



LACTU 2040

Mémoire recherche :

Modélisation de la mortalité à l'aide de
modèles de régression gaussiens et de
l'analyse en composantes principales
fonctionnelle

Année Académique : 2020-2021

Miglietti Raphaël, Noma :6403 19 00

Professeur : M. Hainaut

Table des matières

1	Résumé	1
2	Introduction	2
3	Modèles de régression gaussienne	4
3.1	Présentation générale	4
3.1.1	Idée principale	4
3.1.2	Écriture mathématique	5
3.1.3	Utilisation dans le cadre de la régression à des fins de prédictions .	7
3.1.4	Vraisemblance marginale et optimisation des hyperparamètres du noyau	10
3.2	Application à la modélisation de la mortalité	12
3.2.1	Adaptation du modèle à la mortalité	12
3.2.2	Noyau à mélange spectral	13
3.2.3	Fonction de moyenne	15
3.2.4	Résultats	17
3.2.5	Comparaison avec le modèle de Lee-Carter et sa variante Poisson .	24
3.3	Conclusion	26
4	Analyse en composantes principales fonctionnelle	27
4.1	Présentation générale	27
4.1.1	Intérêt d'utiliser des données fonctionnelles	27
4.1.2	Présentation du cas univarié	28
4.1.3	Présentation du cas multivarié	30
4.2	Application à la modélisation de la mortalité	34
4.2.1	Cas univarié	34
4.2.2	Cas multivarié	40
5	Statistiques démographiques et application financière	47
5.1	Calcul des espérances de vie	47
5.2	Prix d'une rente viagère	51
6	Conclusion	53
7	Bibliographie	55
8	Annexes	56

1 Résumé

Au cours des siècles derniers et encore aujourd'hui, l'espérance de vie continue à augmenter. En effet, au début du XXème siècle la durée de vie était en moyenne de 45 ans pour aujourd'hui atteindre 79 ans pour les hommes et 85 ans pour les femmes en France. Cette tendance s'explique par de meilleures conditions de vie et des progrès de la science. L'allongement de la durée de vie à long terme entraîne un nombre important de défis et de problèmes à différents niveaux sociaux. Ceci conduit à un risque sur le plan économique pour les entreprises, les gouvernements et la population. Ainsi les retraites sont le domaine le plus exposé à ce risque de vieillissement de la population. Les assureurs ont donc dû redéfinir leurs prestations et réévaluer leurs tables de mortalité servant de base de calcul des rentes afin de se protéger de cette augmentation de dépenses. La modélisation de l'évolution de la mortalité est donc d'une importance considérable pour bien gérer ce risque de longévité et de mortalité qui est prépondérant aujourd'hui. L'idée de ce mémoire est de reprendre le travail de Ruhao Wu [1] : Gaussian Process and Functional Data Methods for Mortality Modelling et de l'appliquer sur un jeu de données différent et de confirmer l'efficacité ou non de l'utilisation de différentes méthodes statistiques pour estimer les taux de mortalité future. Parmi les différentes méthodes statistiques étudiées dans ce mémoire, nous nous intéresserons principalement aux processus gaussiens et à l'analyse en composantes principales fonctionnelle univariée et multivariée. La première idée est d'utiliser les champs gaussiens pour la régression des taux de mortalités en fonction de l'âge et de l'année. Cette méthode prend en compte une fonction de la moyenne pondérée et une fonction de covariance. Les fonctions de covariance utilisées seront diverses et nous utiliserons des noyaux classiques mais aussi le noyau à mélange spectral. Dans une seconde partie, nous utiliserons d'abord l'analyse en composantes principales fonctionnelle univariée. Ce qui nous permettra d'introduire les données fonctionnelles. Ensuite, nous nous intéresserons à la modélisation simultanée de plusieurs sous-groupes à l'intérieur d'une même population à l'aide de l'analyse en composantes principales fonctionnelles multivariée. Nous traitons les taux de mortalité des sous-populations (ex : homme et femme du même pays) au sein d'une grande population comme un ensemble de données fonctionnelles à plusieurs niveaux. L'analyse en composantes principales fonctionnelle multivariée sera utilisée pour extraire les informations de base des données fonctionnelles et les analyser à l'échelle de plusieurs populations, de sorte que le modèle intègre à la fois des informations globales de la population et des informations spécifiques des sous-populations. Dans la troisième et dernière partie, nous proposerons une application financière des différents modèles présentés auparavant. Cette application se fera d'abord par le calcul de statistiques démographiques comme l'espérance de vie à la naissance ou résiduelle. Puis par la suite nous calculerons le prix d'une rente viagère en fonction de nos différents modèles.

2 Introduction

L'espérance de vie a connu une forte amélioration depuis maintenant plus d'un siècle si nous excluons les périodes de guerre. Un tel changement de l'évolution de la mortalité a des conséquences directes pour les ménages, les entreprises, les assureurs et les gouvernements qui doivent soit fournir des prestations en cas de décès prématurés ou payer des rentes de retraites à une population qui vit plus longtemps. Pour les assureurs, cette amélioration de la mortalité a un impact significatif sur la tarification et les provisions des contrats. La capacité de distinguer des changements de mortalité à court et à long terme est donc cruciale pour bien gérer le risque de mortalité et de longévité. Nous comprenons donc bien pourquoi la modélisation de la mortalité a toujours été un sujet de recherche pour les démographes ou encore les assureurs. Dès 1662, Graunt s'est intéressé à la mortalité des londoniens. En 1825, Gompertz est passé d'une approche biologique à une approche mathématique en modélisant le taux instantané de mortalité μ_x . Il a supposé que la mortalité à l'âge adulte montre une augmentation presque exponentielle. Les deux paramètres de son modèle sont positifs et varient avec le niveau de mortalité et le taux d'augmentation de la mortalité avec l'âge. Le modèle de Gompertz et sa version modifiée par Makeham (1867), où une constante supplémentaire est ajoutée pour tenir compte de la mortalité de fond due à des causes non liées à l'âge, ont été largement utilisées comme modèle standard pour la mortalité adulte pendant de longues années. Au delà de la modélisation de la mortalité, les actuaires se sont très vite intéressés à la prévision des taux de mortalité future. Jusque dans les années 80, les modèles mathématiques utilisés pour prévoir directement les taux de mortalité ou l'espérance de vie étaient relativement simples et impliquaient un certain degré de jugement subjectif. Au cours des trente dernières années, un certain nombre de nouvelles approches ont été développées pour la prévision de la mortalité à l'aide de modèles stochastiques. Le modèle le plus connu est celui de Lee-Carter [2] et nous nous servons de sa variante Log-Poisson comme référence pour comparer les résultats obtenus à l'aide de la régression gaussienne.

Dans la première partie de ce mémoire, nous présenterons la régression gaussienne de manière générale. Dans cette même régression gaussienne qui est le coeur de notre étude, la mortalité est classée par groupes d'âge spécifiques, elle suit un processus gaussien. Un tel processus incorpore une fonction de moyenne et une fonction de covariance. La fonction de covariance permet de saisir diverses structures de covariance des taux de mortalité au fil du temps pour différents groupes d'âge, et la fonction de moyenne pondérée modélise la tendance à long terme. Nous appliquerons donc cette méthode à différentes fonctions de covariance et notamment à un noyau de mélange spectral. Les résultats obtenus seront comparés en fonction du noyau et au modèle Log-Poisson.

Comme nous l'avons vu, il existe de nombreuses méthodes statistiques pour projeter les taux de mortalité en fonction de l'âge. La méthode Lee-Carter utilise une méthode en composantes principales pour les taux de mortalité par âge et extrait un indice unique variant dans le temps à partir duquel les prévisions sont obtenues par une marche aléatoire avec dérive. La méthode a depuis été étendue et modifiée. Par exemple, Renshaw et Haberman [3] ont proposé le modèle âge période cohorte ; Hyndman et Ullah [4] ont proposé un modèle avec des données fonctionnelles qui utilise le lissage non paramétrique et des composantes principales qui ne sont plus uniques. Ces travaux ont principalement porté sur la prévision de la mortalité pour une seule population ou plusieurs populations individuellement. Cependant, les prévisions individuelles peuvent impliquer une divergence croissante de la mortalité sur le long terme, à contre-courant de la tendance attendue et observée. C'est pour cela, que dans la deuxième section de notre exposé, nous nous intéresserons à la modélisation de la mortalité pour une population donnée mais aussi pour plusieurs populations simultanément. La méthode utilisée sera l'analyse en composante principale fonctionnelle univariée puis multivariée en fonction du nombre de populations étudiées. Une telle modélisation permet d'identifier les corrélations entre deux ou plusieurs populations, faire la distinction entre effets à long terme et à court terme sur l'évolution de la mortalité, et explorer les informations supplémentaires contenues dans l'expérience d'autres populations pour améliorer la précision des prévisions. Ces populations peuvent être regroupées par sexe, état, groupe, statut socio-économique et d'autres attributs. Pour se faire, nous utiliserons l'analyse en composante principale fonctionnelle et multivariée. Comme pour la section précédente nous expliquerons cette méthode d'abord de manière générale puis nous l'appliquerons à nos données de mortalité. Enfin un des rôles de l'actuaire étant de quantifier l'impact financier d'un risque, nous proposerons une application financière reprenant l'ensemble des modèles traités dans les sections précédentes. Cette application financière se fera en deux temps. D'abord par le calcul des probabilités de survie qui permettent le calcul notamment de l'espérance de vie résiduelle ou à la naissance mais aussi d'autres statistiques intéressantes. Par la suite, nous calculerons le prix d'une rente viagère pour des caractéristiques données.

3 Modèles de régression gaussienne

Dans cette partie nous présenterons dans un premier temps les processus gaussiens utilisés dans un but de régression puis nous détaillerons comment cette technique d'apprentissage supervisé peut être utilisée pour la modélisation des taux de mortalité.

3.1 Présentation générale

3.1.1 Idée principale

Les processus gaussiens sont une technique d'apprentissage supervisé, dont le but est de faire un lien entre des données d'entrées et de sorties à partir de données empiriques (l'ensemble de données d'entraînement). Il s'agit dans notre cas d'un problème de régression. Nous donnerons les données d'entrée X et les sorties ou valeurs cibles Y . Pour ce problème, nous possédons un jeu de données d'entrée $\mathcal{D}$ qui contient n observations $\mathcal{D}=\{\{x_i, y_i\}|i = 1, ..., n\}$. A l'aide de nos données d'entraînement nous voulons prédire de nouvelles sorties pour de nouvelles données X^* . Nous devons donc passer d'un problème fini avec nos n données d'entrées à une fonction f qui nous permettra de prédire n'importe quelle nouvelle donnée d'entrée. Pour ce faire, nous devons faire des hypothèses sur les caractéristiques de la fonction sous-jacente, sinon toute fonction qui est en accord avec les données d'apprentissage serait également valide. Une grande variété de méthodes a été proposée pour traiter le problème d'apprentissage supervisé ; nous décrivons ici deux approches courantes. La première approche consiste à restreindre la classe de fonctions que nous considérons, par exemple en ne considérant que les fonctions linéaires de nos entrées. La deuxième approche consiste à donner une probabilité a priori à chaque fonction possible, où des probabilités plus élevées sont données à des fonctions que nous considérons comme plus probables, par exemple parce qu'elles sont plus lisses que d'autres fonctions. Nous voyons bien qu'avec la première approche cela pose un problème évident, si nous utilisons un modèle basé sur une certaine classe de fonctions (par exemple des fonctions linéaires) et que la fonction cible n'est pas bien modélisée par cette classe, alors les prédictions seront médiocres. La seconde approche pose également un problème, en ce qu'il existe sûrement un ensemble infini de fonctions. C'est là que le processus Gaussien nous est utile. Un processus gaussien est une généralisation de la distribution de probabilité gaussienne. Alors qu'une distribution de probabilité décrit des variables aléatoires qui sont des scalaires ou des vecteurs (pour les distributions multivariées), un processus stochastique régit les propriétés des fonctions. Nous pouvons vaguement considérer une fonction comme un vecteur très long, chaque entrée dans le vecteur spécifiant la valeur de la fonction au point $f(x)$. Pour traiter par calcul ces objets de dimension infinie nous nous intéresserons aux propriétés de la fonction en un nombre fini de points. L'un des

principal avantage des processus gaussien est précisément qu'il associe une vision sophistiquée qui peut être résolues par des algorithmes informatiques (voir Rasmussen C. K. I. Williams)[5].

3.1.2 Écriture mathématique

Commençons par donner la définition d'un processus gaussien et d'un vecteur gaussien.

Définition : Un vecteur aléatoire $X = (X_1, \dots, X_d)$ est dit gaussien si toute combinaison linéaire de ses composantes est gaussienne réelle, i.e, $\forall a \in \mathbb{R}^d, \langle a, X \rangle$ est une variable aléatoire réelle gaussienne.

Définition : Un processus $(X_t)_{t \geq 0}$ est appelé processus gaussien si pour tout $n \geq 1$, tout $t_1 \leq \dots \leq t_n$, le vecteur $(X_{t_1}; \dots; X_{t_n})$ est un vecteur gaussien.

Un processus gaussien est défini par sa fonction de moyenne et sa fonction de covariance. Nous définissons la fonction moyenne du processus $m(x)$ et la fonction de covariance $k(x; x')$ d'un processus réel $f(x)$ de la manière suivante.

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x; x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$$

Dans le cas de la régression par processus gaussien, nous faisons l'hypothèse que la fonction $f(x)$ est distribuée comme un processus gaussien :

$$f(x) \sim GP(m(x); K(x; x'))$$

Souvent la fonction de moyenne est considérée comme nulle mais nous verrons que pour notre application la fonction de moyenne a son importance et ne sera donc pas considérée comme nulle. La moyenne $m(x)$ représente la moyenne de toutes les fonctions considérées au point x .

La fonction k est communément appelée le noyau du processus gaussien (Jakel, Scholkopf et Wichmann)[6]. Le choix du noyau est basé sur des hypothèses telles que la régularité et les modèles probables attendus dans les données. Une hypothèse raisonnable est généralement que la corrélation entre deux points décroît avec la distance entre les points. Cela signifie que les points plus proches devraient se comporter de manière plus similaire que les points qui sont plus éloignés les uns des autres. Nous étudierons quatre noyaux différents dans la prochaine section. A titre d'exemple, nous allons introduire dès

maintenant un noyau qui est très souvent utilisé, le noyau à fonction de base radiale. Il se définit de la manière suivante :

$$k(x; x') = \sigma_f^2 * \exp\left(\frac{-|x - x'|^2}{2\lambda^2}\right)$$

La fonction de base radiale fournit une expression pour modéliser les fonctions régulières stationnaires. Les deux hyper-paramètres λ (appelés échelle de longueur) et σ_f^2 (la variance du signal) peuvent être modifiés pour augmenter ou réduire la corrélation a priori entre les points et par conséquent la variabilité de la fonction résultante. Une fois qu'une fonction moyenne et un noyau sont choisis, nous pouvons utiliser les processus gaussiens pour estimer des valeurs de fonction a priori, ainsi que des valeurs de fonctions a posteriori conditionnelles aux observations précédentes. Nous verrons dans notre application numérique comment le choix d'une fonction moyenne et d'un noyau a un impact sur la qualité de nos prédictions.

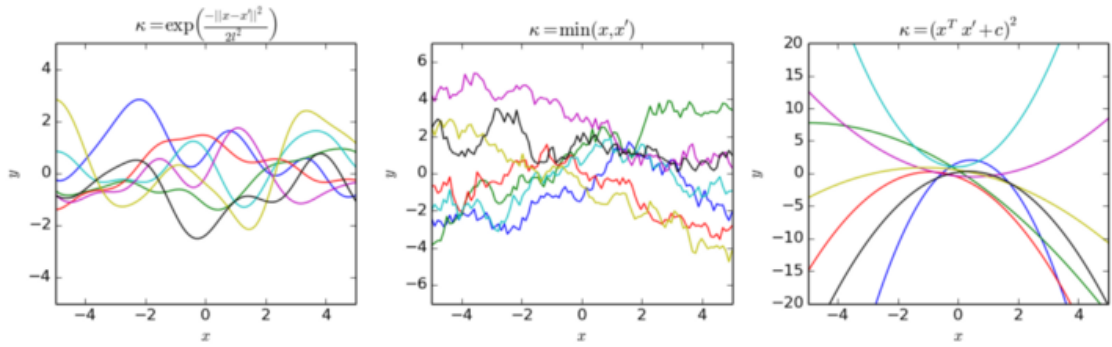


FIGURE 1: Représentation de différents processus gaussien en fonction du noyau

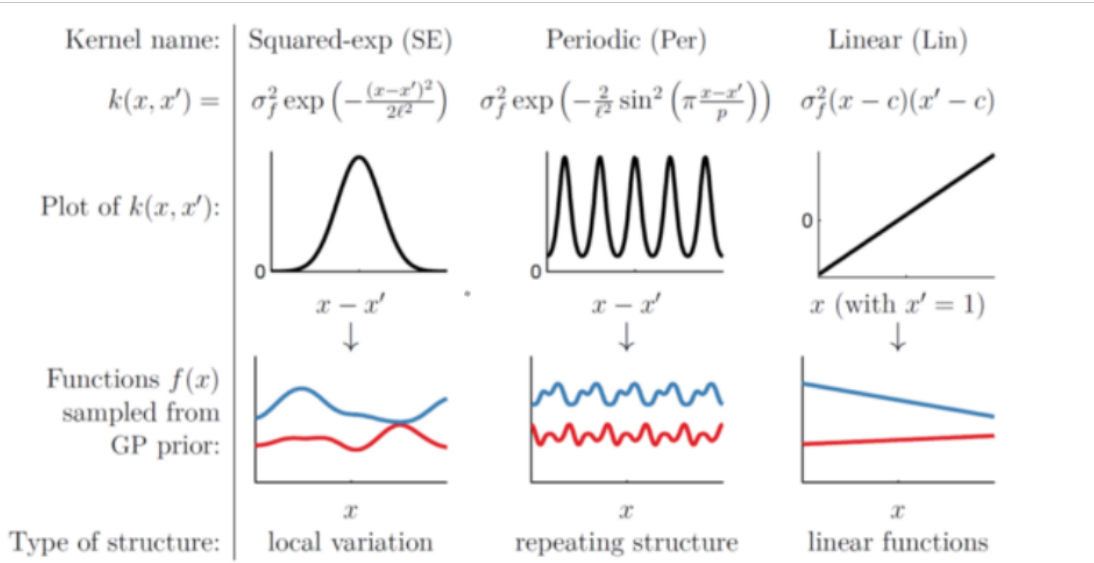


FIGURE 2: Fonction échantillonnée à partir d'un processus gaussien

3.1.3 Utilisation dans le cadre de la régression à des fins de prédictions

Pour rappel, l'objectif de la régression en mathématiques est de prédire une variable quantitative en utilisant une méthode statistique. Comme nous l'avons vu précédemment, les processus gaussiens sont continus. Néanmoins, nous ne possédons que des échantillons de notre processus gaussien en entrée. En effet, nous avons seulement les valeurs de notre fonction pour certains points donnés. Théoriquement, une fonction peut être représentée comme un vecteur de taille infinie, comme nous devons faire des prédictions pour un nombre fini de points, nous pouvons prédire les sorties pour ces points en utilisant une distribution normale multivariée avec une matrice de covariance générée par un noyau. Pour cela, nous allons utiliser les connaissances sur la fonction recherchée f qui nous sont données par notre jeu d'entrée $\mathcal{D} = \{\{x_i, y_i\} | i = 1, \dots, n\}$. Si dans un premier temps nous considérons des données sans bruit, la distribution jointe a priori sur nos données d'entraînement noté f et sur nos données de test f^* peut s'écrire de la manière suivante :

$$\begin{pmatrix} f \\ f^* \end{pmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right)$$

Nous considérons que nous avons n données d'entrée et n^* prédictions à réaliser.

— $K(X, X)$ est une matrice de taille $n \times n$ qui peut être calculée de manière explicite

$$K(X, X) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix}$$

Il en est de même pour les matrices $K(X, X^*)$, $K(X^*, X)$ et $K(X^*, X^*)$ en fonction du noyau k choisi. Afin d'obtenir la distribution a posteriori, nous devons restreindre la distribution conjointe a priori pour qu'elle ne contienne que les fonctions qui concordent avec les points de données observées. En termes probabilistes, cette opération est simple. Il s'agit de conditionner la distribution gaussienne conjointe sur nos données d'observations. Mathématiquement cela s'écrit :

$$f^* | X^*, X, f \sim N(K(X^*, X)K(X, X)^{-1}f, K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*))$$

Les valeurs de la fonction f^* pour nos données test X^* peuvent ainsi être déterminées avec la distribution jointe ci-dessus et en calculant la matrice de la moyenne et de covariance.

