Méthodes numériques en actuariat

Mortalité

Students

Fatine YORO fatine.yoro@eleves.enpc.fr

14 février 2025

Table des matières

1 Présentation du Projet

Dans le cadre de ce projet, nous allons travailler sur la table de mortalité **fm_t67_2019.xls**, qui présente le nombre de décès survenus au sein d'un groupe de 1000 individus pour chaque année, entre 1962 et 2019. Ces données sont collectées pour trois catégories : les hommes, les femmes, et l'ensemble de la population étudiée. Chaque feuille du fichier correspond à l'une de ces catégories, et les données sont structurées de manière à faciliter leur analyse :

- Les lignes représentent les années (1962-2019), permettant d'analyser le nombre de décès.
- Les colonnes correspondent aux classes d'âges (0-110 ans), organisées en tranches pour observer les tendances de mortalité.
- Chaque valeur indique le nombre de décès pour une cohorte de 1000 individus, facilitant l'estimation des taux de mortalité et de survie.

Ces données de mortalité, qui indiquent le nombre de décès pour chaque groupe d'âge et chaque année, sont cruciales pour estimer des indicateurs démographiques tels que les taux de mortalité et les probabilités de survie. Elles permettent ainsi de modéliser l'espérance de vie et d'analyser les tendances de santé publique au fil des décennies. Dans ce projet, nous nous concentrerons spécifiquement sur les données relatives aux hommes et aux femmes.

Cette étape de modélisation joue un rôle fondamental, notamment pour les assureurs, en offrant une meilleure compréhension et prédiction des risques de mortalité sur le long terme. La capacité à estimer précisément la probabilité de décès en fonction de facteurs démographiques tels que l'âge et le sexe est essentielle pour l'élaboration des contrats d'assurance vie, la fixation des primes et la gestion des risques associés.

Dans ce projet, nous avons d'abord exploré des méthodes non paramétriques, suivies de modèles paramétriques et stochastiques. Chaque approche a offert des perspectives différentes sur l'estimation des fonctions de survie et des taux de mortalité, contribuant à une modélisation plus précise et adaptée aux besoins des professionnels du secteur assurantiel. Les modèles stochastiques, qui intègrent les variations aléatoires et les comportements observés dans les tables de mortalité, apportent une approche robuste pour anticiper les évolutions futures. Cette capacité d'anticipation permet aux assureurs d'optimiser la gestion de leurs portefeuilles et d'ajuster leurs stratégies face à des incertitudes démographiques.

```
df_T67h.head()
```

	Année	Tous âges	moins d'un an (a)	1 à 4 ans	5 à 9 ans	10 à 14 ans	15 à 19 ans	20 à 24 ans	25 à 29 ans	30 à 34 ans	35 à 39 ans	40 à 44 ans	45 à 49 ans	50 à 54 ans	55 à 59 ans	60 à 64 ans	65 à 69 ans	70 à 79 ans	80 à 89 ans	90 à 110 ans
0	1962	12.1	24.4	2.46	0.52	0.41	0.94	1.34	1.64	2.14	2.9	4.14	6.76	10.69	16.9	25.5	37.9	71	170	361
1	1963	12.3	23.9	2.39	0.49	0.4	0.93	1.43	1.65	2.04	2.87	4.15	6.82	10.7	17.5	26.6	38.9	73.1	171.4	352
2	1964	11.4	22.6	1.99	0.46	0.43	0.96	1.51	1.69	1.94	2.84	4.06	6.32	10.12	16.2	25	36.6	67.3	155.4	306
3	1965	11.8	21.2	1.81	0.48	0.4	0.99	1.51	1.59	2.01	2.91	4.26	6.42	10.57	16.7	25.7	38.4	69.8	162.3	329
4	1966	11.4	21.3	1.76	0.48	0.37	1.02	1.53	1.67	1.97	2.83	4.22	6.31	10.27	16.1	24.9	37.9	66.2	151.5	305

FIGURE 1 – Extrait du tableau de mortalité chez les hommes

```
df_T67h.head()
```

