PROJECTIONS PROBABILISTES BAYÉSIENNES DE POPULATION POUR LA FRANCE

Vianney COSTEMALLE (*)

(*) Insee, Département des méthodes statistiques, Division des méthodes et référentiels géographiques

vianney.costemalle@insee.fr

Mots-clés : projections probabilistes, inférence bayésienne, séries temporelles, mortalité, fécondité, migrations

Résumé

Les projections de population sont un exercice régulier des instituts nationaux de statistiques. En France, les dernières ont été produites en 2016 par l'Insee en utilisant une approche déterministe basée sur 27 scénarios différents. Dans cet article nous proposons une nouvelle approche en projetant de façon probabiliste la population et en utilisant de plus le paradigme bayésien, afin de quantifier l'incertitude sur le niveau des populations futures sans recourir à des scénarios.

Selon la méthode des composantes, on projette de manière indépendante et par sexe et âge les trois composantes qui déterminent l'évolution d'une population, à savoir le taux de mortalité, le taux de fécondité et le solde migratoire. Ces trois composantes sont modélisées en tenant compte des données d'état-civil (nombre de naissances et de décès) et des séries du solde migratoire.

On suppose que le nombre de naissances et le nombre de décès suivent une loi de Poisson. On projette en premier lieu l'indice conjoncturel de fécondité en utilisant un modèle autorégressif du premier ordre. Connaissant alors le niveau de fécondité projeté, on en déduit une projection des taux de fécondité par âge à l'aide du modèle de Lee-Carter étendu, en réalisant une estimation bayésienne des paramètres.

Les taux de mortalité par âge sont quant à eux projetés grâce à une méthode implémentée dans un package R par l'Institut de Statistiques de Nouvelle-Zélande qui utilise également une approche bayésienne.

Enfin, le solde migratoire est projeté à l'aide d'une série temporelle.

Les estimations bayésiennes des paramètres des différents modèles sont conduites à partir du logiciel libre Stan.

Les résultats montrent que la population de la France continuera de croître pour atteindre un niveau compris entre 66,1 millions et 77,2 millions d'habitants en 2070 avec une probabilité de 95 %.

De plus, il y a une probabilité de 20 % que la population totale en France atteigne un pic en 2050 avant de décliner. Selon nos estimations, la structure globale de la population sera profondément transformée avec en particulier une augmentation significative de la part des personnes âgées de plus de 65 ans.

On souhaite montrer que les projections probabilistes bayésiennes permettent d'éclairer le futur et de quantifier l'incertain en se basant sur les connaissances actuelles. Si les projections probabilistes peuvent au premier abord apparaître plus compliquées à appréhender que les projections déterministes, elles ont l'avantage de résumer toute l'information en une seule distribution de probabilité, alors que les projections déterministes produisent de nombreux indicateurs.

Introduction

Les projections de population sont un exercice régulier des instituts statistiques à travers le monde ainsi que de certaines organisations internationales comme l'Organisation des Nations Unies (ONU) qui tous les deux ou trois ans depuis 1951 publie les *World Population Prospects* (ONU, 2017). L'intérêt des projections de population est multiple et ces utilisateurs sont nombreux. Elles servent avant tout à prévoir, sous certaines hypothèses, ce que pourrait être la population d'une région, d'un pays ou du monde entier, en nombre d'habitants ainsi qu'en structure. À court ou moyen terme, elles sont à la base de la planification économique et sociale comme par exemple le financement des retraites (COR, 2017) ou la construction d'infrastructures publiques. Elles sont aussi un élément essentiel dans certains autres exercices de projections, comme les projections économiques, climatiques ou environnementales.

En ce qui concerne la France, les dernières projections officielles datent de 2016 (Blanpain & Buisson, 2016a ; 2016b) et indiquent quelle sera la population en 2070 si les tendances passées se poursuivent, avec différentes variantes sur ces hypothèses. L'objectif de cet article est d'explorer une nouvelle méthode pour projeter la population de la France : les projections probabilistes. L'approche proposée est dite probabiliste car elle permet de quantifier l'incertitude sur le niveau de la population future, elle se distingue en cela de l'approche traditionnelle qui est un ensemble de projections déterministes basées sur différents scénarios. Ce qui change fondamentalement entre ces deux approches ce ne sont pas tellement les résultats eux-mêmes mais plutôt la façon dont ceux-ci sont interprétés et utilisés.

Les projections probabilistes reposent sur des modèles statistiques, la plupart du temps paramétriques. L'incertitude sur certaines composantes de la population peut être captée par des termes d'erreurs, comme dans le cas des séries temporelles, mais elle peut aussi provenir d'une inférence bayésienne des paramètres du modèle. Tout l'objectif est de quantifier l'incertitude sur la population future. Pour cela, on peut utiliser l'approche stochastique, l'approche bayésienne, ou même une combinaison des deux. Dans cet article, nous utilisons des modèles stochastiques avec inférence bayésienne des paramètres.

Dans une tribune libre du *Journal of Official Statistics*, une série de démographes et d'universitaires de différents pays mentionnent les apports et les défis des projections probabilistes en démographie et appellent à plus de recherche et de pratique dans ce domaine de la part des instituts statistiques (Bijak et al., 2015). Ils soulignent le fait que les projections probabilistes ont déjà été développées et utilisées avec succès dans d'autres disciplines comme la météorologie, la climatologie ou bien l'aviation. Les statistiques bayésiennes mettent également du temps à pénétrer le champ de la démographie. Bien que le théorème de Bayes a été établi il y a plus de 250 ans, ce n'est que récemment, avec l'apparition des algorithmes MCMC (*Markov Chains Monte-Carlo*) à partir de années 1980 et avec l'explosion de la puissance de calculs des ordinateurs, que l'inférence bayésienne est mise en œuvre (Bijak & Bryant, 2016).

Certains instituts statistiques dans le monde ont déjà adopté la démarche visant à produire des projections démographiques probabilistes pour leurs statistiques officielles. C'est le cas en particulier des Pays-Bas et de la Nouvelle-Zélande. Les Pays-Bas ont commencé à produire des projections probabilistes basées sur des méthodes stochastiques dès 1998. La Nouvelle-Zélande communique également depuis 2012 des résultats probabilistes de projections de population (MacPherson, 2016 ; Dunstan & Ball, 2016). Enfin, l'ONU qui réalise des projections pour l'ensemble des pays est passé d'une méthode déterministe à une méthode probabiliste en 2014 (Costemalle, 2015). De plus certaines composantes de ces projections sont basées sur l'inférence bayésienne.

La très grande majorité des projections de population repose sur la méthode des composantes qui consiste à projeter séparément les trois composantes essentielles de la dynamique des populations, à savoir la fécondité, la mortalité et les migrations. La population à une période donnée est décomposée par sexe et catégories d'âges et elle est égale à la population de la période précédente à laquelle on ajoute les naissances et les immigrants et à laquelle on retire les décès est les émigrants. De cette façon on peut faire évoluer, période par période, la population et sa structure par sexe et catégories d'âges. Pour cela, il faut à chaque période, déterminer le nombre de naissances par sexe ainsi que le nombre de décès et le solde migratoire par sexe et catégorie d'âges. En ce qui concerne les naissances et les décès, les méthodes les plus répandues reposent sur la projection des taux de fécondité et des taux de mortalité.

Les projections probabilistes de population sont encore un domaine de recherche actif. Il n'existe pas de méthode unique, il y a au contraire presque autant d'approches que de données qui diffèrent d'un pays à un autre.

Dans un premier temps, nous soulignerons les différences essentielles entre les projections déterministes et les projections probabilistes, puis nous présenterons quelques-unes des différentes approches qui ont été développées en démographie en ce qui concerne les projections probabilistes de population. La troisième partie est consacrée à la description des données françaises de mortalité, de fécondité et de solde migratoire. Dans les parties quatre et cinq nous présentons et validons les modélisations retenues pour chacune des trois composantes. Enfin nous présentons les résultats des projections probabilistes ainsi obtenues pour la France, avant de discuter les hypothèses des modèles.

