^2 = 1}{\operatorname{argmax}\, \operatorname{cov}} \left(l, Y^{(2)}\right)$$

En faisant le même raisonnement que dans l'étape 1, on a :

$$w_2 = \left(Y^{(2)}\right)^T X^{(2)} / \left\| \left(Y^{(2)}\right)^T X^{(2)} \right\|$$

Nous effectuons ensuite la régression simple de $Y^{(2)}$ sur $l^{(2)}$; $Y^{(2)} = r_2 l^{(2)} + \hat{\varepsilon}_2$ où $r_2 \in \mathbb{R}$ est le coefficient de la régression estimée par les moindres carrés et $\hat{\varepsilon}_2 = P_{l^{(2)}\perp} Y^{(2)}$ est le vecteur des résidus de la régression simple sans constante.

Nous poursuivons le même calcul, jusqu'à la $k^{ième} - 1$ étape.
Etape k : On pose $Y^{(k)} = \hat{\varepsilon}_{k-1} = P_{l^{(k-1)}\perp} Y^{(k-1)}$ la partie non encore expliquée de Y et $X^{(k)} = P_{l^{(k-1)}\perp} X^{(k-1)}$ la partie de $X^{(k-1)}$ n'ayant pas servi à expliquer la composante PLS $l^{(k)}$.
La composante PLS k; $l^{(k)} = X^{(k)} w_k \in \mathbb{R}^n$ est choisie telle que :

$$l^{(k)} = \underset{l = X^{(k)} w_k, w_k \in \mathbb{R}^P, \|w_k\|^2 = 1}{\operatorname{argmax}\, \operatorname{cov}} \left(l, Y^{(k)}\right)$$

Soit par le même raisonnement que dans les étapes précédentes, on a :

$$w_k = \left(Y^{(k)}\right)^T X^{(k)} / \left\| \left(Y^{(k)}\right)^T X^{(k)} \right\|$$

Ensuite, nous effectuons la régression simple de $Y^{(k)}$ sur $l^{(k)}$: $Y^{(k)} = r_k l^{(k)} + \hat{\varepsilon}_k$ où $r_k \in \mathbb{R}$ est le coefficient de la régression estimée par les moindres carrés et $\hat{\varepsilon}_k = P_{l^{(k)}\perp} Y^{(k)}$.
En utilisant le postulat que les composantes PLS sont orthogonales entre elles, le modèle PLS s'écrit sous la forme suivante :

$$Y = P_{l^{(1)}} Y^{(1)} + \cdots + P_{l^{(k)}} Y^{(k)} + \hat{\varepsilon}_k$$

$$Y = r_1 l^{(1)} + \cdots + r_k l^{(k)} + \hat{\varepsilon}_k \text{ , avec } \hat{\varepsilon}_k = P_{l^{(k)}\perp} Y^{(k)} = P_{B\left(l^{(1)}, \ldots, l^{(k)}\right)^\perp} Y$$

9.2. Améliorations du modèle Logit-PCR en utilisant la régression PLS

La finalité ici est d'améliorer le modèle Logit-PCR en utilisant la régression PLS et en choisissant de manière optimale (MSEP par validation croisée) les composantes principales PLS pour chaque tranche d'âge (ou âge) modélisée. On appellera ce modèle ***logit-PLS*** .

A l'image de la régression sur composantes principales, l'objectif ici est de trouver pour chaque âge x le meilleur modèle à k_x composantes PLS.
Nous rappelons que ces composantes PLS sont notées $l^{(1)}, l^{(2)}, \dots, l^{(k)}$, combinaisons linéaires des variables de départ $l^{(j)} = Xc_j$, et qui soient orthogonales entre elles et classées par ordre d'importance, c'est-à-dire leurs liens avec la variable à expliquer. Cependant, le calcul de ces composantes $l^{(j)}$ se fait par leur lien avec la variable à expliquer et non par la variabilité qu'elles représentent parmi les variables explicatives originales (comme en régression sur composantes principales).
Et le modèle sera donc défini par :

$\text{logit}\, q_x\,(t) = r_1 l^{(1)} + \cdots + r_{k_x} l^{(k_x)} + \varepsilon_{k_x}$, qui sera appelé **Logit-PLS**

où $l^{(j)} = \underset{l=X^{(j)}w, w\in\mathbb{R}^P, \|w\|^2=1}{\text{argmax cov}} \left(l, Y^{(j)}\right)$ et $r_j \in \mathbb{R}\ \forall\ 1 \le j \le k_x$ est le coefficient de la régression estimé par les moindres carrés et $\hat{\varepsilon}_{k_x} = P_{l^{(k_x)}} \perp Y^{(k_x)}$. Pour plus de détails sur la procédure itérative des composantes PLS .

Tableau 2 : Comparaison théorique des modèles Logit-PCR-Optimal et Logit-PLS

Etapes	**Modèle Logit-PCR-Optimal**	**Modèle Logit-PLS**
Critère du choix du modèle	MAPE, MSE par validation croisée	MAPE, MSE par validation croisée
Procédure de sélection du modèle	Recherche exhaustive ou *stepwise*	Méthode emboîtée (intégration progressive des composantes selon la variance expliquée)
Modèles à estimer	$\text{logit}\, q_x\,(t) = a_x + \beta^*_{c_1} x^*_{c_1} + \beta^*_{c_2} x^*_{c_2} + \cdots + \beta^*_{c_{k_x}} x^*_{c_{k_x}} + \varepsilon_{x,t}$, $x = x_m, \dots, x_M$	$\text{logit}\, q_x\,(t) = r_1 l^{(1)} + \cdots + r_{k_x} l^{(k_x)} + \varepsilon_{k_x}$, $x = x_m, \dots, x_M$

10. Modèles mixtes cliométriques PCR-Optimal et PLS

Certains pays sont plus « *précoces »* dans la transition démographique que d'autres qui sont alors considérés comme étant en situation de « *rattrapage »* relatif. Une originalité de cette partie consiste à utiliser l'histoire quantitative des mortalités des pays transitionnellement « plus âgés », pour tenter d'améliorer les prévisions à long terme de mortalité dans les pays transitionnellement « moins âgés ». Nous faisons cela au travers d'une série *temporelle cliométrique de mortalité* (*en abrégé série cliométrique de mortalité*).
En pratique, ces améliorations pourraient s'appliquer aux études prospectives de la mortalité dans les pays émergents ou en développement.

Nous avons défini le *temps transitionnel* comme étant une série temporelle qui a une « forte » concordance statistique avec le temps. De même, nous avons défini l'*âge transitionnel* d'un pays, comme étant la différence entre le temps transitionnel et sa valeur minimale (quel que soit le pays) dans la période de transition. Cette valeur minimale est un seuil de passage, quel que soit le pays.

Le temps transitionnel, permet de faire un appariement temporel entre le pays à modéliser et le pays transitionnellement « plus âgé ».

La série *cliométrique de mortalité à l'âge* x du pays R par rapport au pays A ($C_{R,A,x}$) est une série créée à partir des données historiques du pays A transitionnellement « plus âgé », pour jouer un rôle de série exogène dans les modèles de prévisions de mortalité du pays R. Cette série cliométrique exogène de mortalité sera construite de sorte à avoir la même allure temporelle que dans les pays « transitionnellement plus âgés », mais adaptée au rythme du pays à modéliser.

La série cliométrique sera ainsi utilisée comme série explicative supplémentaire dans les modèles prospectifs de mortalité du pays R. Nous nommons les modèles utilisant la série cliométrique, modèles cliométriques : nous en proposons deux types qui sont les modèles PCR-optimal-cliometric et PLS-optimal-cliometric.

10.1. Temps transitionnel

Une originalité et apport de notre approche consiste à développer une notion de temps transitionnel différent du temps calendaire, qui s'écoule au rythme des phases et évènements transitionnels et est mesuré à partir d'une série temporelle spécifique.
Ce temps transitionnel nous permet de comparer la trajectoire d'entités (géographiques ou autres) différentes traversées par un même phénomène transitionnel, mais qui se déroule à des périodes historiques (donc calendaires) différentes.

Techniquement, nous définissons le temps transitionnel, comme une série temporelle qui a une dépendance monotone forte avec le temps. Nous mesurons cette dépendance monotone (concordante ou discordante) par le coefficient de corrélation des rangs de Kendall τ, permettant ainsi de tester l'hypothèse qu'une série temporelle donnée puisse être utilisée comme temps transitionnel.

Nous définissons l'âge transitionnel d'un pays, comme différence entre le temps transitionnel et sa valeur minimale internationale dans la période de transition.
Ces notions sont détaillées dans la suite de cette section.

10.1.1. Propriété de concordance du subordinateur du temps

Une série temporelle X est un subordinateur du temps si et seulement si X est un processus croissant (à variations bornées).
Si X est un subordinateur du temps, alors X est parfaitement *concordant* avec le temps, i.e. que le coefficient de corrélation des rangs de Kendall (entre X et le temps) est égal à 1. Car les rangs (du temps et de son subordinateur X) suivent le même ordre que les valeurs du temps et de son subordinateur X.
Rappel : le coefficient de corrélation des rangs de Kendall, mesure le niveau de *concordance* entre deux classements (rangs).
Le coefficient de corrélation de rangs est également un coefficient de dépendance monotone car il est invariant pour toute transformation monotone croissante des variables.

10.1.2. Coefficient de corrélation des rangs de Kendall et tendance temporelle monotone

Lecoutre, Tassi [1987], pp. 151-154, ont montré que la tendance temporelle monotone d'une série temporelle X, peut se mesurer par le coefficient de corrélation des rangs de Kendall τ. Ils ont développé également un test de tendance monotone d'une série temporelle X, basée sur la loi asymptotique (normale centrée) du coefficient de corrélation des rangs de Kendall τ.

Pour savoir si deux séries temporelles X et Y définies sur un même intervalle de temps, varient dans le même sens ou en sens contraire, on peut étudier le signe du produit $\left(X_{\mathrm{i}} - X_{\mathrm{j}}\right)\left(Y_{\mathrm{i}} - Y_{\mathrm{j}}\right)$ où $(X_{\mathrm{i}}, Y_{\mathrm{i}})\ et\ \left(X_{\mathrm{j}}, Y_{\mathrm{j}}\right)$ sont deux réalisations indépendantes du couple (X, Y). Avec *i* et *j* des entiers naturels non-nuls distincts.

Si $\left(\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right) = 1/2$, il y a autant de chances d'observer une variation de X et Y dans le même sens (concordance) que dans des sens contraires (discordance).

Saporta [2011], pp. 138-139, définit théoriquement le coefficient de corrélation des rangs de Kendall τ par la différence entre la probabilité de concordance (p_c) et la probabilité de discordance (p_d):

$$\tau = P\left(\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right) - P\left(\left(X_i - X_j\right)\left(Y_i - Y_j\right) < 0\right) = p_c - p_d$$

Donc : $\tau = p_c - (1 - p_c) = 2.p_c - 1$

Ce coefficient est donc compris entre -1 (discordance parfaite) et +1 (concordance parfaite) et s'annule lorsque X et Y sont indépendantes (mais pas seulement dans ce cas).

Essayons de mesurer à présent la monotonie (croissante ou décroissante) de X (continue, sans exæquos) dans le temps.
Cette dépendance monotone entre X et la variable déterministe de temps (que nous supposons discret) peut se mesurer par le coefficient de corrélation des rangs de Kendall τ (Lecoutre, Tassi [1987], pp. 151-154).

Le temps est une variable déterministe monotone croissante par définition.
Donc toute série temporelle concordante avec le temps, est monotone croissante.
Nous pouvons donc écrire :
Probabilité de concordance $p_c = P\left(\left(X_i - X_j\right) > 0\right)$, $\forall\, i > j$
Et probabilité de discordance $p_d = P\left(\left(X_i - X_j\right) < 0\right)$, $\forall\, i > j$

Or $\left\{\left(X_i - X_j\right) > 0, \forall\, i > j\right\} = \{(X_i - X_{i-1}) > 0, \forall\, i\}$
Et $\left\{\left(X_i - X_j\right) < 0, \forall\, i > j\right\} = \{(X_i - X_{i-1}) < 0, \forall\, i\}$

Donc
$p_c = P\left((X_i - X_{i-1}) > 0\right), \forall\, i$
$p_d = P\left((X_i - X_{i-1}) < 0\right), \forall\, i$

D'où :

$$\tau = p_c - p_d = P\left((X_i - X_{i-1}) > 0\right) - P\left((X_i - X_{i-1}) < 0\right) = 2.P\left((X_i - X_{i-1}) > 0\right) - 1$$

Empiriquement, pour un échantillon de taille n on a : $p_c = \frac{\sum_{i=1}^{n} \mathbb{1}_{R_+^*}(X_i - X_{i-1})}{n}$
Donc :

$$\tau = 2.\frac{\sum_{i=1}^{n} \mathbb{1}_{R_+^*}(X_i - X_{i-1})}{n} - 1$$

10.1.3. Temps transitionnel

L'objectif ici est de définir un *temps transitionnel* universel, qui permet de situer chaque pays dans son processus de transition (démographique, ou économique par exemple), à un moment (temps) donné de son histoire.

Nous rappelons la définition du *temps transitionnel* comme étant une série temporelle caractérisée par une forte concordance (dépendance monotone) statistique avec le temps calendaire.
Nous faisons l'hypothèse statistique d'une concordance *forte*, pour les cas où le coefficient de corrélation des rangs de Kendall τ entre la série temporelle testée (pour jouer le rôle de temps transitionnel) et le temps, est significatif (au seuil de 5%) et supérieur ou égal à 50% (en nous inspirant de l'échelle de Cohen, Cohen [1988], p. 227).

Remarque : Toute transformée symétrique (par rapport à l'axe horizontal du temps, et à une transformation monotone près) d'une série temporelle discordante au temps, est concordante au temps, donc peut jouer le rôle de *temps transitionnel* (si la discordance est forte).
Dans notre quête d'un *temps transitionnel*, nous utiliserons la propriété ci-dessus pour rechercher les séries temporelles ayant une concordance ou une discordance forte (donc une monotonie forte) avec le temps.

10.1.4. Age transitionnel

Nous définissons l'âge transitionnel d'un pays, comme différence entre le temps transitionnel et sa valeur minimale (quel que soit le pays) dans la période de transition. Cette valeur minimale [notée $Min(X)$] est

un seuil de passage, i.e. un évènement définissant le début théorique de la période transitoire, quel que soit le pays.
Soit Δ la série temporelle mesurant l'âge transitionnel d'un pays (qui varie donc dans le temps). On a :

$$\Delta = X - Min(X)$$

où X est une série temporelle de temps transitionnel.

Remarque : Si $X^{(s)}$ est une transformée symétrique (par rapport à l'axe horizontal du temps, et à une transformation monotone près) du temps transitionnel X, alors il est possible de donner une définition d'âge transitionnel par :

$$\Delta = -X^{(s)} - \left(-Max\left(X^{(s)}\right)\right) = Max\left(X^{(s)}\right) - X^{(s)}$$

$Max\left(X^{(s)}\right)$ étant un seuil indépassable durant la période de transition.

10.2. Concept de série temporelle cliométrique

Nous cherchons à modéliser le taux de mortalité d'un pays R, pour un âge donné x : $\mathrm{m}_{\mathrm{R,x}}$.

Pour cela nous allons considérer une série temporelle créée à partir des données historiques d'un pays A transitionnellement « plus âgé » que le pays R, pour jouer un rôle de série exogène dans les modèles de prévisions de mortalité du pays R. Nous allons dénommer cette série : série temporelle cliométrique de mortalité à l'âge x du pays R par rapport au pays A et allons la noter $\mathcal{C}_{R,A,x}$.

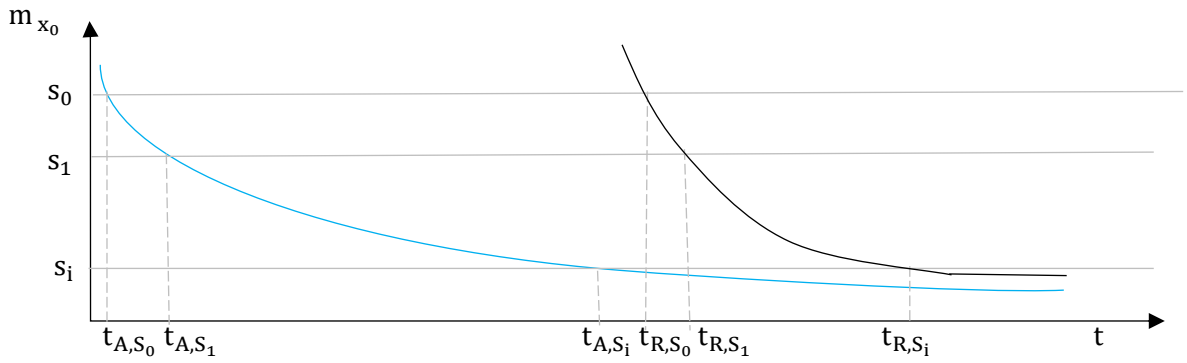
L'idée est d'utiliser l'histoire quantitative des mortalités du pays A pour tenter d'améliorer les prévisions de mortalité dans le pays R. Une telle série peut être utilisée comme série explicative supplémentaire dans les modèles prospectifs de mortalité du pays R.
Nous présentons ici la méthodologie d'estimation de cette série *cliométrique*.

10.3. Méthodologie de création d'une série explicative cliométrique

Pour créer cette série explicative cliométrique, nous allons chercher d'abord à estimer une fonction qui associe chaque point temporel du processus de transition du pays A et chaque point temporel du processus de transition du pays R.

Soient $\mathrm{m}_{\mathrm{A},\,\mathrm{x}_0} = \left(\mathrm{m}_{\mathrm{A},\,\mathrm{x}_0,t}\right), \mathrm{t} \in \mathbb{N}$ et $\mathrm{m}_{\mathrm{R},\,\mathrm{x}_0} = \left(\mathrm{m}_{\mathrm{R},\,\mathrm{x}_0,t}\right), \mathrm{t} \in \mathbb{N}$ deux séries temporelles de mortalités à l'âge x_0 respectivement des pays A et R. Les symétriques (par rapport à l'axe horizontal du temps, et à une transformation monotone près) de ces séries temporelles sont des temps transitionnels.

Figure 19 : Courbes de mortalité des pays A (bleue) et R (noire)



Les s_i sont des valeurs seuils du taux de mortalité m_{x_0},
$t_{A,s_i} = \inf\{t \in \mathbb{N}, m_{A,\, x_0,t} \leq s_i\}$ et $t_{R,s_i} = \inf\{t \in \mathbb{N}, m_{R,\, x_0,t} \leq s_i\}$.
Soient $t_A = (t_{A,s_i}), i \in \mathbb{N}$ et $t_R = (t_{R,s_i}), i \in \mathbb{N}$ deux séries des dates calendaires d'atteinte ou de franchissement (à la baisse) des valeurs de mortalité à l'âge x_0, $(s_i), i \in \mathbb{N}$, respectivement dans les pays A et R.

Les séries temporelles m_{A,x_0} et $m_{R,\, x_0}$ ont les propriétés suivantes :

- La monotonie en tendance temporelle (nécessaire au rôle de temps transitionnel tel que préalablement défini).
- Des valeurs dans un même intervalle dans les deux pays (pour permettre les appariements entre les points temporels des 2 processus liés aux deux pays).

Du fait de leur monotonie statistique par rapport au temps, ces séries nous permettent de faire un appariement entre les points du processus transitionnel du pays A et celui du pays R.

Dans ce qui suit, nous proposons deux méthodes distinctes pour la détermination de la fonction d'appariement des points temporels calendaires des processus de transition des deux pays A et R.
Nous utiliserons cette fonction d'appariement pour créer la série *cliométrique de mortalité à l'âge x* du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$).

Nous comparerons les performances de ces deux méthodes par la suite.

10.3.1. Estimation de la série cliométrique par une approche empirique des taux de mortalité

Pour estimer la série cliométrique, nous utilisons les données sans faire l'hypothèse d'une fonction explicite du taux de mortalité à l'âge x_0.

Soient $m_{A,x} = (m_{A,x,t}), t \in \mathbb{N}$ et $m_{R,x} = (m_{R,x,t}), t \in \mathbb{N}$ deux séries temporelles de mortalités à l'âge x respectivement des pays A et R.
Nous avons le tableau initial suivant pour le pays A :

Tableau 3 : Tableau du pays A. Cas de l'approche empirique

$m_{A,\,x_0}$	t_A	$m_{A,x}$
s_0	t_{A,s_0}	$m_{A,x,\,t_{A,s_0}}$
s_1	t_{A,s_1}	$m_{A,x,\,t_{A,s_1}}$
⋮	⋮	⋮
s_i	t_{A,s_i}	$m_{A,x,\,t_{A,s_i}}$
⋮	⋮	⋮
s_d	t_{A,s_d}	$m_{A,x,\,t_{A,s_d}}$
s_{d+1}	$t_{A,s_{d+1}}$	$m_{A,x,\,t_{A,s_{d+1}}}$
⋮	⋮	⋮
s_r	t_{A,s_r}	$m_{A,x,\,t_{A,s_r}}$

d : Nombre d'observations disponibles de la série temporelle $m_{R,\,x_0}$ après le temps t_{R,s_0} .
$m_{A,\,x_0} = (s_i\,, 0 \leq i \leq r)^T$, avec $r, d \in \mathbb{N}^*,\ r > d$ et $s_r < s_d$ (car la symétrique de $m_{A,\,x_0}$ par rapport à l'axe horizontal du temps [à une transformation monotone près] est un temps transitionnel).

Dans le cas présent t_{R,s_i} est le premier temps d'atteinte ou de franchissement (à la baisse) de la valeur s_i par la série m_{R,x_0} :

$$t_{R,s_i} = \inf\{t \in \mathbb{N}, m_{R,\,x_0,t} \leq s_i\},\ i = 0, \dots, d.$$

Pour estimer t_{R,s_i} sur l'intervalle $[t_{R,s_{d+1}}; t_{R,s_r}]$ nous estimons un modèle linéaire entre t_R et t_A en utilisant les valeurs connues de ses deux variables.
L'écart entre les deux séries temporelles t_R et t_A n'étant pas constant dans le temps pour une même valeur de s_i, on peut écrire :

$$t_{R,s_i} - t_{A,s_i} = \alpha_0 + \alpha_1 t_{A,s_i} + \mathcal{E}_{s_i}, \text{ avec } \alpha_0 > 0 \text{ et } \alpha_1 < 0.$$

Le modèle final est donné par :

$$t_{R,s_i} = \alpha_0 + (\alpha_1 + 1) t_{A,s_i} + \mathcal{E}_{s_i}, \text{ avec } \alpha_0 > 0 \text{ et } \alpha_1 < 0.$$

Les $\mathcal{E}_{s_i}$ sont des erreurs *i.i.d.* qui doivent respecter les conditions suivantes : espérance nulle, homoscédasticité, non-autocorrélation ; autrement dit ce sont des bruits blancs.
Si les résidus suivent les conditions usuelles du modèle linéaire, alors nous utilisons le modèle fourni par la régression pour réaliser les prévisions. Sinon, les modèles ne sont pas valides et il sera nécessaire de compléter le modèle en modélisant les résidus par un ARMA (cf. Boutaharet Royer-Carenzi [2019], p.148-158).

Puisque les t_{R,s_i} sont des entiers (années), nous prendrons la valeur arrondie à l'entier le plus proche.

Contrainte de calcul : $t_{A,s_i} \leq t_{R,s_i}$. Dans le cas où : $t_{R,s_i} - t_{A,s_i} \leq 0$, on fixe $t_{R,s_i} = t_{A,s_i}$.
Interprétation de la contrainte de calcul ci-dessus : On n'autorise pas le pays R à atteindre la valeur s_i ($\forall\ i$) avant le pays A, car le pays R est supposé en rattrapage transitionnel par rapport au pays A.
Cette contrainte pourra être relâchée dans une approche alternative.

Nous souhaitons à présent apparier chaque observation du pays A avec une observation du pays R.
A chaque observation au temps t_{A,s_i} du pays A, nous associons une observation au temps t_{R,s_i} du pays R.