	Année	Tous âges	moins d'un an (a)	1 à 4 ans	5 à 9 ans	10 à 14 ans	15 à 19 ans	20 à 24 ans	25 à 29 ans	30 à 34 ans	35 à 39 ans	40 à 44 ans	45 à 49 ans	50 à 54 ans	55 à 59 ans	60 à 64 ans	65 à 69 ans	70 à 79 ans	80 à 89 ans	90 à 110 ans
0	1962	12.1	24.4	2.46	0.52	0.41	0.94	1.34	1.64	2.14	2.9	4.14	6.76	10.69	16.9	25.5	37.9	71	170	361
1	1963	12.3	23.9	2.39	0.49	0.4	0.93	1.43	1.65	2.04	2.87	4.15	6.82	10.7	17.5	26.6	38.9	73.1	171.4	352
2	1964	11.4	22.6	1.99	0.46	0.43	0.96	1.51	1.69	1.94	2.84	4.06	6.32	10.12	16.2	25	36.6	67.3	155.4	306
3	1965	11.8	21.2	1.81	0.48	0.4	0.99	1.51	1.59	2.01	2.91	4.26	6.42	10.57	16.7	25.7	38.4	69.8	162.3	329
4	1966	11.4	21.3	1.76	0.48	0.37	1.02	1.53	1.67	1.97	2.83	4.22	6.31	10.27	16.1	24.9	37.9	66.2	151.5	305

FIGURE 2 – Extrait du tableau de mortalité chez les femmes

Les tableaux de mortalité présentés ci-dessus illustrent le nombre de décès pour 1 000 hommes (8) et pour 1 000 femmes (2) dans chaque groupe d'âges pour une année considérée. Une première problématique réside dans le fait que le taux de mortalité est fourni sur une base annuelle pour des groupes d'individus répartis sur des intervalles d'âges spécifiques. Dans un premier temps, nous aborderons

l'interpolation de ces données afin d'obtenir une estimation précise de la mortalité pour un âge donné. Ensuite, nous nous intéresserons à la modélisation de la mortalité en fonction de l'âge et du sexe de l'individu, en utilisant des modèles stochastiques.

Cette étape de modélisation est cruciale, notamment pour les assureurs, car elle permet de mieux comprendre et prédire le risque de mortalité sur le long terme. En effet, la capacité à estimer de manière précise la probabilité de décès en fonction de ces facteurs démographiques (âge, sexe) est essentielle pour établir des contrats d'assurance vie, fixer des primes et gérer les risques associés. Ces modèles stochastiques, en tenant compte des variations aléatoires et des comportements observés dans les tables de mortalité, offrent une approche robuste pour anticiper les évolutions futures et permettre aux assureurs de mieux gérer leurs portefeuilles de contrats d'assurance.

2 Interpolation du taux de mortalité pour obtenir des valeurs par âge

Nous avons développé plusieurs fonction en Python permettant de passer du nombre de décès dans une classe d'individus à une discrétisation des décès par âge, en raison de la disponibilité limitée des données. Dans ce contexte, nous avons observé les tendances suivantes : le taux de mortalité, tant chez les femmes que chez les hommes, augmente avec l'âge. En outre, pour une grande majorité des classes d'âge, le nombre de décès est naturellement plus élevé chez les hommes que chez les femmes.