De légères modifications doivent cependant être ajoutées dans notre cas. Comme nous travaillons avec des données bruitées. Nous n'avons pas réellement connaissance de la valeur de notre f en certains points mais plutôt à des observations bruitées du type $y = f(x) + \epsilon$. ϵ représente un bruit blanc gaussien. C'est à dire un bruit blanc qui suit une loi normale de moyenne nulle et variance σ_n^2 . Ainsi, les équations vues auparavant sont légèrement modifiées et la fonction de covariance s'écrit alors :

$$cov(y_q, y_p) = k(x_q, x_p) + \sigma_n^2 \delta_{p,q}$$

$\delta_{p,q}$ représente le delta de Kronecker qui est une fonction de deux variables est égale à 1 si celles-ci sont égales, et 0 sinon.

$$\delta_{ij} = \delta_i^j = \delta^{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

Ainsi, nous pouvons modifier l'écriture de la distribution jointe sur nos données d'entraînement f et sur nos données test f^* :

$$\begin{pmatrix} f \\ f^* \end{pmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right)$$

Nous retrouvons alors la distribution jointe de nos données d'entraînement et de test :

$$f^* | X, y, X^* \sim N(\bar{f}^*, cov(f^*))$$

Avec :

- $\bar{f}_* \triangleq \mathbb{E}[f_* | X, y, X_*] = K(X, X_*)[K(X, X) + \sigma_n^2]^{-1}y$
- $Cov(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2]^{-1}K(X, X_*)$

A ce stade, nous pouvons faire plusieurs remarques sur la distribution prédictive de nos données. La moyenne $\bar{f}_*$ est une combinaison linéaire de nos données d'entrée y . Il est possible de réécrire :

$$\bar{f}(X_*) = \sum_{i=1}^n \alpha_i k(X, X_*)$$

Avec $\alpha = (K + \sigma_n^2 I)^{-1}y$ Notons que la somme ne va pas à l'infini car nous nous intéressons à nos n données d'entrée.

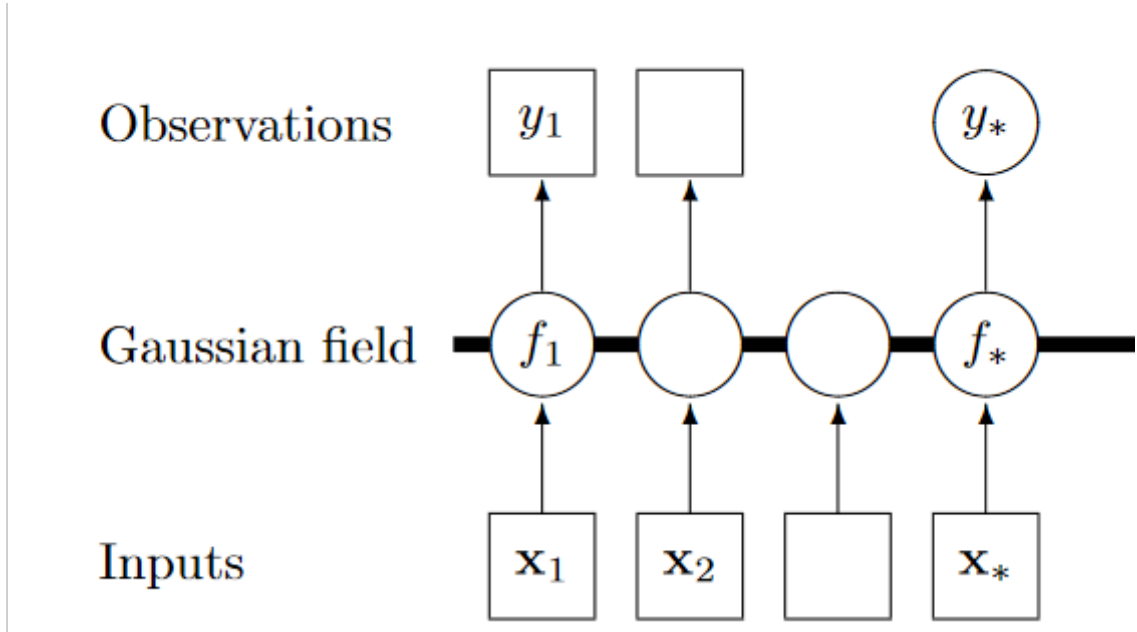


FIGURE 3: Schéma représentant un modèle de régression gaussienne

Pour conclure pour cette sous-partie nous illustrerons les résultats mathématiques énoncés infra de manière graphique. Sur la figure 3, nous pouvons voir que chaque variable y_i est indépendante des autres conditionnellement à f_i . La ligne qui relie les fonctions f peut être vue comme la distribution multivariée. L'ajout de données X_* ne modifie en aucun cas la distribution de nos données d'observations.

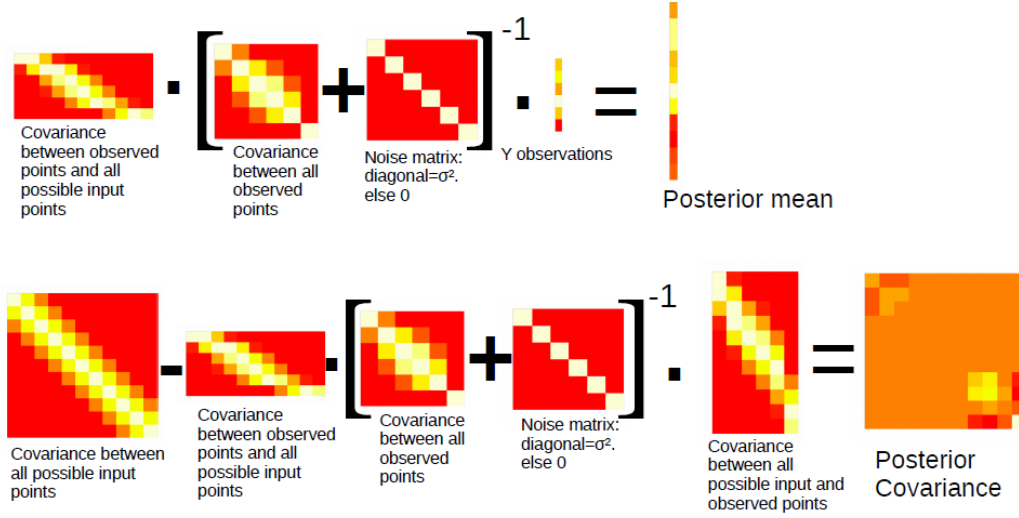


FIGURE 4: Visualisation des formules de moyenne et de covariance pour nos prédictions

Le schéma 4 schématise les formules trouvées auparavant pour obtenir la distribution sur nos données test. La moyenne pour une donnée à prédire correspond à la covariance entre tous nos points et le point à prédire multiplié par l'inverse de la somme entre la covariance de toutes nos données en entrée et la matrice de bruit, le tout multiplié par les sorties de nos observations d'entrée. La covariance de nos données à prédire est calculée comme la covariance entre tous nos points à prédire moins le produit entre la covariance entre tous nos points et le point à prédire avec l'inverse de la somme entre la covariance de toutes nos données en entrée et la matrice de bruit multipliée par la covariance entre nos entrées et le point à prédire.

Ces formules seront d'autant plus explicitées dans notre application à nos données sur la mortalité.

3.1.4 Vraisemblance marginale et optimisation des hyperparamètres du noyau

Comme nous l'avons vu précédemment, les différents noyaux contiennent généralement des hyper-paramètres. En reprenant l'exemple du noyau à fonction de base radiale, les hyperparamètres sont l'échelle de longueur λ , la variance du signal σ_f^2 et la variance du bruit σ_n , qui sont inconnus et doivent être déduits à partir des données. De manière générale toutes les fonctions de covariance ont des hyperparamètres à estimer comme nous le verrons dans notre application numériques sur les données de mortalité. Comme la distribution postérieure sur les hyper-paramètres n'est pas facile à obtenir, l'inférence bayésienne complète des hyper-paramètres n'est pas fréquemment utilisée dans la pratique. Au lieu de cela, la pratique courante consiste à obtenir des estimations à certains

points des hyper-paramètres en maximisant la log vraisemblance marginale. Néanmoins, nous allons rappeler quelques principes de l'inférence bayésienne et montrer comment nous arrivons à l'expression de la logvraisemblance marginale qui nous permet de trouver la valeur de nos hyperparamètres. Moyennant, une distribution initiale qui synthétise les informations autres que les données, la formule de Bayes conduit à une distribution finale qui permet de calculer les probabilités des hypothèses. La règle de Bayes nous donne :

$$p(w|y, X, \theta, \mathcal{H}_i) = \frac{p(y|X, w, \mathcal{H}_i)p(w|\theta, \mathcal{H}_i)}{p(y|X, \theta, \mathcal{H}_i)}$$

Avec :

- w : les paramètres du modèle
- θ : les hyperparamètres
- $\mathcal{H}_i$: l'ensemble des structures des modèles possibles

Le terme $p(y|X, w, \mathcal{H}_i)$ représente la vraisemblance et $p(w|\theta, \mathcal{H}_i)$ les paramètres de la distribution a priori. Ce dernier terme est une distribution de probabilité qui encode nos connaissances sur les paramètres grâce à nos données. Ainsi, la distribution postérieure prend des informations à la fois sur les données grâce à la vraisemblance et sur la distribution a priori. Le terme au dénominateur est $p(y|X, \theta, \mathcal{H}_i)$ est indépendant des paramètres et représente la vraisemblance marginale et nous pouvons noter :

$$p(y|X, \theta, \mathcal{H}_i) = \int p(y|X, w, \mathcal{H}_i)p(w|\theta, \mathcal{H}_i)dw$$

Ainsi de manière analogue, il est possible d'écrire :

$$p(\theta|y, X, \mathcal{H}_i) = \frac{p(y|X, \theta, \mathcal{H}_i)p(\theta|\mathcal{H}_i)}{p(y|X, \mathcal{H}_i)}$$

Avec $p(\theta|\mathcal{H}_i)$ la distribution a priori des hyperparamètres.

Les principes bayésiens fournissent un cadre convaincant et cohérent pour l'inférence. En pratique, en particulier l'évaluation des intégrales peut être difficile, et à titre d'approximation il est plus facile de maximiser la vraisemblance marginale dans l'équation en fonction des hyperparamètres, Θ . Dans le modèle de régression gaussien la distribution a priori est gaussienne. $f|X \mathcal{N}(0, K)$. Par calcul, il est possible de trouver la logvraisemblance marginale :

$$\log p(y|X) = -\frac{1}{2}y^t(K + \sigma_n^2 I)^{-1}y - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi$$

La logvraisemblance marginale peut être vue comme une mesure d'ajustement du modèle avec un terme de pénalisation. Le premier terme $\frac{1}{2}y^t(K + \sigma_n^2 I)^{-1}y$ mesure l'ajustement

à notre jeu de données et le terme $\frac{1}{2}\log|K + \sigma_n^2 I|$ est un terme de pénalisation de la complexité du modèle. Le terme final $\frac{n}{2}\log 2\pi$ est une constante de normalisation. La vraisemblance marginale est normalement maximisée grâce à un outil d'optimisation basé sur la descente de gradient ou alors un ensemble de valeurs initiales est sélectionné au hasard dans une plage possible et celles qui donnent la probabilité marginale la plus élevée sont utilisées comme estimations (Ruhao Wu[1]).

3.2 Application à la modélisation de la mortalité

3.2.1 Adaptation du modèle à la mortalité

Dans cette partie nous appliquerons la théorie de la régression à l'aide des processus gaussiens pour modéliser et prévoir les taux de mortalité. Nous noterons $y_x(t)$ le logarithme des taux de mortalité pour l'âge x l'année t . Pour appliquer la théorie de la partie précédente, nous considérons qu'il existe une fonction $f_x(t)$ dont nous observons les valeurs bruitées au point t . Nous observons les points $\{t_i; y_x(t_i)\}$, $x = 0, \dots, m$, $i = 1, \dots, n$ où m représente l'âge maximum considéré et n le nombre d'année. Notons :

$$y_x(t_i) = f_x(t_i) + \epsilon_{i,x}$$

Avec $\epsilon_{i,x}$ qui représente le bruit. Nous considérons le bruit comme une variable aléatoire indépendante et identiquement distribuée. De plus, le bruit est une variable aléatoire distribuée normalement tel que $\epsilon_{i,x} \sim \mathcal{N}(0, \sigma_x^2)$. Notre objectif est de prévoir les taux de mortalité (log) pour tous les âges étudiés à un horizon h c'est à dire pour les années $t = t_{n+1}, t_{n+2}, t_{n+h}$. Pour un âge x donné, nous considérons que la fonction $f_x(\cdot)$ est un processus gaussien qui a pour fonction moyenne $\mu_x(\cdot)$ et pour fonction de covariance $k_x(\cdot; \cdot)$. Il est alors possible de construire un modèle de régression gaussien pour la fonction $f_x(\cdot)$ comme nous l'avons vu dans la partie précédente. Nous allons toutefois rappeler et adapter certaines formules précédentes car comme nous allons le voir et l'expliquer la fonction de moyenne $\mu_x(\cdot)$ n'est pas nulle. Il sera possible alors d'estimer la moyenne et la variance des taux de mortalité à une date future t^* avec les formules suivantes :

$$\mathbb{E}[f(t^*)] = \mu(t^*) + K(t^*, t)[K(t, t) + \sigma_n^2 * I]^{-1}(Y - \mu)$$

$$Var(f(t^*)|\mathcal{D}) = K(t^*, t^*) - K(t^*, t)[K(t, t) + \sigma_n^2 * I]^{-1}K(t, t^*)$$

Comme nous l'avons dit cette adaptation des formules prend en compte une fonction de moyenne $\mu_x(\cdot)$. Nous devons donc choisir une fonction de moyenne. Comme les logs des taux de mortalité ont tendance à être linéairement décroissants pour la plupart des

âges, nous utilisons un modèle linéaire pour notre moyenne.

$$\mu_x(t) = \alpha_x + \beta_x t$$

Les coefficients α_x et β_x sont des constantes à estimer sur nos données et qui varient en fonction de l'âge x d'étude. Comme nous l'avons vu précédemment et comme nous pouvons le voir dans nos formules la fonction de covariance joue un rôle primordial pour prédire la moyenne et la variance de nouveaux points d'entrée, ici des années. Les fonctions de covariance contiennent nos présomptions sur la fonction que nous souhaitons trouver et définissent la proximité et la similitude entre nos données d'entrée. Le choix de la fonction de covariance a un impact profond sur les performances des modèles de régressions gaussiennes. Il existe un large éventail de fonctions de covariance dans la littérature. Voir par exemple Rasmussen et Williams [5] ou David Duvenaud [7]. Dans notre application, nous considérerons trois fonctions de covariance stationnaire très couramment utilisées, à savoir le noyau à fonction de base radiale (SE), Matern (MA) (avec un degré de liberté égal à 3/2) et quadratique rationnel (RQ), et un noyau à mélange spectral (SM). Ces noyaux sont définis de la manière suivante :

$$\begin{aligned} K_{SE}(t, t') &= \sigma_{SE}^2 \exp\left\{-\frac{(t - t')^2}{2l_{SE}^2}\right\} \\ K_{MA}(t, t') &= \sigma_{MA}^2 \left(1 + \frac{\sqrt{3}(t - t')}{l_{MA}}\right) \exp\left\{-\frac{\sqrt{3}(t - t')}{l_{MA}}\right\} \\ K_{RQ}(t, t') &= \sigma_{RQ}^2 \left(1 + \frac{(t - t')^2}{2\alpha l_{RQ}^2}\right)^{-\alpha} \end{aligned}$$

Bien sûr les hyperparamètres sont à calculer pour chaque âge considéré. Le noyau à mélange spectral étant un peu moins familier et plus complexe nous allons le décrire de façon plus complète dans la partie suivante.

3.2.2 Noyau à mélange spectral

Les noyaux à mélange spectral nous permettent de tirer des informations sur nos données en étudiant leur densité spectrale. Plus précisément un noyau à mélange spectral nous permet de connaître toutes les fréquences de nos données par l'étude de la densité spectrale. La capacité d'apprentissage de notre modèle via la densité spectrale est la même que via l'utilisation du noyau lui-même. Grâce au théorème de Bochner, qui relie les fonctions de covariance à leur transformée de Fourier. Ce théorème est l'idée clé derrière l'utilisation de noyau à mélange spectral. Le théorème implique que tout noyau stationnaire

$k(t, t') = k(|t - t'|)$ est la transformée de Fourier de sa densité spectrale $S(s)$:

$$k(t, t') = \int S(s) e^{2\pi i s^T (t - t')}$$

L'inverse est aussi vrai c'est à dire :

$$S(s) = \int k(t, t') e^{2\pi i s^T (t - t')}$$

En d'autres mots, si nous connaissons la densité spectrale d'un noyau nous pouvons connaître le noyau. Ainsi, ici nous supposons que la densité spectrale est gaussienne :

$$S(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(s - \mu)^2\right\}$$

En appliquant le théorème de Bochner et en posant $\tau = t - t'$

$$\begin{aligned} k(t, t') &= \int \exp(2\pi i s(t - t')) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(s - \mu)^2\right\} ds \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(2\pi i \tau s - \frac{1}{2\sigma^2}(s^2 - \mu s + \mu^2)\right) ds \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-\frac{1}{2\sigma^2}s^2 + \left(2\pi i \tau + \frac{\mu}{\sigma^2}\right)s - \frac{\mu^2}{2\sigma^2}\right) ds \end{aligned}$$

On pose $a = \frac{1}{\sigma^2}, b = 2\pi i \tau + \frac{\mu}{\sigma^2}$ et $c = -\frac{\mu^2}{2\sigma^2}$

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp\left(-a\left(s - \frac{b}{2a}\right)^2\right) \exp\left(\frac{b^2}{4a} + c\right) ds \\ &= \exp\left(\left(-4\pi^2\tau^2 + 4\pi i \tau \frac{\mu}{\sigma^2} + \frac{\mu^2}{\sigma^4}\right) \frac{\sigma^2}{2} - \frac{\mu^2}{2\sigma^2}\right) \\ &= \exp(-2\pi^2(t - t')^2\sigma^2) [\cos(2\pi(t - t')\mu) + i \sin(2\pi(t - t')\mu)] \end{aligned}$$

De plus si nous avons une densité spectrale symétrique avec

$$S(s) = [\phi(s) + \phi(-s)]/2$$

$$\phi(s; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(s - \mu)^2\right\}$$

Nous obtenons alors :

$$k(\tau) = \exp\{-2\pi^2\tau^2\sigma^2\} \cos(2\pi\tau\mu)$$

Si $\phi(s)$ est un mélange de Q gaussienne sur $\mathbb{R}^P$, avec le q^{ieme} composant qui a comme vecteur moyenne $\mu_q = (\mu_q^1, \dots, \mu_q^P)$ et comme matrice de covariance $M_q = diag(v_q^{(1)}, \dots, v_q^{(1)})$. Avec τ_p le p^{ieme} composant du vecteur P -dimensionnelle $\tau = t - t'$, alors notre noyau devient :

$$k(\tau) = \sum_{q=1}^Q w_q \prod_{p=1}^P exp(-2\pi^2 \tau_p^2 v_q^{(p)}) cos(2\pi \tau_p \mu_q^{(p)})$$

Les poids w_q déterminent la contribution de chaque mélange et μ_q et v_q sont les hyperparamètres à estimer. Dans le cas de la mortalité $P=1$ et nous poserons $Q=4$ (Ruhao Wu [1]). En outre, il ressort des exemples numériques présentés plus loin que les taux de mortalité pour différents groupes d'âge présentent des tendances très différentes au fil du temps. Par conséquent, différentes fonctions de covariance peuvent être nécessaires pour différents groupes d'âge, ce qui peut demander du temps de calcul. Le noyau à mélange spectral est censé remédier à ce problème car il peut prendre en charge une large classe de fonctions de covariance stationnaire et permet de capturer diverses structures de covariance dans les taux de mortalité pour différents groupes d'âges. Le noyau à mélange spectral doit donc atténuer les différences de résultats en fonction du noyau.