Après association des observations du pays R à celles du pays A, nous obtenons le tableau suivant :

Tableau 4 : Tableau d'appariement des tables du pays A et du pays R

m_{A,x_0}	t_A	$m_{A,x}$	t_R	$m_{R,x}$
s_0	t_{A,s_0}	$m_{A,x,\, t_{A,s_0}}$	t_{R,s_0}	$m_{R,x,\, t_{R,s_0}}$
s_1	t_{A,s_1}	$m_{A,x,\, t_{A,s_1}}$	t_{R,s_1}	$m_{R,x,\, t_{R,s_1}}$
⋮	⋮	⋮	⋮	⋮
s_i	t_{A,s_i}	$m_{A,x,\, t_{A,s_i}}$	t_{R,S_i}	$m_{R,x,\, t_{R,s_i}}$
⋮	⋮	⋮	⋮	⋮
s_d	t_{A,s_d}	$m_{A,x,\, t_{A,s_d}}$	t_{R,S_d}	$m_{R,x,\, t_{R,s_d}}$
s_{d+1}	$t_{A,s_{d+1}}$	$m_{A,x,\, t_{A,s_{d+1}}}$	$t_{R,S_{d+1}}$	NA
⋮	⋮	⋮	⋮	⋮
s_r	t_{A,s_r}	$m_{A,x,\, t_{A,s_r}}$	t_{R,S_r}	NA

La mention NA désigne les valeurs manquantes.

Les points de la série *explicative cliométrique à l'âge x* du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) sont définis par les couples (t_{R,S_i} , $m_{A,x,\, t_{A,s_i}}$), $i = 0, \dots, r$:

$$\mathcal{C}_{R,A,x,\, t_{R,S_i}} : t_{R,S_i} \mapsto m_{A,x,\, t_{A,s_i}}$$

Si les arrondis de t_{R,S_i} donnent des valeurs identiques pour des valeurs différentes de $m_{A,x}$, la moyenne des valeurs de $m_{A,x}$ est associée aux valeurs identiques de t_{R,S_i} , lors de la création des points de la série explicative cliométrique.

Pour éviter les valeurs manquantes dans la série explicative cliométrique, dues aux valeurs manquantes dans la série temporelle t_R (qui devrait contenir des valeurs consécutives des années), une interpolation linéaire (ou moyenne mobile) pourra être utilisée. Les tests avec l'interpolation spline se sont révélés non-concluants car elle générait des valeurs atypiques.
Notons :

T le temps (année) présent (e).

H l'horizon maximal de prévision envisagé.

Si l'horizon $T + H > t_{A,s_r}$ alors nous aurons recours aux prévisions par séries temporelles pour estimer les valeurs de la série cliométrique sur la période $[\, t_{A,s_r} + 1; T + H]$.

10.3.2. Estimation de la série cliométrique par une approche déterministe des taux de mortalité

10.3.2.1. Description de la méthode

Dans cette méthode, nous cherchons à déterminer deux fonctions déterministes monotones qui modélisent au mieux les courbes des séries temporelles m_{A,x_0} et m_{R,x_0}. A partir de ces fonctions déterministes, il sera alors possible d'établir une relation entre les points temporels dans les deux pays A et R car m_{A,x_0} et m_{R,x_0} sont des fonctions bijectives respectivement de t_A et t_R.

Soit s_0 un seuil arbitraire (des séries temporelles m_{A,x_0} et m_{R,x_0}) dont l'atteinte ou le franchissement (à la baisse) indique une présence du pays (A ou R) dans son processus de transition démographique.
En vertu de l'hypothèse d'universalité de l'allure des courbes de mortalité à l'âge x_0 durant la transition démographique, les courbes du taux de mortalité à l'âge x_0 des pays A et R (m_{A,x_0} et m_{R,x_0}) ont même allure après le passage du seuil s_0.

Soit t_{A,s_0} le premier temps d'atteinte ou de franchissement (à la baisse) de la valeur s_0 par la série temporelle $m_{A,\,x_0}$:

$$t_{A,s_0} = \inf\{t \in \mathbb{N}, m_{A,\,x_0,t} \leq s_0\}$$

De la même manière t_{R,s_0} est le premier temps d'atteinte ou de franchissement (à la baisse) de la valeur s_0 par la série temporelle m_{R,x_0} :

$$t_{R,s_0} = \inf\{t \in \mathbb{N}, m_{R,\,x_0,t} \leq s_0\}$$

Soit f_{A,x_0} la fonction déterministe monotone modélisant la série temporelle $m_{A,\,x_0}$
Soit f_{R,x_0} la fonction déterministe monotone modélisant la série temporelle $m_{R,\,x_0}$

On peut donc écrire :

$$m_{A,x_o,t_{A,s_i}} = f_{A,x_0}(t_{A,s_i}) + \mathcal{E}_{t_{A,s_i}} \Rightarrow s_i = f_{A,x_0}(t_{A,s_i}) + \mathcal{E}_{t_{A,s_i}} \quad \textbf{(a),}$$

avec f_{A,x_0} une fonction strictement monotone, c'est-à-dire f_{A,x_0} est bijective.

De même, $m_{R,x_o,t_{R,s_i}} = f_{R,x_0}(t_{R,s_i}) + \mathcal{E}_{t_{R,s_i}} \Rightarrow s_i = f_{R,x_0}(t_{R,s_i}) + \mathcal{E}_{t_{R,s_i}}$ **(b)**,
avec f_{R,x_0} une fonction strictement monotone c'est-à-dire f_{R,x_0} est bijective.

Puisque ces deux fonctions prennent des valeurs dans le même intervalle, nous avons la relation suivante déduite de **(a)** et **(b)** : $f_{A,x_0}(t_{A,s_i}) = f_{R,x_0}(t_{R,s_i}) \Rightarrow t_{R,s_i} = f_{R,x_0}^{-1}\left(f_{A,x_0}(t_{A,s_i})\right)$

Les $\mathcal{E}_{t_{A,s_i}}$ et $\mathcal{E}_{t_{R,s_i}}$ sont des erreurs *i.i.d.* qui doivent respecter les conditions suivantes : espérance nulle, homoscédasticité, non-autocorrélation : Autrement dit $\left(\varepsilon_{t_{A,s_i}}\right)_i$ et $\left(\varepsilon_{t_{R,s_i}}\right)_i$ sont des bruits blancs.
Si les résidus suivent les conditions usuelles du modèle linéaire, alors nous utilisons le modèle fourni par la régression pour réaliser les prévisions. Sinon, les modèles ne sont pas valides et il sera nécessaire de compléter le modèle en modélisant les résidus par un ARMA (cf. Boutaharet Royer-Carenzi [2019], p.148-158).

Nous avons le tableau initial suivant pour le pays A :

Tableau 5 : Tableau du pays A Cas de l'approche déterministe

$m_{A,\,x_0}$	t_A	$m_{A,x}$	$f_{A,\,x_0}$
s_0	t_{A,s_0}	$m_{A,x,\,t_{A,s_0}}$	$f_{A,\,x_0}(\,t_{A,s_0})$
s_1	t_{A,s_1}	$m_{A,x,\,t_{A,s_1}}$	$f_{A,\,x_0}(\,t_{A,s_1})$
⋮	⋮	⋮	⋮
s_i	t_{A,s_i}	$m_{A,x,\,t_{A,s_i}}$	$f_{A,\,x_0}(\,t_{A,s_i})$
⋮	⋮	⋮	⋮
s_d	t_{A,s_d}	$m_{A,x,\,t_{A,s_d}}$	$f_{A,\,x_0}(\,t_{A,s_d})$
s_{d+1}	$t_{A,s_{d+1}}$	$m_{A,x,\,t_{A,s_{d+1}}}$	$f_{A,\,x_0}(\,t_{A,s_{d+1}})$
⋮	⋮	⋮	⋮
s_r	t_{A,s_r}	$m_{A,x,\,t_{A,s_r}}$	$f_{A,\,x_0}(\,t_{A,s_r})$

d : Nombre d'observations disponibles de la série temporelle $m_{R,\,x_0}$ après le temps t_{R,s_0}
$m_{A,\,x_0} = (S_i\,, 0 \le i \le r)^T$, avec $r > d$ et $S_r < S_d$ (car leurs symétriques par rapport à l'axe horizontal du temps, et à une transformation monotone près sont des temps transitionnels).

Nous souhaitons apparier chaque observation du pays A avec une observation du pays R.
Nous ferons l'appariement entre les valeurs de $f_{A,\,x_0}$ et de $f_{R,\,x_0}$.
Pour chaque observation du pays A, nous cherchons à associer une observation du pays R, telle que :
$f_{A,\,x_0}(t_{A,s_i}) = f_{R,\,x_0}(t_{R,s_i}) = s_i$
D'où :

$t_{R,s_i} = f_{R,x_0}^{-1}\left(f_{A,x_0}(t_{A,s_i})\right)$, si la réciproque existe.

Puisque les t_{R,s_i} sont des entiers, nous prendrons la valeur arrondie à l'entier le plus proche.
Après association des observations du pays R à celles du pays A, nous obtenons le tableau suivant :

Tableau 6 : Tableau d'appariement des tables du pays A et du pays R

m_{A,x_0}	t_A	$m_{A,x}$	f_{A,x_0}	$t_R = f_{R,x_0}^{-1}\left(f_{A,x_0}(t_A)\right)$	$m_{R,x}$
s_0	t_{A,s_0}	$m_{A,x,\,t_{A,s_0}}$	$f_{A,\,x_0}(t_{A,S_0})$	t_{R,s_0}	$m_{R,x,\,t_{R,s_0}}$
s_1	t_{A,s_1}	$m_{A,x,\,t_{A,s_1}}$	$f_{A,\,x_0}(t_{A,S_1})$	t_{R,s_1}	$m_{R,x,\,t_{R,s_1}}$
⋮	⋮	⋮	⋮	⋮	⋮
s_i	t_{A,s_i}	$m_{A,x,\,t_{A,s_i}}$	$f_{A,\,x_0}(t_{A,S_i})$	t_{R,S_i}	$m_{R,x,\,t_{R,s_i}}$
⋮	⋮	⋮	⋮	⋮	⋮
s_d	t_{A,s_d}	$m_{A,x,\,t_{A,s_d}}$	$f_{A,\,x_0}(\,t_{A,s_d})$	t_{R,S_d}	$m_{R,x,\,t_{R,s_d}}$
s_{d+1}	$t_{A,s_{d+1}}$	$m_{A,x,\,t_{A,s_{d+1}}}$	$f_{A,\,x_0}(\,t_{A,s_{d+1}})$	$t_{R,S_{d+1}}$	NA
⋮	⋮	⋮	⋮	⋮	⋮
s_r	t_{A,s_r}	$m_{A,x,\,t_{A,s_r}}$	$f_{A,\,x_0}(\,t_{A,s_r})$	t_{R,S_r}	NA

La mention NA désigne les valeurs manquantes.

Les points de la série *explicative cliométrique à l'âge x* du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) sont définis par les couples ($\mathrm{t_{R,S_i}}$, $\mathrm{m_{A,x,\,t_{A,S_i}}}$), $i = 0, \dots, r$:

$$\mathcal{C}_{R,A,x,\,\mathrm{t_{R,S_i}}} : \mathrm{t_{R,S_i}} \mapsto \mathrm{m_{A,x,\,t_{A,S_i}}}$$

Si les arrondis de $\mathrm{t_{R,S_i}}$ donnent des valeurs identiques pour des valeurs différentes de $\mathrm{m_{A,x}}$, la moyenne des valeurs de $\mathrm{m_{A,x}}$ est associée aux valeurs identiques de $\mathrm{t_{R,S_i}}$, lors de la création des points de la série explicative cliométrique.

Pour éviter les valeurs manquantes dans la série cliométrique, dues aux valeurs manquantes dans la série temporelle $\mathrm{t_R}$ (qui devrait contenir des valeurs consécutives des années), une interpolation linéaire (ou moyenne mobile) pourra être utilisée. Les tests avec l'interpolation spline se sont révélés non-concluants car elle générait des valeurs atypiques.

Notons :

T le temps (année) présent(e).

H l'horizon maximal de prévision envisagé.

Si l'horizon $\mathrm{T + H} > \mathrm{t_{A,s_r}}$ alors nous aurons recours aux prévisions par séries temporelles pour estimer les valeurs de la série cliométrique sur la période $[\,\mathrm{t_{A,s_r}} + 1; T + H]$.

10.3.2.2. Application : Choix de la fonction déterministe à utiliser

a. Modélisation du taux de mortalité à l'âge x_0 par une fonction exponentielle déterministe

Nous faisons l'hypothèse que les courbes de taux de mortalité à l'âge x_0 peuvent être estimées par une fonction exponentielle déterministe, c'est-à-dire la courbe de taux de mortalité x_0 pour le pays A ($m_{A,\,x_0}$) en fonction de t_{A,s_i} peut être estimée par :

$$m_{A,\,x_0,t_{A,s_i}} = ae^{-bt_{A,s_i}} + \mathcal{E}_{t_{A,s_i}},\ a, b > 0$$

et la courbe de taux de mortalité x_0 pour le pays R ($m_{R,\,x_0}$) en fonction de t_{R,s_i} est donnée par :

$$m_{R,\,x_0,t_{R,s_i}} = ce^{-dt_{R,s_i}} + \mathcal{E}_{t_{R,s_i}},\ c, d > 0.$$

Nous effectuons les régressions Log-linéaires des taux de mortalité à l'âge x_0 sur le temps (année) de chaque pays (pays A et pays R) en nous basant sur leurs courbes de taux de mortalité.

Si nous désignons par $\widehat{m}_{A,\,x_0,t_{A,s_i}}$ et $\widehat{m}_{R,\,x_0,t_{R,s_i}}$ des estimateurs respectifs de $m_{A,\,x_0,t_{A,s_i}}$ et $m_{R,\,x_0,t_{R,s_i}}$, alors en appliquant la Log à ces taux de mortalité, nous obtenons des régressions Log-linéaires données par :

$$\begin{cases} \widehat{m}_{A,\,x_0,t_{A,s_i}} = \hat{a}e^{-\hat{b}t_{A,s_i}} \\ \widehat{m}_{R,\,x_0,t_{R,s_i}} = \hat{c}e^{-\hat{d}t_{R,s_i}} \end{cases} \Longrightarrow \begin{cases} log(\widehat{m}_{A,\,x_0,t_{A,s_i}}) = log(\hat{a}) - \hat{b}t_{A,s_i} \\ log(\widehat{m}_{R,\,x_0,t_{R,s_i}}) = log\,(\hat{c}) - \hat{d}t_{R,s_i} \end{cases}$$

On a : $s_i = m_{A,\,x_0,t_{A,s_i}}$ et $s_i = m_{R,\,x_0,t_{R,s_i}}$

$$\widehat{m}_{A,\,x_0,t_{A,s_i}} = \widehat{m}_{R,\,x_0,t_{R,s_i}} \Longrightarrow log(\hat{a}) - \hat{b}t_{A,s_i} = log(\hat{c}) - \hat{d}t_{R,s_i}$$

$$\Rightarrow\ t_{R,s_i} = \frac{1}{\hat{d}} log\left(\frac{\hat{c}}{\hat{a}}\right) + \frac{\hat{b}}{\hat{d}} t_{A,s_i}$$

avec $\hat{a}$, $\hat{b}$, $\hat{c}$ et $\hat{d}$ des estimateurs respectifs des paramètres a, b, c et d.
Finalement t_{R,s_i}est sous la forme $t_{R,s_i} = \widehat{D} + \hat{B} t_{A,s_i}$ où $\widehat{D}, \hat{B} \in \mathbb{R}_+$, avec $\widehat{D} = \frac{1}{\hat{d}} log\left(\frac{\hat{c}}{\hat{a}}\right)$, $\widehat{B} = \frac{\hat{b}}{\hat{d}}$
Cette relation est utilisée pour déterminer les valeurs $t_{R,s_{d+1}}$ à t_{R,s_r}.

Les points de la série *explicative cliométrique à l'âge x* du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) sont définis par les couples ($\mathrm{t_{R,S_i}}$, $\mathrm{m_{A,x,\,t_{A,s_i}}}$), $i = 0, \dots, r$:

$$\mathcal{C}_{R,A,x,\,\mathrm{t_{R,S_i}}} :\ \mathrm{t_{R,S_i}} \mapsto \mathrm{m_{A,x,\,t_{A,s_i}}}$$

b. Modélisation du taux de mortalité à l'âge x_0 par une fonction logistique déterministe décroissante
En faisant l'hypothèse que les courbes de taux de mortalité à l'âge x_0 peuvent être estimées par une fonction logistique déterministe, la courbe de taux de mortalité à l'âge x_0 pour le pays A ($m_{A,\,x_0}$) en fonction de t_{A,s_i} est définie par :
$m_{A,\,x_0,t_{A,s_i}} = \frac{S_0}{1+e^{-(a_0+a_1 t_{A,s_i})}} + \mathcal{E}_{t_{A,s_i}}$, $a_0\ et\ a_1$ sont des paramètres à estimer avec $a_1 \leq 0$;
avec $\widehat{m}_{A,\,x_0,t_{A,s_i}}$ un estimateur de $m_{A,\,x_0,t_{A,s_i}}$

Celle de taux de mortalité à l'âge x_0 pour le pays R ($m_{R,\,x_0}$) en fonction de t_{R,s_i} est définie par :
$m_{R,\,x_0,t_{R,s_i}} = \frac{S_0}{1+e^{-(b_0+b_1 t_{R,s_i})}} + \mathcal{E}_{t_{R,s_i}}$, $b_0\ et\ b_1$ sont des paramètres à estimer avec $b_1 \leq 0$;
avec $\widehat{m}_{R,\,x_0,t_{R,s_i}}$ un estimateur de $m_{R,\,x_0,t_{R,s_i}}$

Nous travaillons avec les estimateurs $\widehat{m}_{A,\,x_0,t_{A,s_i}}$ et $\widehat{m}_{R,\,x_0,t_{R,s_i}}$. Pour les linéariser, nous procédons à leur transformation, puis nous effectuons les régressions Log-linéaires des taux de mortalité à l'âge x_0 sur le temps (année) de chaque pays (pays A et pays R).
Pour le pays A, nous avons :

$$\frac{\widehat{m}_{A,\,x_0,t_{A,s_i}}}{S_0} = \frac{1}{1+e^{-\left(\widehat{a_0}+\widehat{a_1}t_{A,s_i}\right)}} \Rightarrow \frac{S_0}{\widehat{m}_{A,\,x_0,t_{A,s_i}}} = 1 + e^{-\left(\widehat{a_0}+\widehat{a_1}t_{A,s_i}\right)}$$

$$\Rightarrow \frac{S_0}{\widehat{m}_{A,\,x_0,t_{A,s_i}}} - 1 = e^{-\left(\widehat{a_0}+\widehat{a_1}t_{A,s_i}\right)}$$

$$\Rightarrow \frac{1}{\frac{S_0}{\widehat{m}_{A,\,x_0,t_{A,s_i}}} - 1} = e^{\left(\widehat{a_0}+\widehat{a_1}t_{A,s_i}\right)}$$

$$\Rightarrow log\left(\frac{1}{\frac{S_0}{\widehat{m}_{A,\,x_0,t_{A,s_i}}}-1}\right) = \widehat{a_0} + \widehat{a_1} t_{A,s_i}$$

Avec $\widehat{a_0}$ et $\widehat{a_1}$ des estimateurs respectifs des paramètres a_0 et a_1
De même pour le pays R nous obtenons :

$$\frac{\widehat{m}_{R,\,x_0,t_{R,s_i}}}{S_0} = \frac{1}{1+e^{-\left(\widehat{b_0}+\widehat{b_1}t_{R,s_i}\right)}} \Rightarrow \frac{S_0}{\widehat{m}_{R,\,x_0,t_{R,s_i}}} = 1 + e^{-\left(\widehat{b_0}+\widehat{b_1}t_{R,s_i}\right)}$$

$$\Rightarrow \frac{S_0}{\widehat{m}_{R,\,x_0,t_{R,s_i}}} - 1 = e^{-\left(\widehat{b_0}+\widehat{b_1}t_{R,s_i}\right)}$$

$$\Rightarrow \frac{1}{\frac{S_0}{\widehat{m}_{R,\,x_0,t_{R,s_i}}} - 1} = e^{\left(\widehat{b_0}+\widehat{b_1}t_{R,s_i}\right)}$$

$$\Rightarrow log\left(\frac{1}{\frac{S_0}{\widehat{m}_{R,\,x_0,t_{R,s_i}}}-1}\right) = \widehat{b_0} + \widehat{b_1}t_{R,s_i}$$

Avec $\widehat{b_0}$ et $\widehat{b_1}$ des estimateurs respectifs des paramètres b_0 et b_1

Soit s_i une valeur de m_{A,x_0} interne au processus de transition quel que soit le pays.
$\forall\, i \in [0\,;\, T]$ on a :

$s_i = m_{A,\,x_0,t_{A,s_i}} = m_{R,\,x_0,t_{R,s_i}} \Longrightarrow \widehat{m}_{A,\,x_0,t_{A,s_i}} = \widehat{m}_{R,\,x_0,t_{R,s_i}}$
Donc :

$$log\left(\frac{1}{\frac{S_0}{\widehat{m}_{A,\,x_0,t_{A,s_i}}}-1}\right) = log\left(\frac{1}{\frac{S_0}{\widehat{m}_{R,\,x_0,t_{R,s_i}}}-1}\right).$$

Par conséquent :

$$\widehat{a_0} + \widehat{a_1}t_{A,s_i} = \widehat{b_0} + \widehat{b_1}t_{R,s_i} \Rightarrow t_{R,s_i} = \frac{\widehat{a_0} - \widehat{b_0}}{\widehat{b_1}} + \frac{\widehat{a_1}}{\widehat{b_1}}t_{A,s_i}$$

D'où :

$$t_{R,s_i} = \hat{\alpha} + \hat{\beta}t_{A,s_i}, \text{ avec } \hat{\alpha} = \frac{\widehat{a_0}-\widehat{b_0}}{\widehat{b_1}}\ et\ \hat{\beta} = \frac{\widehat{a_1}}{\widehat{b_1}}$$

Cette relation est utilisée pour déterminer les valeurs $t_{R,s_{d+1}}$ à t_{R,s_r}.

Les points de la série *explicative cliométrique à l'âge x* du pays R par rapport au pays A ($\mathcal{C}_{R,A,x}$) sont définis par les couples ($\mathrm{t_{R,S_i}}$, $\mathrm{m_{A,x,\,t_{A,s_i}}}$), $i = 0, \dots, r$:

$$\mathcal{C}_{R,A,x,\,\mathrm{t_{R,S_i}}} : \mathrm{t_{R,S_i}} \mapsto \mathrm{m_{A,x,\,t_{A,s_i}}}$$

10.4. Modèle mixte cliométrique Logit PCR-optimal

Nous souhaitons améliorer le modèle Logit-PCR, en y intégrant la série cliométrique correspondante à chaque tranche d'âge (ou âge) modélisée. On appellera ce modèle **Logit-PCR-cliométrique**.
Cette intégration de la série cliométrique peut se faire selon deux (2) modalités :

- **Avant** le calcul et la sélection des composantes principales (donc prise en compte dans le calcul des composantes principales) : cette option présuppose une prévision des mortalités par âge. Or ces prévisions sont précisément l'objectif à atteindre. Cette option n'est donc pas pertinente.

- **Après** le calcul et la sélection des composantes principales (donc non-prise en compte dans le calcul des composantes principales). La série cliométrique est visible dans le modèle final, mais il y a un risque de colinéarité avec les composantes principales. Ce risque de colinéarité obligerait à recourir à une méthode d'estimation adaptée (régressions de Ridge, Lasso, elastic-net, sélection *stepwise* ou *recherche exhaustive* des variables, PCR, PLS), le cas échéant. Nous adopterons cette approche.

Choix d'un temps transitionnel $\mathbf{m_{x_0}}$

En application, nous utiliserons le taux de mortalité infantile lissé comme temps transitionnel m_{x_0}. En effet le taux de mortalité infantile lissé est une série temporelle qui peut empiriquement être utilisée comme temps transitionnel, dans une étude en coupe transversale, du fait de sa tendance baissière généralisée depuis au moins 2 siècles (exception faite des crises sanitaires).

Notons également que la pente moyenne du taux de mortalité lissé est de -1.17 (médiane de -1.06), ce qui est assez proche en valeur absolue de 1 : ce qui renforce son rôle potentiel de temps transitionnel. Précisons toutefois que le taux de mortalité est lui-même une statistique généralement lissée avant publication officielle, du moins dans la période d'après-Guerre. (Gaba [2021], p. 187).

10.5. Modèle mixte cliométrique Logit-PLS

Il s'agit ici d'une adaptation du modèle mixte Logit-PCR cliométrique précédent, à la différence que nous utilisons une regression PLS à la place d'une regression PCR optimale.