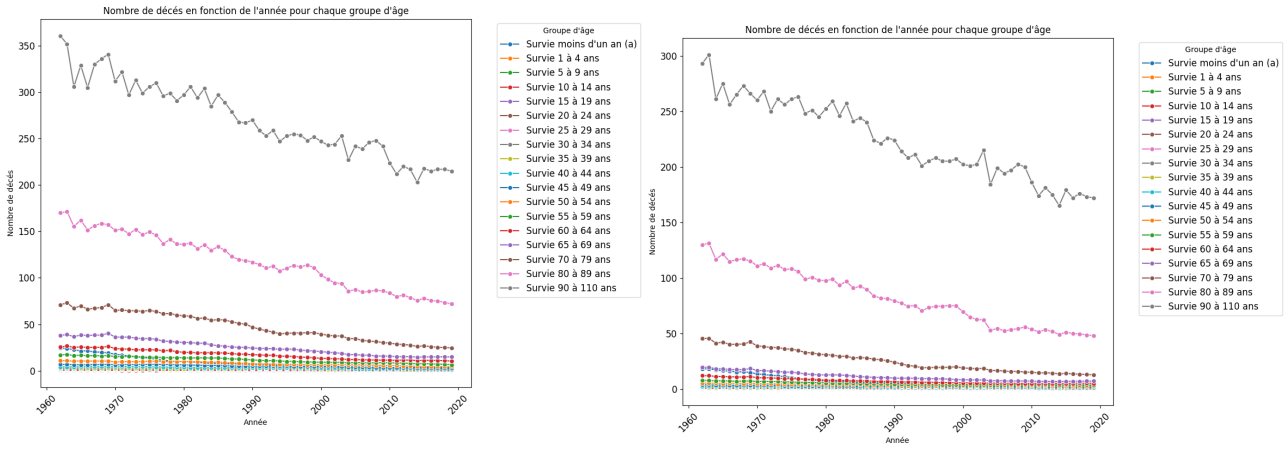

FIGURE 3 – Graphique de la mortalité chez les hommes et femmes

Dans un premier temps, nous allons explorer les méthodes non paramétriques permettant d'estimer le taux de survie, ainsi que leurs limitations dans le cadre de notre problème. Dans un second temps, nous présenterons les méthodes paramétriques, qui supposent une structure fonctionnelle spécifique pour le taux de mortalité. Parmi elles, les processus gaussiens, une approche stochastique, permettent de modéliser les incertitudes et de capturer la variabilité des données, offrant une grande flexibilité pour prédire des intervalles de confiance des taux de survie. Nous évoquerons aussi la méthode de Gompertz, une approche paramétrique non stochastique, qui suppose une évolution déterministe des taux de mortalité. Bien que couramment utilisée, elle peut être moins précise dans des contextes de forte variabilité ou d'influences externes modifiant rapidement la dynamique de survie.

2.1 Approche naive : méthode non paramétrique

Les methodes non paramétrique ne nécessitent pas de suppositions sur la forme de la distribution des données. Pour cette méthode, nous supposons que le taux de mortalité augmente de manière exponentielle. Nous allons calculer le taux de survie à l'âge t qui est défini par la formule suivante : $S_{(t)} = 1 - \mu(t)$, où $\mu(t)$ est le taux de mortalité à l'âge t.

Soit une série de tranches d'âge définies par $[a_1, a_2], [a_3, a_4], \ldots, [a_{n-1}, a_n]$, avec des taux de mortalité annuels associés $\mu(a_1), \mu(a_3), \ldots, \mu(a_{n-1})$. La **survie cumulative** pour une personne qui commence à l'âge a_1 est obtenue en multipliant les taux de survie annuels pour chaque tranche d'âge. La survie cumulative à un âge $a_i \in [a_j, a_{j-1}]$ pour une personne qui débute à a_j est donc calculée comme suit :

$$S_{\text{cumulative}}(a_i) = \prod_{a \in [a_j, a_{j-1}]} (1 - \mu(a_i))$$

Cela permet de modéliser la probabilité qu'un individu survive au-delà de plusieurs tranches d'âge successives en fonction des taux de mortalité spécifiques à chaque tranche. Par exemple, nous appliquerons cette méthode pour les années 1983,1995, 2017 et 2019 en utilisant les taux de mortalité associés à chaque tranche d'âge pour ces années.

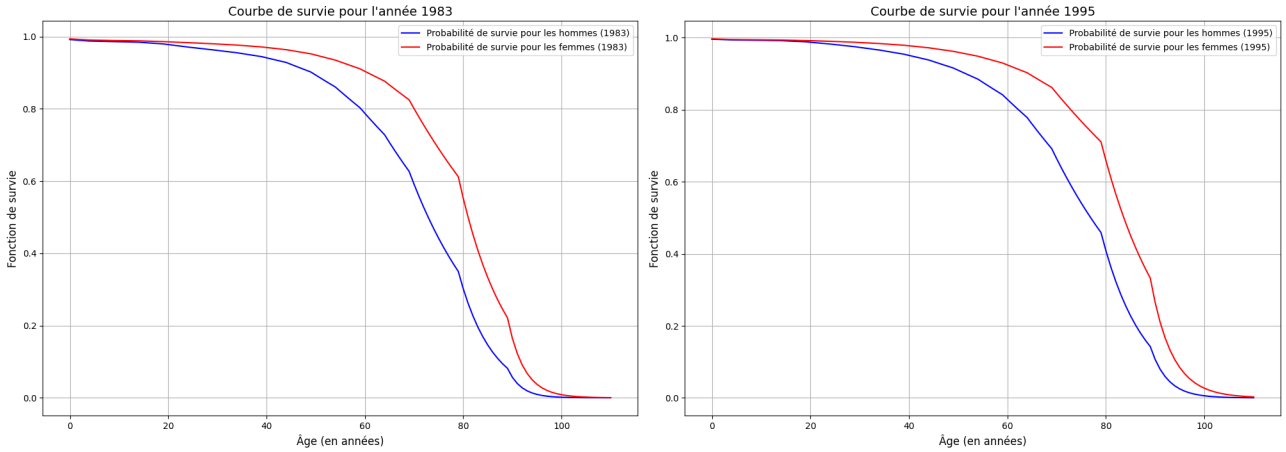

FIGURE 4 – Graphique de la fonction de survie non paramétrique homme vs femme pour l'année 1983 et 1995