3.2.3 Fonction de moyenne

Dans les modèles de régression gaussienne, la fonction de moyenne a priori joue un rôle non négligeable dans la prédiction de nos données puisque sur le long terme nos extrapolations tendent vers la moyenne. Comme nous nous intéressons à la mortalité il est logique de considérer que les données récentes ont un impact plus important sur nos prévisions que les données plus anciennes. Autrement dit, la mortalité estimée pour une population de 20 ans en 2025 ressemblera davantage à la mortalité de 2018 que celle de 1960 pour cette même population. Plus notre point d'entrée est récent et plus il aura un poids important pour nos prévisions. Comme nous l'avons vu notre fonction de moyenne est modélisée par une fonction linéaire. Pour obtenir nos coefficients α_x et β_x nous pouvons procéder en appliquant une régression linéaire classique. Mais en procédant de la sorte chaque donnée aurait le même poids, ce qui n'est pas le cas comme nous venons de l'expliquer. Il est donc préférable de modéliser la fonction de moyenne à l'aide d'un modèle de régression linéaire avec poids. Nous utiliserons la méthode des moindres carrés avec poids. Ainsi pour un âge x donné, les paramètres à estimer (α_x et β_x) doivent minimiser l'erreur :

$$e = \sum_{i=1}^n w_i [y_x(t_i) - \alpha_x - \beta_x t_i]^2$$

Où w_i représente le poids attribué pour l'observation de l'année i . Pour calculer ces poids, nous considérons comme dans le travail de Ruhao Wu, qu'ils sont égaux à l'inverse de la

distance temporelle jusqu'à la première année à prévoir, à savoir t_0 dans la partie suivante nous poserons $t_0 = 2000$. Ainsi $w_i = \frac{1}{t_0 - t_i}$. Ainsi, nous pouvons calculer nos paramètres de la manière suivante :

$$W = (w_1, \dots, w_n)$$

$$X = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

$$y_x = (y_x(t_1), \dots, y_x(t_n))^T$$

Puis α_x et β_x sont estimés par : $\begin{pmatrix} \hat{\alpha}_x \\ \hat{\beta}_x \end{pmatrix} = (X^T W X)^{-1} X^T W y_x$

3.2.4 Résultats

Nous proposons d'utiliser la méthode de régression gaussienne pour modéliser et prévoir les taux de mortalité. Nous rappelons que contrairement à la plupart des méthodes existantes de modélisation de la mortalité qui traitent les taux de mortalité pour tous les âges dans une année dans son ensemble et étudient leur évolution dans le temps, nous considérons les taux de mortalité pour tous les âges comme des séries temporelles. Nous travaillerons sur des données issues du site Human Mortality Database. Notre population d'étude sera la population belge dans son ensemble (hommes et femmes). Parmi l'ensemble des données extraites de ce site nous avons récupéré les taux de mortalité pour chaque âge pour les années allant de 1950 à 2018. Les données pendant la seconde guerre mondiale n'étant pas représentatives des conditions de mortalité passée. Les séries temporelles des taux de mortalité de la population belge pour les âges 0, 10, 20, 30, 40, 50, 100 ont été représentées ci-dessous. Nous observons que la mortalité pour tous les âges suit une tendance à la baisse et que cette tendance est linéaire. Ceci vient confirmer le choix de la fonction de moyenne modélisée à l'aide d'une régression linéaire.

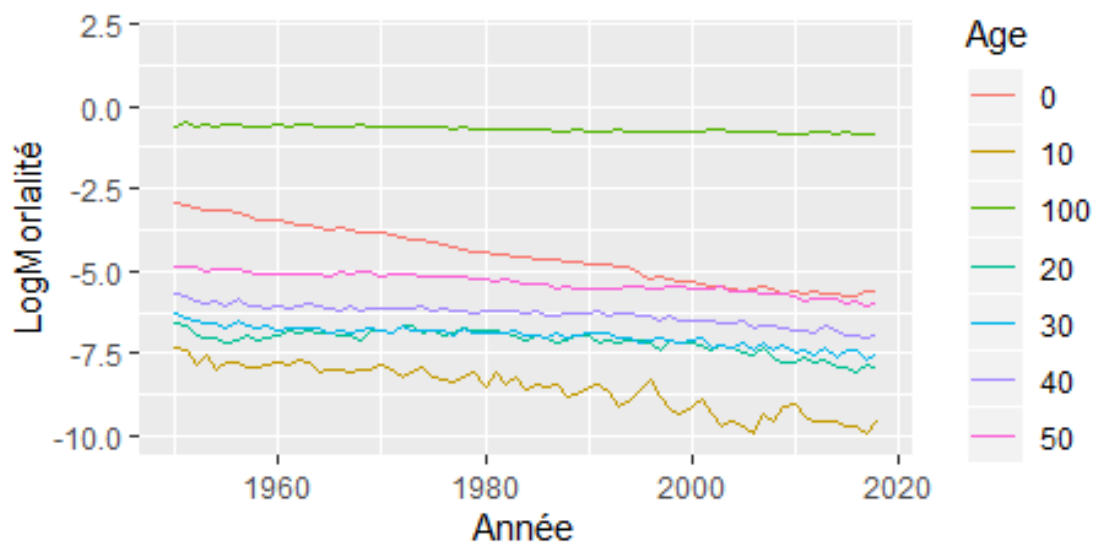


FIGURE 5: Représentation des logarithmes des taux de mortalité de 1950 à 2018 pour les âges 0, 10, 20, 30, 40, 50, 100

Dans un premier temps, nous allons entraîner notre modèle sur les âges suivants : 0, 10, 20, 30, 40, 50 et 100 ans. Par la suite nous étendrons le nombre d'âges choisis pour pouvoir reconstruire des tables de mortalité et construire différentes statistiques. Nous allons ajuster le modèle de régression gaussien aux quatre noyaux décrits précédemment combinés avec notre fonction de moyenne. Pour tester l'ajustement de nos modèles nous avons séparé notre jeu de données en deux échantillons. Un jeu de données d'entraînement qui contient les taux de mortalité de 1950 à 1999. C'est sur ce jeu ci que le modèle va itérer

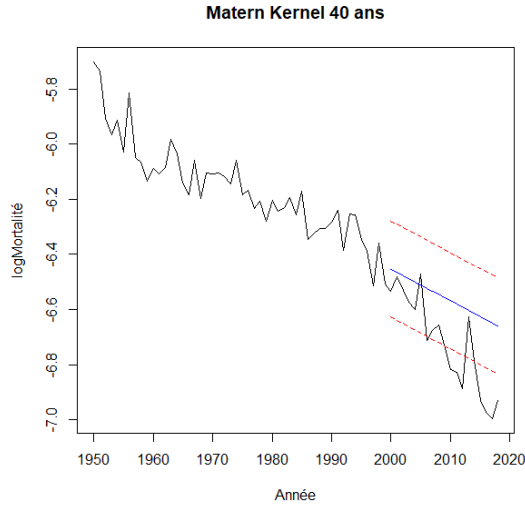


FIGURE 6: Matern Kernel (MA)

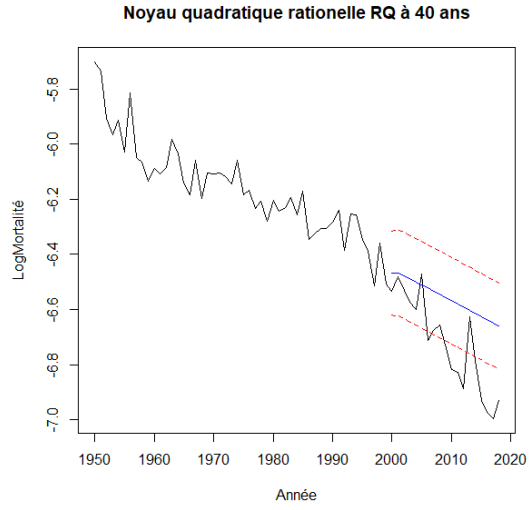


FIGURE 7: Quadratique Rationnel(RQ)

durant la phase d'entraînement pour pouvoir s'appropriier des paramètres, et les ajuster au mieux. Puis un jeu de données de test allant de 2000 à 2018. L'ensemble de données de test est utilisé pour évaluer le modèle final sélectionné au cours du processus de validation. Pour mesurer la qualité de notre modèle nous utiliserons la racine de l'erreur quadratique moyenne (REQM). Elle est une mesure fréquemment utilisée des différences entre les valeurs (valeurs d'échantillon ou de population) prédites par un modèle et les valeurs observées (ou vraies valeurs). La REQM représente la racine carrée du deuxième moment d'échantillonnage des différences entre les valeurs prédites et les valeurs observées. La REQM agrège les erreurs de prédiction de différents points de données en une seule mesure de puissance prédictive accrue. La REQM est une mesure de précision, qui sert à comparer les erreurs de différents modèles prédictifs pour un ensemble de données particulier. Les hyperparamètres de chaque noyau sont estimés à l'aide d'un algorithme d'optimisation afin de maximiser la logvraisemblance marginale. Ensuite, nous projetons les taux de mortalité sur dix-neuf ans de 2000 à 2018 et nous comparons aux valeurs originales. Ci-dessous nous avons représenté les prédictions de nos différents modèles. Pour le groupe de population âgé de 40 ans. La ligne bleue représente la moyenne de nos prédictions et la ligne rouge en pointillé l'intervalle de confiance à 95% de nos estimations.

La racine de l'erreur quadratique pour les âges 0, 10, 20, 30, 40, 50 et 100 ans et pour les différents modèles est donnée dans le tableau ci-dessous. Pour rappel la REQM est calculée de la façon suivante :

$$REQM = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

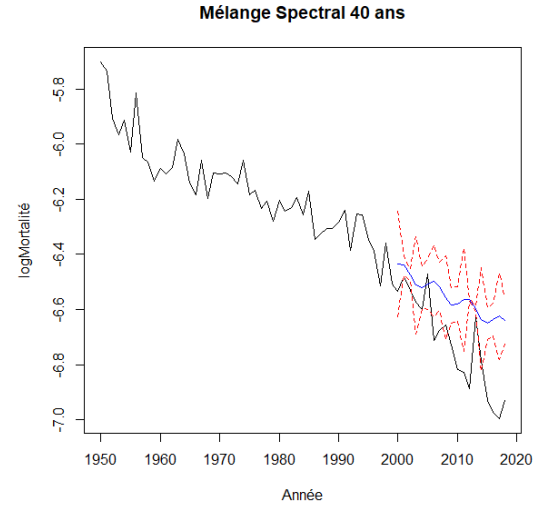
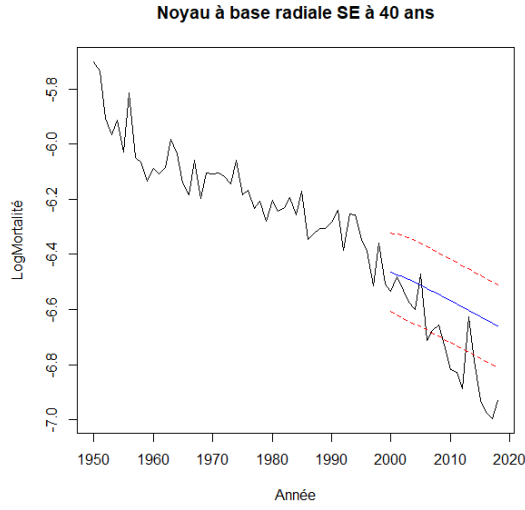


FIGURE 8: Noyau à base radiale(SE) FIGURE 9: Noyau mélange spectral(SM)

Avec ici :

- $\hat{y}_i$ le taux de mortalité projeté l'année i
- y_i le taux de mortalité réel de l'année i
- $n = 19$ allant de 2000 à 2018.

Age	Noyau			
	Matern	RQ	SE	SM
0	0,2545	0,2545	0,2546	0,2567
10	0,4164	0,2754	0,2768	0,3323
20	0,4021	0,4015	0,4030	0,4027
30	0,1965	0,1674	0,1627	0,1652
40	0,2012	0,2010	0,2008	0,2051
50	0,1416	0,1412	0,1419	0,1430
100	0,0515	0,0515	0,0515	0,0512
Moyenne	0,2377	0,2132	0,2130	0,2152

TABLE 1: Calcul de la REQM pour tous les noyaux

Nous observons à travers ces résultats que les différents noyaux performant de manière similaire. Pour certains âges, le choix du noyau est significatif notamment à 10 ans. Néanmoins, sur les âges considérés le noyau à mélange spectral n'est pas le meilleur modèle en terme de REQM. Cette différence de conclusion sur ce noyau comparé au travail de Ruhao Wu peut s'expliquer par le fait que nous avons utilisé une fonction de moyenne pour tous les noyaux et pas uniquement pour le noyau à mélange spectral. Ce qui améliore considérablement les prévisions de ces mêmes noyaux.

Pour approfondir notre analyse des résultats nous allons étendre notre étude à un plus large choix d'âges. Ceci nous sera utile afin de construire des tables et courbe de mortalité, notamment pour une application financière ou pour calculer des informations démographiques. Nous considérons cette fois les âges suivants : 0, 1, 2, 5, 10, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90 et 100 ans. Cette sélection s'explique par le fait que nous avons besoin de plus de points d'échantillons pour les âges qui ont connu les plus grandes variations de mortalité et que nous avons besoin de moins de points pour les zones à plus faibles variations. Nous considérons comme dans la partie précédente les mêmes noyaux et les mêmes ensembles pour nos données d'entraînement et de test. Une fois nos prévisions réalisées sur tous ces âges, nous allons interpoler les âges manquants à l'aide de spline cubique. Lorsqu'on a déterminé les éléments du spline, c'est-à-dire les paramètres des polynômes qui joignent chacun des points donnés (les nœuds), on peut calculer l'ordonnée d'un point situé entre deux nœuds : c'est la valeur du polynôme pour l'abscisse correspondante. La spline permet donc d'interpoler les valeurs comprises entre les nœuds. Nous recalculerons l'erreur quadratique moyenne afin de comparer la performance de nos noyaux sur un échantillon plus large. Dans cette partie nous arrêterons la projection à 2018 qui est la dernière année pour laquelle nous possédons des données complètes. L'idée est de confirmer l'ensemble de nos modèles pour pouvoir par la suite construire des tables de mortalité prospectives. Comme dans les modèles classiques de projection de la mortalité nous considérons plus seulement les taux de mortalité comme des séries temporelles pour un âge donné mais plutôt ici l'ensemble des taux de mortalité pour tous les âges une année donnée.

Age	Noyau			
	Matern	RQ	SE	SM
0	0,2545	0,2545	0,2546	0,2567
1	0,3124	0,2982	0,2981	0,3199
2	0,2631	0,3558	0,3726	0,3650
5	0,3978	0,4561	0,4599	0,4276
10	0,4164	0,2754	0,2768	0,3323
12	0,3231	0,3844	0,3852	0,3859
15	0,3528	0,3877	0,3898	0,3860
18	0,5629	0,3980	0,3749	0,4018
20	0,4021	0,4015	0,4030	0,4027
22	0,3756	0,3921	0,3921	0,3935
25	0,2576	0,2618	0,2617	0,2625
28	0,2017	0,1613	0,1668	0,1624
30	0,1965	0,1674	0,1627	0,1652
40	0,2012	0,2010	0,2008	0,2051
50	0,1416	0,1412	0,1419	0,1430
60	0,0376	0,0980	0,1018	0,0910
70	0,2072	0,0925	0,0900	0,1010
80	0,1683	0,0998	0,0997	0,1033
90	0,0395	0,0401	0,0430	0,0426
100	0,0515	0,0515	0,0515	0,0512
Moyenne	0,2582	0,2459	0,2463	0,2499

TABLE 2: Calcul de la REQM pour tous les noyaux et les âges étendus

Nous observons à travers ces résultats élargis qu'encore une fois certains noyaux prédisent très bien pour des âges donnés mais un peu moins pour d'autres. Par exemple, le noyau Matern prédit le mieux à deux ans mais beaucoup moins bien à 18 ans. Il est intéressant de voir que le meilleur noyau est cette fois-ci le noyau quadratique exponentiel. Encore une fois, les résultats en moyenne sont semblables. Nous constatons aussi que certains âges sont plus difficiles à estimer et ceci conforte notre choix de sélectionner plus de points d'échantillon pour les jeunes âges pour construire les courbes de mortalité.

Une fois la projection effectuée pour tous ces âges nous utilisons l'interpolation par spline cubique pour les âges allant de 0 à 100 ans. Nous avons représenté ci-dessous la courbe de mortalité pour l'année 2018 avec donc une projection sur dix-neuf années avec les mêmes données d'entraînement que précédemment (1950 à 1999).

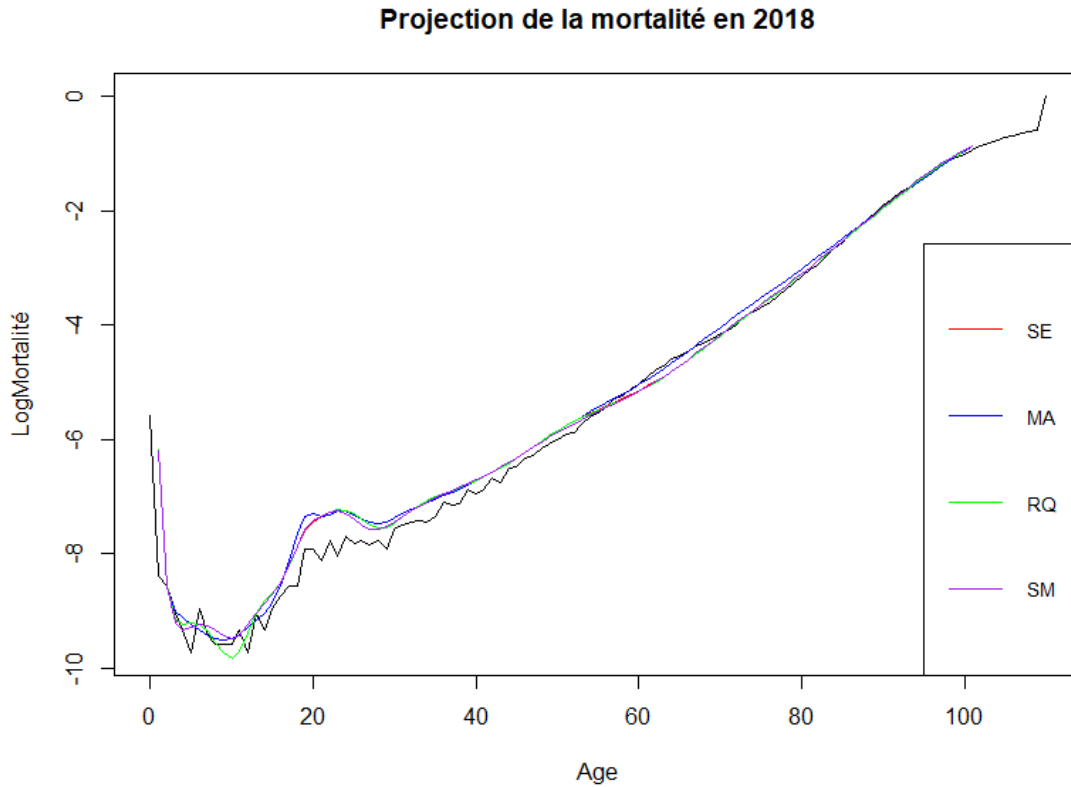


FIGURE 10: Représentation de la log mortalité pour tous les âges et l'ensemble des noyaux en 2018.

Grâce à cette représentation nous pouvons confirmer que nos noyaux prédisent bien la mortalité notamment pour les grands âges. L'ensemble des modèles ont tendance à légèrement surestimer les taux de mortalité aux alentours des vingt ans. Cette bosse de mortalité est souvent décrite par les démographes comme un élément distinct de la mortalité humaine qui s'ajoute aux autres processus dictant la mortalité générale. On parle de surmortalité dans la mesure où elle s'additionne au niveau attendu du fait des contraintes biologiques et des circonstances épidémiologiques. Les modèles décrivent très bien cette tendance générale de la mortalité. Avec le risque de décès élevé à la naissance et durant les premières années de vie, puis celui-ci diminue pendant l'enfance jusqu'à atteindre un minimum aux alentours de 10 ans. Il augmente ensuite au cours de la vie adulte. En nous penchant de plus près sur chaque modèle indépendamment nous observons que le noyau matern surestime d'autant plus la mortalité à 20 ans. Le noyau quadratique rationnel lui sous-estime la mortalité aux alentours des dix ans. Par la suite les courbes sont superposées et très proches des taux réels. L'intérêt du noyau à mélange spectral se reflète davantage ici que sur les autres exemples. C'est le noyau qui prédit le mieux à un horizon plus lointain, ici à une projection de 19 ans dans le temps. Comme nous pouvons le voir

en terme de REQM dans le tableau 3.