10.6. Modèle mixte cliométrique composite à facteurs PCR-O/PLS

En capitalisant sur les apports originaux des sections précédentes, nous proposons le type de modèle suivant : modèle **mixte cliométrique et composite** à facteurs **PCR-optimal/PLS.**

Comme précédemment évoqué, le caractère composite de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles diversifiées (par exemple via le choix du ou des pays A, facteurs PCR-O ou PLS).

Ces modèles seront développés et testés empiriquement sur des pays émergents dans la suite de ce mémoire.

11. Modèles relationnels (à référence externe) cliométriques

Les modèles relationnels cliométriques que nous développons sont des solutions à la problématique des pays ayant un historique insuffisant pour les modèles internes de prospective.

11.1. Modèles externes à appariement temporel cliométrique

Comme indiqué dans les problématiques, ces techniques d'appariement seront développées dans la suite de ce mémoire, mais nécessitent d'un point de vue statistique au moins une vingtaine d'années d'observations pour le pays d'expérience R.

Ainsi l'appariement entre le pays d'expérience R et le pays de référence A (voir les définitions dans la section précédente), permet :

1. de disposer de couples appariés de temps calendaires du pays d'expérience R et ceux du pays de référence A, via le temps transitionnel qui leur est commun

2. de créer plusieurs séries cliométriques pour chaque âge du pays R (par rapport à un ou plusieurs pays A, avec divers paramétrages de sexe ou d'âge). Ces séries cliométriques qui ont par construction des valeurs dans le passé et dans le futur, sont des séries explicatives potentielles dans les modèles externes cliométriques pour le pays d'expérience R.

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le premier résultat ci-dessus, nous proposons les 3 classes de modèles externes cliométriques suivants, adaptés des modèles classiques externes :

1. Modèle **externe cliométrique** adapté du modèle externe de BRASS
2. Modèle **externe cliométrique** adapté du modèle externe de COX
3. Modèle **externe cliométrique** adapté du modèle externe de TGH05-TGF05

Les trois modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** mélangeant les adaptations de BRASS/COX/TGH05-TGF05 .

En capitalisant sur les apports originaux des sections précédentes, et en exploitant le deuxième résultat ci-dessus, nous proposons les 2 classes de modèles externes cliométriques suivants, adaptés des modèles à facteurs PCR-Optimal et des modèles à facteurs PLS :

1. Modèle **externe cliométrique** à facteurs PCR-Optimal (construits à partir des séries cliométriques)
2. Modèle **externe cliométrique** à facteurs PLS (construits à partir des séries cliométriques)

De même, les deux modèles ci-dessus peuvent être combinés dans le modèle composite suivant : Modèle **externe cliométrique composite** à facteurs PCR-O/PLS .

Comme précédemment évoqué, le caractère **composite** de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles différentes.

11.2. Hypothèses et notations

Nous rappelons les hypothèses suivantes :

- Le pays A est un pays de référence, le pays R est un pays d'expérience plus précisément dans ce contexte un pays ayant un historique insuffisant pour les modèles internes de prospective
- $x \in [x_m,\ x_M]$ c'est-à-dire $x_m = min(x)$ et $x_M = max(x)$
- $\mathrm{m}_{\mathrm{A},\,\mathrm{x}_0} = \left(\mathrm{m}_{\mathrm{A},\,\mathrm{x}_0,t}\right), \mathrm{t} \in \mathbb{N}$ et $\mathrm{m}_{\mathrm{R},\,\mathrm{x}_0} = \left(\mathrm{m}_{\mathrm{R},\,\mathrm{x}_0,t}\right), \mathrm{t} \in \mathbb{N}$ sont deux séries temporelles de mortalités à l'âge x_0 ($\mathrm{x}_0 \in [x_m,\ x_M]$) respectivement des pays A et R. Les symétriques (par rapport à l'axe horizontal du temps, et à une transformation monotone près) de ces séries temporelles sont des temps transitionnels
- Les $\mathrm{s_i}$ sont des valeurs seuils du taux de mortalité $\mathrm{m}_{\mathrm{x}_0}$
- $\mathrm{t}_{\mathrm{A},\mathrm{s_i}} = \inf\{\mathrm{t} \in \mathbb{N}, \mathrm{m}_{\mathrm{A},\,\mathrm{x}_0,t} \leq \mathrm{s_i}\}$ et $\mathrm{t}_{\mathrm{R},\mathrm{s_i}} = \inf\{\mathrm{t} \in \mathbb{N}, \mathrm{m}_{\mathrm{R},\,\mathrm{x}_0,t} \leq \mathrm{s_i}\}$
- $\mathrm{t_A} = \left(\mathrm{t}_{\mathrm{A},\mathrm{s_i}}\right), \mathrm{i} \in \mathbb{N}$ et $\mathrm{t_R} = \left(\mathrm{t}_{\mathrm{R},\mathrm{s_i}}\right), \mathrm{i} \in \mathbb{N}$ deux séries temporelles des dates calendaires d'atteinte des valeurs de mortalités $(\mathrm{s_i}), \mathrm{i} \in \mathbb{N}$ à l'âge x_0, respectivement dans les pays A et R

11.3. Modèle externe cliométrique adapté du modèle externe de BRASS

Nous nous inspirons du modèle classique de BRASS pour formuler notre modèle externe logistique cliométrique à paramètres indépendants de l'âge.
Dans le modèle externe cliométrique logistique, les logits des probabilités de décès d'expérience $q^{exp}_{x,t_{R,s_i}}$ sont liés aux probabilités de décès de référence $q^{ref}_{x,t_{A,s_i}}$ à partir d'une relation linéaire de la forme :

$$logit\left(q^{exp}_{x,t_{R,s_i}}\right) = \theta_1 + \theta_2 \times logit\left(q^{ref}_{x,t_{A,s_i}}\right)$$

11.4. Modèle externe cliométrique adapté du modèle externe de COX

Le modèle externe cliométrique proportionnel est une adaptation du modèle classique de COX dans un cadre cliométrique.
Le modèle externe cliométrique proportionnel est défini par :

$$\mu^{exp}_{x,t_{R,s_i}} = \theta \times \mu^{ref}_{x,t_{A,s_i}}, \theta > 0$$

11.5. Modèle externe cliométrique adapté du modèle externe de TGH05-TGF05

Le modèle externe cliométrique composite logistique à paramètres dépendants de l'âge que nous proposons est une adaptation du modèle logistique TGH05-TGF05.
Dans le modèle externe cliométrique composite logistique, nous faisons pour chaque âge x une régression entre les logits des probabilités de décès d'expérience $q^{exp}_{x,t_{R,s_i}}$ et les logits des probabilités de décès de référence $q^{ref}_{x,t_{A,s_i}}$

$$\forall\, x, \quad logit\left(q^{exp}_{x,t_{R,s_i}}\right) = a_x logit\left(q^{ref}_{x,t_{A,s_i}}\right) + b_x$$

11.6. Modèles externes cliométriques et composites à facteurs PCR-optimal/PLS

Au sein des modèles externes cliométriques et composites à facteurs PCR-optimal/PLS, chaque âge x est modélisé par l'un des deux types de modèles suivants :

- Les modèles externes cliométriques à facteurs PCR-optimal
  ou
- Les modèles externes cliométriques à facteurs PLS

Rappelons que le caractère composite joue sur le paramétrage de l'aspect cliométrique (le ou les pays A de référence, leurs sexes par exemple) ainsi que sur le type de facteurs orthogonaux utilisés (PCR-optimal ou PLS).

11.6.1. Modèles externes cliométriques à facteurs PCR-optimal

Les modèles externes cliométriques à facteurs PCR-optimal que nous proposons sont de la forme :

$$logit\, q_x\,(t) = a_x + \beta^*_{c_1} x^*_{c_1}(t) + \beta^*_{c_2} x^*_{c_2}(t) + \cdots + \beta^*_{c_{k_x}} x^*_{c_{k_x}}(t) + \varepsilon_{x,t}$$

où :

$q_x(t)$ est la probabilité de décès à l'âge x, au temps t ;

a_x est la moyenne des $\text{logit}\, q_x\,(t)$ pour l'âge x calculée sur l'ensemble de la période considérée ;

$x^*_{c_i}, 1 \leq i \leq k_x$ **est une composante principale de la matrice $\boldsymbol{X}$ des séries explicatives cliométriques (il ne s'agit pas de la $\boldsymbol{i}^{\text{ème}}$ composante principale) ;**

$p = M - m + 1$ et k_x $(k_x \leq p)$ est le nombre de premières composantes principales optimales sélectionnées ;

$\varepsilon_{x,t}$ est le résidu du modèle (les $\varepsilon_{x,t}$ sont supposés *i.i.d.*, d'espérance nulle et de variance σ^2_ε).

Le critère et la procédure de sélection pour ces k_x composantes principales optimales introduites dans notre modèle sont :

- Critère de sélection du modèle : MSEP.
- Procédure de sélection du modèle : recherche exhaustive.

11.6.2. Modèles externes cliométriques à facteurs PLS

Les modèles externes cliométriques à facteurs PLS que nous proposons sont de la forme :

$$logit\, q_x\,(t) = r_1 l^{(1)}(t) + r_2 l^{(2)}(t) + \cdots + r_{k_x} l^{(k_x)}(t) + \varepsilon_{k_x,t}$$

où $l^{(j)}$ désigne la composante PLS d'ordre j

Partie 3. Résultats empiriques et discussions

12. Choix des modèles classiques internes de référence pour nos travaux empiriques

Notre objectif est de choisir les modèles classiques à utiliser pour nos travaux empiriques, en tenant compte de leurs limites et de notre contexte.

12.1. Problèmes d'erreurs de prévision du modèle CBD

Bien que l'ajustement du CBD ne pose pas de problème particulier en termes de convergences, il souffre parfois de graves erreurs en termes de prévisions.

Diaz et al. (2018, pp. 9-18) en ajustant le modèle CBD sur des données abrégées de taux de mortalité (taux de mortalités par intervalle d'âges) de la Colombie de 1973 à 2005 ont argumenté que le modèle CBD suppose que la mortalité est linéaire sur l'échelle logit, et par conséquent il ne fonctionne bien que sur les âges avancés pour les deux sexes.

Villegas et al. (2015, pp. 15-16) confirment cela également dans leurs travaux sur les données de taux de mortalité de l'Angleterre et du Pays de Galles de la période 1961-2011 : ils ne s'en sont limités qu'aux âges de 55 à 89 ans en prenant soin de bien spécifier que le modèle CBD ajuste bien uniquement sur les âges élevés.

12.2. Problèmes de convergence des modèles GAPC avec effet de cohorte

Des problèmes de convergence ont été détectés pour les modèles APC, RH et M8 selon certains types de données utilisées et des fois liés aux contraintes d'identifiabilité de ces modèles.
Nous pouvons citer Diaz et al. (2018, pp. 9-18) où après avoir ajusté les modèles Lee-Carter Log-Poisson, M8 et RH sur des taux mortalités abrégées pour la Colombie de 1973 à 2005 ; ils se sont rendus compte des faits suivants :

- Les modèles RH et M8 pour les données des hommes n'ont pas convergé
- Le modèle APC bien qu'il converge pour les deux sexes, est difficile à ajuster sur des tables de mortalité abrégées (en effet, des valeurs élevées de RMSE et de MAPE par rapport au Lee-Carter Poisson furent obtenues).

Villegas et al. (2015, pp. 15-16) exposent également des constats similaires à ceux de Diaz et al. (2018), en énonçant que l'ajustement des extensions de cohortes des modèles de Lee-Carter est problématique. Ces problèmes de convergence sont également exposés par Currie (2014), qui à son niveau a rencontré des difficultés à ajuster le modèle RH.
Hunt et al. (2015) ont étudié la convergence des modèles RH et Lee-Carter Log Poisson en testant spécialement deux procédures d'ajustement pour le modèle RH et en formulant de plus une contrainte supplémentaire pour ce dernier. A la fin, ils se sont rendus compte qu'indépendamment des procédures d'ajustement et de l'ajout de la nouvelle contrainte supplémentaire qu'ils ont défini, le modèle RH est instable aux changements dans les données. Ils finalisent en disant que le modèle RH arrive difficilement à répartir la structure des données entre les interactions âge/période et âge/cohorte.

Concernant cette instabilité aux changements dans les données, Cairns et al. (2009) indiquent que les estimations des paramètres du modèle RH passent à une solution qualitativement différente lorsque moins de données sont utilisées.

Kennes (2017, p 17) rejoint également la conclusion de Hunt et al. (2015) en précisant que le modèle RH est confronté à un problème d'identification approximative ou exacte, c'est-à-dire qu'il existe des régions approximativement (ou exactement) plates dans sa fonction de vraisemblance qui rendent sa convergence lente ou infructueuse.
Kennes (2017, p. 17) conclut sa réflexion sur les modèles de Lee-Carter avec extension à effet de cohorte en disant que le mécanisme sous-jacent exact lié à leurs problèmes de convergence est complexe et qu'il convient d'accorder de l'attention à la nature des données ajustées.

12.3. Choix des modèles internes classiques de référence

Les conclusions des auteurs évoqués dans les deux sections ci-dessus sont également confirmées par nos résultats d'ajustement des taux de mortalité de l'Equateur et de l'Inde où seuls les modèles de Lee-Carter et de Lee-Carter Log-Poisson ont pu converger *et* donner des résultats convenables.

Outre les modèles de Lee-Carter et Lee-Carter Log-Poisson, les autres modèles testés dans notre cas furent le modèle CBD, le modèle CBD Log-Poisson, le modèle APC, le modèle RH, le modèle M7 et le modèle Plat. Les données de taux de mortalité utilisées sont des données abrégées (par intervalle d'âge : [0-1[, [1-5], [5-10], et le reste en groupes d'âge de 5 ans jusqu'à 80 ans).
Les BIC des modèles CBD, CBD Log-Poisson et RH étaient de l'ordre des millions tandis que ceux du Lee-Carter et Lee-Carter Log Poisson étaient de l'ordre des milliers. Quant aux modèles restants (APC, M7 et Plat), ils n'ont pas pu converger.

En conséquence des analyses et résultats ci-dessus, nous retiendrons comme modèles internes classiques de référence dans nos travaux empiriques les modèles suivants : Lee-Carter, Lee-Carter Log-Poisson et CBD Log-Poisson.

13. Modélisation des mortalités pour la table prospective de l'Inde

Les modèles dont les performances sont comparées sont : Lee-carter (LC), LC Log-Poisson, CBD Log-Poisson, modèle mixte cliométrique composite à facteurs PCR-Optimal/PLS (Cliometric composite model).

13.1. Paramétrages associés au pays « $\mathbf{R}$ » transitionnellement « moins âgé » (ou pays d'expérience) à modéliser

- Pays « R » : Inde
- Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).
- Variable étudiée : Probabilité de décès (femmes)
- Liste des transformations préalables (moyennes mobiles) : "Interpolation + CMA5"
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans

Itérations :

- Itération 1 : Détection **a posteriori** de la meilleure classe de modèle (Best Model Class) pour chaque âge modélisé
    - Début de l'historique étudié : 1930
    - Année actuelle supposée : 1965
    - Horizon de prévision : 1990 (25 ans)
- Itération 2 : Utilisation **a priori** de la meilleure classe de modèle (Best Model Class) pour chaque âge modélisé
    - Début de l'historique étudié : 1930
    - Année actuelle supposée : 1990
    - Horizon de prévision : 2015 (25 ans)

13.2. Paramétrage et performances du meilleur modèle obtenu

13.2.1. Paramétrages d'exploration associés aux pays de référence A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.
Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Italy"), ("Spain")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans
- Liste de vecteurs de sexes testés : ("M"), ("F"), ("MF")
- Liste de vecteurs d'âges d'appariement testés : Age 0
- Liste des transformations préalables (moyennes mobiles) : "Interpolation + CMA5"

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

13.2.2. Performances du meilleur modèle obtenu

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 20 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

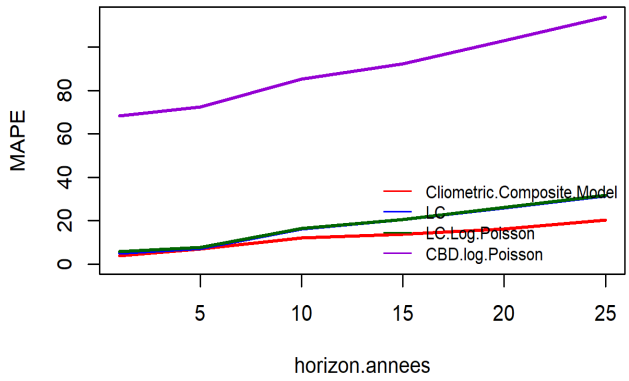
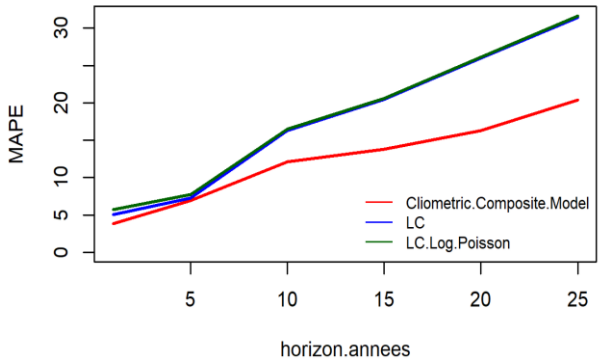


Figure 21 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Lors de cette première itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié.

Tableau 7 : MAPE (%)

MAPE (%)	Horizons					
Models	1	5	10	15	20	25
CBD.log.Poisson	68,3	72,5	85,3	92,3	103,0	113,9
Cliometric.Composite.Model	3,8	6,9	12,1	13,8	16,3	20,4
LC	5,1	7,3	16,3	20,5	26,0	31,4
LC.Log.Poisson	5,7	7,7	16,5	20,6	26,1	31,6

Pour chaque horizon, les prévisions de toutes les années précédentes sont incluses dans le calcul du MAPE (et cela en considérant tous les âges). Chaque MAPE est donc calculé sur une fenêtre dont l'étendue augmente au fur et à mesure que l'horizon augmente : cela peut expliquer l'allure lissée des courbes MAPE.

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 22 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

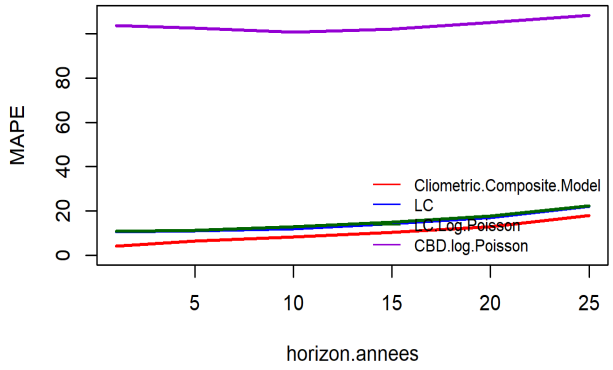
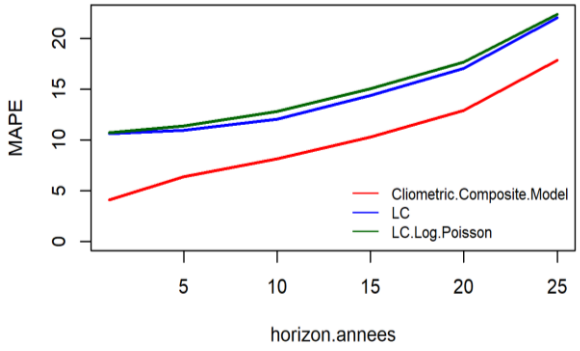


Figure 23 : Performance des modèles selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette 2ième itération, nous constatons un avantage systématique et conséquent pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié.
Nous pouvons toutefois conclure à un très léger effet de surapprentissage sur le long terme.

Tableau 8 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,7	102,6	100,8	102,2	105,1	108,3
Cliometric.Composite.Model	4,1	6,4	8,2	10,3	12,9	17,9
LC	10,6	11,0	12,0	14,4	17,0	22,1
LC.Log.Poisson	10,7	11,4	12,8	15,0	17,6	22,4

Vérifions à présent si les constatations sur les horizons de prévisions sont également avérées au niveau des âges modélisés.

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 24 : MAPE tous horizons confondus, selon les différents âges modélisés

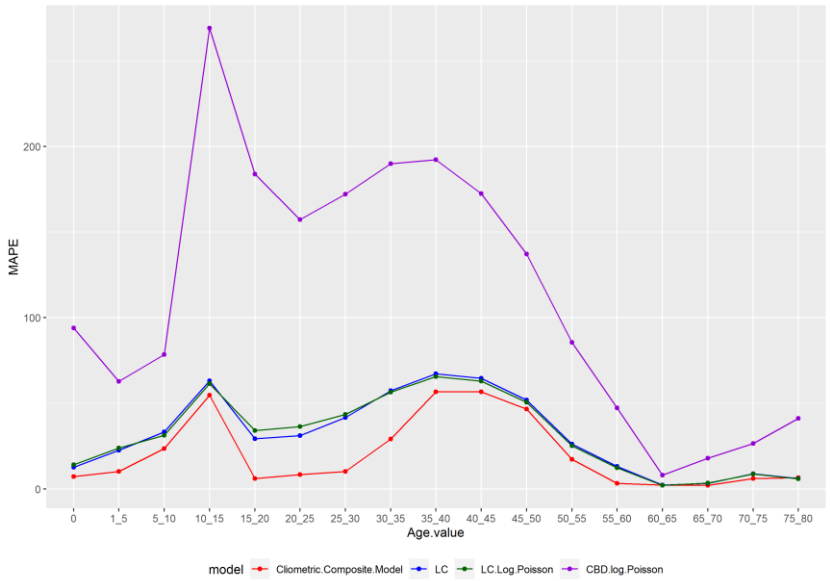
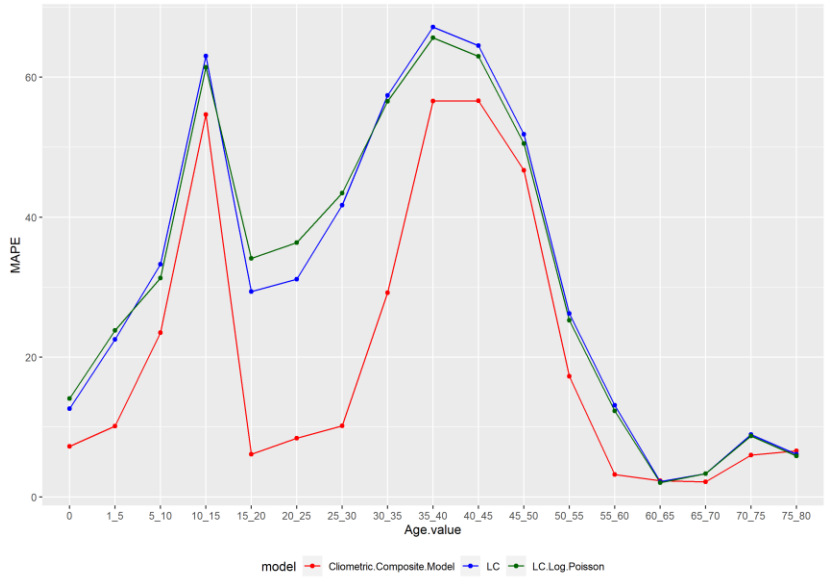


Figure 25 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette 1ère itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Sauf pour les tranches 60-65 ans et 75-80 ans où il y a un léger écart défavorable au modèle mixte cliométrique.