1. Projections déterministes et projections probabilistes : différentes façons dôaborder lôincertitude.

Prévoir l'avenir est un exercice difficile et de nombreuses méthodes se sont développées au cours des siècles. Les méthodes les plus récentes et sophistiquées se basent sur des modèles mathématiques qui tentent de détecter certains motifs ou invariants dans les données et de prolonger les tendances observées, tout en respectant certaines contraintes qu'on peut s'imposer. Les projections déterministes et probabilistes font toutes deux appel à un certain degré de modélisation des données observées, elles ne diffèrent que sur la nature des prévisions. Dans le premier cas, ce qu'on cherche à projeter dépend de façon déterministe de certains paramètres. On se donne alors un scénario d'évolution de ces paramètres, qu'on juge le plus probable au vu des connaissances accumulées, des avis des experts et de l'intuition. Un scénario donné correspond à une et une seule projection possible, et le rapport de l'un à l'autre est déterministe. Dans le cas où le scénario se réaliserait, la projection serait certaine. Les projections déterministes répondent donc à la question « que se passerait-t-il dans l'avenir dans le cas de l'avènement d'un tel scénario ? ». On peut ainsi formuler des scénarios extrêmes pour voir comment se comporterait alors le futur dans le cas de leur réalisation. Les projections déterministes sont donc un formidable outil pour explorer l'avenir à partir de scénarios préétablis. Toute l'incertitude de la projection repose alors sur la réalisation du scénario. On formule des scénarios possibles, mais on n'est pas en mesure de savoir à quel degré de probabilité ils pourront se réaliser. On peut même affirmer que la probabilité de leur réalisation est nul (si les grandeurs sont continues) ou très faible (si les grandeurs sont discrètes). Le degré de probabilité est estimé de façon intuitive et se reflète dans les termes utilisés pour décrire ces scénarios : on parle de scénario « central », pour le scénario considéré comme le plus plausible compte tenu des connaissances actuelles et de scénarios « extrêmes ».

Au contraire, les projections probabilistes sont basées sur des modèles qui essayent de tenir compte de l'incertitude résultant de l'ignorance de certains aspects des projections. Ces modèles reposent sur des hypothèses qui sont le fruit des jugements d'experts et des intuitions. Les hypothèses sous-jacentes des modèles dans les projections probabilistes sont l'équivalent des scénarios dans les projections déterministes. L'avantage des projections probabilistes est qu'elles permettent de quantifier l'incertitude à partir des évolutions observées par le passé et de la propager dans le futur afin d'avoir des intervalles de confiance des projections. Ainsi l'interprétation et l'utilisation des projections probabilistes diffèrent de celles des projections déterministes.
Les prévisions météorologiques utilisent par exemple depuis longtemps des projections probabilistes : on ne nous dit pas seulement s'il va pleuvoir ou non le lendemain, mais avec quelle probabilité il risque de pleuvoir (Raftery, 2014). Les événements futurs étant par nature incertains, donner la probabilité de leur réalisation, étant donné les connaissances actuelles, donne ainsi plus d'information qu'une projection déterministe basée sur un scénario. Les séries temporelles sont, en sciences économiques en particulier, un moyen de produire des projections probabilistes : dans le cas d'une marche aléatoire simple par exemple on sait que la variance augmente avec la racine carrée du temps.
En ajoutant des termes d'erreurs dans les modèles, on peut donc créer des projections probabilistes stochastiques. Une autre manière de quantifier l'incertitude est de s'appuyer sur le paradigme bayésien. Dans ce dernier, les paramètres des modèles sont considérés comme des variables aléatoires, au même titre que les termes d'erreurs dans les modèles stochastiques. L'inférence bayésienne consiste alors à estimer la distribution *a posteriori* de ces paramètres, c'est-à-dire après l'observation des données. Cette distribution donne des valeurs possibles des paramètres et leur degré de probabilité. Elle diffère de la distribution *a priori* qui est la distribution donnée par le modélisateur et qui est censée refléter la connaissance du problème avant toute observation des données.

2. Les projections probabilistes en démographie : des modélisations en pratique très variées

On peut classer les techniques de projections de population en trois catégories (Booth, 2006). La première regroupe les méthodes basées sur l'extrapolation des tendances, qui cherchent à prolonger, de façon linéaire le plus souvent, les tendances détectées dans le passé. Elles se basent uniquement sur les données passées et ne cherchent pas à expliquer les mécanismes sous-jacents d'évolution. Elles se révèlent souvent efficaces. La deuxième façon de projeter la population est de se fixer des tendances de long terme. Ces méthodes sont basées sur le fait qu'on s'attend à ce que l'avenir se déroule d'une certaine manière. Cela peut être justifié par des avis d'experts, qui évaluent ce qu'on pourrait attendre pour le futur compte tenu de leurs connaissances actuelles, ou par des intentions de personnes, comme celles mesurées par les enquêtes d'intentions de fécondité (Régnier-Loilier et Vignoli, 2011). Enfin, la dernière catégorie de projection comprend les modèles structurels, qui essayent d'expliquer l'évolution de la population avec des variables extérieures. Il faut alors projeter ces variables extérieures selon l'une des trois catégories de projection. Souvent, les approches proposées mêlent plusieurs de ces techniques et les techniques utilisées diffèrent selon les composantes (mortalité, fécondité et migration) qu'on souhaite projeter.

Une méthode classique de projection de la mortalité a été développée par Lee et Carter (1992) et consiste à décomposer l'évolution du logarithme des taux de mortalité en un effet de l'âge et un effet du temps, spécifique à chaque âge. L'effet temporel est ensuite considéré comme une série temporelle dont on estime les paramètres. Par calculs ou par la simulation un très grand nombre de

fois des valeurs futures de cet effet temporel en faisant appel à la modélisation retenue, il est possible d'avoir une projection probabiliste. L'idée essentielle de cette approche est de capter dans les données les évolutions régulières et d'extrapoler ces régularités. La méthode de Lee-Carter a depuis été utilisée très fréquemment pour projeter la mortalité, mais aussi pour projeter la fécondité et les migrations. Wiśniowski et al. (2015) en proposent une version plus étendue, en ajoutant un effet de génération, qu'on peut appliquer aux trois composantes des variations de la population. Ces auteurs proposent de plus de réaliser ces projections dans un cadre entièrement bayésien. Le modèle de Lee-Carter a également été généralisée par Hyndman et Ullah (2007) qui décomposent le logarithme des taux de mortalité, ou de fécondité, en composantes principales et qui prolongent les coefficients de chacune de ces composantes à l'aide de séries temporelles. Hyndman et Booth (2006) suggèrent de plus de réaliser une transformation de Box et Jenkins sur les taux étudiés afin de généraliser la transformation logarithmique. Cette approche est entièrement stochastique.

Tout l'intérêt des projections probabilistes est de pouvoir quantifier le degré de probabilité des projections futures. C'est ainsi qu'en 2001, Lutz et al. (2001) annoncent qu'il est probable que la population mondiale cesse de croître d'ici la fin du siècle. Plus précisément, leurs modèles stochastiques et leurs calculs prévoient que la population mondiale pourrait commencer à décroître d'ici la fin du siècle avec une probabilité de 85%. L'ONU, qui publie régulièrement des projections de population, à commencer à utiliser une méthode probabiliste et bayésienne à partir de 2014. Leurs résultats donnent un aperçu différent de l'évolution de la population à long terme. Ils jugent en effet que la fin de la croissance mondiale de population est improbable d'ici à 2100 (Gerland et al., 2014). Leur méthodologie est différente de celle de Lutz et al. (2001). Les grandeurs agrégées que sont l'espérance de vie à la naissance et l'indice conjoncturel de fécondité sont projetés directement dans un premier temps, avant de décomposer ces indicateurs en taux de mortalité par sexe et âge et en taux de fécondité par âge. Pour projeter l'espérance de vie, le gain d'espérance de vie tous les 5 ans est modélisé par une double fonction logit, dépendant de l'espérance de vie actuelle et de nombreux paramètres. Ces paramètres sont estimés par inférence bayésienne, ce qui conduit à avoir une distribution *a posteriori* des gains d'espérance de vie, et donc une distribution *a posteriori* de l'espérance de vie elle-même à l'horizon 2100 (Raftery et al., 2013). Ceci est l'exemple d'une projection probabiliste ne faisant pas intervenir de termes stochastiques, mais étant uniquement basée sur une modélisation paramétrique et une inférence bayésienne. L'indicateur conjoncturel de fécondité est quant à lui modélisé selon un processus d'évolution en trois phases : phase de haute fécondité, phase de déclin rapide de la fécondité jusqu'en dessous du seuil de renouvellement des générations, et phase de stagnation de la fécondité avec une convergence à long terme vers un niveau à 2,1 enfants par femmes (Alkema et al., 2010).