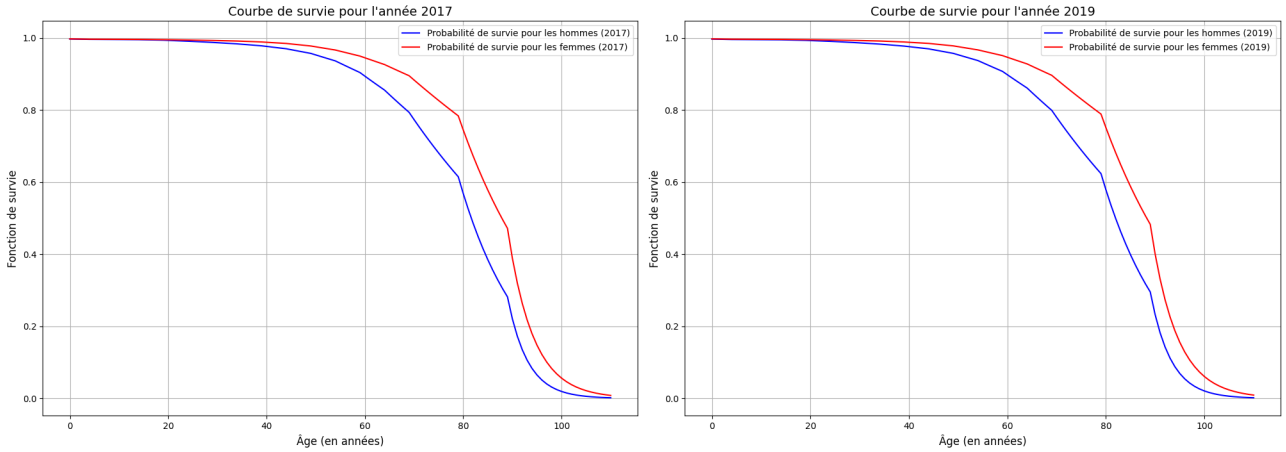

FIGURE 5 – Graphique de la fonction de survie non paramétrique homme vs femme pour l'année 2017 et 2019

Nous constatons que le modèle rencontre des difficultés à capturer la corrélation entre les différentes tranches d'âge, ce qui se manifeste par l'apparition de formes irrégulières et non lisses sur les courbes de survie. Cette non-lissité indique que les relations entre les groupes d'âge ne sont pas bien modélisées, ce qui peut être dû à une mauvaise gestion des transitions entre les tranches d'âge ou à l'absence de prise en compte de dynamiques plus complexes, comme les effets de dépendance entre les âges voisins. De plus, cette forme discontinue de la courbe reflète une des limitations des méthodes utilisées : leur incapacité à intégrer pleinement l'incertitude ou la variabilité des données. Par conséquent, les méthodes non paramétriques peinent à obtenir une modélisation fluide et cohérente, essentielle pour une estimation fiable du taux de survie à long terme. Bien qu'elles soient capables de fournir des estimations locales relativement précises, elles manquent de la flexibilité nécessaire pour prendre en compte les influences stochastiques ou les tendances temporelles, ce qui peut affecter négativement la qualité des prédictions globales, surtout lorsqu'il s'agit de situations dynamiques et complexes.

2.2 Par méthode paramétrique

Par le modèle de Gompertz-Makeham

Le modèle de Gompertz-Makeham (modèle exponentielle) est largement utilisé en actuariat et en démographie pour modéliser la mortalité. Il repose sur une **force de mortalité** combinant une composante indépendante de l'âge et une composante liée au vieillissement :

$$\mu(x) = A + B \cdot \exp(C \cdot x) \tag{1}$$

où :

- A représente la composante constante, indépendante de l'âge (ex. accidents),
- $B \cdot \exp(C \cdot x)$ modélise une mortalité croissante avec l'âge x.

La fonction de survie S(x) est donnée par :

$$S(x) = \exp\left(-A \cdot x - \frac{B}{C}\left(\exp(C \cdot x) - 1\right)\right) \tag{2}$$

Modèle puissance

Le modèle puissance suit la forme :

$$\mu(x) = A + B \cdot C^x \tag{3}$$

où :

- A joue le même rôle qu'auparavant (risques constants),
- $B \cdot C^x$ modélise une mortalité croissant exponentiellement avec l'âge.

La fonction de survie, définie comme la probabilité de survie au-delà d'un âge donné, est donnée par :

$$S(x) = \exp\left(-A \cdot x - B \cdot \frac{\ln C}{C^x}(C^x - 1)\right) \tag{4}$$

Nous avons implémenté une fonction Python permettant d'estimer de manière paramétrique la fonction de survie. Cette fonction prend en entrée une année spécifique, un jeu de données donné, ainsi que le type de modèle de Gompertz-Makeham choisi (exponentiel ou puissance). Elle retourne alors une estimation de la fonction de survie correspondant aux paramètres ajustés.