Tableau 9 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	94,0	7,2	12,6	14,1
1_5	62,8	10,2	22,5	23,8
5_10	78,4	23,5	33,2	31,3
10_15	269,1	54,6	63,0	61,4
15_20	183,9	6,1	29,4	34,1
20_25	157,2	8,4	31,1	36,4
25_30	172,1	10,2	41,7	43,4
30_35	190,0	29,2	57,4	56,6
35_40	192,2	56,6	67,2	65,6
40_45	172,5	56,6	64,5	63,0
45_50	137,2	46,7	51,8	50,5
50_55	85,6	17,3	26,2	25,3
55_60	47,3	3,2	13,1	12,3
60_65	8,1	2,3	2,2	2,0
65_70	17,9	2,2	3,3	3,3
70_75	26,6	6,0	8,9	8,7
75_80	41,2	6,6	6,1	5,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 26 : MAPE tous horizons confondus, selon les différents âges modélisés

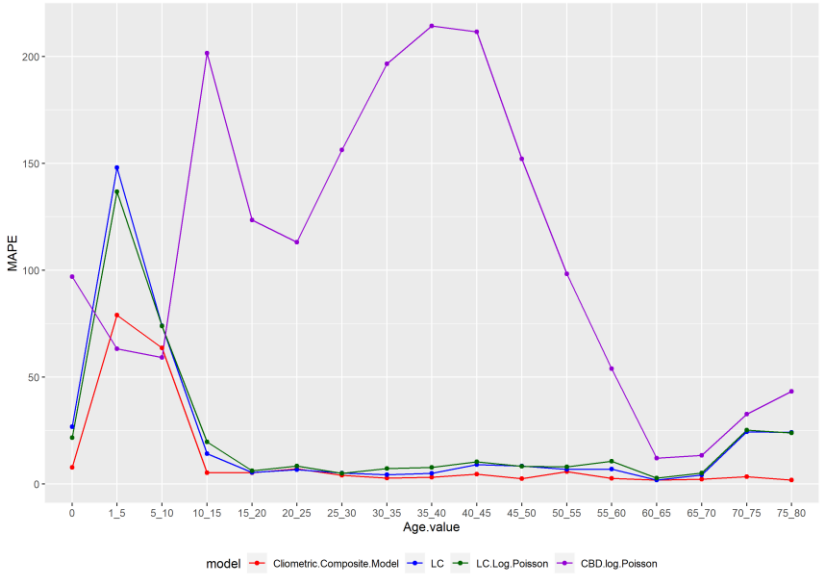
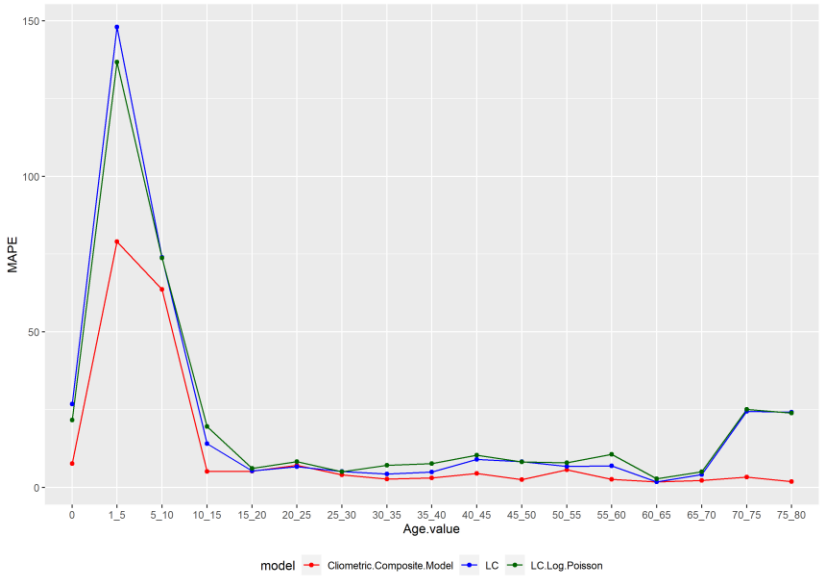


Figure 27 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette 2ème itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Sauf pour les tranches 1-5, 20-25 ans où il y a un très léger écart défavorable au modèle mixte cliométrique.
Pour l'essentiel, la hiérarchie de la première itération semble conservée.
Nous pouvons toutefois conclure à un léger effet de surapprentissage, car la deuxième itération a une performance relative moindre.

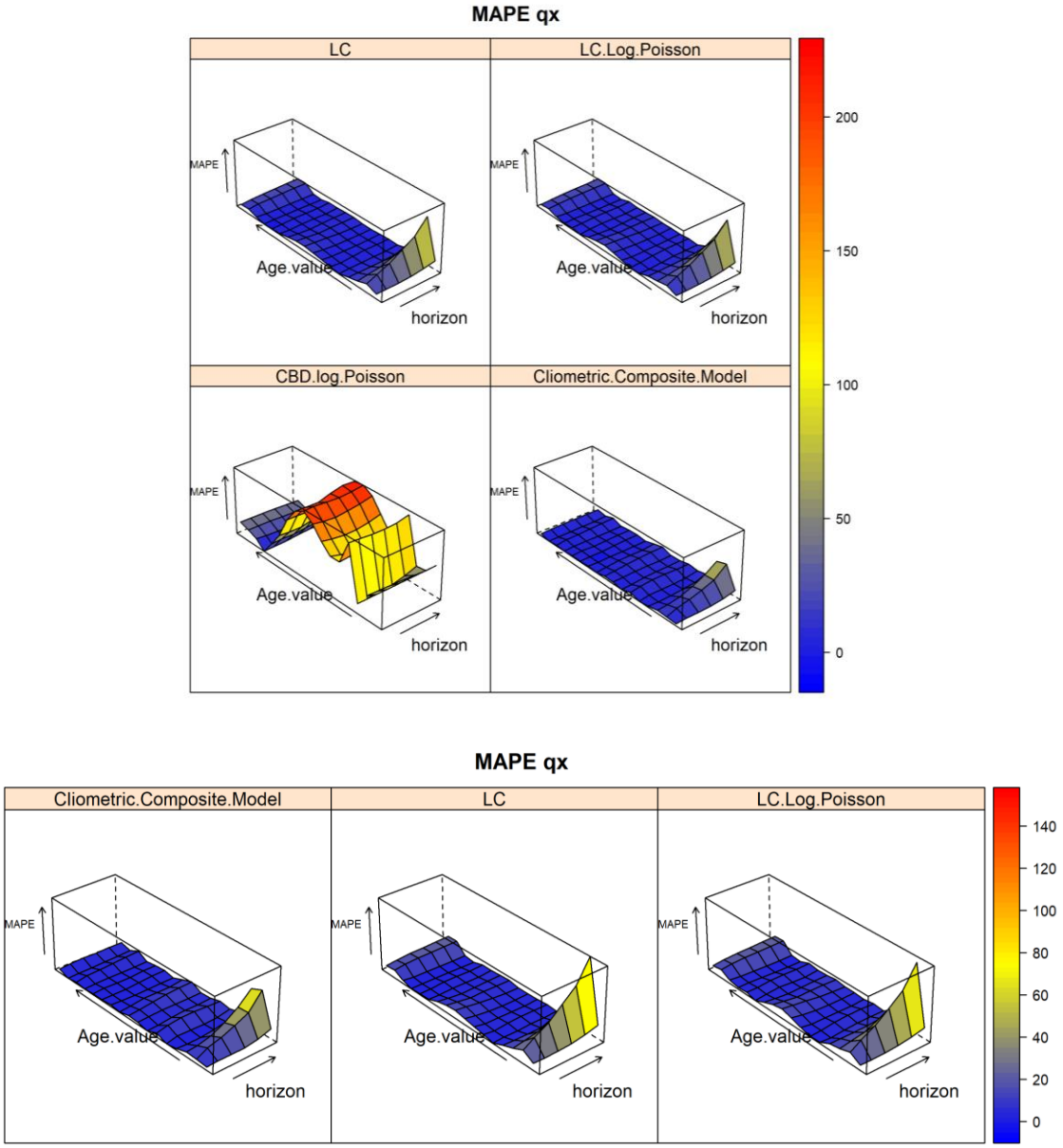
Tableau 10 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	97,0	7,7	26,8	21,7
1_5	63,3	79,0	148,0	136,7
5_10	59,1	63,7	74,0	73,8
10_15	201,5	5,2	14,1	19,6
15_20	123,5	5,2	5,4	6,1
20_25	113,1	7,1	6,7	8,3
25_30	156,3	4,1	5,2	5,0
30_35	196,6	2,7	4,3	7,2
35_40	214,2	3,1	4,9	7,7
40_45	211,6	4,5	9,1	10,4
45_50	152,0	2,5	8,4	8,3
50_55	98,2	5,7	6,8	8,0
55_60	53,9	2,6	6,9	10,7
60_65	12,1	1,8	1,8	2,8
65_70	13,4	2,3	4,2	5,1
70_75	32,6	3,4	24,4	25,1
75_80	43,4	1,9	24,2	23,9

MAPE des probabilités de décès selon les âges et les horizons (itération 2)

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1.

Figure 28 : MAPE (%) des probabilités de décès selon les âges et les horizons

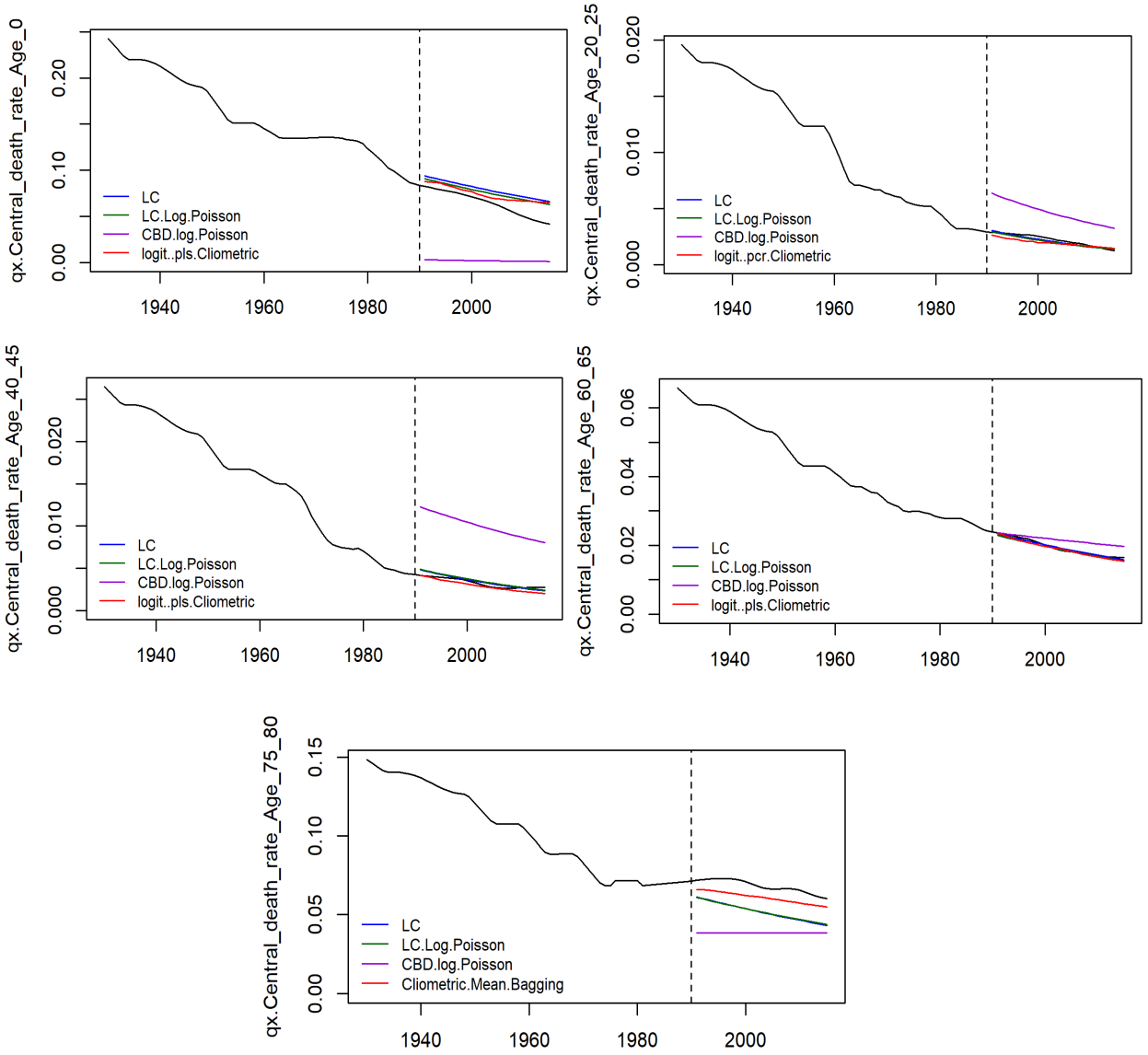


Prévisions comparées pour quelques âges (itération 2)

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1.

A titre illustratif, nous fournissons ci-après, les prévisions comparées pour quelques tranches d'âge : 0, [20 ; 25[, [40 ; 45[, [60 ; 65[, [75 ; 80[

Figure 29 : Prévisions comparées des modèles, pour quelques âges



13.3. Tests de sensibilité

13.3.1. Scénario alternatif 1

Paramétrage d'exploration alternative du pays A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.

Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Danemark"), ("Spain"), ("France"), ("Italy"), ("Sweden")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans

- Liste de vecteurs de sexes testés : ("M"), ("F"), ("MF")
- Liste de vecteurs d'âges d'appariement testés : Age 0
- Liste des transformations préalables (moyennes mobiles) : ("Interpolation + CMA5","Interpolation + CMA7")

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

Le nombre de pays A à explorer est passé de 2 pour le meilleur modèle à 5 dans le scenario 1 de test de sensibilité.
D'autre part le nombre de transformations à explorer est passé de 1 pour le meilleur modèle à 2 dans le scenario 1 de test de sensibilité.
Les autres paramètres sont restés inchangés entre le meilleur modèle et le scenario 1 de test de sensibilité.
Il y a donc une augmentation importante du nombre de choix paramétriques à explorer.

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 30 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

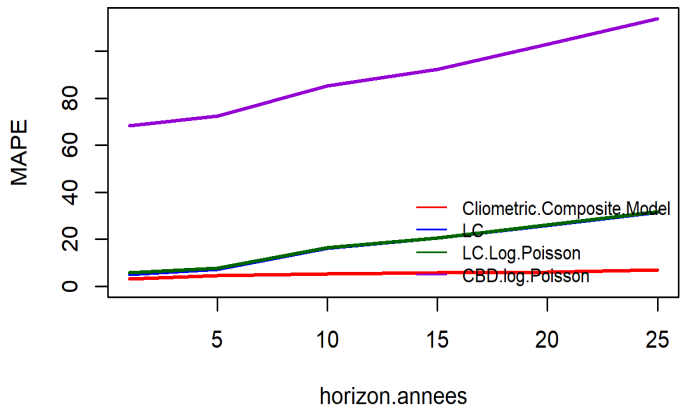
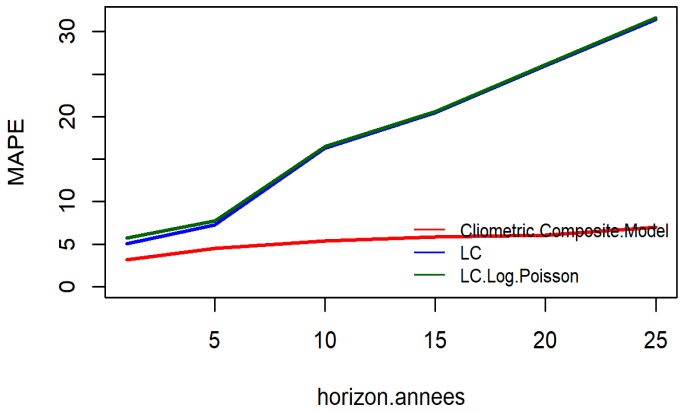


Figure 31 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette 1ère itération, nous constatons un avantage systématique et écrasant pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision.
Ces écarts très importants laissent présager un phénomène de surapprentissage dû à la hausse des choix de paramétrage exploratoire.

Tableau 11 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	68,3	72,5	85,3	92,3	103,0	113,9
Cliometric.Composite.Model	3,1	4,5	5,4	5,8	6,1	7,0
LC	5,1	7,3	16,3	20,5	26,0	31,4
LC.Log.Poisson	5,7	7,7	16,5	20,6	26,1	31,6

Pour chaque horizon, les prévisions de toutes les années précédentes sont incluses dans le calcul du MAPE (et cela en considérant tous les âges). Chaque MAPE est donc calculé sur une fenêtre dont l'étendue augmente au fur et à mesure que l'horizon augmente : cela peut expliquer l'allure lissée des courbes MAPE.

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 32 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

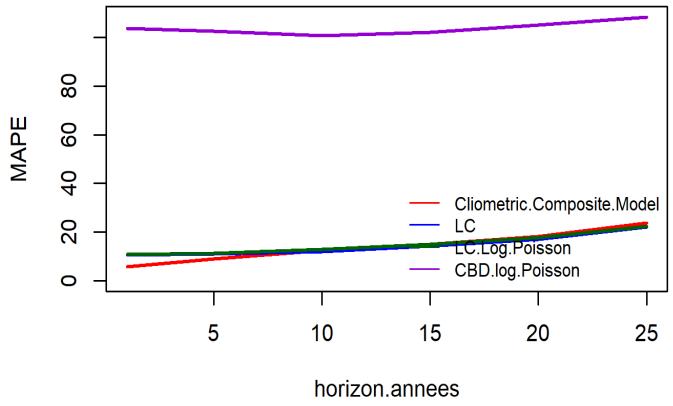
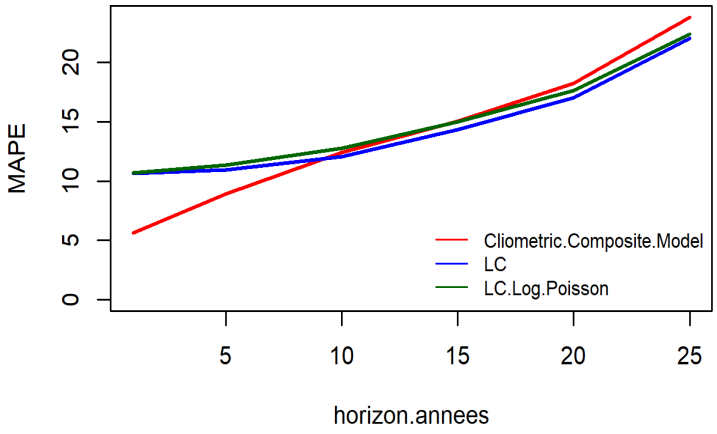


Figure 33 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette deuxième itération, nous constatons un **désavantage** systématique pour le modèle mixte cliométrique composite pour tous les horizons de prévision, à l'exception des horizons 1 et 5 ans.
Cette dégradation spectaculaire des performances du modèle mixte cliométrique composite, confirme l'hypothèse d'un phénomène de surapprentissage dû à la hausse des choix de paramétrage exploratoire.

Tableau 12 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,7	102,6	100,8	102,2	105,1	108,3
Cliometric.Composite.Model	5,6	8,9	12,4	15,0	18,2	23,8
LC	10,6	11,0	12,0	14,4	17,0	22,1
LC.Log.Poisson	10,7	11,4	12,8	15,0	17,6	22,4

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 34 : MAPE tous horizons confondus, selon les différents âges modélisés

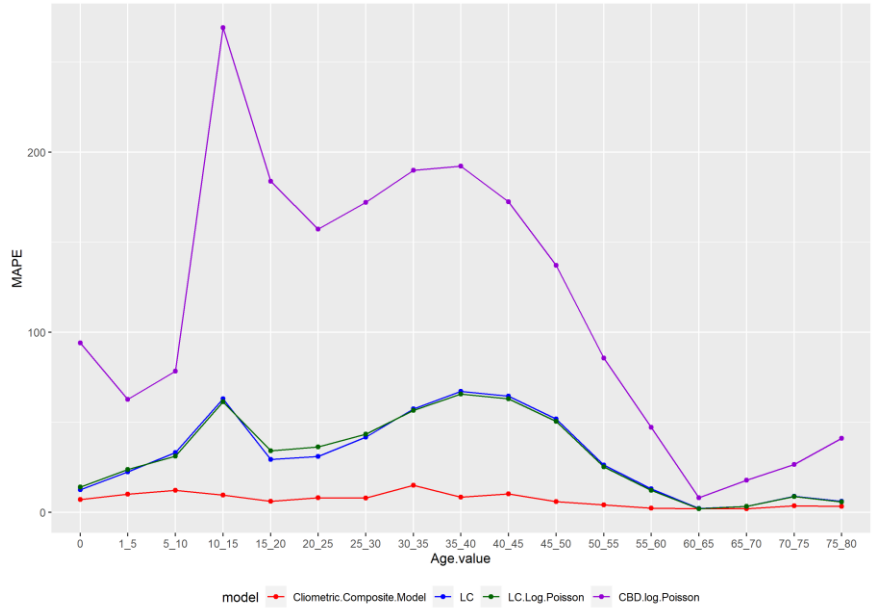
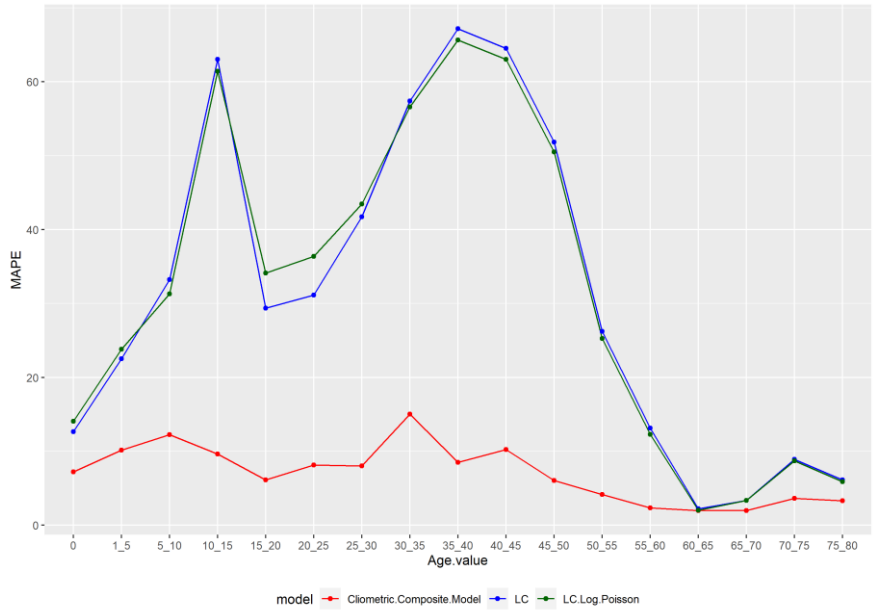


Figure 35 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette première itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié.

Tableau 13 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	94,0	7,2	12,6	14,1
1_5	62,8	10,2	22,5	23,8
10_15	269,1	9,6	63,0	61,4
15_20	183,9	6,1	29,4	34,1
20_25	157,2	8,1	31,1	36,4
25_30	172,1	8,0	41,7	43,4
30_35	190,0	15,0	57,4	56,6
35_40	192,2	8,5	67,2	65,6
40_45	172,5	10,2	64,5	63,0
45_50	137,2	6,1	51,8	50,5
5_10	78,4	12,2	33,2	31,3
50_55	85,6	4,2	26,2	25,3
55_60	47,3	2,4	13,1	12,3
60_65	8,1	2,0	2,2	2,0
65_70	17,9	2,0	3,3	3,3
70_75	26,6	3,6	8,9	8,7
75_80	41,2	3,3	6,1	5,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 36 : MAPE tous horizons confondus, selon les différents âges modélisés

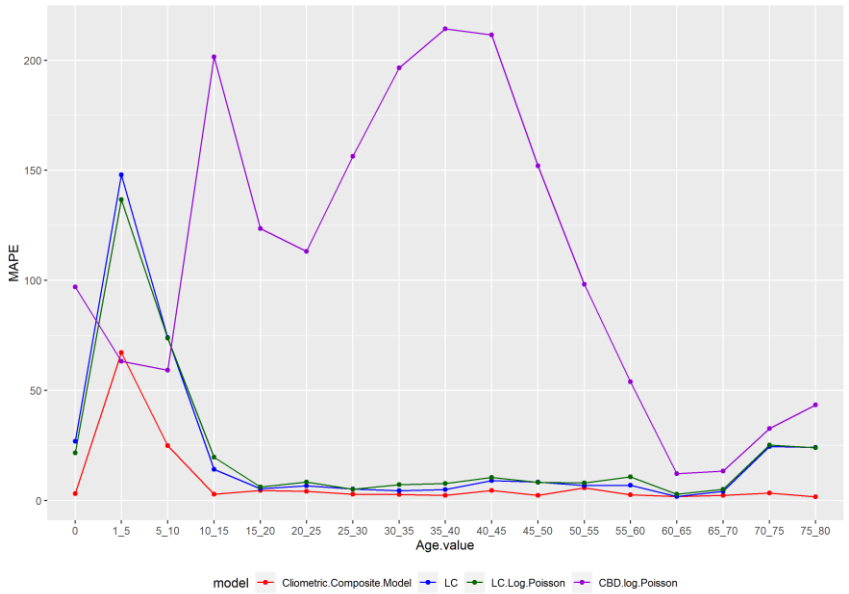
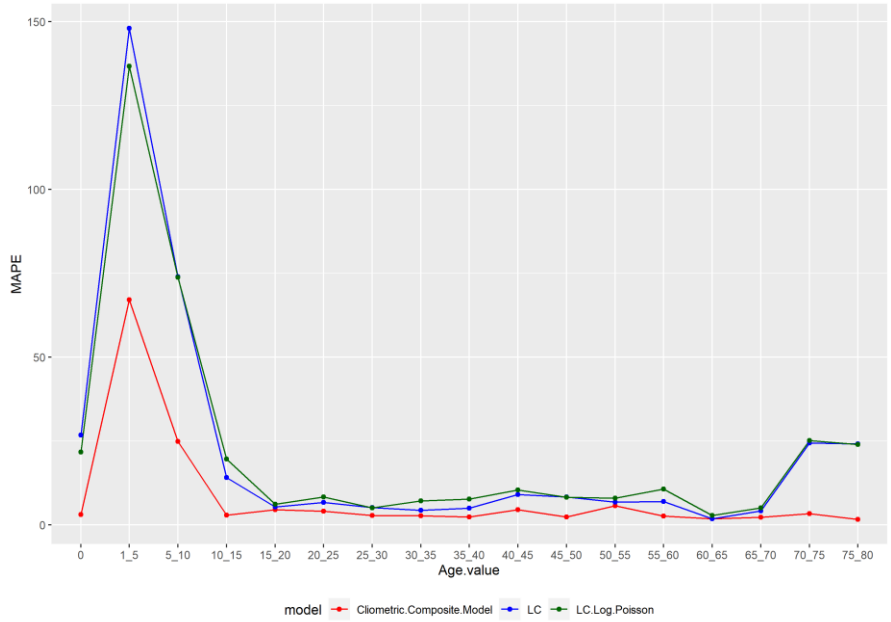


Figure 37 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Mais l'écart avec les modèles classiques s'est fortement réduit. Cette baisse relative de performance est cohérente avec celle observée sur les horizons de prévision, sans pour autant inverser la hiérarchie des modèles (contrairement à ce qui a été observé sur les horizons de prévisions).