Afin de comparer le pouvoir prédictif de nos modèles sur un plus court terme, nous avons représenté la même courbe sur un horizon plus restreint de 5 ans donc en 2004.

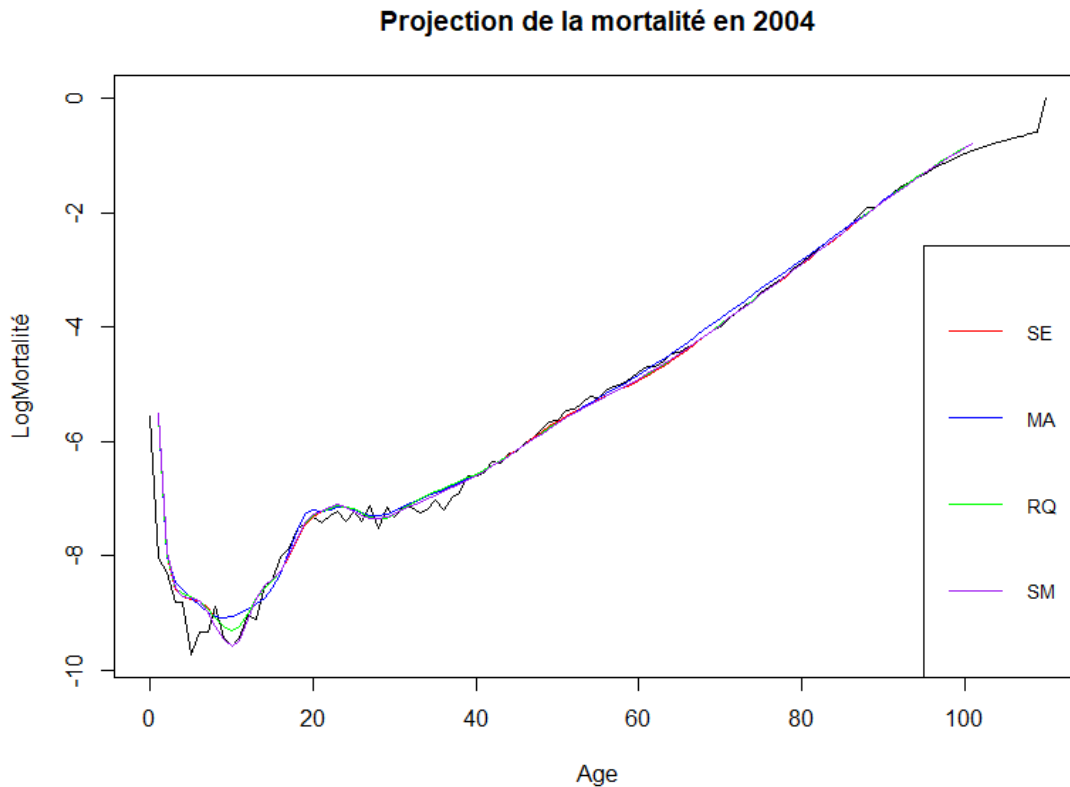


FIGURE 11: Représentation de la log mortalité pour tous les âges et l'ensemble des noyaux en 2004.

Nous observons que pour les projections à 5 ans sont logiquement meilleures. La sur-estimation de la bosse de mortalité à 20 ans est beaucoup plus faible. Toutes les courbes sont superposées sauf pour les âges inférieurs à 20 ans ou certaines prédictions s'éloignent de la réalité. Dans l'ensemble tous les résultats sont très bons et reproduisent bien les données originales.

REQM	2004	2018
Matern	0,0693	0,1961
RQ	0,0643	0,1864
SE	0,0654	0.1855
SM	0,0657	0.1843

TABLE 3: Erreur de prédictions pour tous les noyaux à différents horizon

Le tableau des erreurs de prédictions confirme nos conclusions sur les courbes de mortalité. Les résultats de à 5 ans sont très bons et très proches de la réalité. Les noyaux performant de manière assez similaire. Pour les prévisions à 20 ans, les résultats sont logiquement plus éloignés qu'à 20 ans mais ils restent très proches de la réalité. Il est intéressant de remarquer que le noyau à mélange spectral est celui qui prédit le mieux à long terme.

3.2.5 Comparaison avec le modèle de Lee-Carter et sa variante Poisson

Dans cette partie nous nous intéresserons au modèle log-Poisson qui est une expansion du modèle de Lee-Carte. L'idée n'est pas d'explorer en détail ce modèle bien connu des actuaires et démographes mais d'ajuster le modèle sur les mêmes données que précédemment pour comparer les résultats obtenus et évaluer notre modèle.

Rappel sur le modèle de Poisson

Comme nous l'avons dit précédemment le modèle de Poisson est une extension du modèle de Lee-Carter. Le modèle de Lee-Carter nous dit que le taux de mortalité instantané $\mu_x(t)$ se modélise de la façon suivante :

$$\ln(\mu_x(t)) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}$$

Avec :

- α_x : la moyenne des log de la mortalité définie pour tous les âges
- κ_t : un processus stochastique qui décrit l'évolution de la mortalité
- β : la sensibilité de la mortalité instantanée à l'âge x par rapport à l'évolution générale des κ_t
- $\epsilon_{x,t}$ un bruit gaussien avec une volatilité σ constante

L'idée du modèle de Poisson est de modéliser le nombre de décès à l'âge x de l'année t, $D_{x,t}$, par une loi de Poisson de paramètre $L_{x,t}\mu_{x,t}$. Le modèle de Poisson corrige l'hypothèse d'homoscédasticité du modèle de Lee-Carter en permettant à la volatilité de $\epsilon_{x,t}$ de varier en fonction de l'âge. Néanmoins le taux de mortalité instantanée $\mu_x(t)$ dans le modèle de Poisson garde la même structure que dans le modèle de Lee-Carter.

Afin d'ajuster les paramètres du modèle et obtenir des paramètres uniques nous transformons les variables de la façon suivante :

$$(\alpha_x, \beta_x^1, \kappa_t^1) \rightarrow (\alpha_x + c_1 \beta_x^1, c_2(\kappa_t^1 - c_1))$$

De la même façon nous transformons nos contraintes de modèle :

$$\sum_x \beta_x^1 = 1, \sum_t \kappa_t^1 = 0 \rightarrow c_1 = \frac{1}{n} \sum_t \kappa_t^1, c_2 = \sum_x \beta_x^1$$

Comme auparavant, nous ajustons le modèle sur un jeu de données d'entraînement qui contient les taux de mortalité de 1950 à 1999. Puis un jeu de données de test allant de 2000 à 2018. Nous projetons les taux de mortalité pour tous les âges et au même horizon (2018) pour comparer à nos modèles de régressions gaussiennes. Nous avons représenté ci-dessous plusieurs trajectoires des taux de mortalité pour l'ensemble de la population belge et le groupe d'âge de 40 ans. Nous avons aussi calculé la racine de l'erreur quadratique moyenne pour les âges suivants : 0, 1, 2, 5, 10, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90 et 100 ans. Le modèle ne demande pas d'interpolation car il projette pour tous les âges mais encore une fois cette sélection est utilisée dans un objectif de comparaison.

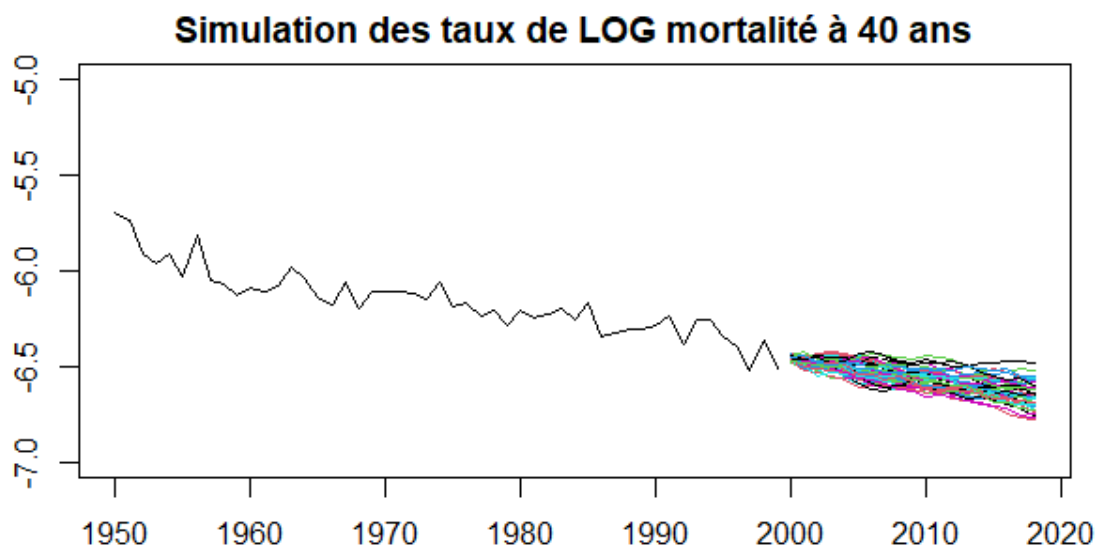


FIGURE 12: Représentation de la log mortalité à 40 ans avec le modèle Log-Poisson

Age	0	1	2	5	10	12	15	18	20	22	25	28	30	40	50	60	70	80	90	100
REQM	0.4948	0.5964	0.3539	0.3592	0.3572	0.2942	0.2980	0.5601	0.4839	0.3497	0.2393	0.1846	0.1722	0.2063	0.1060	0.0466	0.1445	0.1310	0.0455	0.1776
Moyenne	0.28																			

L'erreur de notre modèle de Lee-Carter sur les âges considérés précédemment est de 0.28. Par comparaison avec les erreurs obtenues sur les noyaux le modèle de Lee-Carter Poisson est le moins bon. Il est possible de tracer la courbe de ce dernier modèle pour l'année 2018 et de comparer à notre modèle avec le mélange spectral. Sur le graphique, nous pouvons voir que le modèle de Poisson évalue moins bien la mortalité pour les jeunes âges et c'est ce qui explique une erreur moyenne plus élevée. Par la suite, la performance des deux modèles est similaire et très proche de la courbe réelle de mortalité.

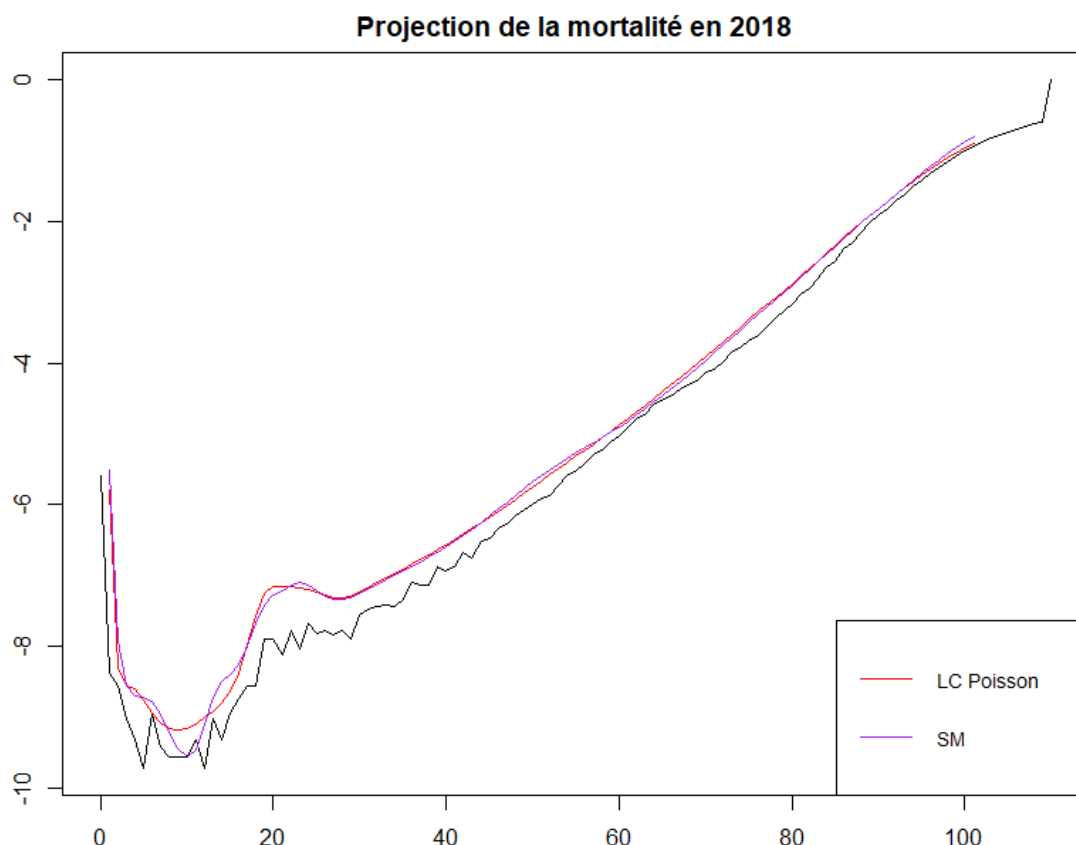


FIGURE 13: Représentation de la log mortalité pour tous les âges avec le modèle de Lee-Carter Poisson et le noyau à mélange Spectral

3.3 Conclusion

Dans cette première partie, nous avons introduit les modèles de régression gaussienne d'abord de manière générale puis dans le but de prévoir la mortalité future pour la population belge sans distinction de sexe. Nous avons étudié différentes fonctions de covariance classiques et celle à mélange spectral. Nous avons obtenu de très bons résultats avec l'ensemble des noyaux, résultats qui comme nous l'avons vu surpassent ceux du modèle de Lee-Carter et de sa variante Poisson. Nous avons vu que le choix du noyau peut sensiblement jouer sur les performances de nos modèles. L'introduction du noyau à mélange spectral est intéressant car il est celui qui prédit le mieux sur un horizon lointain et avec une stabilité de résultats intéressante. Cette première partie permet de valider l'ensemble de nos modèles qui parviennent bien à reproduire les tendances de mortalité. Dans une autre partie nous utiliserons l'ensemble des données de mortalité disponibles afin de projeter la mortalité à des dates futures et de calculer des statistiques sur l'espérance de vie et d'utiliser ces dernières pour une application financière comme pour une rente viagère.

4 Analyse en composantes principales fonctionnelle

Dans cette partie, nous présenterons l'analyse en composantes principales fonctionnelle. D'abord, nous indiquerons l'intérêt d'utiliser des données fonctionnelles. Ensuite, nous présenterons de manière générale l'analyse en composantes principales fonctionnelle avec le cas univarié puis multivarié. Enfin, nous montrerons comment ces techniques peuvent être utilisées afin de modéliser la mortalité et la projeter pour les années futures.

4.1 Présentation générale

4.1.1 Intérêt d'utiliser des données fonctionnelles

L'analyse des données fonctionnelles est une branche des statistiques qui analyse les données fournissant des informations sur les courbes, les surfaces ou tout autre élément variant en continu. Ainsi, chaque échantillon de données est considéré comme une fonction. L'idée est de ne plus travailler sur des dimensions finies mais en dimensions infinies avec l'utilisation de fonction. Cette extension s'explique notamment par les progrès en matière d'accès à l'information et de stockage de ces mêmes informations. Dès lors que nous disposons de données mesurées à différents instants nous pouvons considérer que nos observations sont des courbes aléatoires dépendant du temps. Par conséquent, ces mêmes courbes sont des échantillons d'un processus stochastique. L'analyse de données fonctionnelles permet d'extraire davantage d'informations sur les dérivées ou la régularité de ces fonctions. L'analyse en composante principale qui est une technique particulière de l'analyse fonctionnelle a pour but de réduire les données en dimensions finies ce qui permet de produire des fonctions dites composantes principales de dimension minimale qui reproduisent un maximum d'informations sur les données étudiées.

Construction de données fonctionnelles

Avant de présenter en détail l'analyse en composantes principales fonctionnelle nous allons expliquer brièvement comment construire des fonctions à partir des données de mortalité que nous possédons. En effet, nous possédons les taux de mortalité à des points discrets, âges et années entières et nous voulons travailler sur des fonctions donc avec des dimensions infinies.

Pour créer des fonctions, il faut :

- 1. Définir un ensemble de blocs fonctionnels ϕ_k appelés fonctions de base.
- 2. Ensuite, nous définissons un vecteur, une matrice ou un tableau de coefficients pour définir la fonction comme une combinaison linéaire de ces fonctions de base.

Il y a généralement deux types de fonctions, les fonctions périodiques et les fonctions non-périodiques. Pour la première catégorie, une base de Fourier est utilisée alors que

pour les fonctions non-périodiques, notre cas, nous allons utiliser une base de fonctions constituée de splines et plus particulièrement de bsplines. Notre fonction $x(t)$ peut s'écrire mathématiquement :

$$x(t) = \sum_{k=1}^K c_k \phi_k(t)$$

Les paramètres $c_1, c_2, \dots, c_K$ sont les coefficients d'expansion. Ici, nos fonctions de base sont des splines. Les splines sont des polynômes par morceaux. Les bases de spline sont plus flexibles et donc plus compliquées que les séries de Fourier. Elles sont définies par le domaine de validité, les nœuds, et l'ordre. Les splines sont construites en divisant l'intervalle d'observation en sous-intervalles, avec des limites aux points appelés points de rupture. Sur n'importe quel sous-intervalle, la fonction spline est un polynôme de degré ou d'ordre fixe, mais la nature de chaque polynôme change au fur et à mesure que l'on passe au sous-intervalle suivant.

Pour résumer, la création de données fonctionnelles se fait à partir observations discrètes sur une ou plusieurs courbes et pour une ou plusieurs variables qui sont ajustées avec un ensemble de courbes lissées, chacune définie par une expansion en terme de fonctions de base. Le critère d'ajustement est la méthode des moindres carrés pondérés, et le lissage peut être défini avec une pénalité. Les systèmes de fonctions de base sont les splines pour les courbes non périodiques et les séries de Fourier pour les courbes périodiques, voir J.O. Ramsay, Giles Hooker et Spencer Graves [8].

4.1.2 Présentation du cas univarié

L'analyse en composantes principales fonctionnelle (FPCA) est une méthode statistique qui étudie la variation de plusieurs courbes en réduisant la dimension de nos données d'une dimension infinie à des dimensions finies par une représentation en composantes principales de la variation. L'idée est de réduire la dimension de nos données et de rendre l'information moins redondante. Le FPCA fait essentiellement la même chose que le PCA mais lors du cas classique nous traitons de vecteurs, tandis qu'en FPCA nous traitons de fonctions. Cela signifie que nos composants principaux sont également des fonctions ou des courbes et non plus des vecteurs.

Considérons un processus stochastique $Y(x) \in \mathbb{L}^2$ et défini sur le domaine $\mathcal{X}$. Nous définissons :

- Une fonction de moyenne $\mu(x) = \mathbb{E}(Y(x)) \forall x \in \mathcal{X}$
- Une fonction de covariance $K(X, X') = Cov(Y(x), Y(x')) \forall x \in \mathcal{X}$

En admettant l'existence d'un opérateur de covariance $\Lambda : L^2(\mathcal{X}) \rightarrow L^2(\mathcal{X})$ pour toute fonction $f \in L^2(\mathcal{X})$

$$(\Lambda f)(x) = \int_{\mathcal{X}} K(x, x') f(x') dx' \quad \forall x \in \mathcal{X}$$

Avec l'opérateur de covariance λ défini, nous pouvons faire l'analyse spectrale de la fonction de covariance $K(x, x')$ tel que :

$$(\Lambda \phi)(x) = \int_{\mathcal{X}} K(x, x') \phi(x') dx' = \lambda \phi(x)$$

De tel sorte que nous avons un ensemble de fonctions propres $\{\phi_n(x)\}_{n=1}^{\infty}$ avec les valeurs propres correspondantes $\{\lambda_n\}_{n=1}^{\infty}$, avec λ_n représentant la variabilité de $Y(x)$ expliquée par la n_{ieme} fonction propre. Notre processus $Y(x)$ peut alors être représenté comme une combinaison linéaire infinie de fonction orthonormée et par la représentation de Karhunen-Loeve :

$$Y(x) = \mu(x) + \sum_{n=1}^{\infty} \beta_n \phi_n(x) \quad , \forall x \in \mathcal{X}$$

Avec β_n le score de la composante principale :

$$\beta_n = \int_{\mathcal{S}} (Y(x) - \mu(x)) \phi_n(x) dx$$

Les scores des composantes principales $\{\beta_n\}_{n=1}^{\infty}$ sont des variables aléatoires non corrélées de moyenne zéro et de variance $\{\lambda_n\}_{n=1}^{\infty}$. β_n sert de poids et de projection au processus stochastique centré $(Y(x) - \mu(x))$ dans la direction de la n_{ieme} fonction propre $\phi_n(x)$ dans la représentation de Karhunen-Loeve de $Y(x)$. Comme les valeurs propres décroissent rapidement vers zéro, seules les premières fonctions propres sont nécessaires pour représenter les caractéristiques les plus importantes du processus $Y(x)$. La version tronquée de la représentation de Karhunen-Loeve avec les N premières composantes, s'écrit :

$$Y(x) = \mu(x) + \sum_{n=1}^N \beta_n \phi_n(x), \quad \forall x \in \mathcal{X}$$

Ainsi, il est possible de réduire la dimension infinie initiale de nos données fonctionnelles vers une dimension finie où les variations principales sont expliquées par les n premières composantes.