Il apparaît donc que de nombreux modèles existent pour projeter chacune des trois composantes. Partant du principe qu'aucun modèle ne peut à lui seul rendre compte de l'ensemble des hypothèses possibles sur l'évolution de la mortalité, surtout lorsque ces hypothèses ne sont pas cohérentes entre elles, Kontis et al. (2017) se sont servis de 21 modèles différents de projections probabilistes, dont les résultats ont ensuite été pondérés selon la performance de chacun des modèles, pour au final déboucher sur une seule distribution de probabilité pour les indicateurs souhaités.

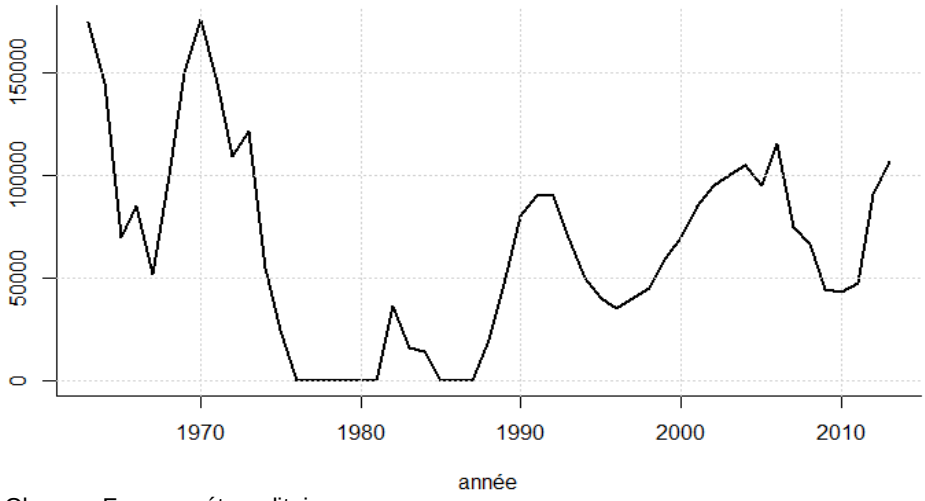
3. Les données pour la France

Afin de disposer de séries longues, on se restreint au champ de la France métropolitaine. On dispose ainsi de 1962 à 2013, de façon détaillée par sexe et âge, du total de la population au 1er janvier de chaque année, du solde migratoire annuel, du nombre de décès et du nombre de naissances selon l'âge de la mère. La dernière année pour lesquelles toutes ces données sont définitives est 2013. En

effet, le solde migratoire n'est pas encore disponible pour l'année 2014. Nous n'utiliserons pas les données provisoires qui sont disponibles jusqu'en 2016, mais qui sont révisées d'une année à l'autre avant de devenir définitives, et sont donc de natures différentes des données définitives. De plus, nous choisissons le même horizon de projection que celui qui a été retenu pour les dernières projections officielles de la France (Blanpain & Buisson, 2016b). L'objectif est donc de projeter la population de 2014 à 2070. De 1962 jusqu'à 1998, les données ne sont pas détaillées par âge au-delà de 100 ans. À partir de 1999 elles sont ventilées en détail jusqu'à 110 ans. On choisit alors de rester sur des catégories d'âges d'un an car les données sont disponibles et on créer une catégorie d'âges supérieure, correspondant aux personnes de 100 ans et plus (en différence de millésimes). Dans la suite de cette partie, on va décrire les données du solde migratoire, de la mortalité et de la fécondité, pour en dégager les invariants, les tendances, et aussi les irrégularités.

Le solde migratoire est, une année donnée, le nombre de personnes vivant en dehors de France métropolitaine venant habiter en France, quelles que soient leurs nationalités, moins le nombre de personnes vivant en France métropolitaine et allant vivre en dehors du territoire. C'est sans doute la composante la plus difficile à mesurer, car bien qu'on puisse estimer les entrées à l'aide du recensement de la population (Brutel, 2014), on ne connaît pas les sorties. Le solde migratoire peut se déduire comme la différence entre l'évolution de la population et le solde naturel. En 1962, en raison du retour des français d'Algérie, le solde migratoire a été exceptionnellement très important, de l'ordre de 860 000 personnes, et depuis 1963 le solde migratoire atteint des niveaux toujours positifs, mais bien plus faibles : il vaut en moyenne 64 000 sur la période 1963-2013. Le solde migratoire présente une grande stabilité au fil des ans à partir des années 1990, même si de fortes fluctuations apparaissent (figure 1), notamment dues aux différentes politiques menées, mais aussi au contexte économique et international. En moyenne sur la période 1990-2013, le solde est de 72 000 et sur les dix dernières années disponibles (2004-2013) il est de 79 000.

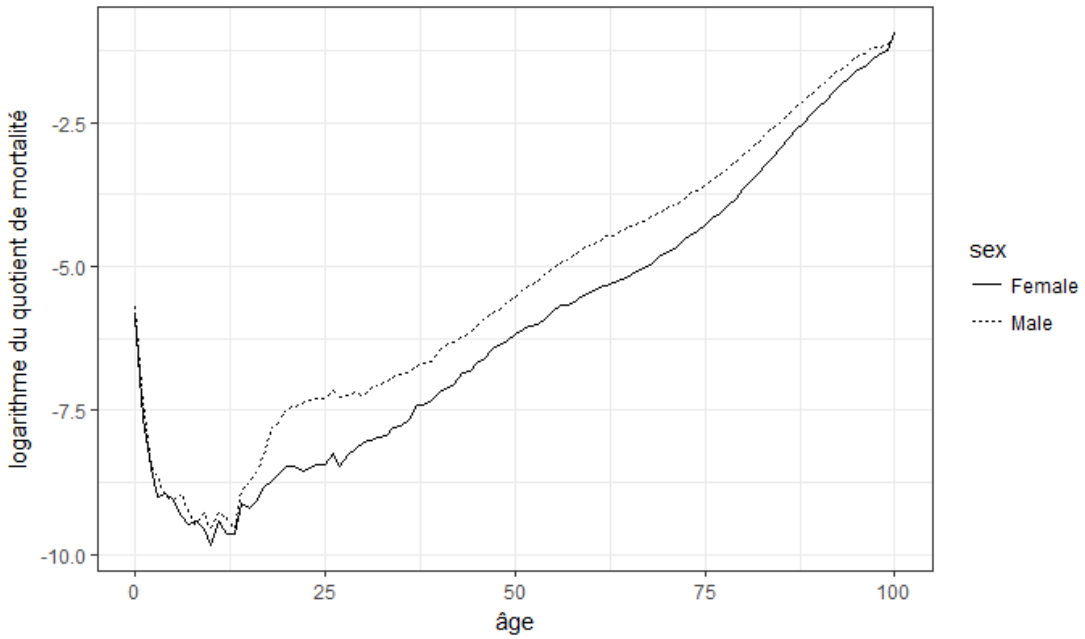
Figure 1 : Évolution du solde migratoire de 1963 à 2013



Champ : France métropolitaine
Source : Insee, estimations de population et statistiques de l'état-civil.