Tableau 14 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	97,0	3,1	26,8	21,7
1_5	63,3	67,1	148,0	136,7
10_15	201,5	2,9	14,1	19,6
15_20	123,5	4,5	5,4	6,1
20_25	113,1	4,1	6,7	8,3
25_30	156,3	2,8	5,2	5,0
30_35	196,6	2,7	4,3	7,2
35_40	214,2	2,4	4,9	7,7
40_45	211,6	4,5	9,1	10,4
45_50	152,0	2,3	8,4	8,3
5_10	59,1	24,8	74,0	73,8
50_55	98,2	5,7	6,8	8,0
55_60	53,9	2,6	6,9	10,7
60_65	12,1	1,8	1,8	2,8
65_70	13,4	2,3	4,2	5,1
70_75	32,6	3,4	24,4	25,1
75_80	43,4	1,6	24,2	23,9

13.3.2. Scénario alternatif 2

Paramétrage d'exploration alternative du pays A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.

Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Danemark"), ("Spain"), ("France"), ("Italy"), ("Sweden")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans
- Liste de vecteurs de sexes testés : ("F","M","MF")
- Liste de vecteurs d'âges d'appariement testés : Age 0
- Liste des transformations préalables (moyennes mobiles) : ("Interpolation + CMA5","Interpolation + CMA7")

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

Le nombre de pays A à explorer est passé de 2 pour le meilleur modèle à 5 dans le scenario 2 de test de sensibilité.

D'autre part le nombre de transformations à explorer est passé de 1 pour le meilleur modèle à 2 dans le scenario 2 de test de sensibilité.

Au niveau des paramètres du sexe, les 3 modalités sont intégrées simultanément et systématiquement dans tous les modèles testés, et non plus une seule à la fois comme dans le scenario de paramétrage du meilleur model.
Les autres paramètres sont restés inchangés entre le meilleur modèle et le scenario 2 de test.
Il y a donc une augmentation du nombre de choix paramétriques à explorer, mais avec une plus grande quantité d'information par combinaison paramétrique (dû à l'usage simultané des 3 modalités de sexe).

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 38 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

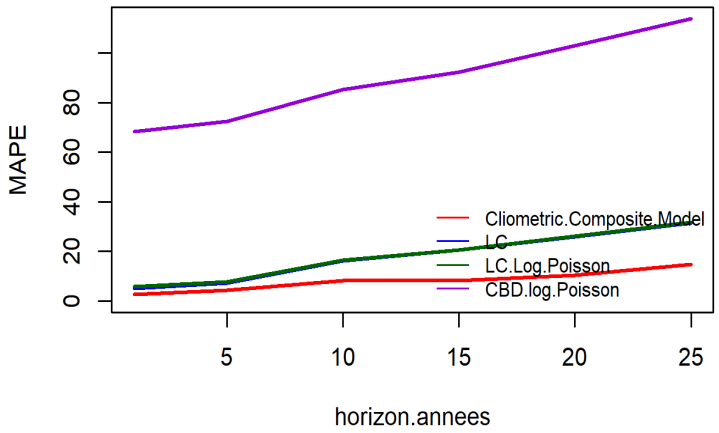
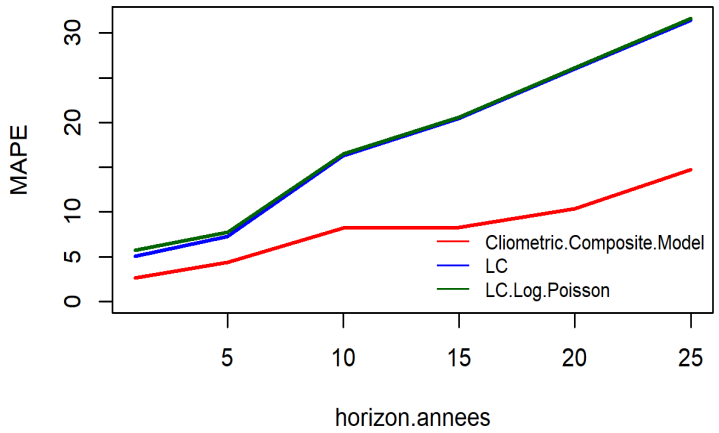


Figure 39 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette première itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Mais cette domination est nettement moins forte que lors de l'inflation précédente des choix paramétriques à explorer. Le phénomène de surapprentissage n'est pas exclu au vu des grands écarts de performance associés à une hausse des choix paramétriques (même si cette hausse de choix est moins forte que pour le scénario 1 de test).

Tableau 15 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	68,3	72,5	85,3	92,3	103,0	113,9
Cliometric.Composite.Model	2,6	4,4	8,2	8,3	10,4	14,8
LC	5,1	7,3	16,3	20,5	26,0	31,4
LC.Log.Poisson	5,7	7,7	16,5	20,6	26,1	31,6

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 40 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

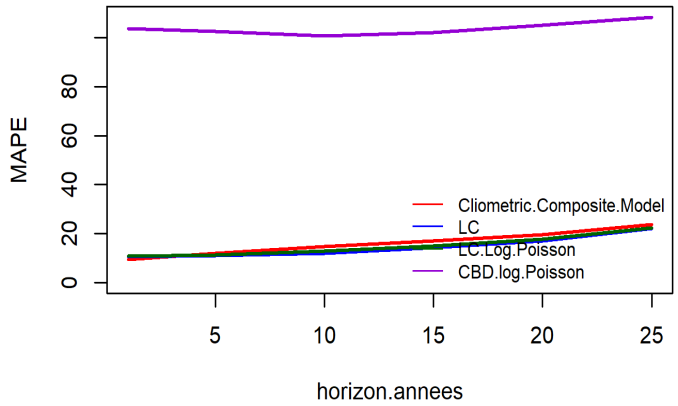
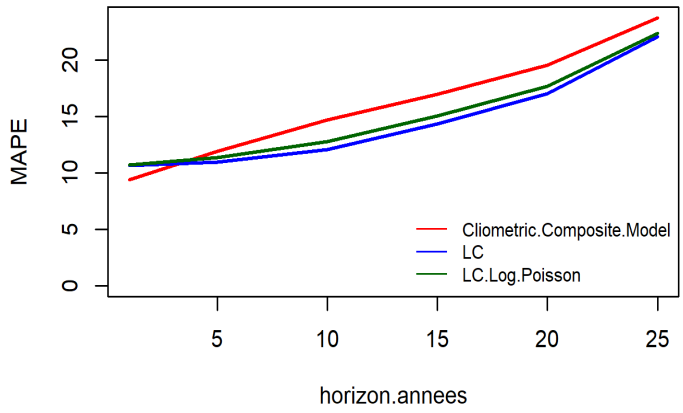


Figure 41 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette deuxième itération, nous constatons un **désavantage** systématique pour le modèle mixte cliométrique composite pour tous les horizons de prévision, à l'exception de l'horizon de 1 an.
Cette dégradation spectaculaire des performances du modèle mixte cliométrique composite, confirme l'hypothèse d'un phénomène de surapprentissage dû à la hausse des choix de paramétrage exploratoire.

Tableau 16 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,7	102,6	100,8	102,2	105,1	108,3
Cliometric.Composite.Model	9,4	11,9	14,7	17,0	19,5	23,7
LC	10,6	11,0	12,0	14,4	17,0	22,1
LC.Log.Poisson	10,7	11,4	12,8	15,0	17,6	22,4

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 42 : MAPE tous horizons confondus, selon les différents âges modélisés

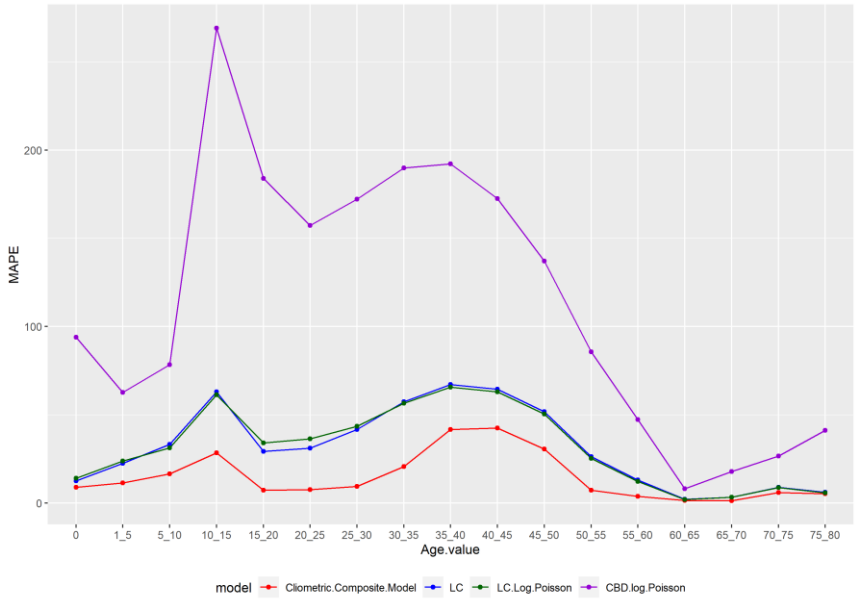
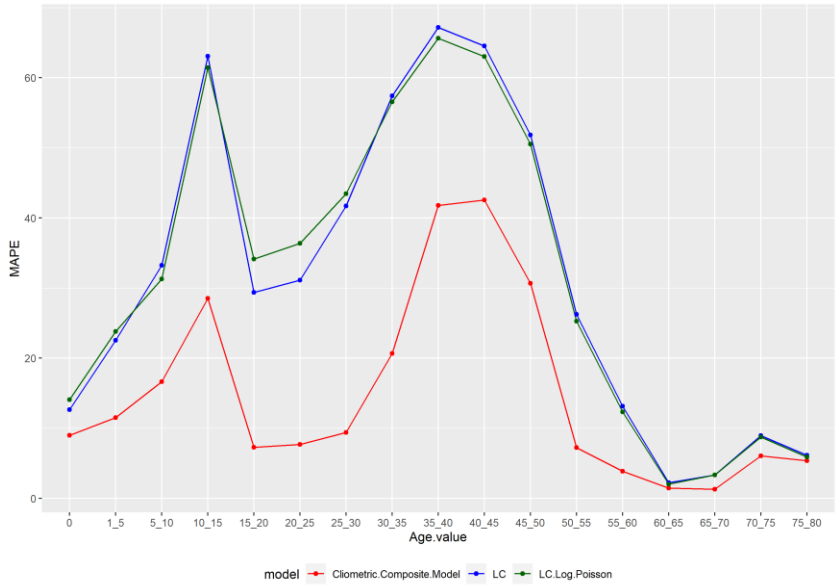


Figure 43 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette première itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié.

Tableau 17 : MAPE (%)

MAPE5 (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	94,0	9,0	12,6	14,1
1_5	62,8	11,5	22,5	23,8
10_15	269,1	28,5	63,0	61,4
15_20	183,9	7,3	29,4	34,1
20_25	157,2	7,7	31,1	36,4
25_30	172,1	9,4	41,7	43,4
30_35	190,0	20,7	57,4	56,6
35_40	192,2	41,8	67,2	65,6
40_45	172,5	42,6	64,5	63,0
45_50	137,2	30,7	51,8	50,5
5_10	78,4	16,6	33,2	31,3
50_55	85,6	7,2	26,2	25,3
55_60	47,3	3,8	13,1	12,3
60_65	8,1	1,5	2,2	2,0
65_70	17,9	1,3	3,3	3,3
70_75	26,6	6,0	8,9	8,7
75_80	41,2	5,4	6,1	5,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 44 : MAPE tous horizons confondus, selon les différents âges modélisés

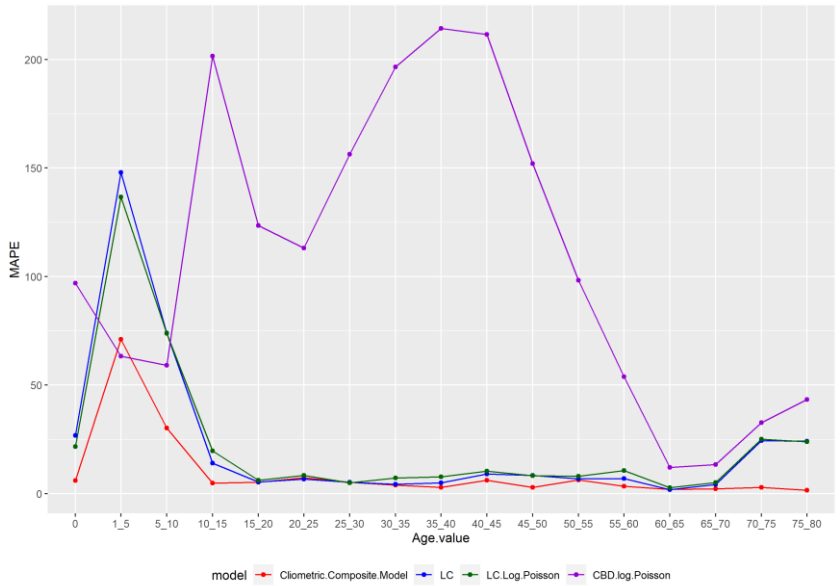
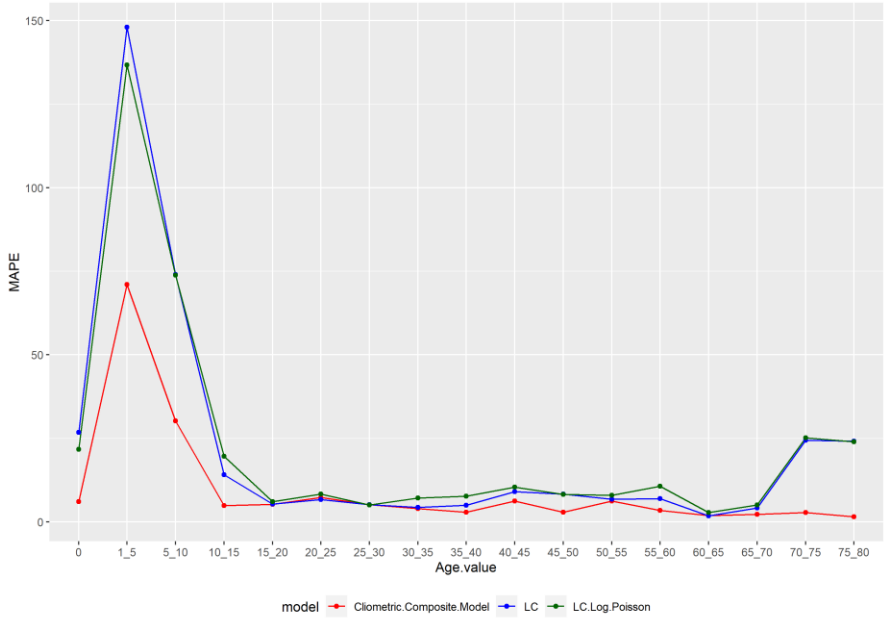


Figure 45 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié (à l'exception mitigée de la tranche 20-25 ans). Mais l'écart avec les modèles classiques s'est fortement réduit.

Cette baisse relative de performance est cohérente avec celle observée sur les horizons de prévision, sans pour autant inverser la hiérarchie des modèles (contrairement à ce qui a été observé sur les horizons de prévisions).

Toutefois, cette baisse des performances du modèle mixte cliométrique composite, confirme l'hypothèse d'un surapprentissage dû à l'inflation des choix paramétriques à explorer.

Tableau 18 : MAPE (%)

MAPE5 (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	97,0	6,1	26,8	21,7
1_5	63,3	71,0	148,0	136,7
10_15	201,5	4,8	14,1	19,6
15_20	123,5	5,2	5,4	6,1
20_25	113,1	7,4	6,7	8,3
25_30	156,3	5,2	5,2	5,0
30_35	196,6	4,0	4,3	7,2
35_40	214,2	2,9	4,9	7,7
40_45	211,6	6,2	9,1	10,4
45_50	152,0	2,9	8,4	8,3
5_10	59,1	30,2	74,0	73,8
50_55	98,2	6,2	6,8	8,0
55_60	53,9	3,4	6,9	10,7
60_65	12,1	1,9	1,8	2,8
65_70	13,4	2,2	4,2	5,1
70_75	32,6	2,8	24,4	25,1
75_80	43,4	1,5	24,2	23,9

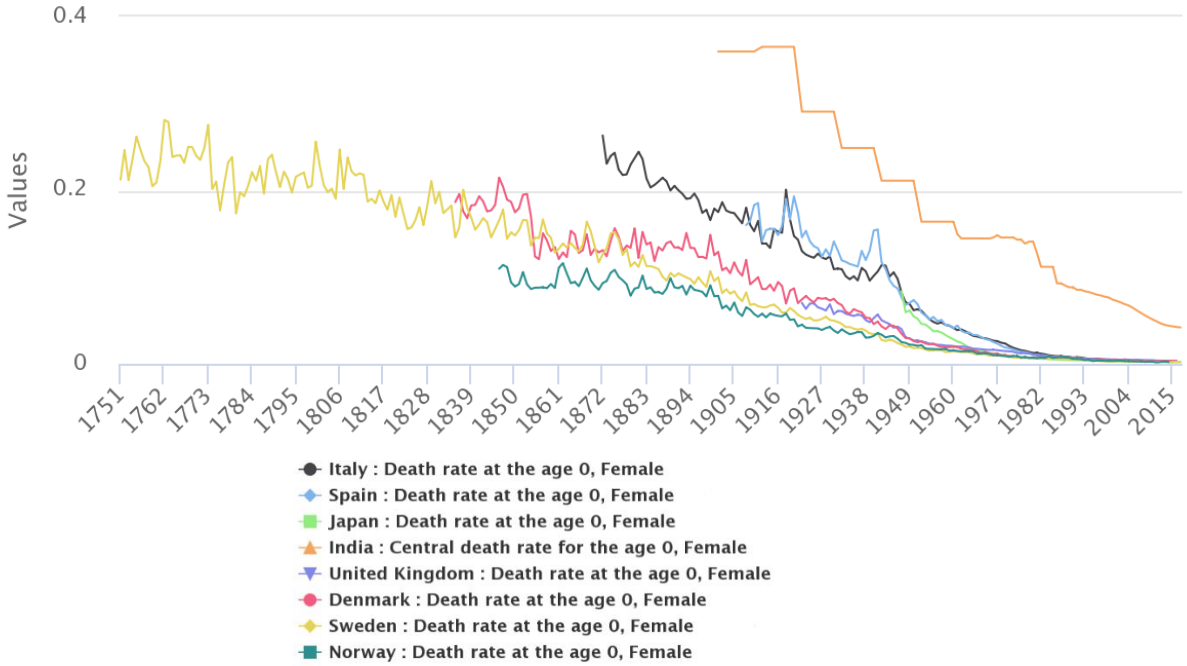
13.3.3. Conclusion sur les tests de sensibilité

Nous sommes dans le contexte d'un meilleur modèle (modèle mixte cliométrique composite) qui présente un très léger effet de surapprentissage sur le long terme. Le dimensionnement du modèle est donc à la lisière du surapprentissage.

Au vu des deux scénarios alternatifs étudiés, l'inflation du nombre de choix paramétriques, crée un risque de surapprentissage. Ce phénomène de surdimensionnement dégrade plus fortement les performances par horizon de prévision, que les performances par âge.
L'impact sur les performances par horizon est si dégradé que la hiérarchie des modèles en est vite inversée en défaveur du modèle mixte cliométrique composite.

Un dosage délicat et une série de tests sur le nombre et les modalités paramétriques à explorer, doivent donc être effectués pour tirer le meilleur parti du modèle mixte cliométrique composite.
C'est ce que nous avons effectué en privilégiant les pays A de référence, à transition plus récente en Europe, notamment l'Italie et l'Espagne.
Effectivement, l'Italie et l'Espagne font partie des derniers pays européens (donc transitionnellement « âgés ») à entrer en transition démographique, ils font donc partie des pays européens les plus « proches » transitionnellement de l'Inde, comme l'illustre le graphique ci-après :

Figure 46 : Taux de mortalité à l'âge 0 des femmes. Quelques pays européens vs Inde



14. Modélisation des mortalités pour la table prospective de l'Equateur

Les modèles dont les performances sont comparées sont : Lee-carter (LC), LC Log-Poisson, CBD Log-Poisson, modèle mixte cliométrique composite à facteurs PCR-Optimal/PLS (Cliometric composite model)

14.1. Paramétrages associés au pays « R » transitionnellement « moins âgé » (ou pays d'expérience) à modéliser

- Pays « R » : Equateur
- Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).
- Variable étudiée : Probabilité de décès (femmes)
- Liste des transformations préalables (moyennes mobiles) : Aucune
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans

Itérations :

- Itération 1 : Détection **a posteriori** de la meilleure classe de modèle (Best Model Class) pour chaque âge modélisé
    - Début de l'historique étudié : 1950
    - Année actuelle supposée : 1969
    - Horizon de prévision : 1994 (25 ans)
- Itération 2 : Utilisation **a priori** de la meilleure classe de modèle (Best Model Class) pour chaque âge modélisé
    - Début de l'historique étudié : 1950
    - Année actuelle supposée : 1994
    - Horizon de prévision : 2019 (25 ans)

14.2. Paramétrage et performances du meilleur modèle obtenu

14.2.1. Paramétrages d'exploration associés aux pays de référence A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.
Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Danemark"), ("Spain"), ("France"), ("Italy"), ("Sweden")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans
- Liste de vecteurs de sexes testés : ("M"), ("F"), ("MF")
- Liste de vecteurs d'âges d'appariement testés : ('0'), ('1-5'), ('5-10'), ('0', '1-5', '5-10')
- Liste des transformations préalables (moyennes mobiles) : ("Interpolation","Interpolation + CMA5","Interpolation + CMA7", "Interpolation + CMA9")

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

14.2.2. Performances du meilleur modèle obtenu

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 47 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

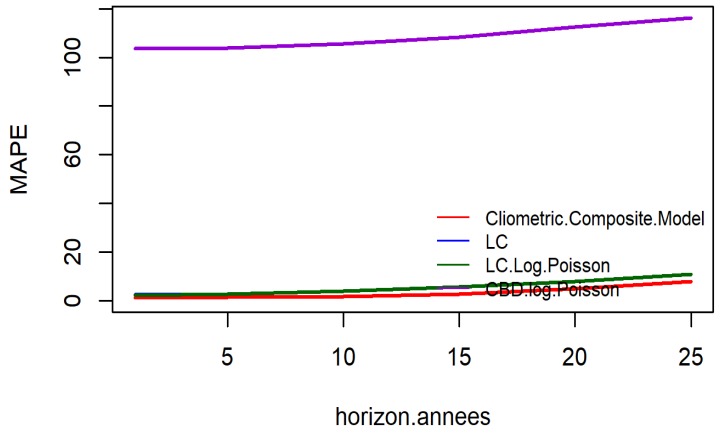
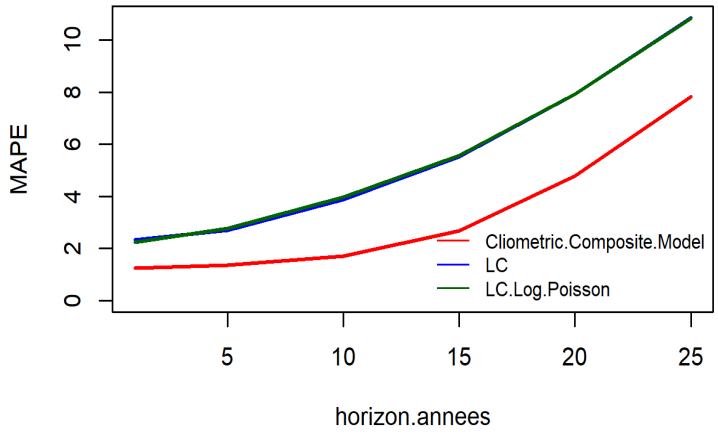


Figure 48 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette première itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié.