4.1.3 Présentation du cas multivarié

Dans cette partie, nous nous intéresserons au cas multivarié de l'analyse en composantes principales fonctionnelle. Nous ferons un lien direct avec la représentation de Karhunen-Loeve et le cas univarié. Par la suite, nous utiliserons le cas multivarié pour modéliser la mortalité de plusieurs sous-populations par exemple les hommes et les femmes en Belgique. Nous présenterons les mathématiques du cas multivarié puis un algorithme d'implémentation numérique de cette méthode. Nous pouvons considérer un échantillon aléatoire de $p \geq 2$ d'un ensemble de fonctions $Y^1(x), \dots, Y^p(x)$ sur le domaine $\mathcal{X}$ pour tout $x \in \mathcal{X}$. Dans le cadre de ce mémoire, nous nous intéresserons surtout au cas bivarié ($p = 2$). Nous pouvons représenter toutes ces fonctions dans un unique vecteur $Y(x)$,

$$Y(x) = (Y^1(x), \dots, Y^p(x))^T \in \mathbb{R}^p$$

Avec comme fonction de moyenne :

$$\mu(x) = (\mathbb{E}(Y^1(x)), \dots, \mathbb{E}(Y^p(x)))^T = (\mu^1(x), \dots, \mu^p(x))^T$$

Et comme fonction de covariance :

$$K_{i,j}(x, x') = Cov(Y^i(x), Y^j(x')) = \mathbb{E}[Y^i(x) - \mu^i(x)(Y^j(x') - \mu^j(x'))]$$

Comme dans le cas univarié, nous considérons cette fois un ensemble de fonctions $f = (f^1, \dots, f^p)$ indexé par $i = 1, \dots, p : f^i \in L^2(\mathcal{X})$. En admettant l'existence d'un opérateur de covariance $\Lambda : L^2(\mathcal{X}) \rightarrow L^2(\mathcal{X}) \forall f \in L^2(\mathcal{X})$ avec le i_{eme} élément de (Λf) ,

$$(\Lambda f)^i(x) = \sum_{j=1}^p \int_{\mathcal{X}} K_{i,j}(x, x') f^j(x') dx', \quad \forall x \in \mathcal{X}$$

.

Une fois l'opérateur de covariance défini, nous pouvons effectuer l'analyse spectrale sur la fonction de covariance $K_{i,j}(x, x')$, avec $\psi(x)$ la base orthonormée des fonctions propres et ν la valeur propre correspondante. Ainsi, pour tout $i, j = 1, \dots, p$ et $x \in \mathcal{X}$ nous pouvons écrire :

$$(\Lambda \psi)^i(x) = \sum_{j=1}^p \int_{\mathcal{X}} K_{i,j} \psi^j(x') dx' = \nu \psi^i(x)$$

En considérant que chaque fonction $Y^1(x), \dots, Y^p(x)$ possède une représentation univariée

de Karhunen-Loève donnée par les N premières composantes, il est possible d'écrire :

$$Y^i(x) = \mu^i(x) + \sum_{n=1}^N \beta_n^i \phi_n^i(x)$$

En admettant que la fonction de covariance est séparable, c'est à dire : $K_{i,j}(x, x') = K_i(x, x').K_j(x, x')$ alors il est possible d'écrire pour $m, l = 1, \dots, N$ et $i, j = 1, \dots, p$ et pour tout $x \in \mathcal{X}$:

$$\begin{aligned} (\Lambda\psi)^i(x) &= \sum_{j=1}^p \sum_{m=1}^N \sum_{l=1}^N \int_{\mathcal{X}} Cov(\beta_m^i \phi_m^i(x), \beta_j^i \phi_j^i(x')) dx' \\ &= \sum_{j=1}^p \sum_{m=1}^N \sum_{l=1}^N Cov(\beta_m^i, \beta_l^j) \phi_m^i(x) \int_{\mathcal{X}} \phi_l^j(x') \psi^j(x') dx' = \nu \psi^i(x) \end{aligned}$$

En notant $Cov(\beta_m^i, \beta_l^j) = Z_{ml}^{ij}$ et $\int_{\mathcal{X}} \phi_l^j(x') \psi^j(x') dx' = c_l^j$ Ainsi :

$$\sum_{j=1}^p \sum_{m=1}^N \sum_{l=1}^N Z_{ml}^{ij} \phi_m^i(x) c_l^j = \nu \psi^i(x)$$

En multipliant de chaque côté par l'intégrale de la fonction propre orthonormée $\phi_n^i(x)$ pour $n = 1, \dots, N$ sur le domaine $\mathcal{X}$,

$$\int_{\mathcal{X}} \sum_{j=1}^p \sum_{m=1}^N \sum_{l=1}^N Z_{ml}^{ij} \phi_m^i(x) c_l^j dx = \nu \int_{\mathcal{X}} \phi_n^i(x) \psi^i(x) dx$$

Nous avons par l'orthonormalité, $\int_{\mathcal{X}} \phi_n^i(x) \phi_m^i(x) dx = 1$ si $m = n$ et zéro sinon. En notant $\int_{\mathcal{X}} \phi_n^i(x) \psi^i(x) dx' = c_n^i$ nous avons le système suivant :

$$\sum_{j=1}^p \sum_{l=1}^N Z^{ij} c_l^j = \nu c_n^i$$

Ou alors sous forme matricielle :

$$Zc = \nu c$$

$$\begin{pmatrix} Z^{1,1} & \dots & Z^{1,p} \\ \vdots & \ddots & \vdots \\ Z^{p,1} & \dots & Z^{p,p} \end{pmatrix} \begin{pmatrix} c^1 \\ \vdots \\ c^p \end{pmatrix} = \nu \begin{pmatrix} c^1 \\ \vdots \\ c^p \end{pmatrix}$$

En substituant le vecteur orthonormal c_n et les valeurs propres correspondantes ν_n , les

fonctions propres $\psi_n^i(x)$ de Λ sont obtenues :

$$\psi_n^i(x) = \frac{1}{\nu_n} \sum_{j=1}^p \sum_{m=1}^N \sum_{l=1}^N Z_{ml}^{ij} [c_n]_n^j \phi_m^i(x) = \sum_{m=1}^N [c_n]_m^i \phi_m^i(x), \quad \forall x \in \mathcal{X}$$

Avec $[c_n]^i$ le i_{eme} bloc du vecteur propre orthonormé c_n de Z correspondant à ses valeurs propres ν_n . La version tronquée multivariée de la représentation de Karhunen-Loève avec les N premières composantes du processus $Y^i(x)$ est alors notée :

$$Y^i(x) = \mu^i(x) + \sum_{n=1}^N \rho_n \psi_n^i(x), \quad \forall x \in \mathcal{X}$$

Avec ρ_n le score de la composante principale multivariée :

$$\begin{aligned} \rho_n &= \sum_{i=1}^p \int_{\mathcal{X}} (Y^i(x) - \mu^i(x)) \psi_n^i(x) dx \\ &= \sum_{i=1}^p \int_{\mathcal{X}} (Y^i(x) - \mu^i(x)) \sum_{m=1}^N [c_n]_m^i \phi_m^i(x) \\ &= \sum_{i=1}^p \sum_{m=1}^N \int_{\mathcal{X}} (Y^i(x) - \mu^i(x)) \phi_m^i(x) \\ &= \sum_i^p \sum_{m=1}^N [c_n]_m^i \beta_m^i \end{aligned}$$

β_m^i est le m_{ieme} score de la composante principale univariée du i_{ieme} élément. Il est possible de calculer la moyenne et la covariance de ρ_n :

$$\mathbb{E}(\rho_n) = \sum_{i=1}^p \int_{\mathcal{X}} \mathbb{E}(Y^i(x) - \mu(x)) \psi_n^i(x) dx = 0$$

$$\begin{aligned} Cov(\rho_n \rho_m) &= \mathbb{E} \left(\sum_{i=1}^p \int_{\mathcal{X}} (Y^i(x) - \mu^i(x)) \psi_n^i(x) dx \cdot \sum_{j=1}^p \int_{\mathcal{X}} (Y^j(x') - \mu^j(x')) \psi_m^j(x') dx' \right) \\ &= \sum_{i=1}^p \int_{\mathcal{X}} \sum_{j=1}^p \int_{\mathcal{X}} K_{ij}(x, x') \psi_m^j(x') dx' \psi_n^i(x) dx \\ &= \sum_{i=1}^p \int_{\mathcal{X}} \nu_m \psi_m^i(x) \psi_n^i(x) dx \\ &= \nu_m \sum_{i=1}^p \int_{\mathcal{X}} \psi_m^i(x) \psi_n^i(x) dx \end{aligned}$$

Implémentation numérique

Dans cette partie, nous reprendrons l'algorithme décrit par HAPP et GREVEN [9] qui permet d'implémenter la méthode décrite précédemment. L'algorithme est constitué de quatre étapes et il utilise des données $x_1, \dots, x_N$ centrées, dont nous avons soustrait la moyenne.

- 1. Calcul des composantes principales fonctionnelles univariées pour chaque élément $j = 1, \dots, p$. Ce qui nous permet d'obtenir les fonctions principales de chaque composante $\phi_1^{(j)}, \dots, \phi_M^{(j)}$, ainsi que leur score $\epsilon_{i,1}^{(j)}, \dots, \epsilon_{i,M}^{(j)}$ pour chaque observation pour un choix du nombre de composante principale bien choisie M_j .
- 2. Combiner tous les coefficients dans une matrice $\Theta \in \mathbb{R}^{N \times M}$ avec $M_+ = M_1 + \dots + M_p$. Les lignes de la matrice sont données par :

$$\Theta_i = (\epsilon_{i,1}^{(1)}, \dots, \epsilon_{i,M_1}^{(1)}, \dots, \epsilon_{i,1}^{(p)}, \dots, \epsilon_{i,M_p}^{(p)})$$

Nous estimons la covariance par la matrice $\hat{Z} = \frac{1}{N-1} \Theta^T \Theta$

- 3. Calcul des vecteurs propres $\hat{e}_m$ et des valeurs propres $\hat{\nu}_m$ de $\hat{Z}$ pour $m = 1, \dots, M$
- 4. Calcul des fonctions $\hat{\psi}_m$ de l'analyse en composantes principales fonctionnelle multivariée et les scores associés $\hat{\rho}_{i,m}$

$$\hat{\psi}_m^j = \sum_{n=1}^{M_j} [\hat{c}_m]_n^j \hat{\phi}_n^j$$

$$\hat{\rho}_{i,m} = \sum_{j=1}^p \sum_{n=1}^{M_j} [\hat{c}_m]_n^j \epsilon_{i,m}^j = \Theta_i, \hat{c}_m \quad m = 1, \dots, M$$

L'intérêt d'utiliser l'analyse en composantes principales fonctionnelle multivariée est la prise en compte de la covariance entre les différents éléments, en utilisant la covariance conjointe des scores de tous les éléments. Pour résumer, la simulation est basée sur la version multivariée du Théorème de Karhunen-Loève pour des données fonctionnelles multivariées $x_i = (x_i^1, \dots, x_i^p)$:

$$x_i(t) = \mu(t) + \sum_{m=1}^M \rho_{i,m} \phi_m(t), \quad i = 1, \dots, N, \quad t = (t_1, \dots, t_p) \in \mathbb{T}_1 \times \dots \times \mathbb{T}_p$$

4.2 Application à la modélisation de la mortalité

Dans cette partie, nous appliquerons l'analyse en composantes principales fonctionnelle univariée et multivariée dans un but de projection et de modélisation de la mortalité. Nous nous intéresserons à la modélisation des taux de mortalité en Belgique en considérant des données spécifiques au sexe. Nous nous intéresserons aux mêmes années d'études que lors de la régression gaussienne c'est à dire de 1950 à 1999 et une projection jusqu'en 2018 pour cette partie. Nous nous intéresserons aussi aux données sans distinction de sexe dans un souci de comparaison pour le cas univarié avec les résultats de la partie précédente. Pour le cas univarié et multivarié, nous fonctionnerons de la même manière avec la création des données fonctionnelles dans un premier temps, puis la modélisation en composantes principales et enfin la projection en séries temporelles des différents scores pour obtenir les taux de mortalité futurs. L'analyse multivariée sera l'occasion d'introduire la modélisation simultanée de plusieurs sous-groupe de population.

4.2.1 Cas univarié

Dans cette partie, nous appliquerons l'analyse en composantes principales fonctionnelle univariée de manière indépendante à la population féminine et masculine en Belgique. Une fois encore nous travaillons avec des données issues du site Human Mortality Database et pour les années allant de 1950 à 2018. La REQM sera encore une fois utilisée pour évaluer nos modèles. Dans un premier temps nous allons créer nos données fonctionnelles à l'aide des splines et des techniques de lissages et d'ajustement décrites dans la partie précédente. Pour rappel la mortalité dans le cas univarié de l'analyse en composantes principales est donnée par :

$$y_t(x_i) = \mu(x_i) + \sum_{n=1}^N \beta_{t,n} \phi_n(x_i)$$

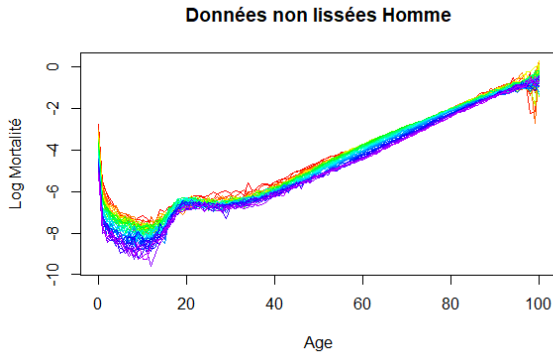


FIGURE 14: Données brutes des hommes

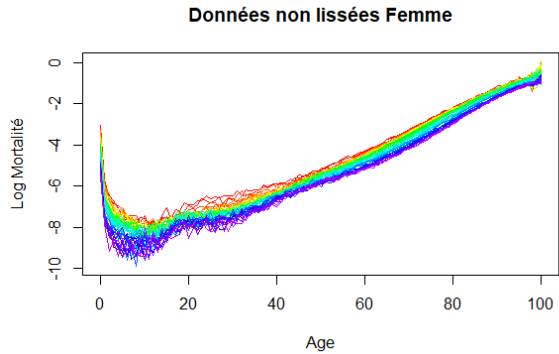


FIGURE 15: Données brutes des femmes

Nous remarquons que sur l'ensemble de nos données, il y a de nombreuses irrégularités que nous allons corriger grâce au lissage ce qui permet d'éliminer au maximum le bruit de nos données.

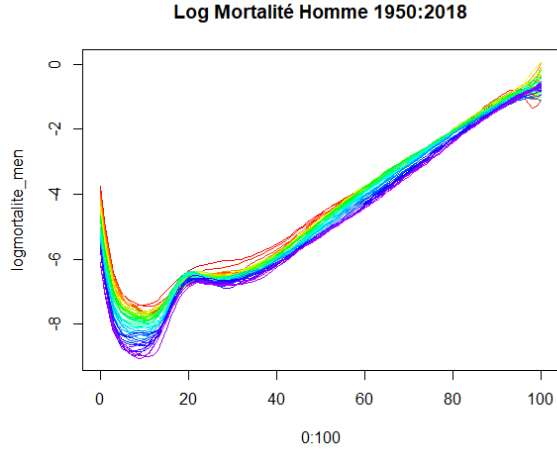


FIGURE 16: Données lissées des hommes

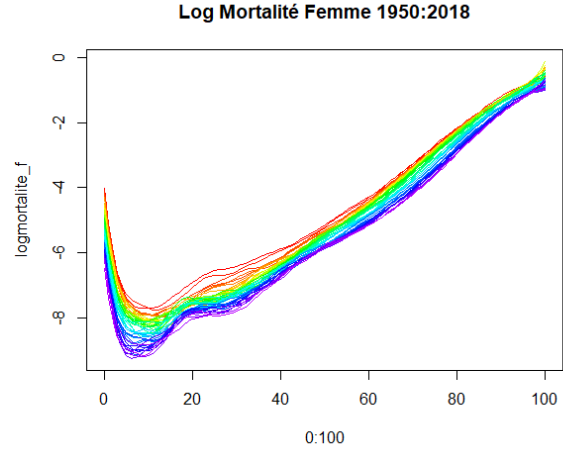


FIGURE 17: Données lissées des femmes

Sur ces deux images, nous observons des différences assez marquées notamment de 0 à 40 ans. Nous soulignerons ces différences dans le cas multivarié. Nous appliquons maintenant l'analyse en composantes principales fonctionnelle multivariée. Pour déterminer le nombre optimum de composantes principales à sélectionner pour notre modèle, nous nous basons sur le pourcentage cumulé de la variance expliquée par chaque composante. En notant N le nombre optimum de composantes à sélectionner, nous pouvons choisir N avec la règle suivante :

$$N = \operatorname{argmin} \left(\frac{\sum_{n=1}^N \hat{\lambda}_n}{\sum_{n=1}^{\infty} \hat{\lambda}_n} \geq P \right)$$

$\hat{\lambda}_n$ représente la valeur propre estimée de l'analyse en composantes principales. Généralement P est au minimum égal à 90%. Nous verrons que dans le cas de la mortalité $N=3$ est largement suffisant pour expliquer plus de 90% de nos données. Pour chaque cas, nous représenterons la moyenne, les trois premières fonctions propres ainsi que leur score associé.

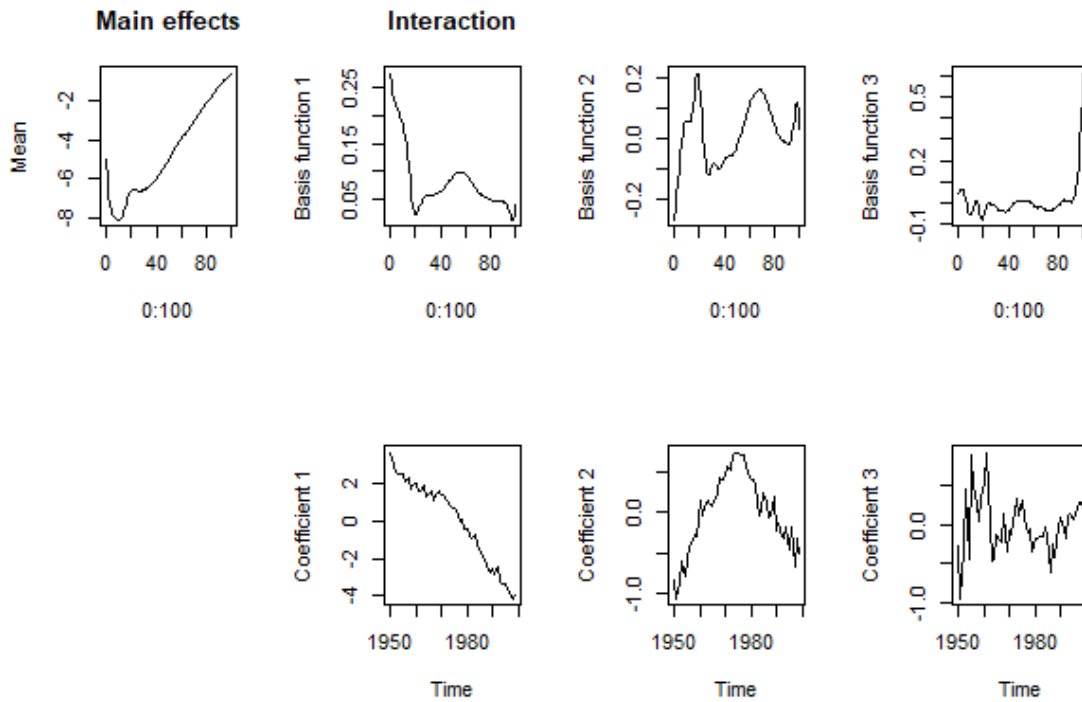


FIGURE 18: Ajustement FPCA sur les hommes Belges

Il peut être intéressant d'interpréter chaque composante principale. La composante 1 présente de grande variation entre 0 et 20 ans et explique principalement la mortalité infantile. La seconde composante connaît deux bosses à 20 ans et à 70 ans, cette composante peut représenter la variation de la mortalité à travers le temps pour ces âges donnés. La troisième composante peut être vue comme un bruit blanc et représente la grande variation de la mortalité aux grands âges. Les 3 premières composantes expliquent 95.25% de la variance des données avec pour la première composante 88.76%, 4% pour la deuxième et 2.6% pour la troisième composante.