Pour décrire la mortalité, il faut rapporter le nombre de décès à la population à risque correspondant. Cette population correspond à la durée totale passée par l'ensemble des personnes résidant en France, et se compte en personne-année. Elle est environ égale à la population présente au 1er janvier, plus la moitié du solde migratoire. En rapportant le nombre de décès à cette population, on obtient alors les taux de mortalité, qu'on peut détailler par sexe, âge et année. Les taux de mortalité évoluent de manière quasi-exponentielle à partir de 25 ans (figure 2). Avant 25 ans, le profil est différent en raison de la mortalité infantile, plus élevée pour les nouveau-nés. Les taux de mortalité diminuent de la naissance jusqu'à 10 ans environ, avant d'augmenter régulièrement. Vers 18 ans, la mortalité des hommes devient nettement plus élevée que celle des femmes, et l'écart reste présent

Figure 2 : Logarithme des taux de mortalité en 2013 selon le sexe et l'âge



Champ : France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

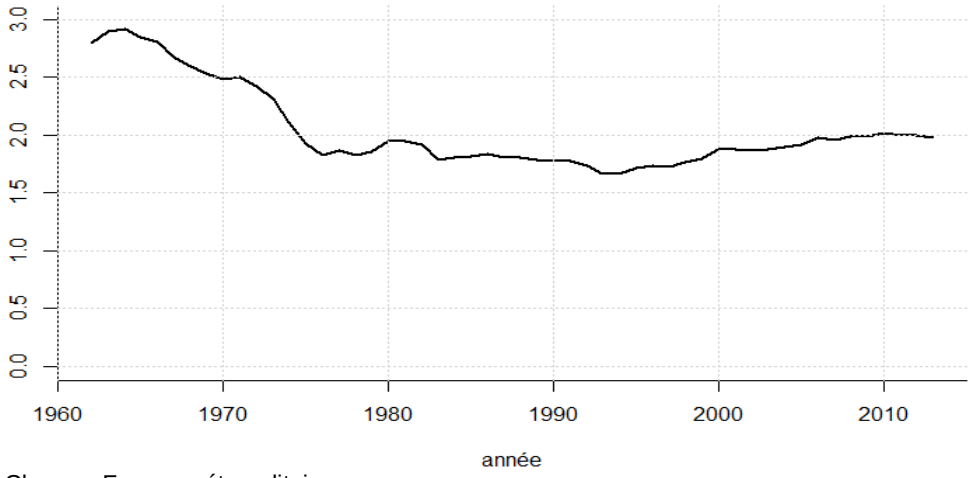
tout au long de la vie, avec une ampleur plus ou moins marquée selon l'âge.

Figure 3 : Évolution des logarithmes des taux de mortalité de 1962 à 2013, pour différents âges

Champ: France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

Le logarithme des taux de mortalité, à âge et sexe fixés, diminuent de façon quasi-linéaire avec le temps. Cela est surtout vrai pour les âges élevés, mais ne semble pas tout à fait le cas pour les âges plus jeunes. Le logarithme du quotient de mortalité à 10 ans diminue de plus en plus vite par exemple. Au contraire, à 30 ans, le logarithme du quotient de mortalité a ralenti sa décroissance, jusqu'à stagner pour les hommes, du début des années 1980 au milieu des années 1990, date à laquelle la mortalité a brusquement diminué pour cet âge, et continue depuis lors à diminuer régulièrement et apparemment linéairement. Cette stagnation de la mortalité chez les jeunes adultes des années 1980 et 1990, alors que la tendance générale est une diminution constante de la mortalité, est à relier à l'épidémie de Sida qui atteignit la France au début des années 1980. De façon générale, les taux de mortalité baissant régulièrement, l'espérance de vie à la naissance augmente chaque année, et ce plus rapidement pour les hommes que pour les femmes (Blanpain, 2016), même s'il arrive parfois que l'espérance de vie baisse d'une année sur l'autre comme c'était le cas en 2015 pour des raisons conjoncturelles (Bellamy & Beaumel, 2016).

Figure 4 : Évolution de l'indice conjoncturel de fécondité, de 1962 à 2013

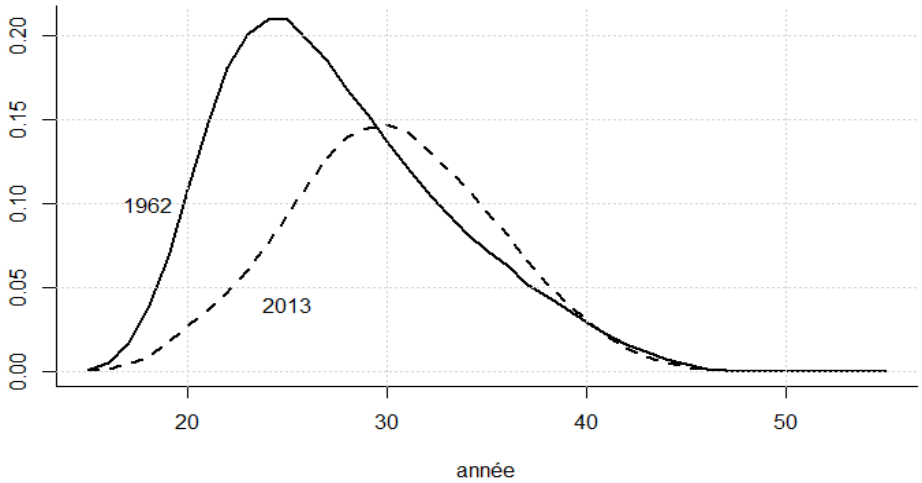


Champ : France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

Depuis le début des années 1970, l'indice conjoncturel de fécondité (ICF) a fortement diminué : il est passé de 2,9 enfants par femme en 1964 à 1,8 enfant par femme en 1976 (figure 4). Depuis lors il s'est stabilisé autour d'une valeur moyenne de 1,85 enfant par femme. On observe toutefois une tendance à l'augmentation de l'ICF depuis le milieu des années 1990.

Le taux de fécondité à un âge donné est défini comme le rapport entre le nombre de naissances de bébés dont la mère est âgée de cet âge et le nombre de femmes du même âge sur l'année considérée. Ce nombre correspond au nombre de femmes au 1er janvier de l'année plus la moitié du solde migratoire correspondante moins la moitié des décès enregistrés pour cette population. Le profil des taux de fécondité par âge suit une courbe en cloche : la probabilité d'avoir un enfant dans l'année augmente avec l'âge à partir de 15 ans jusqu'à atteindre un pic, puis diminue continûment jusqu'à devenir nulle ou presque aux alentours de 50 ans.

Figure 5 : taux de fécondité par âge en 1962 et en 2013



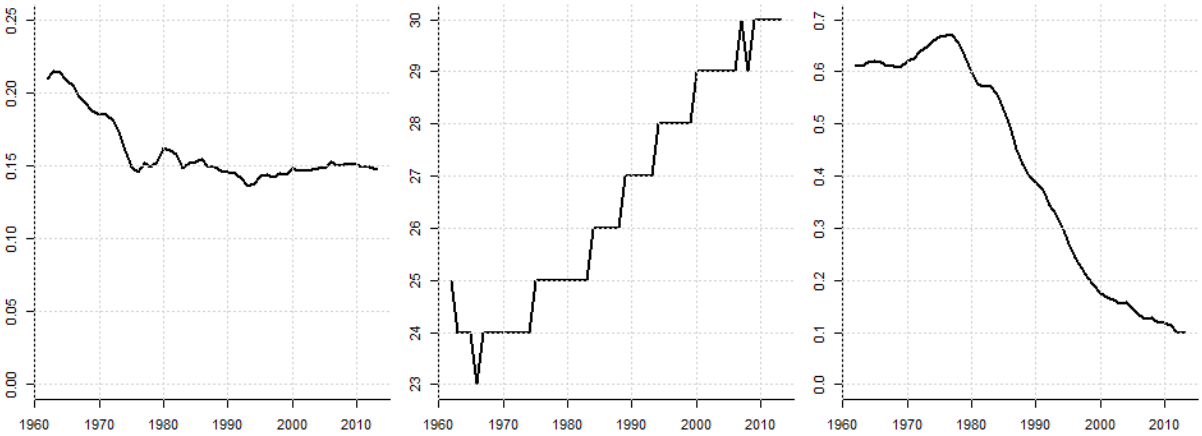
Champ : France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

Avec le temps, cette distribution par âge à tendance à se décaler vers la droite : l'âge auquel le maximum de fécondité est atteint augmente (figure 5). En 1970, le taux de fécondité était maximal à 24 ans alors qu'en 2013, ce maximum est atteint à l'âge de 30 ans. Le niveau du maximum de

fécondité atteint dans l'année quant à lui n'évolue guère depuis le milieu des années 1970 : il fluctue autour de 0,15. Le pic de fécondité se déplaçant vers la droite, la distribution des taux selon l'âge devient de plus en plus symétrique, comme en témoigne la mesure d'asymétrie qui diminue rapidement vers 0 (figure 6).