Tableau 19 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,5	103,9	105,6	108,3	112,5	116,2
Cliometric.Composite.Model	1,2	1,4	1,7	2,7	4,8	7,8
LC	2,3	2,7	3,9	5,5	7,9	10,9
LC.Log.Poisson	2,2	2,8	4,0	5,6	7,9	10,8

Pour chaque horizon, les prévisions de toutes les années précédentes sont incluses dans le calcul du MAPE (et cela en considérant tous les âges). Chaque MAPE est donc calculé sur une fenêtre dont l'étendue augmente au fur et à mesure que l'horizon augmente : cela peut expliquer l'allure lissée des courbes MAPE.

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 49 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

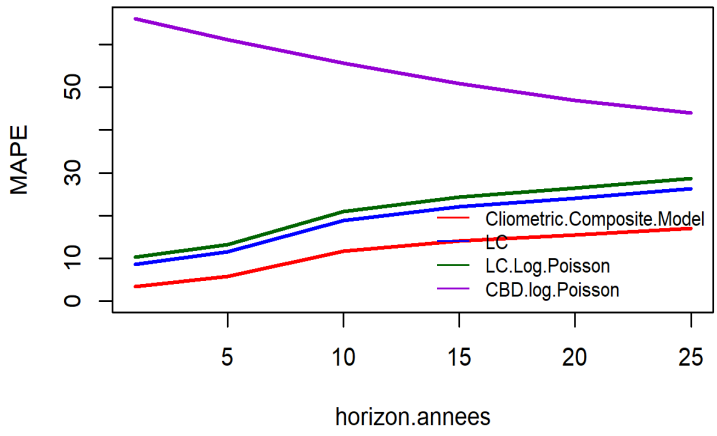
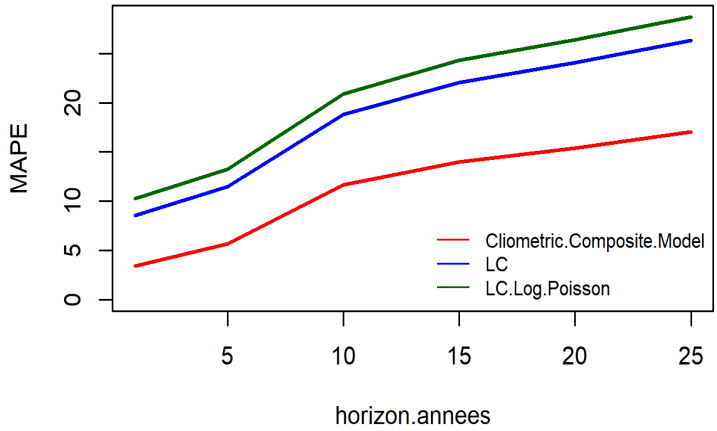


Figure 50 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette 2ième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié. Cela est en cohérence forte avec le schéma de la première itération. Ce qui exclut a priori un phénomène de surapprentissage malgré un fort dimensionnement paramétrique.

Tableau 20 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	66,1	61,1	55,7	50,8	46,9	44,0
Cliometric.Composite.Model	3,4	5,7	11,7	14,0	15,4	17,0
LC	8,6	11,5	18,8	22,1	24,1	26,3
LC.Log.Poisson	10,2	13,2	20,9	24,3	26,4	28,7

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 51 : MAPE tous horizons confondus, selon les différents âges modélisés

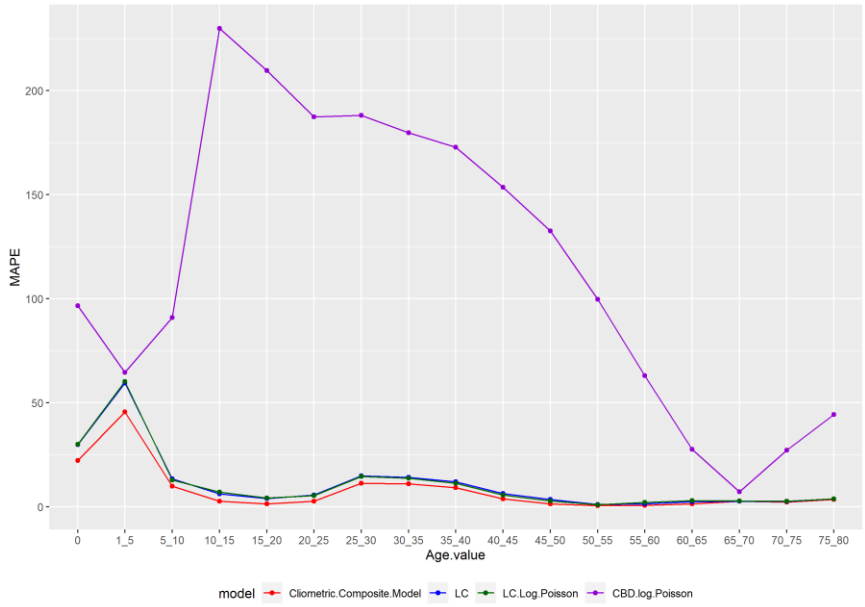
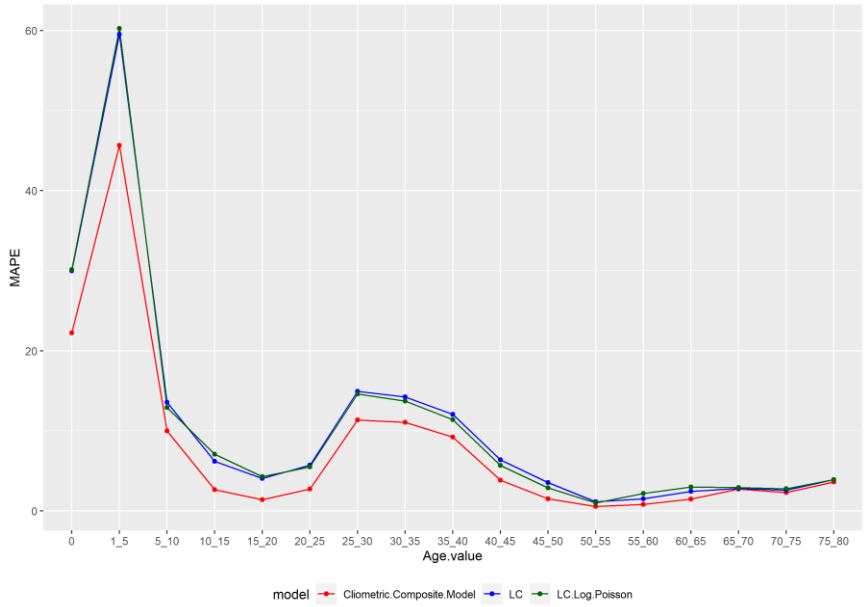


Figure 52 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette 1ère itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié.

Tableau 21 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	96,6	22,2	30,0	30,1
1_5	64,7	45,6	59,5	60,2
5_10	91,0	10,0	13,6	12,9
10_15	229,9	2,7	6,2	7,1
15_20	209,6	1,4	4,1	4,3
20_25	187,5	2,7	5,7	5,5
25_30	188,2	11,4	14,9	14,6
30_35	179,8	11,1	14,2	13,7
35_40	172,9	9,2	12,1	11,4
40_45	153,5	3,9	6,4	5,7
45_50	132,6	1,5	3,5	2,9
50_55	99,8	0,5	1,1	1,0
55_60	63,1	0,8	1,5	2,2
60_65	27,6	1,5	2,4	3,0
65_70	7,2	2,7	2,8	2,9
70_75	27,3	2,3	2,6	2,8
75_80	44,4	3,6	3,9	3,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 53 : MAPE tous horizons confondus, selon les différents âges modélisés

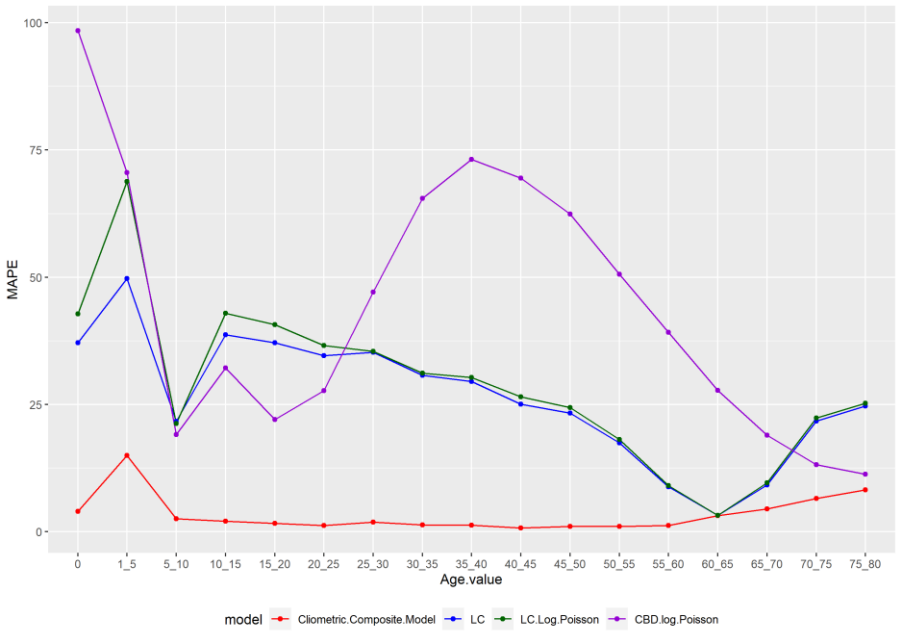
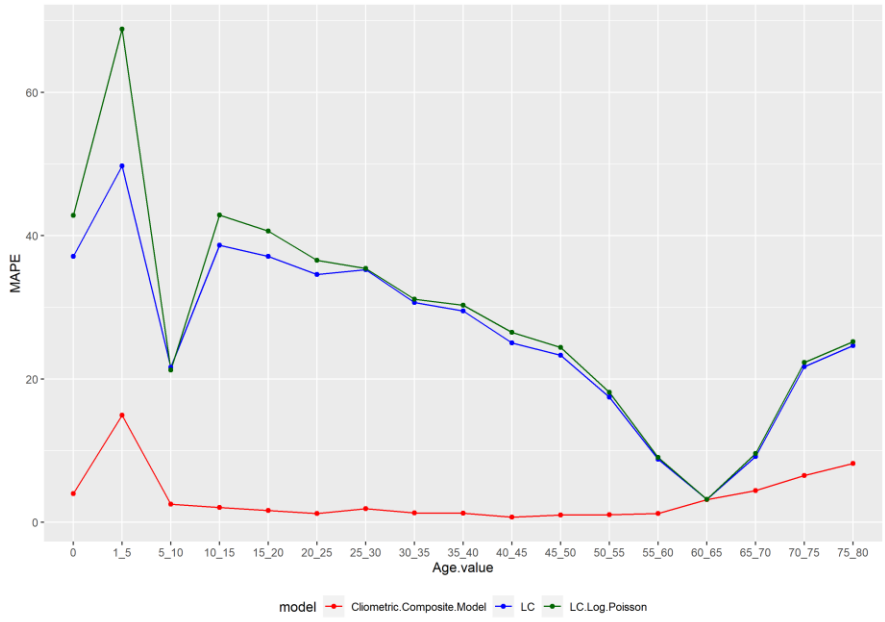


Figure 54 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette 2ème itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Les avantages se sont même largement renforcés.
La hiérarchie de la première itération est conservée et renforcée.

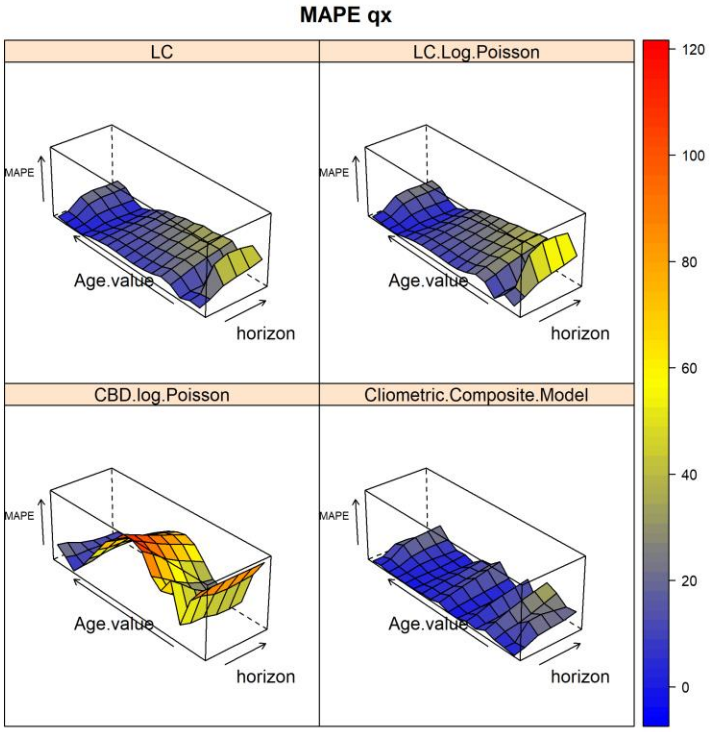
Tableau 22 : MAPE (%)

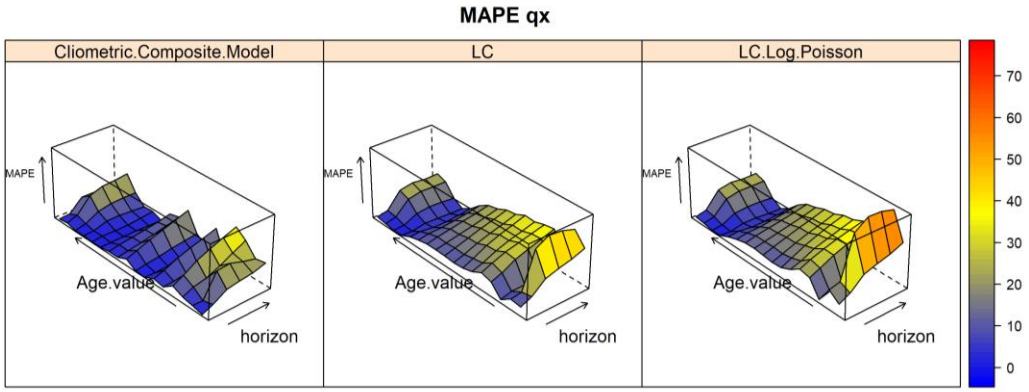
MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	98,4	4,0	37,1	42,8
1_5	70,6	14,9	49,7	68,8
5_10	19,1	2,5	21,6	21,2
10_15	32,2	2,1	38,7	42,9
15_20	22,0	1,7	37,1	40,7
20_25	27,7	1,2	34,6	36,6
25_30	47,1	1,9	35,3	35,4
30_35	65,5	1,3	30,7	31,1
35_40	73,1	1,3	29,5	30,3
40_45	69,5	0,7	25,0	26,5
45_50	62,4	1,0	23,3	24,4
50_55	50,6	1,0	17,5	18,1
55_60	39,2	1,2	8,8	9,0
60_65	27,8	3,1	3,2	3,2
65_70	18,9	4,4	9,2	9,6
70_75	13,2	6,5	21,7	22,3
75_80	11,3	8,2	24,7	25,2

MAPE des probabilités de décès selon les âges et les horizons (itération 2)

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1.

Figure 55 : MAPE (%) des probabilités de décès selon les âges et les horizons

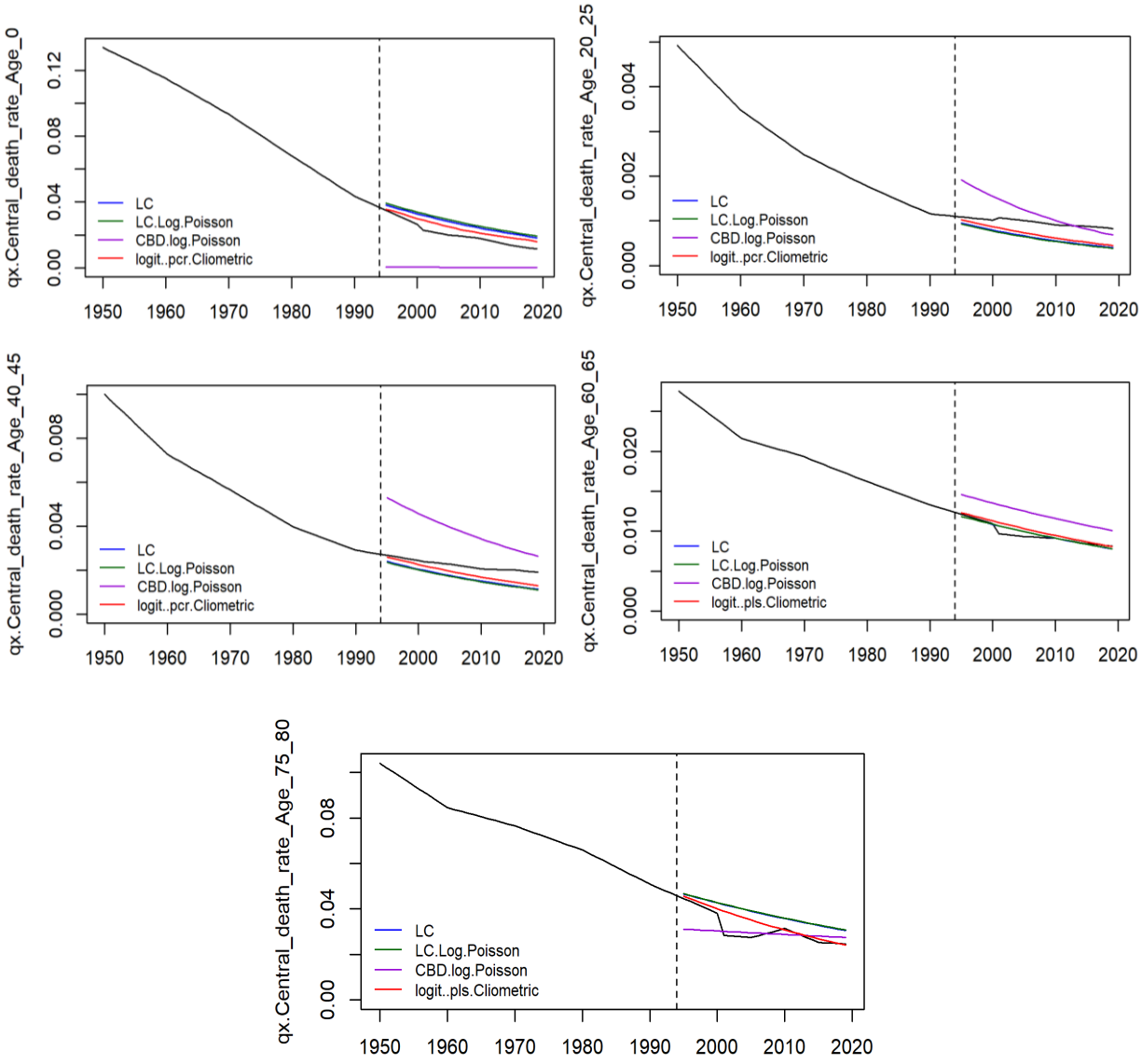




Prévisions comparées pour quelques âges (itération 2)

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1.
A titre illustratif, nous fournissons ci-après, les prévisions comparées pour quelques tranches d'âge : 0, [20 ; 25[, [40 ; 45[, [60 ; 65[, [75 ; 80[

Figure 56 : Prévisions comparées des modèles, pour quelques âges



14.3. Tests de sensibilité

14.3.1. Scénario alternatif 1

Paramétrage d'exploration alternative du pays A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.

Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Italy"), ("Spain")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans
- Liste de vecteurs de sexes testés : ("M"), ("F"), ("MF")
- Liste de vecteurs d'âges d'appariement testés : ('0'), ('1-5'), ('0', '1-5')
- Liste des transformations préalables (moyennes mobiles) : ("Interpolation", "Interpolation + CMA5","Interpolation + CMA7")

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

Le nombre de pays A à explorer est réduit de 5 pour le meilleur modèle à 2 dans le scenario 1 de test de sensibilité.

D'autre part le nombre de transformations à explorer est réduit de 4 pour le meilleur modèle à 3 dans le scenario 1 de test de sensibilité.

Par ailleurs le nombre de modalités d'âges à explorer est réduit de 4 pour le meilleur modèle à 3 dans le scenario 1 de test de sensibilité

Les autres paramètres sont restés inchangés entre le meilleur modèle et le scenario 1 de test de sensibilité.

Il y a donc une **réduction importante** du nombre de choix paramétriques à explorer.

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 57 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

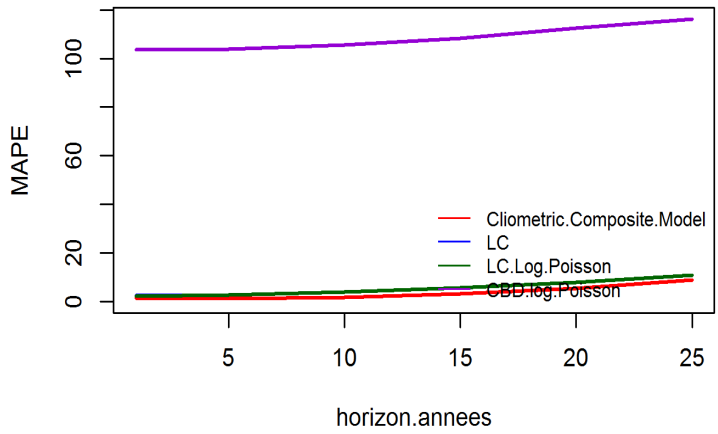
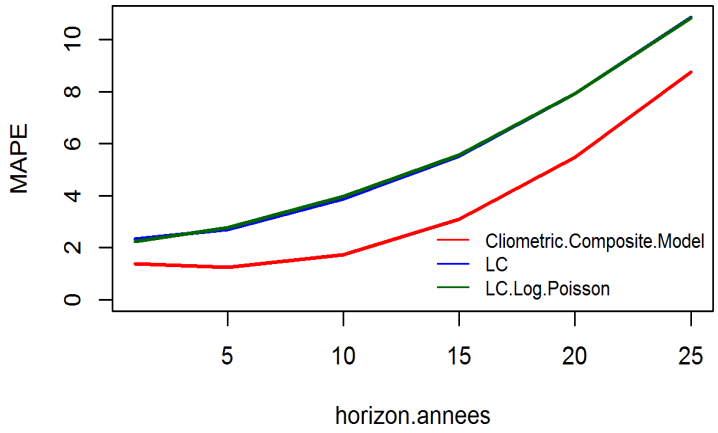


Figure 58 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette 1ère itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision.

Tableau 23 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,5	103,9	105,6	108,3	112,5	116,2
Cliometric.Composite.Model	1,4	1,3	1,7	3,1	5,5	8,8
LC	2,3	2,7	3,9	5,5	7,9	10,9
LC.Log.Poisson	2,2	2,8	4,0	5,6	7,9	10,8

Pour chaque horizon, les prévisions de toutes les années précédentes sont incluses dans le calcul du MAPE (et cela en considérant tous les âges). Chaque MAPE est donc calculé sur une fenêtre dont l'étendue augmente au fur et à mesure que l'horizon augmente : cela peut expliquer l'allure lissée des courbes MAPE.