Nous faisons le même exercice pour les femmes belges.

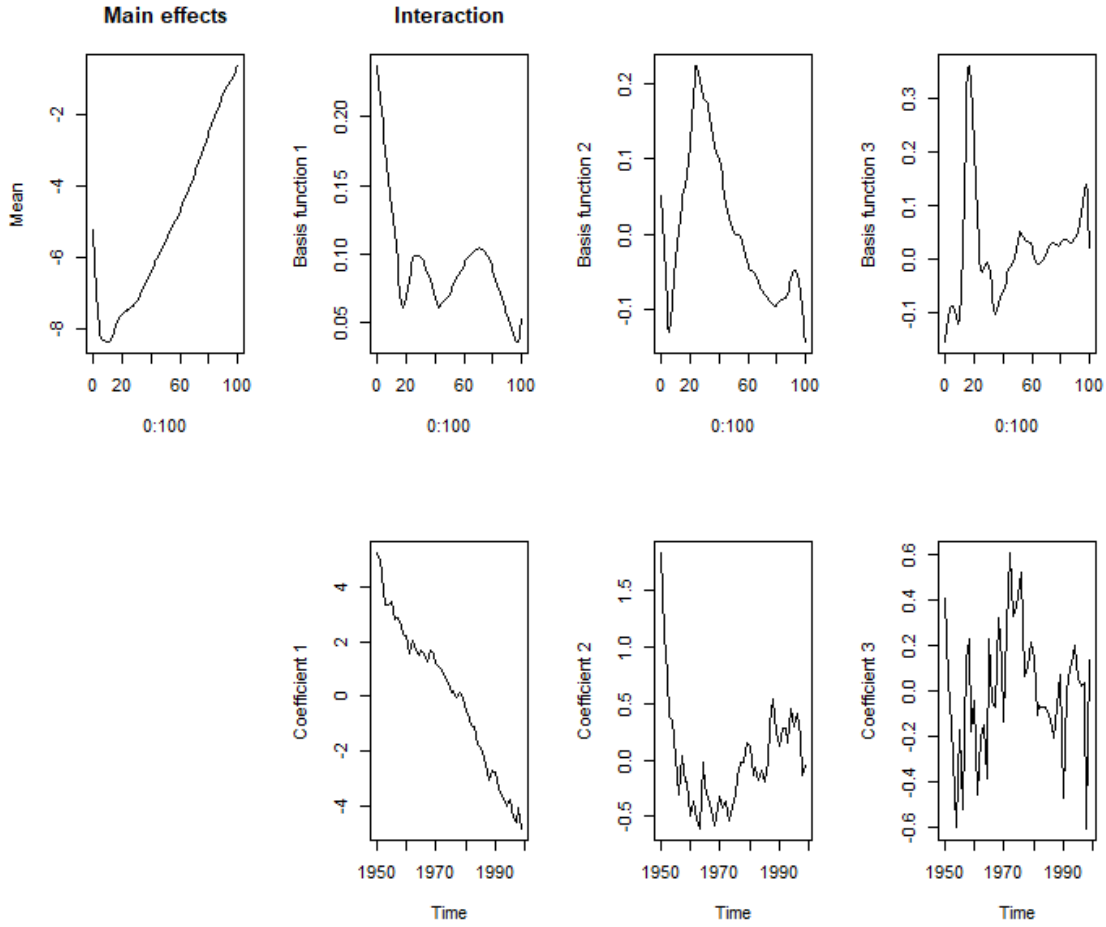


FIGURE 19: Ajustement FPCA sur les femmes Belges

L'interprétation des résultats est moins évidente ici. Il y a toujours une grande part de la mortalité infantile expliquée pour la première composante avec deux bosses à 30 ans et 70 ans. Les composantes deux et trois possèdent une bosse à 20 ans, elles peuvent montrer la variabilité de la mortalité à 20 ans même si la bosse de mortalité est beaucoup moins marquée chez les femmes que les hommes. La dernière composante peut représenter la variation de la mortalité autour de la fonction de moyenne. Les 3 premières composantes expliquent 95.6% de la variance des données avec pour la première composante 93.1%, 2.5% pour la deuxième et moins de 1% pour la troisième composante. La première composante est donc prépondérante.

Une fois que nous avons ajusté nos modèles sur nos données, nous allons devoir projeter les séries temporelles des scores pour obtenir les taux de mortalité future. Ici, nous voulons projeter les taux à un horizon $h = 19$. En notant l'ensemble des courbes étudiées $f(x) = [f_1(x), \dots, f_n(x)]^T$ et les composantes principales $B = [\hat{\phi}_1(x), \dots, \hat{\phi}_N(x)]$ la mortalité

l'année h après la dernière année connue est alors donnée par :

$$y_{t+h}(x_i) = \mathbb{E}[y_{t+h}(x_i)|f(x)|B] = \mu(x_i) + \sum_{n=1}^N \beta_{t+h,n} \phi_n(x_i)$$

Pour projeter nos séries temporelles à savoir les scores, nous utiliserons un modèle ARIMA et le package R "Forecast". Les projections des séries temporelles sont représentées ci-dessous.

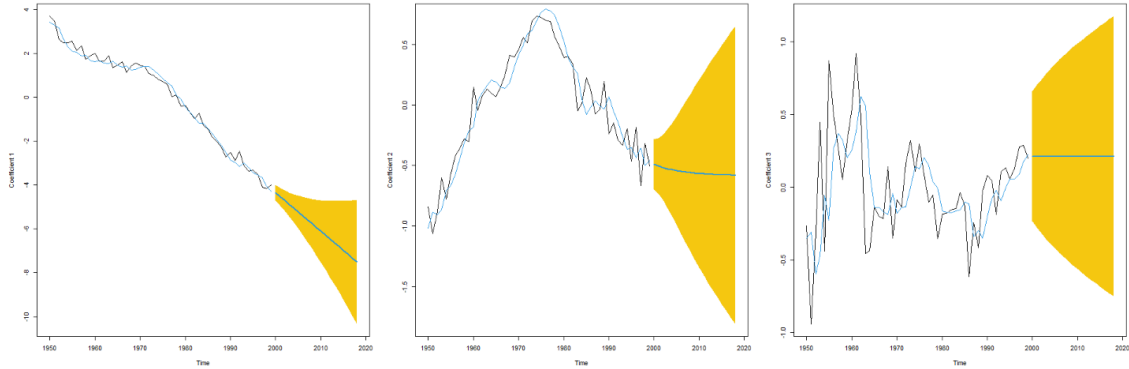


FIGURE 20: Les scores prévisionnels des composantes principales, et leurs intervalles de confiance à 80% (en jaune) - HOMME

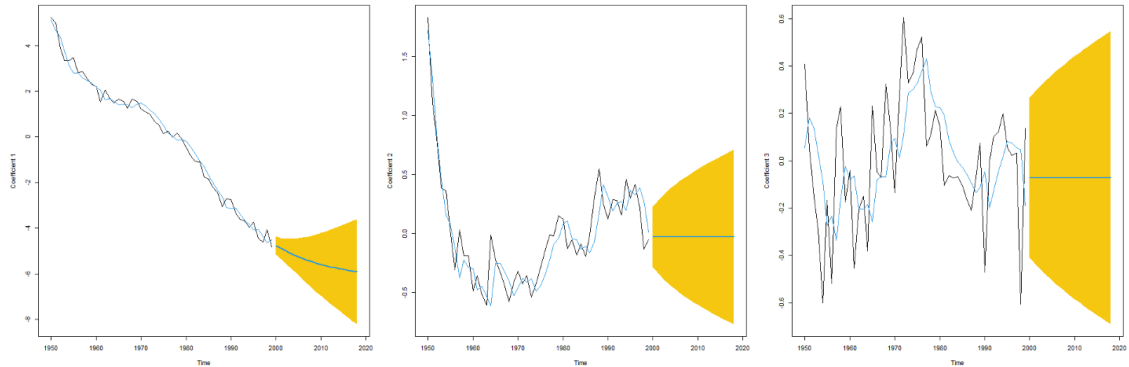


FIGURE 21: Les scores prévisionnels des composantes principales, et leurs intervalles de confiance à 80% (en jaune) - FEMME

Les projections des scores des composantes principales apparaissent se stabiliser rapidement et les intervalles de prédiction s'élargissent très rapidement. Cela peut refléter la difficulté de notre modèle de prévision à moyen ou long terme, en raison de l'augmentation de la variabilité. Nous allons maintenant calculer comme dans la première partie la REQM de nos prévisions de mortalité pour les années de 2000 à 2018 pour évaluer nos modèles. Noter que nous avons aussi calculé l'erreur pour le jeu de données reprenant

l'ensemble de la population belge pour comparer à la partie précédente. Les composantes principales de ce modèle figurent en annexes.

Age	Données		
	Homme	Femme	Total
0	0,4569	0,6317	0,6584
1	0,5334	0,5311	0,5495
2	0,2544	0,4657	0,3829
5	0,3259	0,4485	0,3494
10	0,3362	0,5402	0,4232
12	0,2933	0,4506	0,3245
15	0,2657	0,5804	0,3828
18	0,3543	0,3014	0,3398
20	0,3541	0,4550	0,3663
22	0,3356	0,3228	0,3091
25	0,3191	0,2423	0,2571
28	0,1958	0,1996	0,1681
30	0,2059	0,2228	0,2162
40	0,1150	0,1085	0,1402
50	0,0438	0,0373	0,0618
60	0,0438	0,0373	0,0618
70	0,1997	0,1113	0,1796
80	0,2294	0,1853	0,1798
90	0,0741	0,0830	0,0642
100	0,0988	0,0553	0,0395
Moyenne	0,2518	0,30	0,27

TABLE 4: Calcul de la REQM pour tous les âges

Plusieurs choses sont à remarquer. Les meilleurs résultats obtenus sont pour les données masculines, le FPCA prédit moins bien les données féminines même si les résultats sont bons comme nous pouvons le voir en traçant la courbe de mortalité pour tous les âges en 2018 (voir ci-dessous). Le FPCA sur les données totales prédit moins bien qu'avec les modèles de régression gaussienne mais il surpasse néanmoins le modèle de Lee-Carter. A noter que le calcul de la REQM sur l'ensemble des âges considérés (0 à 100 ans) nous donnent 0,20689, 0,2588 et 0,2433 pour les hommes, les femmes et la population totale. Encore une fois et comme attendu les modèles ont plus de mal à prédire avec certitude les taux de mortalité pour les âges avec les plus fortes variations. Les résultats sont tout de même bon et surpassent le modèle de Lee-Carter.

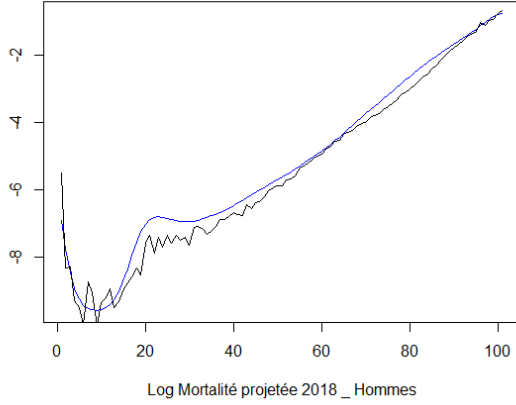


FIGURE 22: Taux de mortalité pour les hommes en 2018

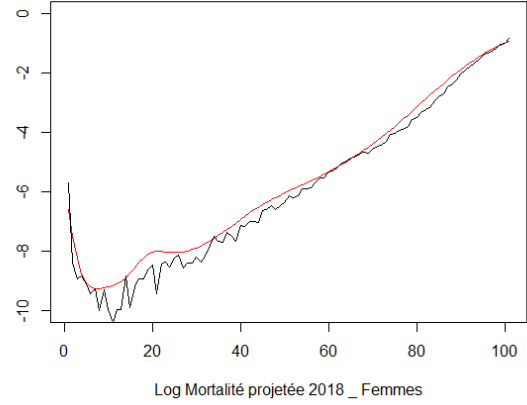


FIGURE 23: Taux de Mortalité pour les femmes en 2018

Sur les différents graphiques en commençant par celui des hommes, nous observons que la courbe réelle est bien répliquée avec une surestimation pour la bosse accident et une légère surestimation pour les grands âges. Pour les femmes le modèle prédit bien mais prédit également une surestimation au moment de l'adolescence de 10 à 20 ans où le modèle a du mal à répliquer les tendances réelles. La même courbe sur les données totales est en annexes. Dans cette partie, nous avons donc vu l'intérêt de l'analyse en composantes principales fonctionnelle univariée, qui apporte de l'information à travers l'analyse de ces composantes et les modes de variations en fonction de l'âge.

4.2.2 Cas multivarié

Dans cette partie, nous appliquerons l'analyse en composantes principales fonctionnelle multivariée, ainsi nous allons modéliser simultanément la population féminine et masculine en Belgique. Une fois encore nous travaillons avec des données issues du site Human Mortality Database et pour les années allant de 1950 à 2018. La REQM sera encore une fois utilisée pour évaluer nos modèles. De la même façon que dans le cas univarié nous allons créer nos données fonctionnelles à l'aide des splines et des techniques de lissages et d'ajustement décrites dans la partie précédente. Pour rappel la mortalité dans le cas multivarié de l'analyse en composantes principales est donnée par :

$$y_t^{(i)}(x_i) = \mu^{(i)}(x_i) + \sum_{n=1}^N \beta_{t,n} \phi_n^i(x_i)$$

La cas multivarié nous permet d'introduire la modélisation de plusieurs sous-groupe de population de manière simultanée. En modélisant la mortalité pour deux ou plusieurs

sous-populations d'une population plus importante simultanément, il est généralement souhaitable que les prévisions soient non divergentes. En effet, dans des conditions socio-économiques très proches il est normal de voir la population d'un même pays suivre une tendance de mortalité similaire. Ici, nous nous intéressons à deux sous-populations, à savoir les hommes et les femmes d'un même pays. Cette analyse peut être élargie à un ensemble de pays, notamment à plusieurs pays de la même zone géographique et avec un nombre de sous-populations supérieur à deux. Il est également intéressant d'utiliser ce modèle pour montrer aussi les divergences de mortalité entre les sous-groupes et d'analyser d'où proviennent ces variations. De telle sorte que notre modèle intègre à la fois à la fois des informations globales de la population et des informations spécifiques des sous-populations. Pour mieux visualiser ces variations nous avons représenté la moyenne des taux de mortalité globale pour les hommes et les femmes ainsi que la déviation de la mortalité par rapport à la moyenne pour les deux sous-groupes considérés.

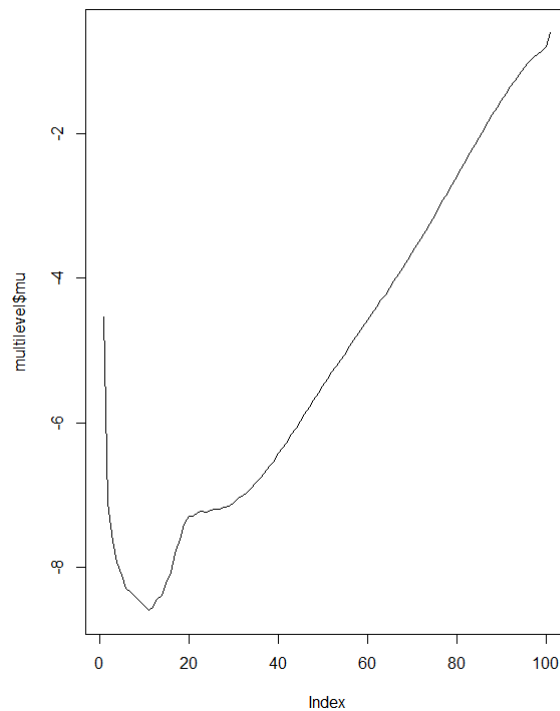


FIGURE 24: Moyenne globale des taux de mortalité

Nous remarquons que logiquement la moyenne suit le même profil que les graphiques 16 et 17. Nous observons toujours la bosse de mortalité mais de façon plus contrastée car les données féminines réduisent son intensité. Il est également possible de représenter l'écartement à la moyenne de chaque sous groupe.

Ces deux graphiques nous apportent beaucoup d'informations quand aux divergences

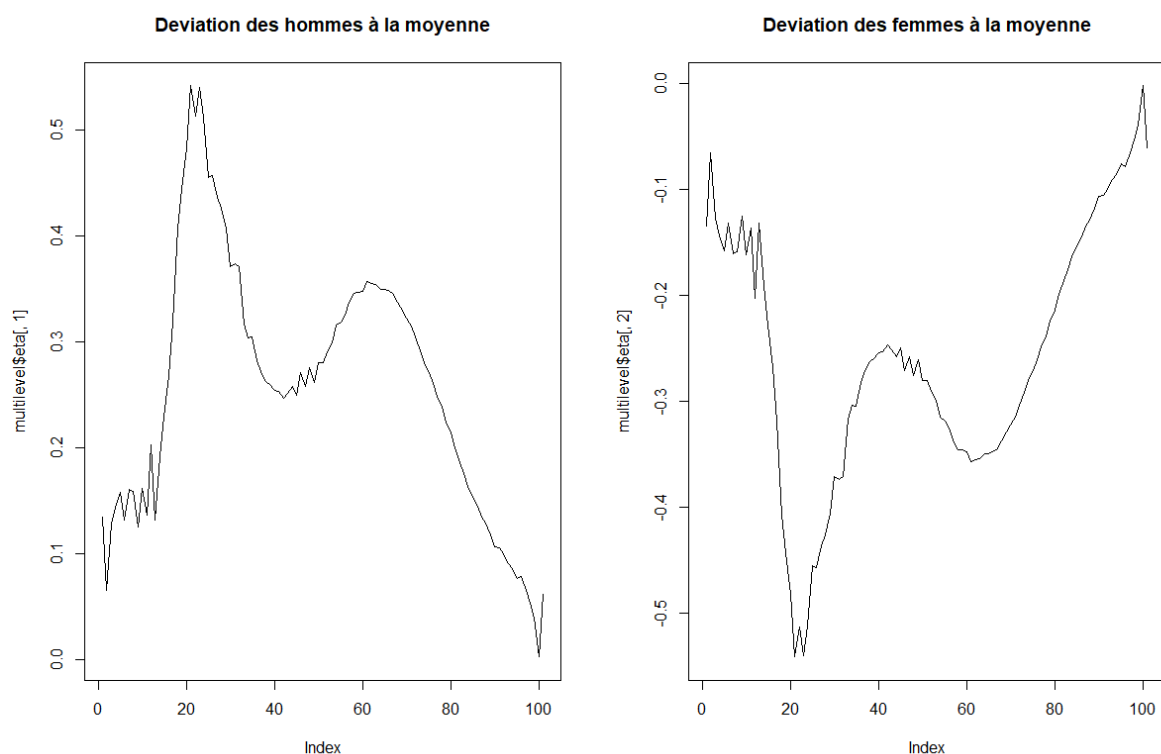


FIGURE 25: Taux de Mortalité Déviation à la moyenne - HOMME

FIGURE 26: Taux de Mortalité Déviation à la moyenne -FEMME

de tendances des deux grandeurs étudiées. Tout d'abord, nous observons que les hommes ont une mortalité qui poussent la moyenne à la hausse car la déviation est toujours positive quelques soient les âges considérés. La tendance inverse est bien sûr observée chez les femmes. Le plus grand écart est observé à 20 ans où la mortalité des hommes est bien plus élevée que celle des femmes. Nous remarquons donc que la bosse de surmortalité est lié au sexe masculin. C'est un phénomène expliqué par les sociologues et démographes par une prise de risque excessive, une impulsivité, et une moindre capacité à considérer les conséquences de ses actes à cette âge là. Ensuite les écarts sont moins marqués avec une légère déviation plus importante aux alentours des 60 ans puis une déviation qui diminue de manière linéaire. Il est intéressant de noter que dans les contrats d'assurance vie la tarification technique se fait en fonction du sexe et ce qui est peut être justifié par ces deux courbes même si l'écart diminue aux grands âges. Nous allons maintenant grâce à la méthode décrite infra calculer nos composantes principales et leurs scores associés. Nous verrons que le choix de 3 composantes principales est largement suffisant pour expliquer nos données. Sur le graphique 27 la première composante principale est représentée en rouge, la seconde en vert et la dernière en bleu.