De plus, nous avons supposé que l'âge correspondant au barycentre des classes d'âge possède le taux de mortalité inscrit dans le fichier de données fourni en argument. Cette hypothèse permet d'associer chaque classe d'âge à un point unique pour l'ajustement du modèle. Cependant, cette approximation reste grossière, car elle suppose une homogénéité du taux de mortalité au sein de chaque classe d'âge.

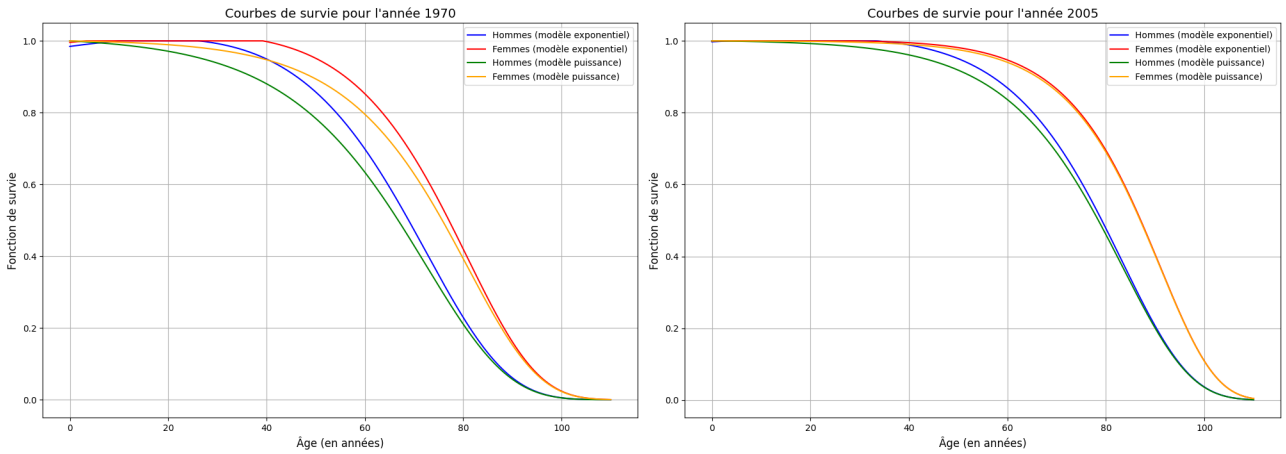

FIGURE 6 – Fonction de survie paramétrique homme vs femme pour l'année 1970 et 2005

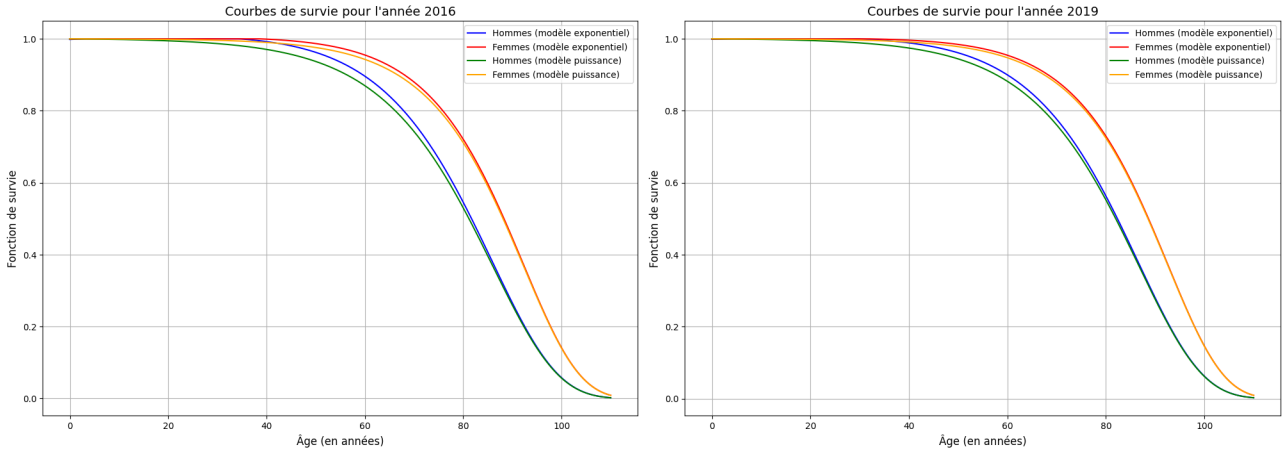

FIGURE 7 – Graphique de la fonction de survie non paramétrique homme vs femme pour l'année 2016 et 2019