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 59 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

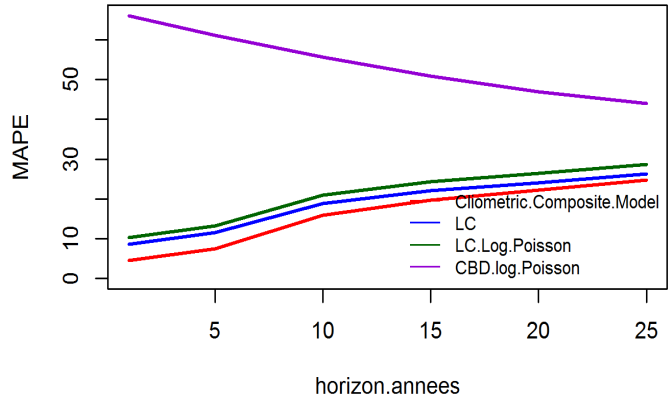
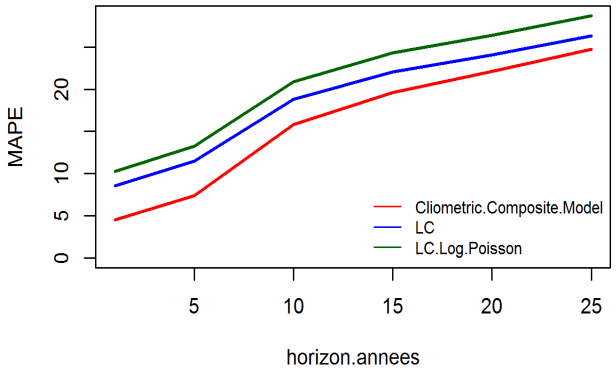


Figure 60 : Performance des modèles selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision.
Mais cela s'accompagne d'une baisse relative systématique des performances pour le modèle mixte cliométrique composite pour tous les horizons de prévision.
Ce constat est paradoxal car cela fait suite à une réduction des choix paramétriques du modèle mixte cliométrique composite.
Il ne suffirait donc pas de réduire le nombre de choix paramétriques pour améliorer les performances mécaniquement. Encore faut-il effectuer les réductions idoines. Cette réduction se révèle dégradante pour la performance face à un meilleur modèle riche en options de paramétrages et ne manifestant pas de signe de surapprentissage.

Tableau 24 : MAPE

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	66,1	61,1	55,7	50,8	46,9	44,0
Cliometric.Composite.Model	4,5	7,4	15,8	19,6	22,1	24,8
LC	8,6	11,5	18,8	22,1	24,1	26,3
LC.Log.Poisson	10,2	13,2	20,9	24,3	26,4	28,7

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 61 : MAPE tous horizons confondus, selon les différents âges modélisés

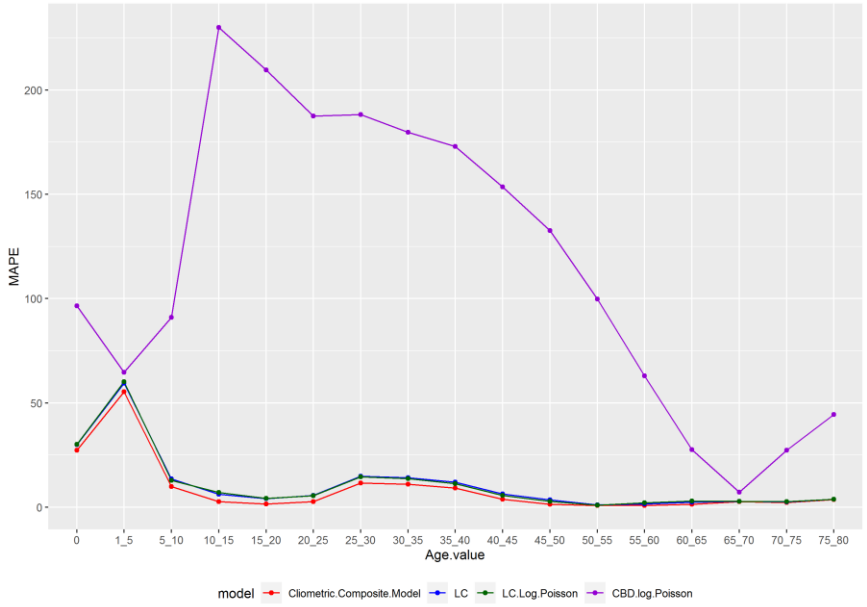
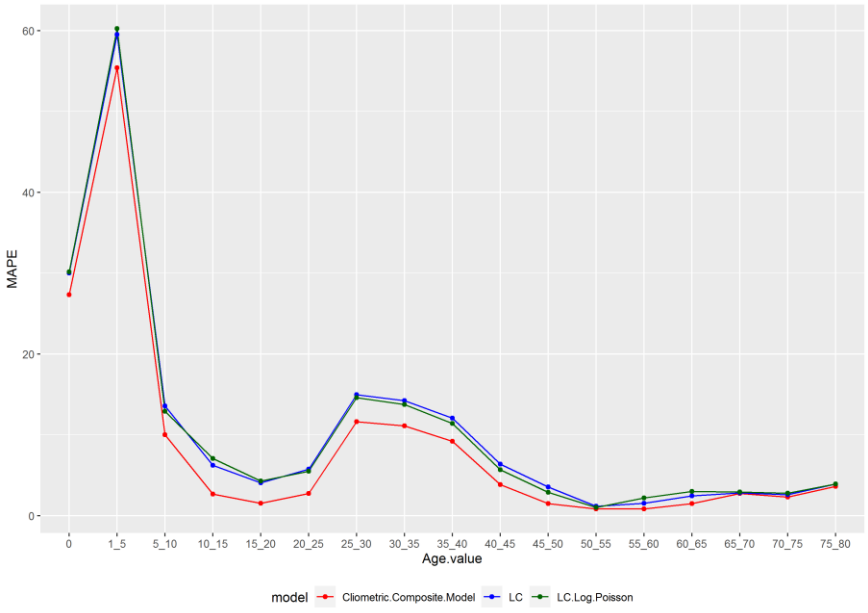


Figure 62 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette première itération, nous constatons un avantage presque systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié.

Tableau 25 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	96,6	27,3	30,0	30,1
1_5	64,7	55,4	59,5	60,2
5_10	91,0	10,0	13,6	12,9
10_15	229,9	2,7	6,2	7,1
15_20	209,6	1,5	4,1	4,3
20_25	187,5	2,7	5,7	5,5
25_30	188,2	11,6	14,9	14,6
30_35	179,8	11,1	14,2	13,7
35_40	172,9	9,2	12,1	11,4
40_45	153,5	3,9	6,4	5,7
45_50	132,6	1,5	3,5	2,9
50_55	99,8	0,8	1,1	1,0
55_60	63,1	0,8	1,5	2,2
60_65	27,6	1,5	2,4	3,0
65_70	7,2	2,7	2,8	2,9
70_75	27,3	2,3	2,6	2,8
75_80	44,4	3,6	3,9	3,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 63 : MAPE tous horizons confondus, selon les différents âges modélisés

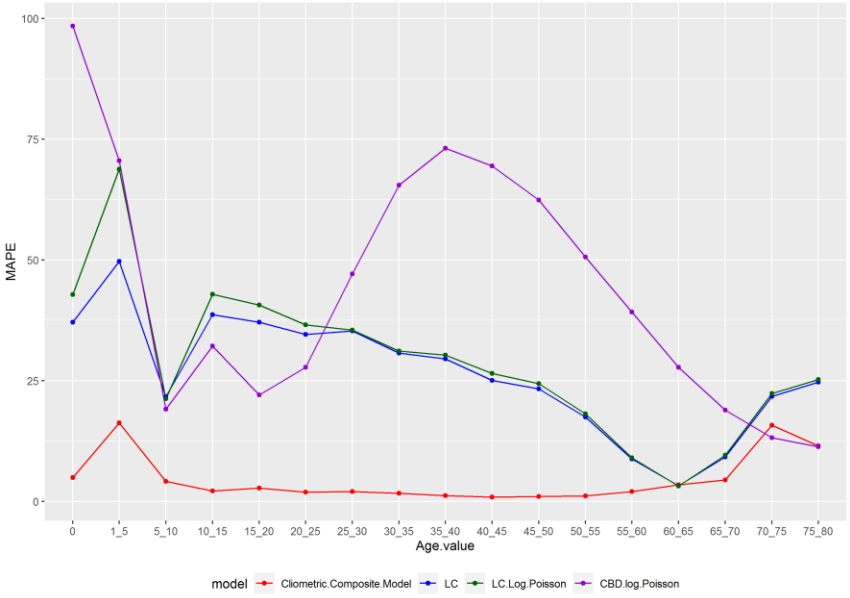
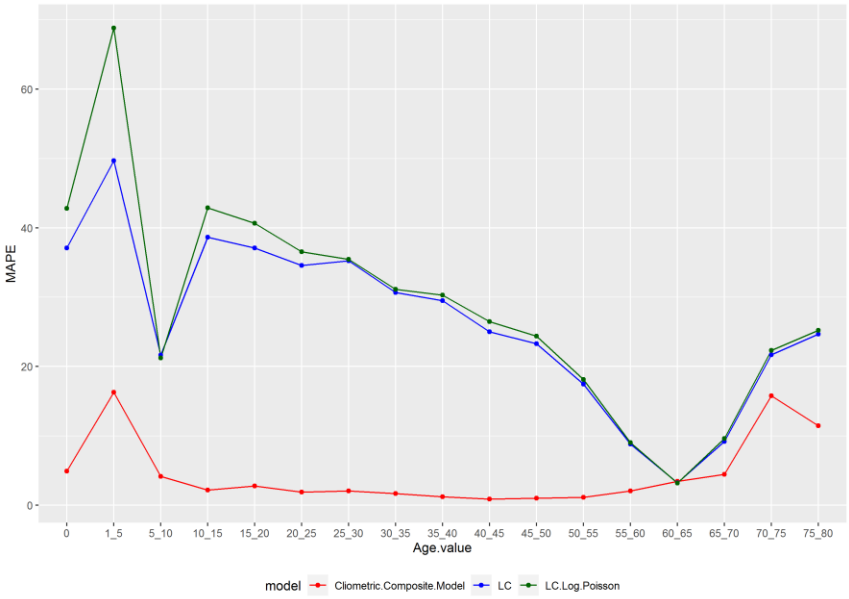


Figure 64 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Mais l'écart avec les modèles classiques s'est fortement agrandi. Ce constat exclut l'hypothèse de surapprentissage pour le modèle le plus performant.

Tableau 26 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	98,4	4,9	37,1	42,8
1_5	70,6	16,3	49,7	68,8
5_10	19,1	4,2	21,6	21,2
10_15	32,2	2,2	38,7	42,9
15_20	22,0	2,8	37,1	40,7
20_25	27,7	1,9	34,6	36,6
25_30	47,1	2,0	35,3	35,4
30_35	65,5	1,7	30,7	31,1
35_40	73,1	1,2	29,5	30,3
40_45	69,5	0,9	25,0	26,5
45_50	62,4	1,0	23,3	24,4
50_55	50,6	1,1	17,5	18,1
55_60	39,2	2,1	8,8	9,0
60_65	27,8	3,4	3,2	3,2
65_70	18,9	4,4	9,2	9,6
70_75	13,2	15,8	21,7	22,3
75_80	11,3	11,5	24,7	25,2

14.3.2. Scénario alternatif 2

Paramétrage d'exploration alternative du pays A

Lors de l'itération 1 de chaque âge modélisé, l'exploration consiste à tester chaque combinaison du produit cartésien des ensembles ci-dessous. Chaque ensemble désigne un paramètre et ses éléments sont les modalités du paramètre.

Pour chaque paramètre les modalités à tester sont :

- Liste des vecteurs de pays « A » : ("Italy", "Spain")
- Variable étudiée : Probabilité de décès
- Ages étudiés : Age 0, et tranches d'âge de 1 à 80 ans
- Liste de vecteurs de sexes testés : ("F","M","MF")
- Liste de vecteurs d'âges d'appariement testés : ('0', '1-5')
- Liste des transformations préalables (moyennes mobiles) : ("Interpolation", "Interpolation + CMA5","Interpolation + CMA7")

Source des données : Les données par tranche d'âge de l'Inde et l'Italie sont issues de la base de données Human Life-table Database (HLD) publiées par la Max Planck Institute for Demographic Research (2023).

Le nombre de pays A à explorer est réduit de 5 pour le meilleur modèle à 2 dans le scenario 2 de test de sensibilité.

D'autre part le nombre de transformations à explorer est réduit de 4 pour le meilleur modèle à 3 dans le scenario 2 de test de sensibilité.

Au niveau des paramètres du sexe, les 3 modalités sont intégrées simultanément et systématiquement dans tous les modèles testés, et non plus une seule à la fois comme dans le scenario de paramétrage du meilleur model.

Par ailleurs le nombre de modalités d'âges à explorer est réduit de 4 pour le meilleur modèle à 1 dans le scenario 2 de test de sensibilité

Les autres paramètres sont restés inchangés entre le meilleur modèle et le scenario 2 de test.

Il y a donc une **réduction importante** (encore plus forte que dans le scenario alternatif 1) du nombre de choix paramétriques à explorer.

MAPE tous âges confondus, selon les horizons testés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 65 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

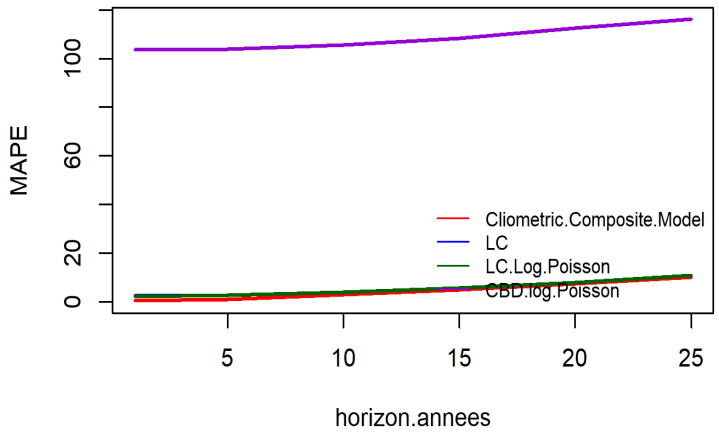
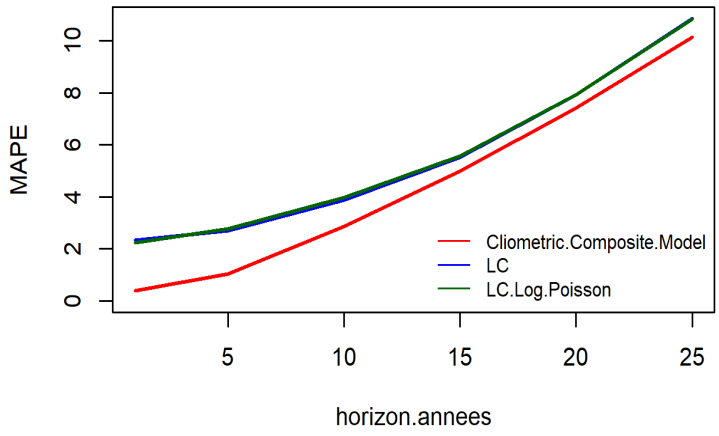


Figure 66 : Performance des modèles selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette 1ère itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision. Domination moins forte que dans le scenario alternatif 1.

Tableau 27 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	103,5	103,9	105,6	108,3	112,5	116,2
Cliometric.Composite.Model	0,4	1,0	2,9	5,0	7,4	10,1
LC	2,3	2,7	3,9	5,5	7,9	10,9
LC.Log.Poisson	2,2	2,8	4,0	5,6	7,9	10,8

Pour chaque horizon, les prévisions de toutes les années précédentes sont incluses dans le calcul du MAPE (et cela en considérant tous les âges). Chaque MAPE est donc calculé sur une fenêtre dont l'étendue augmente au fur et à mesure que l'horizon augmente : cela peut expliquer l'allure lissée des courbes MAPE.

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 67 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision

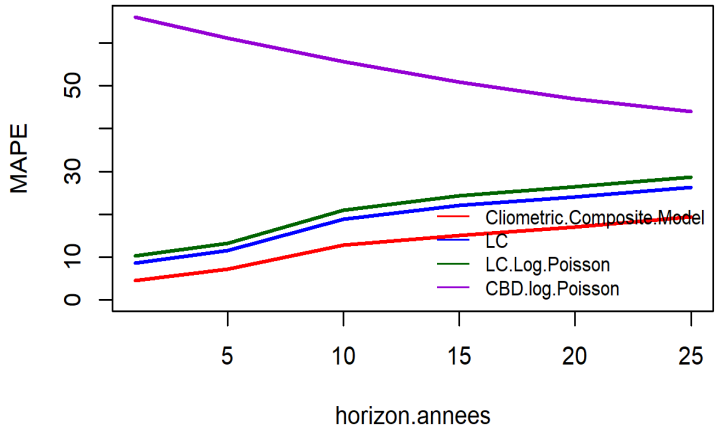
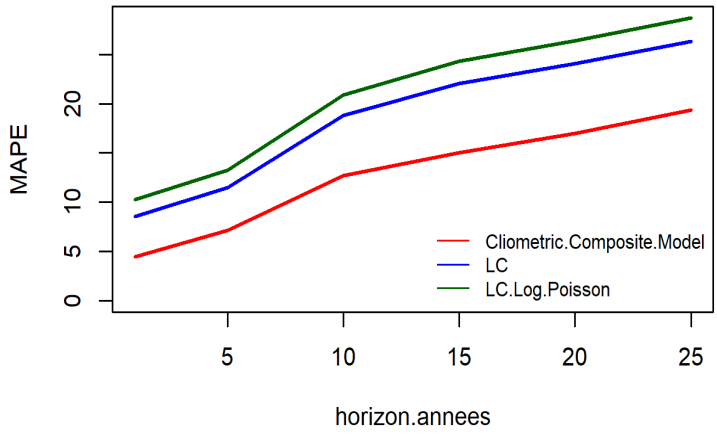


Figure 68 : Performance des modèles (tous âges confondus) selon l'horizon maximal de prévision. Sans le modèle CBD



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision.
De plus cela s'accompagne d'une hausse relative systématique des performances relatives pour le modèle mixte cliométrique composite pour tous les horizons de prévision. Aucun phénomène de surapprentissage n'est suspecté suite à ces constats.

Tableau 28 : MAPE (%)

MAPE (%)	Horizons					
Modèles	1	5	10	15	20	25
CBD.log.Poisson	66,1	61,1	55,7	50,8	46,9	44,0
Cliometric.Composite.Model	4,4	7,1	12,7	15,0	17,0	19,3
LC	8,6	11,5	18,8	22,1	24,1	26,3
LC.Log.Poisson	10,2	13,2	20,9	24,3	26,4	28,7

MAPE tous horizons confondus, selon les différents âges modélisés

Itération 1 : Pour chaque âge, choix a posteriori de sa meilleure classe de modèle (Best Model Class)

Figure 69 : MAPE tous horizons confondus, selon les différents âges modélisés

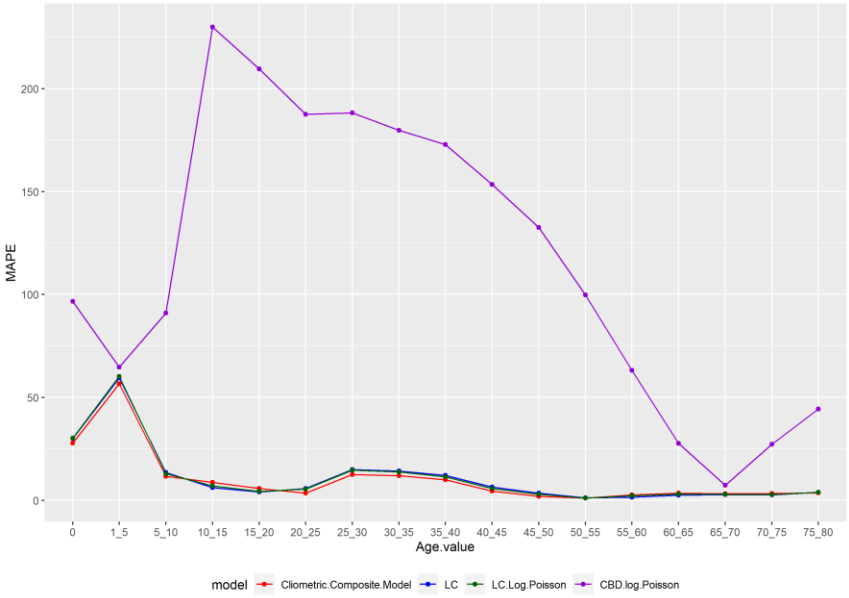
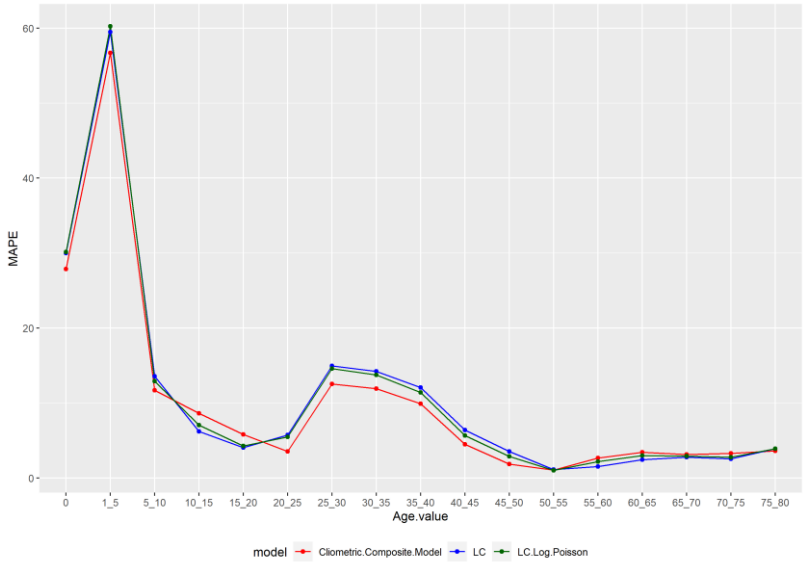


Figure 70 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette première itération, nous ne constatons aucun avantage particulier pour le modèle mixte cliométrique composite, quel que soit l'âge étudié.