Contrairement aux taux de mortalité, les taux de fécondité n'évoluent pas de façon régulière avec le temps. Par exemple, le taux de fécondité à 30 ans a diminué entre le début des années 1960 et le milieu des années 1970, et il augmente depuis lors avec un ralentissement à partir des années 2000. Le taux de fécondité à 20 ans diminue depuis les années 1970, mais à la fin des années 1990 il a connu un léger regain pendant quelques années, avant de diminuer à nouveau à un rythme beaucoup

Figure 6 : Évolution du maximum de fécondité, de l'âge auquel le maximum de fécondité est atteint et de l'asymétrie de la distribution des taux de fécondité par âge, de 1962 à 2013

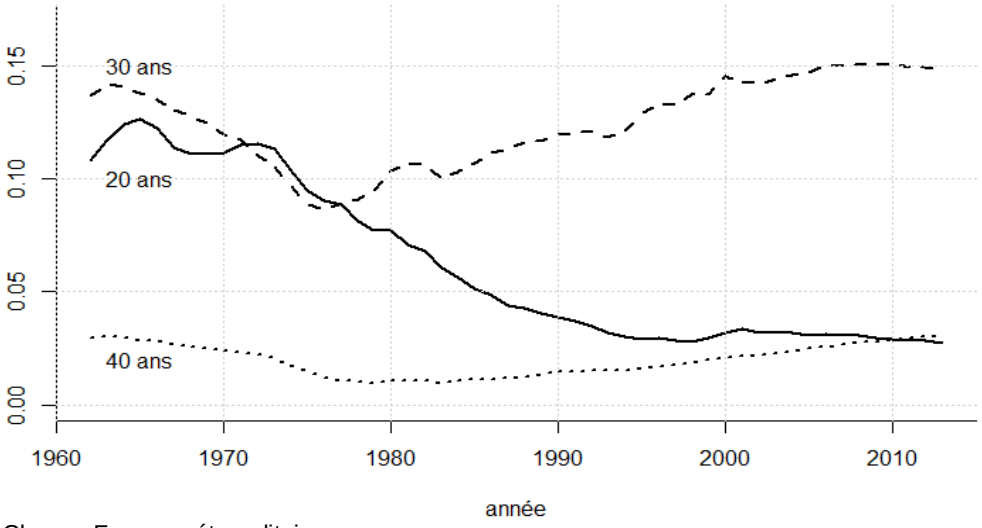


Note : À gauche le maximum du quotient de fécondité atteint dans l'année, au centre l'âge minimum auquel ce maximum est atteint et à droite la mesure d'asymétrie de la répartition des taux selon l'âge.
Champ : France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

plus lent que lors des décennies précédentes. L'évolution n'est ni monotone ni linéaire, ce qui indique la difficulté à prolonger ces courbes dans le futur.

Pour résumer, le solde migratoire de France métropolitaine apparaît stable sur longue période, avec d'importantes fluctuations qui semblent difficiles à prévoir. La mortalité évolue depuis plusieurs décennies dans le même sens, avec une diminution quasi-linéaire du logarithme des taux de mortalité à tous les âges et une diminution de l'écart d'espérance de vie entre les femmes et les hommes. L'évolution récente de la fécondité est plus complexe à cerner, mais il se dégage que l'ICF est stabilisé à un niveau moyen légèrement inférieur à 2 enfants par femme et que la distribution de la fécondité par âge se modifie continûment avec un déplacement du pic de fécondité vers des âges plus élevés et une distribution de plus en plus symétrique. Dans la partie suivante, on propose une modélisation pour chacune des trois composantes de l'évolution de la population prenant en compte ces observations et s'inspirant de modèles déjà développés à l'international et succinctement décris dans la troisième partie.

Figure 7: Évolution des taux de fécondité à différents âges, de 1962 à 2013



Champ: France métropolitaine.
Source : Insee, estimations de population et statistiques de l'état-civil.

4. Méthodes et modélisations

Dans toute la suite de l'article, on utilisera les notations suivantes :

P(a,n,s) : nombre de personnes au premier janvier de l'année *n*, de sexe *s* nées l'année *n-a*.

D(a,n,s) : nombre de décès durant l'année *n*, de personnes nées l'année *n-a* et de sexe *s*.

N(a,n,s) : nombre de bébés de sexe *s* nés vivants durant l'année *n* et dont la mère est née l'année *n-a*.

M(a,n,s) : nombre de personnes entrées moins nombre de personnes sorties de France métropolitaine, durant l'année *n*, de sexe *s* et nées l'année *n-a*. C'est le solde migratoire de l'année *n*, pour les personnes de sexe *s* et nées l'année *n-a.*

De plus, pour faciliter les équations ensuite, on définit *P(0,n,s)* comme le nombre de naissances vivantes l'année *n* de bébés de sexe *s*. Par ailleurs, *D(0,n,s)* et *M(0,n,s)* sont bien définies par la description ci-dessus et correspondent respectivement, pour chaque année *n* et sexe *s*, au nombre de décès de bébés nés durant l'année *n* et au nombre de nouveau-nés entrés moins le nombre de nouveau-nés sortis du territoire. On supposera que les âges de reproduction des femmes se situent entre 15 et 55 ans inclus, et de ce fait on considérera que *N(a,n,s)=0* pour *a*<=14 et *a*>=56.

On définit de plus les populations à risque, pour les décès et les naissances. Les populations à risque sont comptées en personnes-années et dépendent du nombre de personnes observées mais également de la période de temps sur laquelle ces personnes sont présentes. Pour les décès, cela correspond à la population au 1er janvier de l'année plus la moitié du solde migratoire (si on considère que les flux d'entrées et de sorties sont répartis uniformément tout au long de l'année).

$R_D(a,n,s) = P(a,n,s) + 0.5 \cdot M(a,n,s)$, si a>=1

$R_D(0,n,s) = 0.5 \cdot P(0,n,s) + 0.5 \cdot M(0,n,s)$, pour a=0.

Pour les naissances, le nombre de personnes-années à risque est le nombre moyen de femmes sur l'année, en supposant le flux des migrations et des décès uniformes :

$R_N(a,n) = P(a,n,\ femmes) + 0.5 \cdot M(a,n,\ femmes) - 0.5 \cdot D(a,n,\ femmes)$.

On note de plus $M(n) = \sum_{a,s} M(a,n,s)$, $N(a,n) = N(a,n,\ filles) + N(a,n,garçons)$ et $N(n,s) = \sum_{a} N(a,n,s)$. Lorsquon notera des lois normales, on indiquera la moyenne et lécart-type (et non la variance).

Les migrations

On projette directement le solde migratoire total à laide dun modèle autorégressif dordre 1, où M_{lt} représente le solde migratoire de long terme et ε_M un bruit blanc :

$$M(n) = M_{lt} + \rho_M (M(n-1) - M_{lt}) + \varepsilon_M(n)$$

$$\varepsilon_M(n) \overset{i.i.d.}{\sim} N(0, \sigma_M)$$

Afin davoir un processus stationnaire, on impose la contrainte $|\rho_M| \leq 1$. Cette modélisation reflète le fait quon estime que le solde migratoire va continuer à être stable en moyenne et va osciller autour dune tendance de long terme. Lamplitude des oscillations possibles pour lavenir est déterminée par les amplitudes passées. De plus, on fixe un a priori très informatif sur la tendance de long terme en supposant, comme cest le cas dans les travaux de Blanpain et Buisson (2016a), que ce niveau peut être estimé à partir du solde migratoire moyen sur la période récente, à savoir 80 000 personnes. La loi *a priori* pour la tendance de long terme est alors $M_{lt} \sim N(80000,10000)$. Les paramètres M_{lt}, σ_M, μ_M et σ_M sont estimés par inférence bayésienne à partir des données du solde migratoire sur la période 1995-2013.

Pour projeter le solde migratoire total, on tire aléatoirement 1000 fois les paramètres du modèle selon leur distribution *a posteriori* et pour chaque jeu de paramètres on simule lévolution du solde migratoire selon le processus autorégressif dordre 1. Une fois projeté le solde migratoire, on décompose ce dernier par sexe et âge selon des taux fixes calculés à partir de la répartition du solde migratoire par sexe et âge sur la période récente, et lissés, comme dans Blanpain et Buisson (2016a).