L'analyse des courbes pour différentes années révèle les tendances suivantes :

- **Pour les personnes dans les groupes dont l'age est inferieure à 50 ans** : Les deux modèles montrent une faible décroissance de la fonction de survie, indiquant une mortalité faible. Les prédictions des deux modèles sont similaires dans cette tranche d'âge.
- **Pour les personnes dans les groupes dont l'age est compris entre 50 et 80 ans** : Le modèle puissance diminue plus rapidement que le modèle exponentiel, traduisant une mortalité plus élevée à ces âges. Le modèle exponentiel semble sous-estimer légèrement la mortalité dans cette tranche.
- **Pour les individus ayant un âges au desssus de 80 ans**) : Les prédictions des deux modèles sont similaires dans cette tranche d'âge, avec une forte décroissance exponentielle.

Malheureusement, en raison du manque de données précises, il n'est pas possible de déterminer avec certitude quel modèle est le plus performant. De plus, le paramètre p_0 a été choisi de manière aléatoire, ce qui soulève des interrogations quant à la robustesse des estimations obtenues. Pour répondre à ces limites, une approche alternative consiste à introduire des processus Gaussiens dans la modélisation des fonctions de survie. Les modèles déterministes, tels que les modèles exponentiels ou puissance, reposent sur une tendance fixe, ce qui ne permet pas de capturer les incertitudes et les variations aléatoires inhérentes aux trajectoires de mortalité. Les processus Gaussiens offrent une modélisation plus souple et adaptative, capable d'incorporer ces incertitudes.

2.3 Modélisation de la fonction de survie à l'aide des Processus Gaussiens

La modélisation des fonctions de survie $S(x)$ à l'aide de processus Gaussiens repose sur plusieurs considérations théoriques et pratiques. La fonction de survie $S(x)$ est une fonction continue, monotone décroissante, définie sur l'ensemble $[\![0, 130]\!]$ des âges. Ses propriétés fondamentales sont les suivantes :

$$S(0) = 1 \quad \text{et} \quad \lim_{x\to\infty} S(x) = 0.$$

Un processus Gaussien $\mathcal{GP}(m(x), k(x, x'))$ est bien adapté pour modéliser des fonctions continues, en garantissant une estimation lisse et stable de $S(x)$. Le choix d'un noyau $k(x, x')$ approprié permet de contrôler la régularité et la variabilité des estimations.

Pour modéliser les points observés, nous avons supposé que les taux de mortalité correspondent aux barycentres des classes d'âge. Cette hypothèse permet d'utiliser les processus Gaussiens pour ajuster les points observés de manière fluide, avec une estimation conditionnelle donnée par :

$$S(x) \mid \text{Données} \sim \mathcal{N}(\hat{m}(x), \hat{\sigma}^2(x)).$$

Cela présente plusieurs avantages :

- Les estimations passent exactement par les points observés $(x_i, S(x_i))$.
- Les incertitudes autour de la courbe sont quantifiées à l'aide d'intervalles de confiance.

Le noyau $k(x, x')$ joue un rôle clé en contrôlant la structure et la variabilité des estimations de $S(x)$. Différents noyaux ont été considérés :

- **Noyau RBF (Radial Basis Function)** : Convient aux fonctions lisses et sans variations brusques.

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right).$$

  - σ^2 : Amplitude du noyau , ℓ : Échelle de longueur.

- **Noyau Matérn** : Adapté aux fonctions présentant des variations plus abruptes.

$$k(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}|x - x'|}{\ell}\right) \exp\left(-\frac{\sqrt{3}|x - x'|}{\ell}\right), \quad \nu = 1.5.$$

  - ν : paramètre de régularité (fixé ici à $\nu = 1.5$), σ^2 : amplitude du noyau et ℓ : échelle de longueur.

- **Noyau ExpSineSquared** : Idéal pour les fonctions présentant des variations périodiques.

$$k(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2\left(\frac{\pi|x-x'|}{p}\right)}{\ell^2}\right).$$

  - σ^2 : amplitude du noyau , p : périodicité du noyau et ℓ : échelle de longueur, contrôle l'amplitude des variations non périodiques.

- **Noyau Rational Quadratic** : Représente une somme infinie de noyaux RBF avec différentes échelles de longueur.

$$k(x, x') = \sigma^2 \left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

  - σ^2 : amplitude du noyau , ℓ : échelle de longueur et α : Paramètre qui contrôle la distribution des échelles de longueur.