Tableau 29 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	96,6	27,9	30,0	30,1
1_5	64,7	56,7	59,5	60,2
5_10	91,0	11,7	13,6	12,9
10_15	229,9	8,6	6,2	7,1
15_20	209,6	5,8	4,1	4,3
20_25	187,5	3,5	5,7	5,5
25_30	188,2	12,6	14,9	14,6
30_35	179,8	11,9	14,2	13,7
35_40	172,9	9,9	12,1	11,4
40_45	153,5	4,5	6,4	5,7
45_50	132,6	1,9	3,5	2,9
50_55	99,8	1,1	1,1	1,0
55_60	63,1	2,7	1,5	2,2
60_65	27,6	3,4	2,4	3,0
65_70	7,2	3,2	2,8	2,9
70_75	27,3	3,3	2,6	2,8
75_80	44,4	3,6	3,9	3,9

Itération 2 : Pour chaque âge, utilisation a priori de sa Best Model Class issue de l'itération 1

Figure 71 : MAPE tous horizons confondus, selon les différents âges modélisés

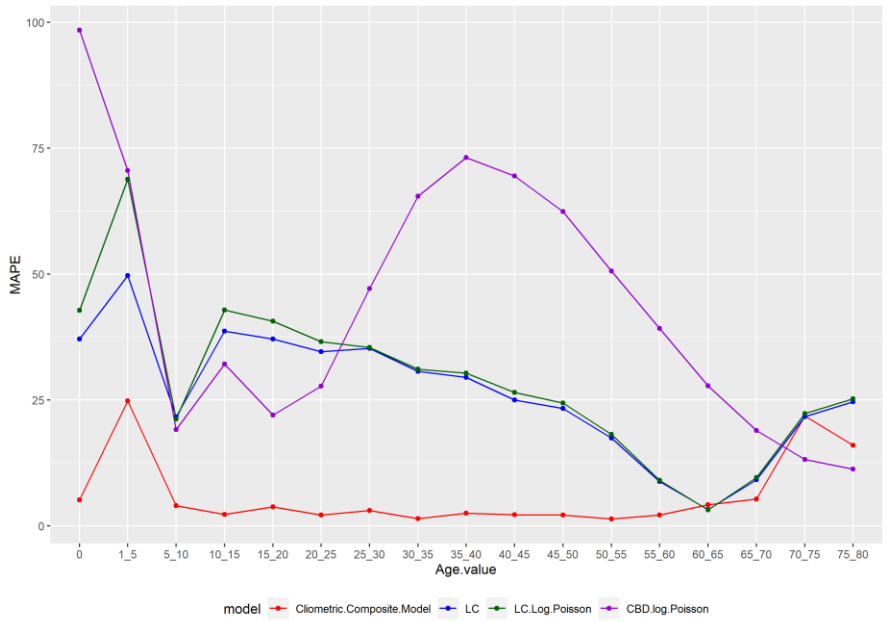
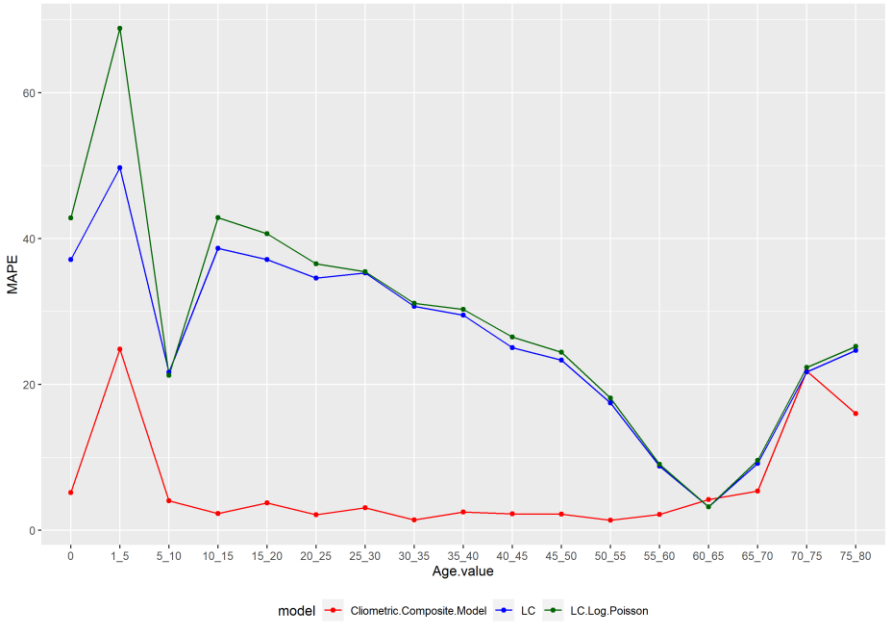


Figure 72 : MAPE tous horizons confondus, selon les différents âges modélisés (sans CBD)



Dans cette deuxième itération, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié, à l'exception notable des âges suivants : 60-65 ans, 70-75 ans et 75-80 ans. Pas de surapprentissage constaté à ce stade.

Tableau 30 : MAPE (%)

MAPE (%)	Modèles			
Ages	CBD.log.Poisson	Cliometric.Composite.Model	LC	LC.Log.Poisson
0	98,4	5,2	37,1	42,8
1_5	70,6	24,8	49,7	68,8
5_10	19,1	4,0	21,6	21,2
10_15	32,2	2,3	38,7	42,9
15_20	22,0	3,8	37,1	40,7
20_25	27,7	2,1	34,6	36,6
25_30	47,1	3,1	35,3	35,4
30_35	65,5	1,4	30,7	31,1
35_40	73,1	2,5	29,5	30,3
40_45	69,5	2,2	25,0	26,5
45_50	62,4	2,2	23,3	24,4
50_55	50,6	1,3	17,5	18,1
55_60	39,2	2,2	8,8	9,0
60_65	27,8	4,2	3,2	3,2
65_70	18,9	5,4	9,2	9,6
70_75	13,2	21,8	21,7	22,3
75_80	11,3	16,0	24,7	25,2

14.3.3. Conclusion sur les tests de sensibilité

Face à un meilleur modèle (mixte cliométrique et composite) performant et riche en options de paramétrages, mais ne présentant aucun signe de surapprentissage, la réduction des choix de paramétrage entraine un risque de réduction des performances.
C'est la principale leçon des deux tests de sensibilité de l'Equateur.

En combinant le résultat ci-dessus avec ceux des tests de sensibilité de l'Inde, nous en déduisons qu'il n'existe pas un dimensionnement paramétrique optimal « absolu » pour les modèles mixtes cliométriques composites. Le dimensionnement paramétrique optimal semble plus dépendre de la structure interne des données traitées que de leur volume : en effet nous constatons que l'Equateur bien que disposant d'un historique plus court (remontant à 1950) a pu supporter un dimensionnement paramétrique plus élevé que l'Inde qui dispose d'un historique plus long (depuis 1930).
Le seuil de surapprentissage est variable selon le pays et ne peut être déterminé qu'après exploration d'un certain nombre de scénarios de dimensionnement paramétrique.

Conclusion et recommandations actuarielles

Notre ambition a été d'améliorer les prévisions actuarielles de long-terme des taux de mortalité par âge dans un contexte de transition démographique par cohortes de pays à l'échelle mondiale. Nous avons sélectionné trois profils de pays ayant des problématiques de données distinctes et avons créé différentes méthodes pour améliorer les modèles classiques, en réponse aux limitations et problèmes identifiés.

Résultats d'exploration des données d'historique des pays selon leurs profils
Les pays développés, qui correspondent aux pays de l'OCDE, disposent d'un historique de taux de mortalité par âges datant de plusieurs générations, soit plus d'un siècle de données en moyenne. Les pays émergents sont définis selon les critères du FMI, mais la disponibilité et la qualité des historiques de taux de mortalité par âge varient considérablement. Les autres pays en voie de développement et les pays les moins avancés ont des historiques de mortalité par âge relativement courts et ne disposent généralement que de statistiques sur la mortalité infantile. Malgré cela, la plupart de ces pays connaissent une évolution démographique, éducative et économique positive.

Développement du modèle interne à facteurs PCR-optimal ou PLS.
Nous avons développé un modèle interne à facteurs PCR-optimal ou PLS pour améliorer la modélisation de la mortalité par âge, en prenant en compte les décalages temporels et les différences d'allure dans les courbes des taux de mortalité selon les âges. Cette approche permet de généraliser les modèles classiques de mortalité en utilisant une ou plusieurs composantes principales sélectionnées de manière optimale pour chaque âge modélisé.

Introduction de la stratégie composite dans la modélisation des taux de mortalité.
Dans cette démarche scientifique, nous abordons les limites des modèles classiques de mortalité qui appliquent une même classe de modèle à toutes les courbes des taux de mortalité par âges. Nous introduisons alors une stratégie composite qui permet de mélanger plusieurs classes de modèles pour modéliser les différents âges d'une population. Cette approche offre la possibilité d'attribuer différentes classes de modèles à différents âges, permettant ainsi une modélisation indépendante de la mortalité de chaque âge. La modélisation est réalisée en deux itérations pour sélectionner d'abord la classe de modèles la mieux adaptée à chaque âge, puis pour estimer les performances sans biais des modèles sélectionnés à l'itération précédente.

Développement du modèle mixte cliométrique et composite à facteurs PCR-optimal/PLS
Nous avons développé une méthode pour améliorer les prévisions à long terme des mortalités dans les pays en rattrapage transitionnel en utilisant l'histoire quantitative des mortalités des pays transitionnellement plus âgés. Cette méthode utilise une série temporelle cliométrique de mortalité créée à partir des données historiques du pays de référence et adaptée au rythme du pays à modéliser. Cette série est utilisée comme série explicative supplémentaire dans les modèles internes prospectifs de mortalité du pays à modéliser, créant ainsi un modèle mixte interne et externe que nous avons nommé modèle cliométrique. Nous avons ensuite développé un modèle mixte cliométrique et composite à facteurs PCR-optimal/PLS en capitalisant sur les apports originaux. Cette méthode permet d'améliorer les prévisions de mortalité dans les pays émergents ou en développement.

Développement des modèles à références externes cliométriques et composites
Nous avons développé une variante cliométrique à référence externe pour les pays les moins avancés, qui souffrent souvent du manque de données historiques pour la mise en place de modèles internes de prospective. Nous avons utilisé les statistiques officielles et les études de terrain régulières sur la mortalité infantile pour réaliser un appariement entre les temps calendaires du pays d'expérience R et ceux des pays de référence A, via le temps transitionnel commun. Cela permet de créer plusieurs séries cliométriques pour chaque âge du pays R, qui peuvent servir de séries explicatives dans les modèles externes cliométriques à développer pour ce pays.

Nous avons développé trois classes de modèles externes cliométriques adaptés de modèles classiques externes de BRASS, COX et TGH05-TGF05, ainsi qu'un modèle composite combinant ces adaptations. Nous avons également développé deux classes de modèles externes cliométriques adaptés de modèles à facteurs PCR-Optimal et PLS, ainsi qu'un modèle composite combinant les deux modèles.
Comme précédemment évoqué, le caractère composite de ces modèles provient du fait que les âges sont modélisés de façon indépendante et peuvent donc recourir à des classes de modèles différentes.
Ces modèles peuvent aider à surmonter le manque de données historiques pour les pays les moins avancés.

Résultats des travaux empiriques
Nous avons mené des travaux empiriques pour tester le modèle mixte cliométrique et composite à facteurs PCR-optimal/PLS en Inde et en Equateur pour réaliser des prévisions sur un horizon de 25 ans.

Pour l'Inde : nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié.
De plus nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Sauf pour les tranches 1-5, 20-25 ans où il y a un très léger écart défavorable au modèle mixte cliométrique composite.
Nous sommes dans le contexte d'un meilleur modèle qui présente un très léger effet de surapprentissage sur le long terme. Le dimensionnement du modèle est donc à la lisière du surapprentissage.
Au vu des deux scénarios alternatifs étudiés dans les *tests sensibilité*, l'inflation du nombre de choix paramétriques, crée un risque de surapprentissage. Ce phénomène de surdimensionnement dégrade plus fortement les performances par horizon de prévision, que les performances par âge.
L'impact sur les performances par horizon est si dégradé que la hiérarchie des modèles en est vite inversée en défaveur du modèle mixte cliométrique composite.
Un dosage délicat et une série de tests sur le nombre et les modalités paramétriques à explorer, doivent donc être effectués pour tirer le meilleur parti du mixte cliometric composite model.
C'est ce que nous avons effectué en privilégiant les pays A de référence, à transition plus récente en Europe, notamment l'Italie et l'Espagne.

Pour l'Equateur : nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'horizon de prévision étudié. Cela est en cohérence forte avec le schéma de la première itération. Ce qui exclut a priori un phénomène de surapprentissage malgré un fort dimensionnement paramétrique. De plus, nous constatons un avantage systématique pour le modèle mixte cliométrique composite, quel que soit l'âge étudié. Les avantages se sont même largement renforcés.
La hiérarchie de la première itération est conservée et renforcée.

Face à un meilleur modèle performant et riche en options de paramétrages, mais ne présentant aucun signe de surapprentissage, la réduction des choix de paramétrage entraine un risque de réduction des performances. C'est la principale leçon des deux tests de sensibilité de l'Equateur.

Au final en combinant les tests de sensibilité de l'Inde et de l'équateur, nous concluons que le seuil de surapprentissage est variable selon le pays et ne peut être déterminé qu'après exploration d'un certain nombre de scénarios de dimensionnement paramétrique. Ce seuil de dimensionnement semble plus dépendre de la structure interne des données que de leur volume.

Résultats de l'enquête qualitative en zone CIMA

Nous nous sommes penchés sur les défis actuariels rencontrés dans les pays d'Afrique subsaharienne francophone regroupés dans la zone CIMA. En plus de la littérature disponible, nous avons mené une enquête qualitative sur le terrain à Lomé (Togo) et Abidjan (Côte d'Ivoire). Les professionnels de l'assurance sur la vie interrogés ont souligné plusieurs problématiques, notamment la modification socio-économique de leur portefeuille et le manque d'historique d'expérience sur les professionnels du secteur informel. Ils ont également signalé un risque de longévité pour les assureurs, lié à l'allongement de l'espérance de vie dans la zone et à la baisse des taux de mortalité. Les assureurs majeurs de la région proposent des produits de retraite complémentaire. Les acteurs de la filière sont conscients de l'importance de prendre en compte l'évolution des taux de mortalité par âge. Les autorités de contrôle ont parfois été critiquées pour leur lenteur et leur prudence excessive, en particulier face aux évolutions démographiques rapides en cours. Les tables règlementaires actuelles ont également été remises en question, notamment en termes de représentativité de l'échantillon d'assureurs utilisés. Les actuaires de la zone ont donc besoin d'une mise à jour urgente des tables règlementaires en vigueur.

Recommandations actuarielles pour la zone CIMA

Notre enquête de terrain ayant été faite en zone CIMA, nous proposons des recommandations techniques par rapport aux problématiques que nous avons identifiées au sein de la zone et en relation avec le sujet du mémoire.

Un des problèmes soulevés par les professionnels de l'assurance santé (et sur la vie) est la modification socio-économique de leur portefeuille due à l'intégration des professionnels du secteur informel.

Une solution serait l'organisation de sondages thématiques auprès des populations de professionnels informels concernés, en attendant l'accumulation d'un historique d'expérience.

Une inquiétude majeure identifiée, concerne le risque avéré de longévité qui pèse sur les assureurs et les caisses de retraite, du fait de l'allongement soutenu de l'espérance de vie dans la zone CIMA.

Notre recommandation serait la construction à brève échéance de tables règlementaires de mortalité prospectives de la zone pour la tarification des rentes viagères. A cet effet, nous suggérons l'utilisation d'un modèle externe cliométrique et composite, proposé dans le mémoire. Un tel changement nécessite la prise en compte de facteurs politiques dus au caractère international de la zone CIMA, ainsi qu'une attention particulière à la qualité des données à utiliser. L'impact sur les tarifications, les provisions et la solvabilité des assureurs devra être également évalué en amont de tout changement règlementaire.

Par ailleurs, l'hétérogénéité des profils nationaux de mortalité au sein de la zone, pourrait être prise en compte par des mécanismes de décalages par rapport à la table agrégée de la zone.

Du fait de la baisse régulière des taux de mortalité due à la transition démographique dans la zone CIMA, les tables règlementaires du *moment* actuelles s'avèrent « obsolètes » car elles ont été établies à partir de données de 2003-2006. Une mise à jour des tables règlementaires du moment en vigueur est donc éminemment souhaitable. Les précautions et les difficultés sont les mêmes que celles évoquées

précédemment pour la construction des tables règlementaires prospectives. Un soin particulier devra être apporté à la question de la représentativité des pays de la zone dans les données à utiliser.

Un dernier point de recommandation serait d'étudier la possibilité de permettre aux assureurs d'utiliser des tables d'expériences certifiées et suivies par des actuaires indépendants agréés. L'agrément des actuaires pourra être délégué à un institut régional des actuaires (à créer) ou aux instituts nationaux, sous réserve d'effectifs et de compétence suffisants.
A cet effet l'autorité de régulation et de contrôle CIMA pourra éditer des *guidelines* pour la qualité des données et des techniques à utiliser. Il s'agit là d'un chantier dont nous ne sous-estimons pas la complexité technique, organisationnelle et politique.

Perspectives et travaux futurs
Des travaux empiriques supplémentaires seront nécessaires pour tester et améliorer les autres modèles théoriques que nous avons développés, en les appliquant dans les contextes adaptés à leur usage. Cela concerne notamment les modèles externes cliométriques composites destinés spécifiquement aux pays les moins avancés.

Le calcul des intervalles de confiance est un autre aspect des modèles développés qui nécessite des recherches et des travaux empiriques complémentaires, notamment via une approche Monte-Carlo.

D'un point de vue stratégique et professionnel, notre démarche est de mettre en place une équipe technique et commerciale pour cibler les problématiques actuarielles identifiées des assureurs, organismes gouvernementaux et autorités de régulation et contrôle dans la zone CIMA.

Bibliographie

ACPR (Autorité de contrôle prudentiel et de résolution), 2023. Conférence interafricaine des marchés d'assurance (CIMA), consulté en ligne le 19/03/2023, https://acpr.banque-france.fr/conference-interafricaine-des-marches-dassurance-cima

Atlas magazine (2023), Zone CIMA : densité et taux de pénétration de l'assurance par pays en 2020, consulté en ligne le 27/01/2023, https://www.atlas-mag.net/category/tags/focus/zone-cima-densite-et-taux-de-penetration-de-l-assurance-par-pays-en-2020

Avdeev A. (2014). « Histoire de la population mondiale et la transition démographique ». Ressources ouvertes de l'Institut de démographie de l'Université de Paris 1, Année académique 2014/2015. (document consulté le 05/06/2015). Disponible sur : http://dmo.econ.msu.ru/teaching/Histpop/

Bocquaire E. (2015), Les grands principes de l'actuariat, L'Argus de l'Assurance Editions

Booth H., Maindonald J., Smith L. (2002). Applying Lee-Carter under conditions of variable mortality decline. Population Studies, 56, 325-336.

Bourbonnais R., Terraza M. (2010). Analyse des séries temporelles, Dunod

Bourbonnais R., Usunier J. C. (2017). Prévision des ventes, Economica

Boutahar M., Royer-Carenzi M. (2019). Méthodes en séries temporelles et applications avec R, Ellipses

Box G. E. P., Tiao G. C. (1975). Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, Vol. 70, No. 349. (Mar., 1975), pp. 70-79

Boyer M., Dorion C., Stentoft L. (2015). Les modèles factoriels et la gestion du risque de longévité. L'Actualité économique, 91(4), 531–565

Brass W. (1971). On the scale of mortality. Biological aspects of mortality., Taylor & Francis Ltd

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

Cairns A. J.G., Blake D., Dowd K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and Calibration, J. Risk Insur., 73, 4, 687-718

Cairns A. J.G., Blake D., Dowd K. (2008). Modelling and management of mortality risk: a review, Scand. Actuar. J., 2008, 2-3, 79-113, (2008) · Zbl 1224.91048

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

Chesnais J-C. (1986), « La transition démographique ». Paris: PUF-INED (Cahier n°113)

CIMA (Conférence interafricaine des marchés d'assurance), 2023. Code CIMA 2019, consulté en ligne le 27/01/2023, https://cima-afrique.org/code-cima-2/

Courbage Y., Todd E. (2007). « Le rendez-vous des civilisations ». Seuil

Clayton D., Schifflers E. (1987). Models for temporal variation in cancer rates. II: Age-period-cohort models. Statistics in Medicine 6(4), 469–481

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

Delwarde A., Denuit M. (2005). Construction de tables de mortalité périodiques et prospectives. Paris : Economica

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

Diaz G., Debón A., Giner-Bosch V. (2018). Mortality forecasting in Colombia from abridged life tables by sex. Genus 74(15)

Droesbeke J. J., Saporta G. (2010). « Analyse statistique des données longitudinales ». Technip

Gaba K. G. (2021), Cliométrie de l'histoire globale des transitions anthropologique, cognitive, institutionnelle, démographique et économique. Thèse de doctorat, Université de Lyon 1.

Gourieroux C., Monfort A. (1999). « Séries temporelles et modèles dynamiques ». Economica

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

Hobcraft J., Menken J., Preston S. (1982). Age, period, and cohort effects in demography: a review. Population Index 48 (1), 4–43

Human Mortality Database (2020), consulté en ligne le 28/09/2020, https://www.mortality.org/cgi-bin/hmd/hmd_download.php

Hunt A., Villegas A. M. (2015). Robustness and convergence in the Lee-Carter model with cohort effects. Insurance: Mathematics and Economics, 64, 186-202

Hyndman R. J. (2019). demography: Forecasting Mortality, Migration, Fertility and Population Data. (R package version 1.21).

Hyndman R. J., Kostenko A. V. (2007). "Minimum Sample Size requirements for Seasonal Forecasting Models," Foresight: The International Journal of Applied Forecasting, International Institute of Forecasters, issue 6, pages 12-15, Spring.

Hyndman R. J., Ullah S., (2007). Robust forecasting of mortality and fertility rates: A functional data approach, Comput. Statist. Data Anal., 51, 10, 4942-4956, (2007) · Zbl 1162.62434

Jaffrelot C. et al. (2008). L'enjeu mondial : Les pays emergents, Les Presses de Sciences Po

Johnston J. (1988), « Méthodes économétriques ». Tome 2, 3e Edition. Economica.

Kamega A. (2011), Outils théoriques et opérationnels adaptés au contexte de l'assurance vie en Afrique subsaharienne francophone : analyse et mesure des risques liés à la mortalité. Gestion et management. Université Claude Bernard - Lyon 1.

Kamega A., Planchet F. (2011). Analyse et comparaison des populations générale et assurée en Afrique subsaharienne francophone pour anticiper la mortalité future, Cahiers de recherche de l'ISFA, 2138 (2011) -38

Kamega A., Pieby F. (2014). Construction de tables de mortalité d'expérience prospectives en zone CIMA. Journée 100% Actuaires, 7 novembre 2014, Institut des Actuaires, Euria Brest

Kennes T. (2017). The convergence and robustness of cohort extensions of mortality models. MaRBLe, 1, 36–53

King G., Hardy D. F. (1880). Notes of the practical application of Mr Makeham's formula to the graduation of mortality tables, Journal of the Institute of Actuaries

Lebart L., Piron M., Morineau A. (2000). « Statistique exploratoire multidimensionnelle ». Dunod

Lee R.D., Carter L.R. (1992). Modeling and forecasting US mortality. Journal of the American Statistical Association, 87, 659-671.

Maheu J. M., McCurdy T. H. (2009). How Useful Are Historical Data for Forecasting the Long-Run Equity Return Distribution? Journal of Business & Economic Statistics, Vol. 27, No. 1 (Jan., 2009), pp. 95-112

Mandzij J. (2011). Risque de longévité - Construction des tables de mortalité prospectives, Modèle Interne Partiel. Mémoire de Master Actuariat, Université de Lyon 1

Max Planck Institute for Demographic Research (2020), Demographic Data, consulté en ligne le 28/09/2020, https://www.demogr.mpg.de/en/research_6120/demographic_data_27/

Millossovich P., Villegas A.M., Kaishev V. K. (2018). StMoMo: An R Package for Stochastic Mortality Modelling. Journal of Statistical Software, 84(3), doi:10.18637/jss.v084.i03

Osmond C. (1985). Using Age, Period and Cohort Models to Estimate Future Mortality Rates. International Journal of Epidemiology 14(1), 124–129

Planchet F. (2006). Tables de mortalité d'éxperience pour les portefeuilles de rentiers (Tables TGH 05 et TGF 05), Notice de présentation, Institut des Actuaires

Planchet F., Lelieur V., (2007). Utilisation des méthodes de Lee-Carter et Log-Poisson pour l'ajustement des tables de mortalité dans le cas de petits échantillons, Bulletin français d'actuariat, Vol. 7, n°14, juillet-décembre 2007, pp. 118-146

Planchet F., Tomas J. (2014). Méthodes de positionnement : Aspects méthodologiques. Note de travail, ISFA Laboratoire SFA

Planchet F. (2021). Tables de mortalité. Cours de Modèles de durée, Université de Lyon 1

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

Safitri L., Mardiyati S., Rahim H. (2018). Estimation of mortality rate in Indonesia with Lee-Carter model, AIP Conference Proceedings 2023, 020210 (2018); https://doi.org/10.1063/1.5064207, Published Online: 23 October 2018

Saporta G. (2011). « Probabilités, Analyse des données et Statistique ». Technip

Todd E. (1999). « La Diversité du monde : Famille et modernité ». Seuil

Tufféry S. (2012). « Data Mining et Statistique Décisionnelle ». Technip.

United Nation, Population Division, (2022). Methodology report. World Population Prospects 2022, United Nations New York

Vidal R., Ma Y., Sastry S., (2016), Generalized Principal Component Analysis. New York, Springer-Verlag

Villegas A. M., Millossovich P., Kaishev V. K. (2015). StMoMo: An R package for stochastic mortality modelling. Journal of Statistical Software 66(3), 1-20

Wandji A. (2015). Construction d'une table de mortalité prospective sur un portefeuille de taille modeste et historique limité. Mémoire de Master Actuariat, Université Paris-Dauphine

Yebouet et al. (2014). Problématique de la mortalité et défis du contrôle des assurances en Afrique subsaharienne : cas de la zone CIMA, CONGRES INTERNATIONAL DES ACTUAIRES, du 30 Mars au 04 Avril 2014, Washington DC, USA