Mortalité

Comme on la vu, le logarithme des taux de mortalité par âges, semblent évoluer linéairement avec le temps. Néanmoins, lévolution des taux de mortalité dans le temps se fait à un rythme différent pour chaque âge. On modélise directement le nombre de décès observés par une loi de Poisson, qui dépend du taux de mortalité et de la population à risque. Cette dernière correspond au nombre de personnes-années présentes en France métropolitaine lannée considérée. La loi de Poisson est couramment utilisée pour modéliser un nombre dévénements arrivant sur une période de temps donné. Elle est souvent retenue pour modéliser le nombre de décès dans les travaux de démographie. On retient la modélisation suivante, où $\varepsilon_D(a,n,s)$ correspond au taux de mortalité lannée n, des personnes de sexe s et dâge a :

$$D(a,n,s) \sim Poisson(\mu_D(a,n,s) \cdot R_D(a,n,s))$$

$$\log(\mu_D(a,n,s)) = \beta^0 + \beta_a^{âge} + \beta_{a,s}^{âge:sexe} + \beta_{a,n}^{âge:année} + \varepsilon_{D,1}(a,n,s)$$

Les $\varepsilon_{D,1}$ sont des termes derreur indépendants et identiquement distribués selon une loi normale centrée et décart-type $\sigma_{D,1}$. Le paramètre β_0 est une constante, le paramètre $\beta^{âge}$ donne la répartition moyenne selon lâge du logarithme des taux de mortalité. Enfin, il y a deux termes croisant deux

dimensions : $\beta^{âge:sexe}$ qui permet d'estimer un effet du sexe spécifique à chaque âge et un paramètre $\beta^{âge:année}$ qui est un effet du temps spécifique à chaque âge. On notera donc que l'évolution temporelle du logarithme des taux de mortalité par âge est le même pour les femmes et les hommes puisqu'on n'a pas spécifié de terme croisant l'année et le sexe. La raison est qu'on a voulu limiter le nombre de paramètres à estimer. En introduisant en plus un terme de croisement année-sexe on s'est aperçu que les distributions *a posteriori* n'étaient pas bien estimées en raison d'une non-convergence des chaînes de Markov. À un troisième niveau, on modélise certains des paramètres par des modèles linéaires dynamiques. Pour le paramètre $\beta^{âge:année}$ cela permet de décomposer l'évolution temporelle, par âge, en un niveau ($\upsilon^{âge:année}$) et une tendance ($\breve{u}^{âge:année}$) :

$$\beta_{a,n}^{âge:année} = \theta_{a,n}^{âge:année} + \eta(a,n)$$
$$\theta_{a,n}^{âge:année} = \theta_{a,n-1}^{âge:année} + \delta_{a,n}^{âge:année} + \upsilon(a,n)$$
$$\delta_{a,n}^{âge:année} = \delta_{a,n-1}^{âge:année} + \omega(a,n)$$

Les termes η, ζ et χ sont des termes d'erreurs indépendants, suivant une loi normale centrée.

Pour projeter dans le futur les taux de mortalité par âge, il suffit donc, une fois estimée la distribution *a posteriori* de l'ensemble des paramètres du modèle, de générer des nouveaux termes de tendance, puis des nouveaux termes de niveaux et enfin des nouveaux paramètres $\beta^{âge:année}$, jusqu'à l'horizon souhaité.

Fécondité

Pour la fécondité, on choisit de procéder en trois étapes. On projette en premier lieu l'indice conjoncturel de fécondité selon un modèle autorégressif d'ordre 1. L'ONU utilise la même méthode pour sa troisième phase d'évolution de la fécondité, en supposant que l'indice conjoncturel de fécondité tend vers 2,1 pour tous les pays (Alkema et al., 2010). Par rapport à la méthode de l'ONU, on choisit de plus d'estimer les paramètres du modèle par inférence bayésienne, et non pas par maximum de vraisemblance. On reste ainsi dans un cadre entièrement bayésien, pour l'ensemble de nos estimations et de nos projections. Le modèle est le suivant :

$$ICF(n) = ICF_{lt} + \rho_F\,(ICF(n-1) - ICF_{lt}) + \varepsilon_F(n)$$

où $ICF(n) = \sum_{a=15}^{55} \frac{N(a,n,\,filles) + N(a,n,garçons)}{R_F(a,n)}$ est l'indice conjoncturel de fécondité de l'année *n*. Comme pour le solde migratoire, on simule, après estimation par inférence bayésienne, 2000 trajectoires possibles d'évolution de cet indice jusqu'à l'horizon souhaité.

La deuxième étape consiste à projeter, indépendamment de la projection de l'indice conjoncturel de fécondité, les taux de fécondité par âge ε_F. Ceux-ci sont définis, comme dans le cas de la mortalité, à travers la modélisation des naissances par un processus de Poisson :

$$N(a,n) \sim Poisson(\mu_F(a,n) \cdot R_F(a,n))$$

où on rappelle que *N(a,n)* correspond au nombre de naissances l'année *n* données par des mères nées l'année *n-a*. Suivant la méthode proposée par Bijak (2015) qui s'est inspiré de la méthode de Lee-Carter, on modélise ensuite le logarithme du taux de fécondité comme la somme d'un effet fixe de l'âge, d'un effet du temps dont l'intensité et la direction sont différents pour chaque âge et d'un effet de génération :

$$\log(\mu_F(a,n)) = \alpha_a + \beta_a \kappa_n + \gamma_{n-a} + \varepsilon_{F,1}(a,n)$$

L'effet temporel κ et l'effet de génération γ évoluent selon des processus autorégressifs d'ordre 1 :

$$\kappa_n = \varphi_0 + \varphi_1 \kappa_{n-1} + \xi(n)$$
$$\gamma_{n-a} = \psi_0 + \psi_1 \gamma_{n-a-1} + \zeta(n)$$

où les termes d'erreurs ξ et ζ suivent des lois normales d'espérances nulles. Cette fois encore, tous les paramètres sont estimés par inférence bayésienne, pour ensuite produire 1000 simulations des taux de fécondité, pour chaque âge et chaque année future. Ces taux projetés prolongent des tendances linéaires, même si les paramètres φ_1 et ψ_1 peuvent, s'ils sont de norme strictement plus petite que 1, amener à annuler l'effet temporel ou l'effet de génération à long terme. Les estimations donnent une distribution *a posteriori* de φ_1 et ψ_1 qui sont très proches de 1. Il en résulte qu'à moyen terme, les taux de fécondité deviennent anormalement élevés pour certains âges, ce qui conduit à des indices conjoncturels de fécondité bien plus élevés que ceux projetés dans la première étape.

La troisième étape, consiste alors à corriger les taux de fécondité par âge, pour chaque année le caler sur l'indice conjoncturel de fécondité projeté en premier lieu. Pour cela on multiplie simplement l'ensemble des taux, une année donnée, par une constante.

Projections par la méthode des composantes

La méthode des composantes permet de faire évoluer la population d'une année sur l'autre en remarquant que la population au 1er janvier d'une année donnée est égale à la population au 1er janvier de l'année précédente, plus le nombre de naissances ayant eu lieu l'année précédente, moins le nombre de décès et plus le solde migratoire. Cela se traduit par les équations suivantes :

$P(a,n,s) = P(a-1,n-1,s) - D(a-1,n-1,s) + M(a-1,n-1,s)$ pour a>=1 et $P(0,n,s) = N(n,s)$.

Le nombre de décès et de naissances sont obtenus chaque année par tirage aléatoire selon une loi de Poisson (voir les modélisations). Pour cela, il faut déterminer les personnes à risque pour les décès et les femmes à risque pour les naissances. On commence alors par calculer les décès pour chaque âge, excepté pour les décès des nouveau-nés. On en déduit ensuite les femmes à risque à chaque âge entre 15 ans et 55 ans (on doit en effet pour cela connaitre le solde migratoire et le nombre de décès). Enfin, on calcule le nombre de décès des nouveau-nés.

La répartition du nombre de naissances une année donnée, entre naissances de garçons et naissances de filles est déterminée par le sexe ratio qu'on fixe à 1,05, conformément aux constats passés.