Parmi les différents modèles testés, le noyau offrant le meilleur ajustement aux données observées a été retenu en utilisant la log-vraisemblance marginale (LML) comme critère de sélection, garantissant une estimation optimale de la fonction de survie. De plus, les hyperparamètres associés aux noyaux ont été ajustés pour maximiser cette mesure de performance. La LML est une mesure clé dans l'ajustement des processus gaussiens. Pour un ensemble de données $\mathcal{D} = \{(x_i, y_i)\}$, elle est définie comme :

$$\log p(y \mid X, \theta) = -\frac{1}{2}\left(y^T K^{-1} y + \log\det(K) + n\log(2\pi)\right),$$

où :

- K est la matrice de covariance calculée avec le noyau $k(x, x')$.
- θ représente les hyperparamètres du noyau.
- n est le nombre de points de données.

Nous obtenons ainsi le modèle présentant le meilleur ajustement qui est indiqué dans la sortie Python ci-dessous :

```
Meilleur noyau sélectionné : 1**2 * RationalQuadratic(alpha=0.1, length_scale=1)
Log-marginal likelihood du meilleur modèle : 49.7767
```

FIGURE 8 – Sortie Python correpond au noyau maximisant LML (et ses hyperparamétres)

Les processus Gaussiens permettent également de calculer la variance conditionnelle $\hat{\sigma}^2(x)$, qui décrit l'incertitude liée à chaque âge x. Les intervalles de confiance à 95% sont donnés par :

$$S(x) \in [\hat{m}(x) - 1.96 \cdot \hat{\sigma}(x), \hat{m}(x) + 1.96 \cdot \hat{\sigma}(x)]\,.$$

Nous avons tracé les fonctions de survie en utilisant les processus gaussiens pour les années 1998, 2009, 2014 et 2018 et pour les différents sexes.

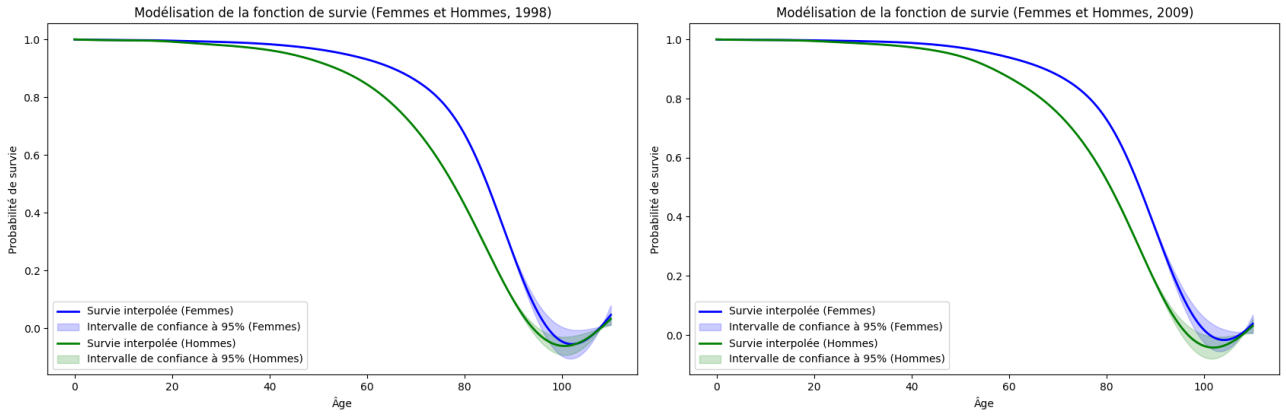

FIGURE 9 – Fonctions de survie approximées par processus gaussiens : Femmes vs Hommes en 1998 et 2009