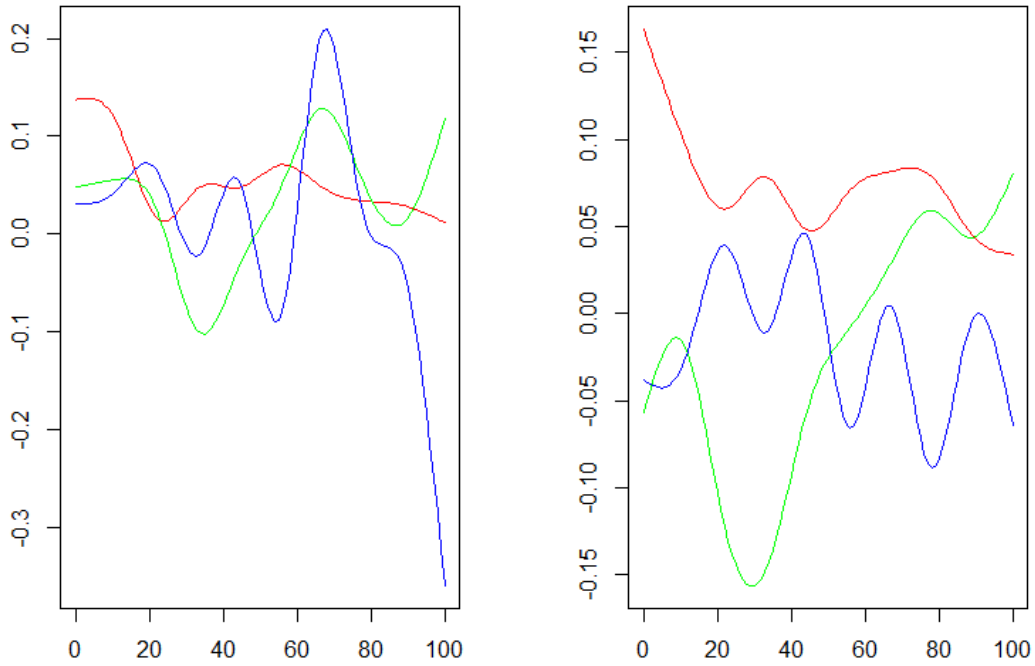


FIGURE 27: Les composantes principales de données fonctionnelles avec à gauche les hommes et à droite les femmes

Nous nous sommes intéressés aux trois premières composantes principales car elles expliquent la totalité de la variation de nos données avec 97% pour la première composante, 2% pour la seconde et le reste pour la dernière.

Comme dans le cas univarié, une fois que nous avons ajusté nos modèles sur nos données, nous allons devoir projeter les séries temporelles des scores pour obtenir les taux de mortalité future. Ici, nous voulons projeter les taux à un horizon $h = 19$. En notant l'ensemble des courbes étudiées $f(x) = [f_1(x), \dots, f_n(x)]^T$ et les composantes principales $B = [\hat{\phi}_1(x), \dots, \hat{\phi}_N(x)]$ la mortalité l'année h après la dernière année connue est alors donnée par :

$$y_{t+h}^{(i)}(x_i) = \mathbb{E}[y_{t+h}(x_i)|f(x)|B] = \mu(x_i) + \sum_{n=1}^N \beta_{t+h,n} \phi_n^{(i)}(x_i)$$

Contrairement au cas univarié les scores projetés pour les hommes et les femmes sont les mêmes.

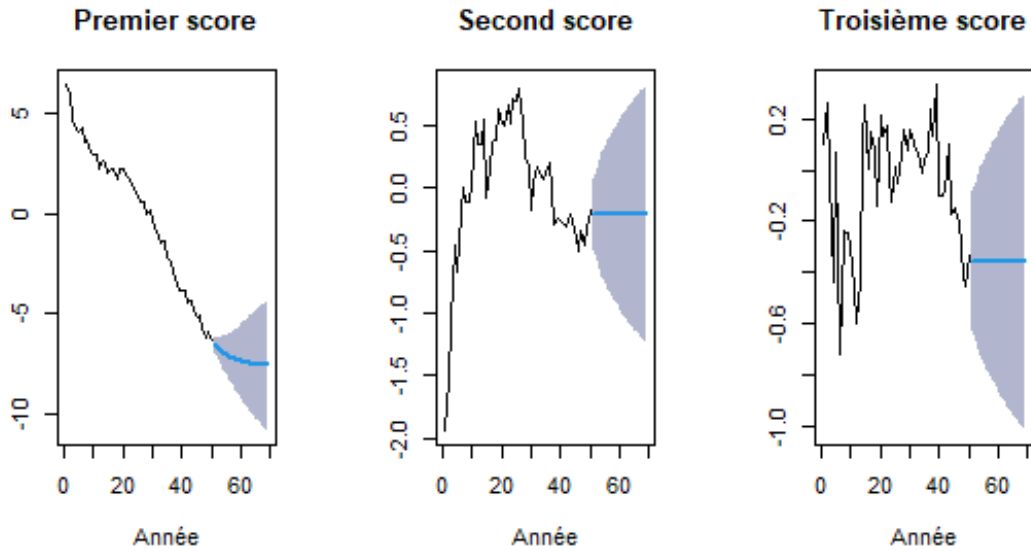


FIGURE 28: Les scores associés aux différentes composantes principales et leur projection

Les intervalles de confiance sont représentés sur le graphique par les bandes grises il s'agit de l'intervalle de confiance à 80%. Comme dans le cas univarié, les projections des scores des composantes principales semblent se stabiliser rapidement et les intervalles de prédiction s'élargissent très rapidement. Il est possible de voir à travers les différents scores et leurs composantes principales associés que le score 1 peut représenter une amélioration à travers le temps de la durée de vie et le troisième score comme un bruit blanc qui tend vers zéro.

Nous allons maintenant calculer comme dans tous les modèles étudiés la REQM de nos prévisions de mortalité pour les années de 2000 à 2018 pour évaluer nos modèles. Bien sûr dans le cas multivarié, il n'est pas possible de comparer notre modèle aux données globales (hommes et femmes simultanément).

Age	Données	
	Homme	Femme
0	0,4569	0,4067
1	0,4300	0,2895
2	0,4630	0,5761
5	0,6682	0,7140
10	0,3727	0,4371
12	0,3769	0,4845
15	0,2835	0,4284
18	0,4317	0,4973
20	0,2742	0,2617
22	0,0969	0,2753
25	0,2611	0,3441
28	0,2511	0,3089
30	0,1347	0,2082
40	0,1459	0,1291
50	0,1538	0,1116
60	0,1560	0,1224
70	0,1368	0,1013
80	0,1023	0,0946
90	0,0859	0,0687
100	0,1676	0,1032
Moyenne	0,272	0,2981

TABLE 5: Calcul de la REQM pour tous les âges

Plusieurs choses sont à remarquer. Les meilleurs résultats obtenus sont une nouvelle fois pour les données masculines, le MFPCA prédit moins bien les données féminines. Le modèle pour les femmes est moins bon que celui de Lee-Carter et quasi équivalent pour les hommes. A noter que le calcul de la REQM sur l'ensemble des âges considérés (0 à 100 ans) nous donne 0,2055 et 0,218 pour les hommes et femmes. Nous avons également représenté les courbes de mortalité pour les femmes et les hommes pour l'année 2018 qui est notre dernière année de projection.

Ces deux courbes montrent les défauts de notre modèle qui a tendance à surestimer les taux de mortalité à différents âges et particulièrement aux âges les plus jeunes. Néanmoins l'analyse fonctionnelle en composantes principales multivariée permet d'analyser les similitudes et les différences entre deux sous-groupes de population. Il peut être intéressant de travailler sur les composantes principales et leur scores pour comprendre l'évolution

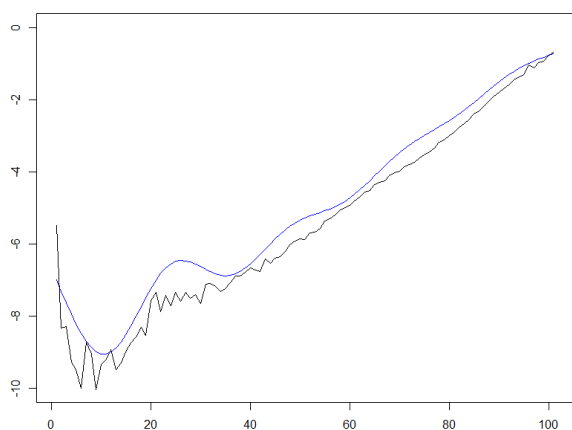


FIGURE 29: Taux de mortalité pour les hommes en 2018

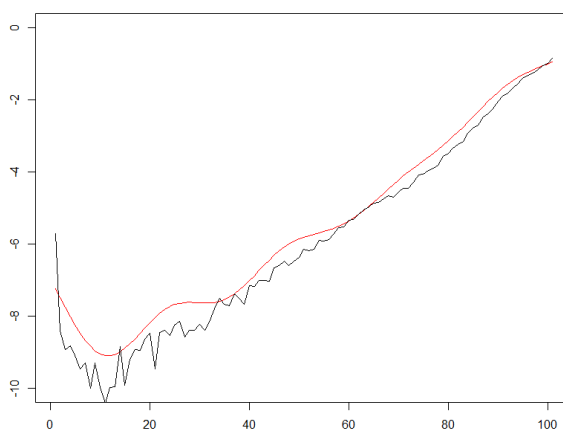


FIGURE 30: Taux de Mortalité pour les femmes en 2018

de la mortalité. Dans un environnement où les différences de mortalité ont tendance à diminuer entre les hommes et les femmes l'analyse de plusieurs sous-groupes de manière simultanée permet d'affirmer ou non cette information utile dans la tarification de certains contrats vie. Même si le modèle a tendance à surestimer la mortalité il pourrait être utilisé en tarification d'assurance décès et permettre de libérer une marge technique pour l'assureur.

5 Statistiques démographiques et application financière

Dans cette dernière partie, nous nous intéresserons dans un premier temps à calculer les espérances de vie à la naissance et résiduelles en fonction de nos différents modèles. L'idée est de se servir des mêmes jeux de données que lors de l'ajustement de nos modèles. Mais cette fois-ci nous utiliserons l'ensemble de nos données, de 1950 à 2018 comme données d'entraînement. Les modèles ayant été validés dans les parties précédentes, cela nous permettra de construire des tables de mortalité prospectives et calculer des espérances de vie future. Dans un second temps nous calculerons le prix d'une rente viagère pour un contrat donné.

5.1 Calcul des espérances de vie

Dans cette partie nous allons calculer l'espérance de vie résiduelle aux âges suivants : 0, 20, 40 et 60. Pour rappel l'espérance de vie résiduelle pour un individu d'âge x est la durée restant à vivre en moyenne. Mathématiquement si on fait l'hypothèse que μ_{x+t} est une constante sur l'intervalle $[t, t+1[$ alors l'espérance de vie à l'âge x s'écrit :

$$\overset{\circ}{e}_x = \sum_{t=1}^{w-x} {}_t p_x$$

Avec :

- w : l'âge maximum de notre jeu de données, ici 100 ans
- ${}_t p_x$: la probabilité pour un individu d'âge x de survivre jusqu'à $x+t$

Pour calculer l'espérance de vie résiduelle il nous faut donc calculer les ${}_t p_x$. Dans nos modèles nous avons estimé les q_x ou μ_x . Nous utilisons la relation suivante pour trouver les ${}_t p_x$: ${}_t p_x = 1 - {}_t q_x$. Nos modèles nous donnent pour chaque âge les q_x qui est la probabilité de décès dans l'année pour un individu d'âge x . Il nous faut donc calculer les probabilités de survie pour un individu d'âge x à horizon 1 an, 2ans, 3 ans,..., $w-x$ ans. Alors ensuite nous pourrons utiliser notre formule de l'espérance de vie résiduelle. Nous rappelons la formule qui nous permet de calculer les ${}_t p_x$:

$${}_t p_x = p_x \times p_{x+1} \times \dots \times p_{x+t-1}$$

Dans le cas où nous avons modélisé la force de mortalité $\mu_x(t)$, avec l'hypothèse que le mortalité est constante sur l'intervalle $[t, t+1[$, nous avons alors les formules suivantes :

$$p_x = \exp(-\mu_x)$$

$${}_tp_x = \prod_{y=x}^{x+t-1} \exp(-\mu_y)$$

Pour chaque modèle nous avons simulé 1000 trajectoires de mortalité pour calculer ensuite l'espérance de vie moyenne, l'écart type et les quantiles à 5% et 95%. Les données utilisées sont exactement les mêmes que celles décrites dans les applications de chaque modèle avec ici les données projetées de 2018 à 2098 donc à 80 ans. Cette projection nous permet de construire des tables de mortalité prospectives et de calculer les espérances de vie à différentes dates. Nous avons décidé de calculer les espérances de vie en 2019, 2040 et 2060. Les statistiques de l'espérance de vie sont représentées dans les tableaux ci-dessous pour le modèle à mélange spectral, le cas univarié de l'analyse en composantes principales fonctionnelle pour les hommes et le cas multivarié pour les femmes. Tous les autres modèles sont en annexes.

2019	Mélange spectral			
	0 an	20 ans	40 ans	60 ans
Moyenne	78,953	59,326	39,45	21,1437
Ecart-type	0,042	0,132	0,42	0,51
Quantile à 5%	78,884	59,109	38,76	20,53
Quantile à 95%	79,022	59,54	40,14	22,01

TABLE 6: Statistiques de l'espérance de vie pour la population belge totale en 2019

2040	Mélange spectral			
	0 an	20 ans	40 ans	60 ans
Moyenne	81,896	61,927	42,19	23,22
Ecart-type	0,122	0,329	0,39	0,65
Quantile à 5%	81,760	61,387	41,52534	22,04
Quantile à 95%	82,193	62,468	42,70	24,11

TABLE 7: Statistiques de l'espérance de vie pour la population belge totale en 2040

2060	Mélange spectral			
	0 an	20 ans	40 ans	60 ans
Moyenne	76,954	63,672	43,10	24,09
Ecart-type	3,149	3,456	3,36	3,99
Quantile à 5%	72,602	57,988	37,58	17,53
Quantile à 95%	83,349	69,356	48,62	30,66

TABLE 8: Statistiques de l'espérance de vie pour la population belge totale en 2060

Nous remarquons que le modèle prend en compte les tendances de la mortalité avec une augmentation de l'allongement de la durée de vie moyenne de plusieurs années de 2019 à 2060. Nous remarquons également que l'écart-type augmente quand la projection est plus lointaine. Ceci est logique car l'incertitude augmente avec l'horizon de nos projections.

2019	FPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	77,501	57,819	38,517	20,551
Ecart-type	0,111	0,120	0,119	0,127
Quantile à 5%	77,319	57,621	38,323	20,346
Quantile à 95%	77,684	58,016	38,720	20,765

TABLE 9: Statistiques de l'espérance de vie pour la population masculine en 2019

2040	FPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	80,53	60,72	41,25	22,83
Ecart-type	0,60	0,60	0,21	0,13
Quantile à 5%	79,55	59,73	40,91	22,62
Quantile à 95%	81,51	61,70	41,59	23,05

TABLE 10: Statistiques de l'espérance de vie pour la population masculine en 2040

2060	FPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	82,91	63,02	43,44	24,69
Ecart-type	3,13	3,34	3,20	3,95
Quantile à 5%	77,76	57,53	38,18	18,20
Quantile à 95%	88,06	68,51	48,71	31,18

TABLE 11: Statistiques de l'espérance de vie pour la population masculine en 2060

L'analyse en composantes principales fonctionnelle suit aussi l'amélioration de la mortalité. Nous pouvons voir qu'un homme né en 2060 peut espérer vivre jusqu'à 83 ans contre 77,5 ans aujourd'hui. Nous voyons aussi par l'analyse de ces espérances que les hommes tirent l'espérance de vie moyenne vers le bas et les femmes vers le haut. Comme le montre le même modèle sur les données féminines en annexes, l'écart entre les hommes et les femmes tend à se réduire dans les années futures.

2019	MFPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	82,13	62,31	42,65	24,12
Ecart-type	0,08	0,08	0,07	0,078
Quantile à 5%	82,01	62,18	42,52	24,00
Quantile à 95%	82,26	62,44	42,79	24,25

TABLE 12: Statistiques de l'espérance de vie pour la population féminine en 2019

2040	MFPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	84,86	64,96	45,21	26,40
Ecart-type	0,90	1,00	0,88	0,79
Quantile à 5%	83,39	63,32	43,77	25,11
Quantile à 95%	86,34	66,61	46,65	27,70

TABLE 13: Statistiques de l'espérance de vie pour la population féminine en 2040

2060	MFPCA			
	0 an	20 ans	40 ans	60 ans
Moyenne	87,22	67,26	47,44	28,40
Ecart-type	1,78	2,21	1,69	1,59
Quantile à 5%	84,29	63,62	44,67	25,79
Quantile à 95%	90,14	70,90	50,22	31,02

TABLE 14: Statistiques de l'espérance de vie pour la population féminine en 2060

Les espérances de vie calculées avec l'analyse en composantes principales fonctionnelle multivariée nous apportent plusieurs conclusions. L'espérance de vie chez les femmes est en augmentation au cours des années. De manière générale, tous les modèles prédisent des espérances de vie assez similaires. Dans chaque modèle, la variabilité des résultats augmente avec l'horizon de nos projections et il est donc plus dur de déterminer de manière certaine les espérances de vie futures. Certains modèles ont une plus forte variabilité que d'autres selon les années et les âges considérés. Le risque de longévité qui est le risque de voir la durée de vie humaine s'allonger de manière non déterministe est très bien retranscrit à travers le calcul des espérances de vie. Nous voyons bien que pour l'assureur il est nécessaire d'avoir de bons modèles de prédiction de la mortalité et que les tables de mortalité officielles ne sont pas suffisantes pour prendre en compte une évolution de la mortalité future.

5.2 Prix d'une rente viagère

Dans cette partie, nous donnerons une application financière de nos différents modèles. Pour se faire, nous allons tarifier une rente viagère payée annuellement à la fin de l'année pour un individu qui a 60 ans en 2040. Nous calculerons le prix de cette rente avec le noyau à mélange spectral et l'analyse en composante principale fonctionnelle univariée pour un homme. Pour rappel une rente viagère est une somme d'argent fixée à l'avance et versée périodiquement (mensuellement, trimestriellement, annuellement) à un bénéficiaire jusqu'à son décès. Les caractéristiques de la rente sont les suivantes :

- Le taux d'intérêt est fixé à $i=2\%$ de manière simplifiée pour toute la durée du contrat
- Le preneur d'assurance est un homme âgé de 60 ans en 2040.
- La somme payée chaque fin d'année est de 10000€.
- Les deux modèles utilisés seront l'analyse en composantes principales fonctionnelle univariée et la régression gaussienne avec le noyau à mélange spectral

La prime unique pure d'une rente viagère notée $\ddot{a}_x$ est égale à :

$$\ddot{a}_x = \sum_{k=0}^{\infty} {}_k p_x v^k$$

Nous produisons 1000 simulations pour obtenir une distribution du prix de la rente viagère dans les différents modèles. Pour les deux modèles étudiés nous avons calculé la moyenne, la variance, le quantile à 5% et le quantile à 95%.

2040	Modèle	
	Noyau à mélange spectral	FPCA
Moyenne	173305.8€	176493.8€
Variance	83646.75€	505813.4€
Quantile à 5%	172863.8€	175307.4€
Quantile à 95%	173795.3€	177648.7€

TABLE 15: Statistiques sur le prix d'une rente viagère

Le calcul de ces différents indicateurs statistiques révèle que le prix dans le modèle utilisant la régression gaussienne est en moyenne inférieur comme la variance. Le prix dans modèle FPCA est plus dispersé. Nous avons représenté ci-dessous, la densité du prix dans chacun des modèles. Les deux graphiques confirment que le prix est plus dispersé avec le modèle FPCA et en moyenne plus élevé. Sur le graphique 31, nous pouvons aussi observer que la queue à droite de distribution est plus épaisse que la queue de distribution à gauche, ce

qui nous indique que les valeurs extrêmes maximales sont plus probables que les valeurs extrêmes minimales.