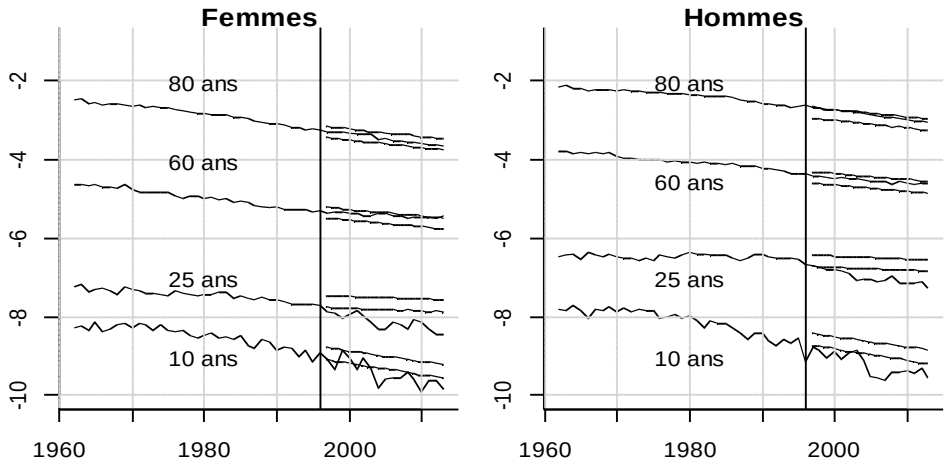
5. Validations des modèles

Une manière de tester les modèles utilisés est de séparer les données portant sur le passé en deux catégories : une partie, environ deux tiers, est utilisée pour estimer les modèles et la partie restante, environ un tiers, est utilisée pour confronter les estimations du modèle à la réalité.

Pour la mortalité, on choisit d'estimer le modèle sur la période 1962-1995 et de comparer les résultats sur la période 1996-2013. Pour la fécondité on estime les modèles sur la période 1975-2000 et on compare les résultats sur la période 2001-2013. On s'aperçoit alors que le logarithme des taux de mortalité est bien projeté aux grands âges (à partir de 35-40 ans environ), mais que la modélisation retenue donne des diminutions moins rapides de ces taux que ce qui est observé en réalité (figure 8). Cela est dû au fait que pour les très jeunes âges, le logarithme du taux de mortalité est légèrement concave et non linéaire. De plus, pour les jeunes adultes, les taux de mortalité ont plus ou moins stagné dans les années 1980 et 1990 avant de baisser fortement. Le modèle n'a pas pu anticiper cette baisse soudaine.

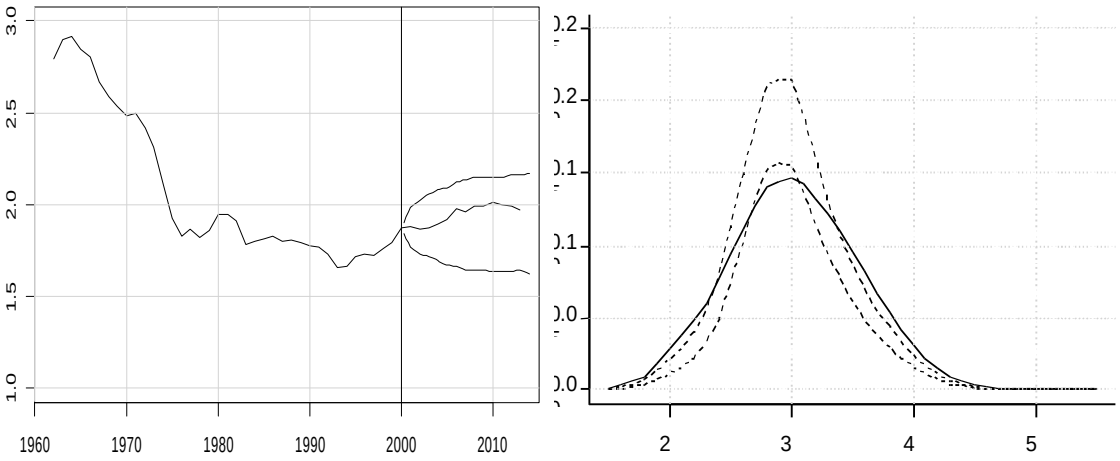
En ce qui concerne la fécondité, l'intervalle de confiance à 95 % des projections probabilistes de l'ICF contient bien l'ICF observé. Mais lorsqu'on regarde la distribution des taux de fécondité par âge, on se rend compte que la méthode utilisée conduit à une distribution plus resserrée que ce qui est réellement observé (figure 9). La déformation de la distribution des taux de fécondité par âge est donc un peu trop forte dans nos projections.

Figure 8 : Logarithme des taux de mortalité, observés (1962-2013) et projetés (1996-2013)



Note : En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % des projections probabilistes et en trait plein les taux de mortalité réels (de 1962 à 2013).
Champ : France métropolitaine.
Sources : Insee estimations de population et statistiques de l'état-civil (taux de mortalité), calculs auteur (projections probabilistes)

Figure 9 : ICF et taux de fécondité par âge, observés (1975-2013) et projetés (2001-2013)



Note : À gauche, l'indicateur conjoncturel de fécondité (ICF) et à droite la distribution des taux de fécondité par âge en 2013. En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % des projections probabilistes et en trait plein l'ICF et les taux de fécondité réels (de 1962 à 2013).
Champ : France métropolitaine.
Sources : Insee estimations de population et statistiques de l'état-civil (taux de fécondité), calculs auteur (projections probabilistes)

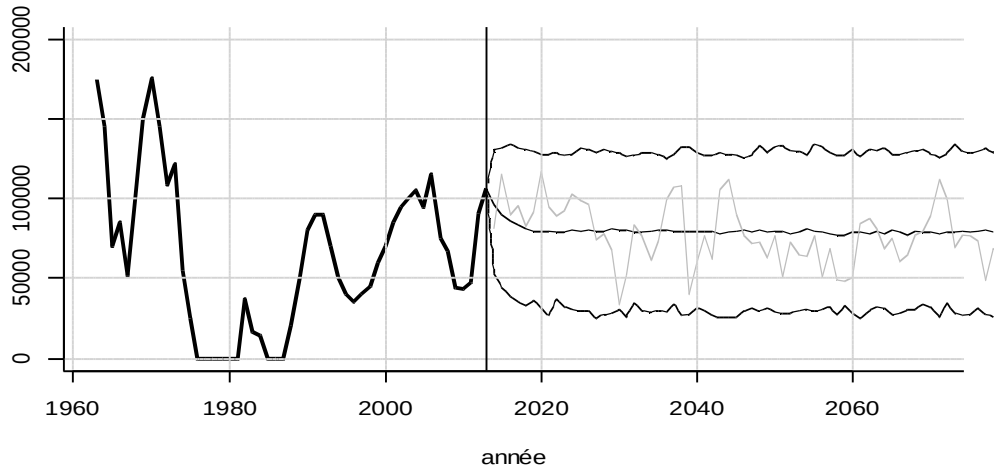
6. Résultats des projections probabilistes bayésiennes : une hausse de la population encore jusqu'en 2070 très probable

Les paramètres des modèles pour le solde migratoire, la mortalité et la fécondité ont été estimés par inférence bayésienne à l'aide du logiciel libre *Stan*. Pour chacun d'eux on a simulé 1000 valeurs selon leur loi *a posteriori*. On a ensuite généré 1000 trajectoires d'évolution possible pour le solde migratoire, les taux de mortalité par sexe et âge et les taux de fécondité par âge. Au final on peut obtenir 1000 estimations de n'importe quel indicateur démographique dérivé de ces trois composantes, dont notamment la taille de la population totale. On en déduit alors des intervalles de confiance à 95 % ou à 80 % qui contiennent respectivement 95 % ou 80 % des estimations.

Projections des migrations : une incertitude forte et constante

Le solde migratoire projeté suit une évolution stable car cela a été spécifié ainsi dans le modèle. La médiane des 1000 trajectoires possibles diminue dans les premières années de projections puis se stabilise rapidement à 79 000 (figure 10). L'intervalle de confiance lui aussi reste constant au cours du

Figure 10 : Solde migratoire passé et projeté



Note : En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % et en trait plein la médiane des distributions *a posteriori*. La courbe en gris clair représente une des 1000 simulations.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070)

temps : avec une probabilité de 95 % le solde migratoire se maintiendra entre 29 000 et 129 000 chaque année. Cette amplitude est due aux larges fluctuations observées dans le passé, dépasse légèrement les minimum et maximum observés respectivement en 1996 et 2006.