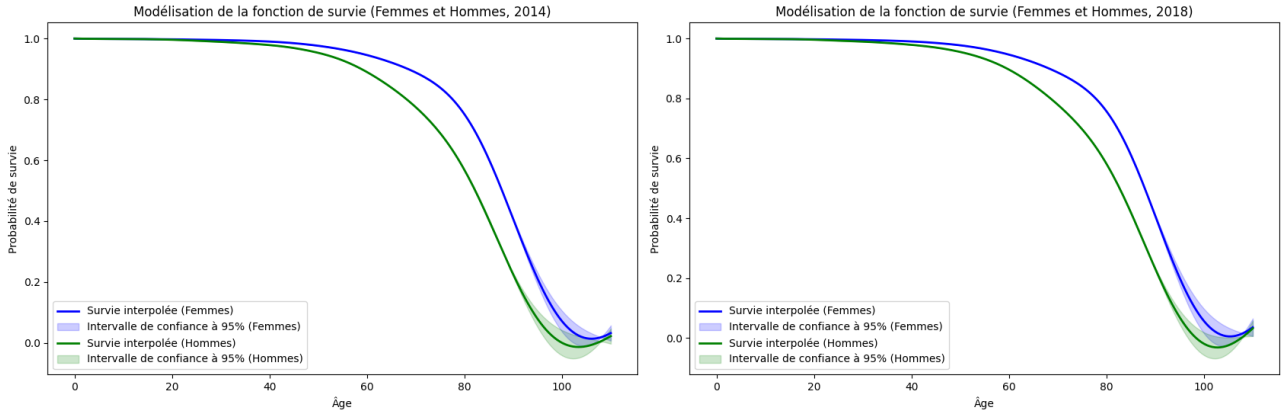

FIGURE 10 – Fonctions de survie approximées par processus gaussiens : Femmes vs Hommes en 2014 et 2018

Il est apparu que les classes d'âge supérieures, notamment [70,79], [80,89] et [90,110], présentent des largeurs plus importantes que les autres classes. Cette différence de largeur induit un élargissement des intervalles de confiance dans ces segments, car la variance conditionnelle du processus Gaussien est

directement influencée par l'espacement entre les points d'observation. Pour améliorer la précision de l'approximation et réduire l'incertitude, il serait pertinent de définir des tranches d'âge plus uniformes et mieux équilibrées sur l'ensemble du domaine d'étude. Cela permettrait une meilleure capture des variations locales de la survie. On observe une légère augmentation de la fonction de survie au-delà de 100 ans. Pour corriger ce biais, il serait nécessaire d'ajouter des données observées supplémentaires afin d'améliorer la précision de l'estimation.

En conclusion, les processus gaussiens constituent une approche puissante et flexible pour la modélisation des fonctions de survie. Grâce au choix judicieux des noyaux, ils garantissent une interpolation précise des données observées tout en offrant la possibilité de quantifier les incertitudes associées à chaque estimation. De plus, leur capacité à optimiser finement les hyperparamètres des noyaux permet de produire des prédictions robustes et fiables, adaptées à des variations locales et des comportements complexes des fonctions de survie.

3 Conclusion

Ce projet a exploré différentes approches pour l'interpolation et la modélisation des taux de mortalité afin d'obtenir des estimations précises des fonctions de survie.

L'approche naïve, basée sur des méthodes non paramétriques, a permis d'obtenir des interpolations simples des données de mortalité, bien que ces méthodes puissent manquer de souplesse et de précision pour des variations complexes. En revanche, le modèle de Gompertz-Makeham a offert une solution paramétrique robuste, particulièrement adaptée à la modélisation de la mortalité humaine sur une large plage d'âges.

Les processus gaussiens appliqués à la modélisation des fonctions de survie offrent une méthode particulièrement flexible et précise. Grâce à l'optimisation des noyaux, ces processus permettent non seulement une estimation fine des fonctions de survie, mais aussi une quantification rigoureuse des incertitudes liées aux prédictions. Cela se révèle particulièrement pertinent dans des contextes de modélisation démographique et actuarielle. Enfin, la quantification des incertitudes, caractéristique de ces modèles, constitue un atout majeur pour les acteurs du secteur assurantiel. Elle contribue à une meilleure évaluation des risques et à une détermination plus précise des primes et des provisions, renforçant ainsi la robustesse et la fiabilité des résultats obtenus.

En ce qui concerne l'utilisation de ces fonctions de survie pour le calcul des rentes, plusieurs éléments méritent d'être soulignés. Tout d'abord, les différences de longévité entre les sexes jouent un rôle important. Les hommes, ayant en moyenne une espérance de vie plus courte que les femmes, se voient attribuer des rentes plus élevées, car la durée de paiement anticipée est réduite.

En conclusion, les processus gaussiens se positionnent comme une méthode puissante et adaptable pour l'estimation des fonctions de survie, particulièrement adaptée au calcul des rentes et à l'évaluation des risques dans un cadre actuariel.

Annexe

— **Modèles de Gompertz-Makeham :** Explication détaillée du modèle de Gompertz-Makeham (Society of Actuaries).

— **Processus Gaussiens (Machine Learning) :** Introduction aux processus gaussiens en machine learning (scikit-learn).

— **Processus Gaussiens (Article) :** Article sur l'utilisation des processus gaussiens en modélisation de données (Management Data Science).

— **Méthodes stochastiques :** Modélisation stochastique des risques en assurance (International Actuarial Association).

— **Interpolation non paramétrique :** Techniques avancées d'interpolation non paramétrique (ScienceDirect).