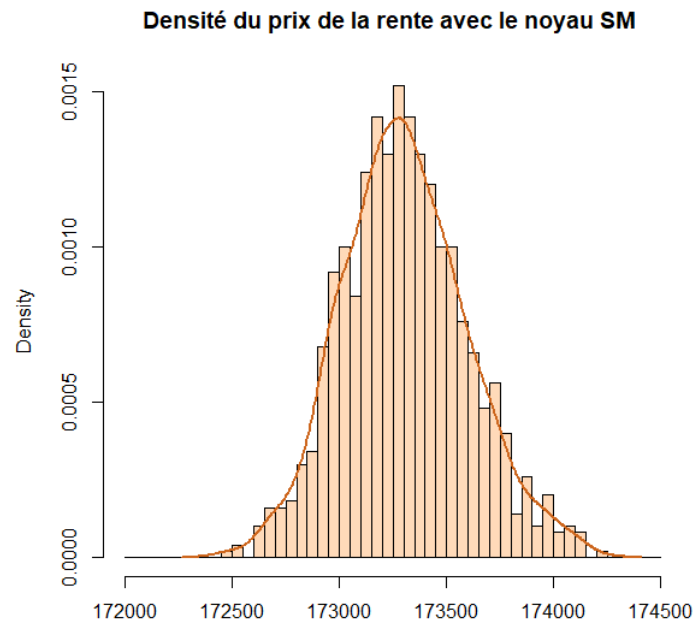


FIGURE 31: Densité du prix d'une rente viagère avec le noyau à mélange spectral

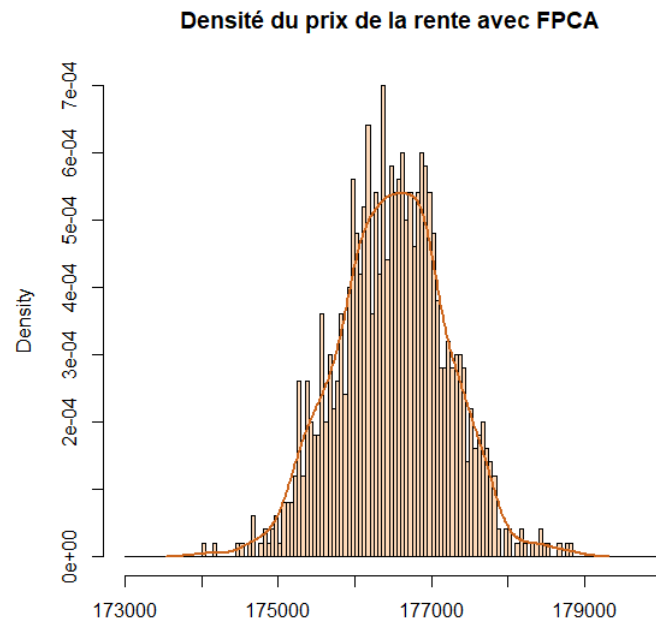


FIGURE 32: Densité du prix d'une rente viagère avec l'analyse en composantes principales fonctionnelle univariée

Cette application numérique nous montre que le prix d'un modèle à un autre peut différer et avoir des conséquences importantes sur la tarification d'un portefeuille d'assurances. Cette application numérique souligne également le rôle de l'actuaire qui est de quantifier des risques initialement non financiers (longévité ou mortalité) en risques financiers. Les modèles employés pour construire des tables de mortalité future sont très utiles dans la mesure où ils nous permettent de réaliser de nombreuses simulations et de calculer la distribution du prix.

6 Conclusion

Ce mémoire nous a permis d'introduire de nombreux modèles de mortalité. Dans la première partie nous avons introduit la régression à l'aide de processus gaussien de manière générale puis ensuite dans un but de modélisation et de projection de la mortalité. Nous avons étudié différentes fonctions de covariance dites classiques et celle à mélange spectral. Nous avons obtenu de très bons résultats pour l'ensemble des noyaux. Les résultats en fonction du noyau et des âges peut varier en fonction de l'horizon de projection et des âges d'études. Dans ce sens, le choix du noyau peut parfois être difficile. Le noyau à mélange spectral présente une stabilité de résultats intéressante et une bonne capacité à prédire à long terme. L'introduction d'une fonction de moyenne non nulle dans l'ensemble de nos modèles de régression gaussiens améliore sensiblement la qualité de prédiction mesurée à l'aide de la REQM. Contrairement à la plupart des méthodes existantes de modélisation de la mortalité qui traitent les taux de mortalité pour tous les âges dans une année dans son ensemble et étudient leur évolution dans le temps, la régression à l'aide processus gaussiens traite les taux de mortalité pour tous les âges comme des séries temporelles. Nous avons également comparé nos résultats avec la variante poisson du modèle de Lee-Carter et nous avons vu que l'ensemble de nos noyaux produisait de meilleurs résultats.

Dans un second temps, nous avons introduit l'analyse en composantes principales fonctionnelle. Ceci, nous a permis d'expliquer l'utilité des données fonctionnelles et comment les construire à partir d'observations à des points discrets. Nous avons ensuite présenté les mathématiques nécessaires pour l'analyse en composantes principales fonctionnelle univariée et multivariée. L'analyse en composantes principales fonctionnelle (FPCA) est une méthode statistique permettant d'étudier les modes dominants de variation des données fonctionnelles. FPCA représente les données fonctionnelles de la manière la plus parcimonieuse possible, avec l'utilisation d'un nombre fixe de fonctions de base. La représentation de Karhunen-Loeve est le coeur de nos modèles. Nous avons par la suite appliqué ces modèles à nos données de mortalité et obtenu de bons résultats surpassant même le modèle de Lee-Carter. La projection des taux de mortalité future s'étant faite par une modélisa-

tion des scores en séries temporelles. L'analyse en composantes principales fonctionnelle univariée et multivariée nous apporte de l'information à travers l'analyse de ces composantes et les modes de variations en fonction de l'âge. Le cas multivarié a aussi permis l'introduction de la modélisation de plusieurs groupes de manière simultanée. La modélisation par des scores en commun pour les hommes et les femmes permet d'obtenir des prévisions non divergentes comme nous pouvons nous y attendre pour des populations soumis aux mêmes contraintes de mortalité.

Dans la dernière partie de ce mémoire, nous avons calculé des statistiques sur l'espérance de vie. Tous les modèles prévoient une augmentation de l'espérance de vie moyenne. Les personnes les plus jeunes sont celles qui peuvent espérer profiter le plus de l'amélioration des conditions de mortalité future. Cette augmentation de l'espérance de vie en moyenne souligne le risque de longévité auquel font face les assureurs. De plus, l'incertitude concernant l'espérance de vie future augmente considérablement avec l'horizon de projection. Néanmoins, une projection à long terme est nécessaire pour pouvoir tarifier des contrats d'assurance vie. Nous avons pu voir que pour un contrat de rente viagère les prix avec deux modèles différents obtenus par simulations peuvent avoir une distribution différente et des conséquences importantes pour la tarification d'un portefeuille d'assurance.

De manière générale, ce mémoire a permis de confronter différents modèles qui doivent être utilisés de manière complémentaire car ils possèdent tous des avantages dans l'analyse de la mortalité et sa projection future. De plus les modèles ne performant pas de manière identique en fonction de nos données d'entrée. Ce mémoire permet également d'introduire des méthodes statistiques qui peuvent être utilisées dans d'autres buts que la modélisation de la mortalité, notamment l'analyse en composantes principales fonctionnelle qui peut être utilisée pour la modélisation des taux d'intérêts.

Enfin ce mémoire nous a permis de faire le lien entre modèles statistiques, démographiques, implémentation de technique de machine learning, quantification du risque et tarification financière. Ce sont ces mêmes liens qui constituent le travail de l'actuaire.

7 Bibliographie

- 1 Ruhao Wu, Gaussian Process and Functional Data Methods for Mortality Modeling, 2016
- 2 Ronald D. Lee and Lawrence R. Carter, Modeling and Forecasting U. S. Mortality, 1992
- 3 Renshaw, A.E., Haberman, a Lee–Carter mortality forecasting : a parallel generalised linear modelling approach for England and Wales mortality projections. Applied Statistics 52, 119–13, 2003
- 4 Hyndman, R.J. Ullah, M.S. . Robust forecasting of mortality and fertility rates : A Functional data approach, Computational Statistics Data Analysis 51 : 4942–4956, (2007)
- 5 C. E. Rasmussen C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006
- 6 Jäkel, F., Schölkopf, B., Wichmann, F. A. A tutorial on kernel methods for categorization. Journal of Mathematical Psychology, 51(6), 343–358, 2007
- 7 David Duvenaud, The Kernel Cookbook : Advice on Covariance functions
- 8 J.O. Ramsay, Giles Hooker,,Spencer Graves, Functional Data Analysis with R and MATLAB, 2009
- 9 Clara Happ, Sonja, Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains, 2018
- 10 M. Ebden, Gaussian Process A Quick Introduction
- 11 Mike Ludkovski, Jimmy Risk, Howard Zail, Gaussian Process Models for Mortality Rates and Improvement Factors 2018 Factors
- 12 Clara Happ-Kurz, Object-Oriented Software for Functional Data, Journal of Statistical Software, Volume 93, Issue 5, April 2020
- 13 Han Lin Shang, Mortality and life expectancy forecasting for a group of populations in developed countries : a multilevel functional data method, 2016
- 14 Arthur Charpentier, Computational Actuarial Science with R

8 Annexes

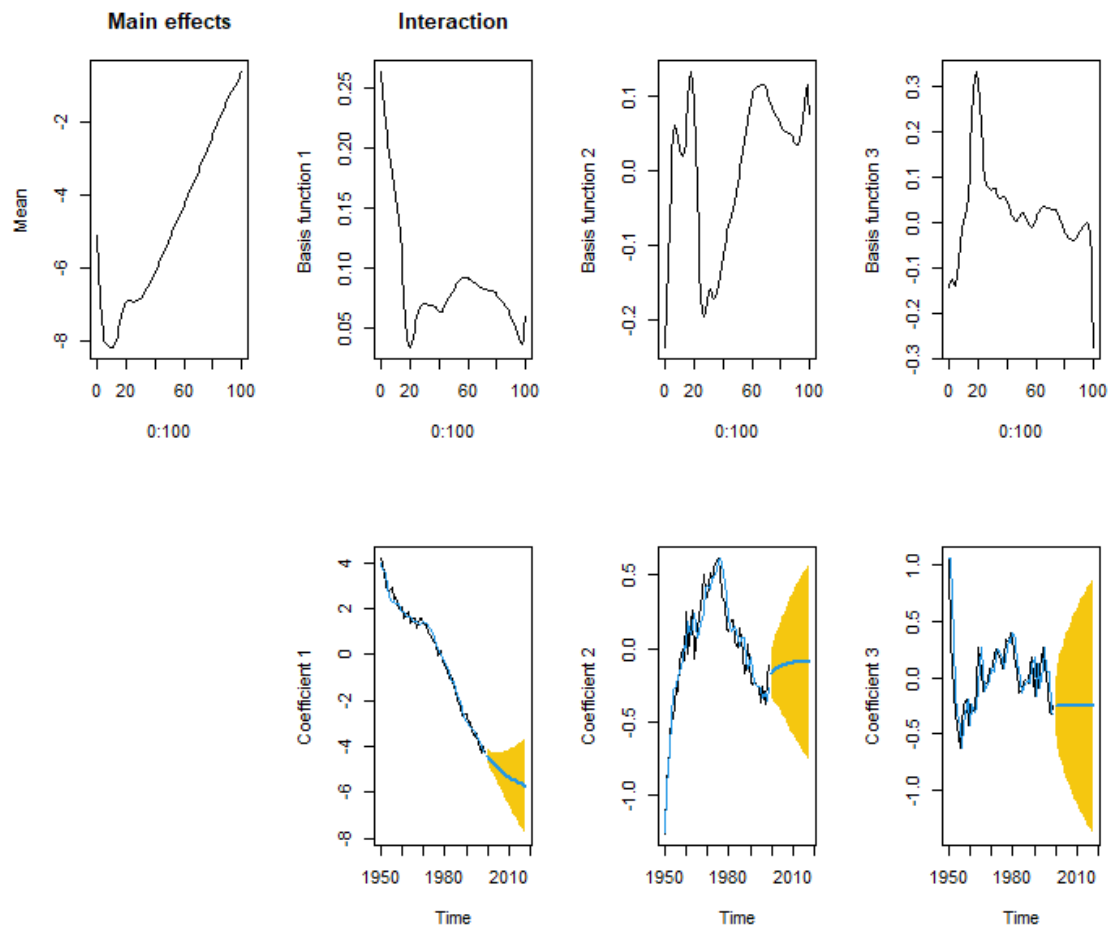


FIGURE 33: Les scores prévisionnels des composantes principales, et leurs intervalles de confiance à 80% (en jaune) - TOTAL

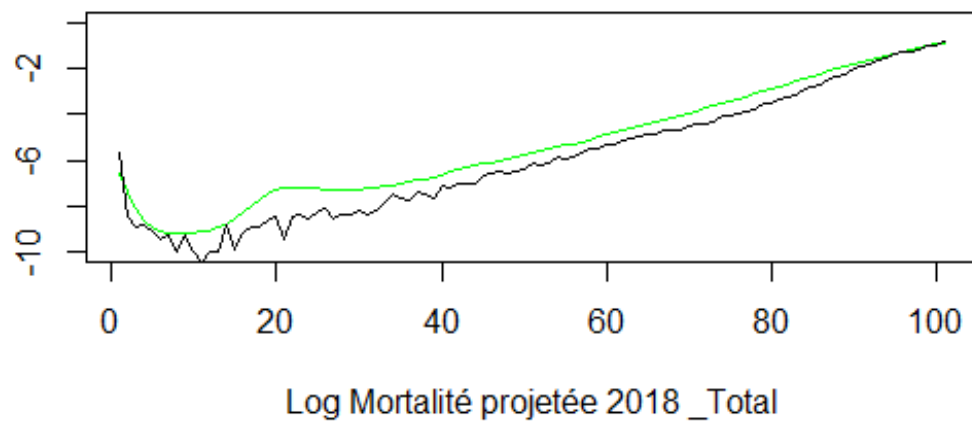


FIGURE 34: Taux de Mortalité pour la population totale en 2018

MATERN				
0	Moyenne	Ecart type	Q5%	Q95%
2019	78,5647	0,0242	78,5250	78,6044
2040	80,7406	0,5919	79,7670	81,7142
2060	82,2188	1,3818	79,9459	84,4917
20	Moyenne	Ecart type	Q5%	Q95%
2019	58,9494	0,0179	58,9199	58,9788
2040	60,9028	0,6168	59,8883	61,9174
2060	64,2218	1,4557	61,8274	66,6161
40	Moyenne	Ecart type	Q5%	Q95%
2019	39,5802	0,0177	39,5512	39,6092
2040	41,4188	0,6045	40,4245	42,4131
2060	43,9664	1,5653	41,3916	46,5411
60	Moyenne	Ecart type	Q5%	Q95%
2019	21,5300	0,0186	21,4993	21,5606
2040	20,8734	1,6581	18,1460	23,6008
2060	24,5247	1,7653	21,6209	27,4284

FIGURE 35: Calcul des statistiques sur l'espérance de vie avec le noyau MATERN-Population Totale

QUADRATIQUE RATIONNEL				
0	Moyenne	Ecart type	Q5%	Q95%
2019	79,2267	0,1779	78,9340	79,5193
2040	80,2361	1,5541	77,6799	82,7924
2060	81,7900	1,9870	78,5217	85,0583
20	Moyenne	Ecart type	Q5%	Q95%
2019	59,6201	0,1539	59,3671	59,8732
2040	61,4992	1,9117	58,3547	64,6437
2060	64,6540	1,6786	61,8929	67,4151
40	Moyenne	Ecart type	Q5%	Q95%
2019	40,1941	0,1595	39,9317	40,4565
2040	41,9780	1,8960	38,8594	45,0966
2060	44,0890	2,3000	40,3058	47,8722
60	Moyenne	Ecart type	Q5%	Q95%
2019	21,5807	0.3312614	20.94813	22.16921
2040	21,5739	1,9684	18,3362	24,8116
2060	24,7800	2,0450	21,4163	28,1437

FIGURE 36: Calcul des statistiques sur l'espérance de vie avec le noyau quadratique rationnel-Population Totale

NOYAU A BASE RADIALE				
0	Moyenne	Ecart type	Q5%	Q95%
2019	78,6896	0,3450	78,1221	79,2570
2040	80,7407	2,4174	76,7645	84,7170
2060	82,5681	2,1192	79,0822	86,0539
20	Moyenne	Ecart type	Q5%	Q95%
2019	58,6800	0,2567	58,2578	59,1022
2040	60,0500	1,9897	56,7772	63,3228
2060	61,8185	2,3448	57,9617	65,6753
40	Moyenne	Ecart type	Q5%	Q95%
2019	37,8700	0,1200	37,6726	38,0674
2040	40,0567	1,5860	37,4480	42,6654
2060	40,9681	2,9411	36,1304	45,8058
60	Moyenne	Ecart type	Q5%	Q95%
2019	20,3560	0,2356	19,9685	20,7435
2040	20,0809	1,9799	16,8242	23,3375
2060	24,7800	2,9670	19,8997	29,6603

FIGURE 37: Calcul des statistiques sur l'espérance de vie avec le noyau à base radiale-
Population Totale

FPCA				
0	Moyenne	Ecart type	Q5%	Q95%
2019	82,1048	0,0670	81,9946	82,2150
2040	84,9225	0,8980	83,4454	86,3995
2060	87,2942	2,0030	83,9996	90,5889
20	Moyenne	Ecart type	Q5%	Q95%
2019	62,3506	0,1238	62,1470	62,5543
2040	65,0514	0,6680	63,9527	66,1502
2060	67,3635	2,4400	63,3501	71,3769
40	Moyenne	Ecart type	Q5%	Q95%
2019	42,6940	0,0450	42,6200	42,7680
2040	45,3075	0,4760	44,5246	46,0905
2060	47,5386	1,0900	45,7457	49,3315
60	Moyenne	Ecart type	Q5%	Q95%
2019	24,1356	0,0500	24,0534	24,2178
2040	26,4662	0,4330	25,7540	27,1784
2060	28,4611	1,9898	25,1882	31,7340

FIGURE 38: Calcul des statistiques sur l'espérance de vie avec l'analyse en composantes principales fonctionnelle univariée-Population Féminine

FPCA				
0	Moyenne	Ecart type	Q5%	Q95%
2019	79,8490	0,0552	79,7800	79,9215
2040	82,8083	0,8698	81,3776	84,2390
2060	85,2771	2,2200	81,6255	88,9287
20	Moyenne	Ecart type	Q5%	Q95%
2019	60,1477	0,0493	60,0804	60,2083
2040	62,9702	0,9866	61,3474	64,5930
2060	65,3611	2,3400	61,5122	69,2101
40	Moyenne	Ecart type	Q5%	Q95%
2019	40,6773	0,0472	40,6070	40,7516
2040	43,3627	0,8790	41,9169	44,8085
2060	45,6783	2,0020	42,3853	48,9713
60	Moyenne	Ecart type	Q5%	Q95%
2019	22,4730	0,0446	22,4012	22,5457
2040	24,7839	0,8990	23,3051	26,2626
2060	26,8009	2,3445	22,9445	30,6572

FIGURE 39: Calcul des statistiques sur l'espérance de vie avec l'analyse en composantes principales fonctionnelle univariée-Population Totale

MFPCA				
0	Moyenne	Ecart type	Q5%	Q95%
2019	77,7909	0,0791	20,3461	20,7655
2040	80,8478	0,8956	79,3747	82,3209
2060	83,2278	1,7790	80,3016	86,1540
20	Moyenne	Ecart type	Q5%	Q95%
2019	58,0302	0,0791	57,9000	58,1604
2040	60,9659	0,9976	59,3250	62,6067
2060	63,2851	2,2124	59,6460	66,9242
40	Moyenne	Ecart type	Q5%	Q95%
2019	38,7496	0,0822	38,6144	38,8849
2040	41,5251	0,8767	40,0830	42,9671
2060	43,7602	1,6860	40,9870	46,5334
60	Moyenne	Ecart type	Q5%	Q95%
2019	20,8515	0,0765	20,7258	20,9773
2040	23,1648	0,7870	21,8703	24,4593
2060	24,9920	1,5898	22,3770	27,6069

FIGURE 40: Calcul des statistiques sur l'espérance de vie avec l'analyse en composantes principales fonctionnelle multivariée-Population Masculine