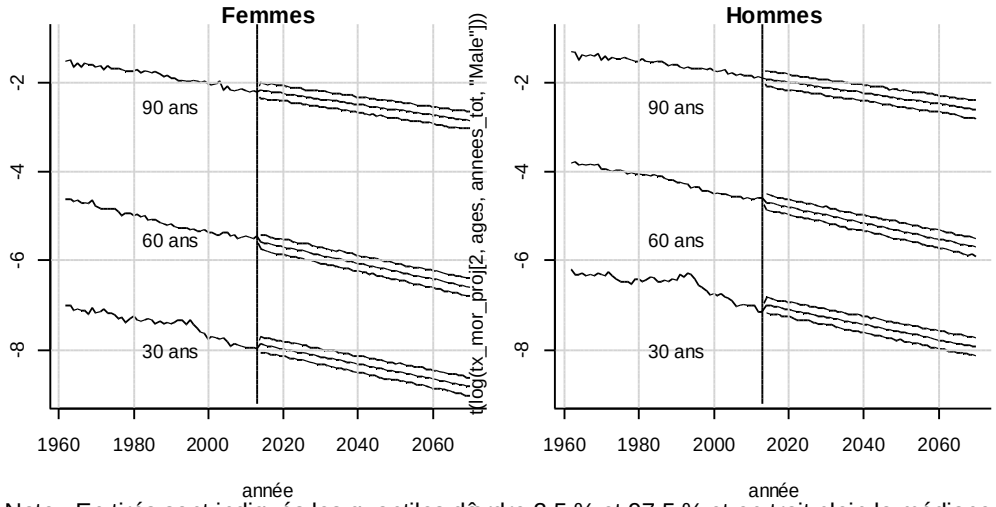
Projections de la mortalité : peu d'incertitude projetée compte tenu des évolutions passées

Le modèle pour la mortalité prévoit que les taux de mortalité par âge continueront à diminuer linéairement, selon la même tendance pour les hommes et pour les femmes (figure 11). L'incertitude sur les taux de mortalité projetés n'augmente presque pas avec le temps. Cela provient du fait la variance des erreurs de niveau et de tendance σ_δ et σ_γ sont très faibles comparées à la variance du terme d'erreur σ_ε. Les erreurs ne s'accumulent donc pas avec le temps. Cela est dû au fait que les tendances observées sont très linéaires.

Du fait de la diminution constante des taux de mortalité, l'espérance de vie continuera à croître dans les décennies qui viennent, pour les hommes comme pour les femmes. Les résultats du modèle indiquent que l'espérance de vie à la naissance des femmes sera avec une probabilité de 95 % comprise entre 91,2 ans et 92,8 ans en 2070 et celle des hommes entre 87,4 ans et 89,4 ans (figue

12). L'écart d'espérance de vie entre les femmes et les hommes va probablement continuer à se résorber pour atteindre 3,6 ans en 2070 (entre 3,3 ans et 3,9 ans avec une probabilité de 95 %).

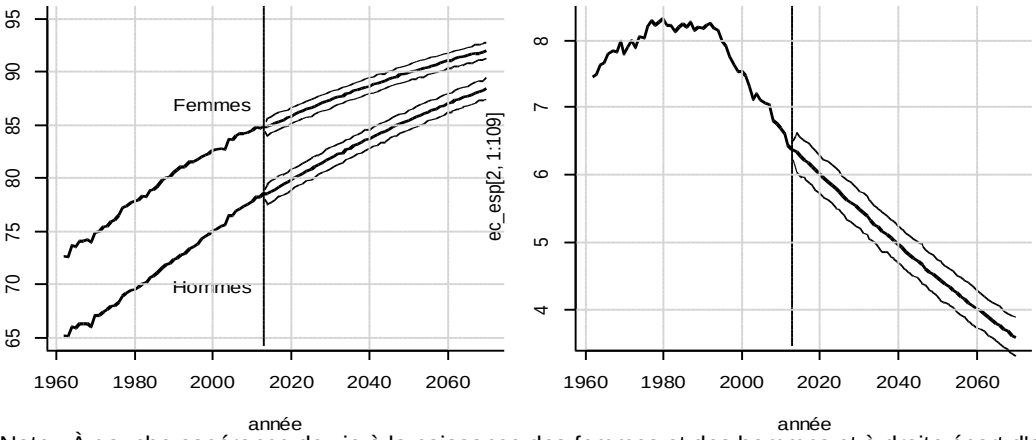
Figure 11 : Évolution du logarithme des taux de mortalité par âge, estimés et projetés



Note : En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % et en trait plein la médiane des distributions *a posteriori*.
Champ : France métropolitaine.
Source : Insee, calcul auteur

Projections de la fécondité : des maternités plus tardives et réparties de façon plus symétrique autour de l'âge modal

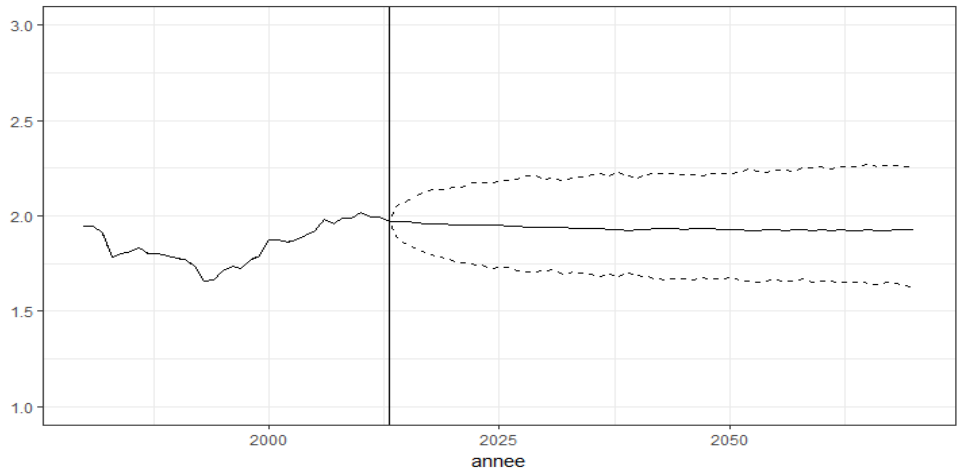
Figure 12 : Évolution de l'espérance de vie estimée et projetée et de l'écart d'espérance de vie entre les femmes et les hommes



Note : À gauche espérance de vie à la naissance des femmes et des hommes et à droite écart d'espérance de vie à la naissance entre les femmes et les hommes. En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

La médiane de l'ICF de long terme est de 1,93, légèrement en-dessous de la moyenne de la loi a priori fixée à 1,95 (figure xx). Selon la modélisation retenue, l'ICF sera compris avec une probabilité de 95 % entre 1,63 et 2,26 enfants par femme en 2070. Contrairement aux projections du solde migratoire et des taux de mortalité, l'intervalle de confiance à 95 % devient de plus en plus large avec le temps. L'incertitude sur la fécondité future devient donc de plus en plus grande malgré le fait d'avoir fixé un ICF de long terme dans la modélisation.

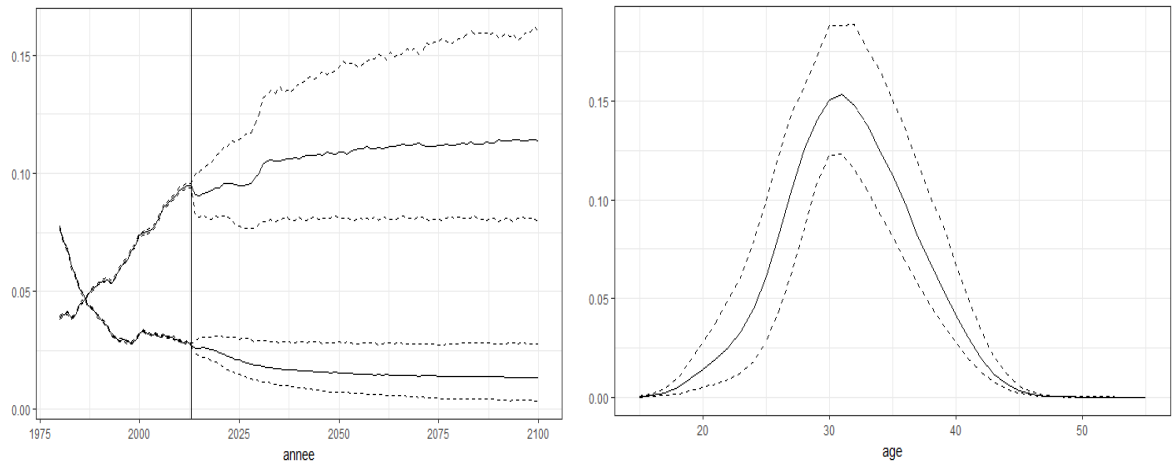
Figure 13 : Évolution de l'indice conjoncturel de fécondité, estimé et projeté.



posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

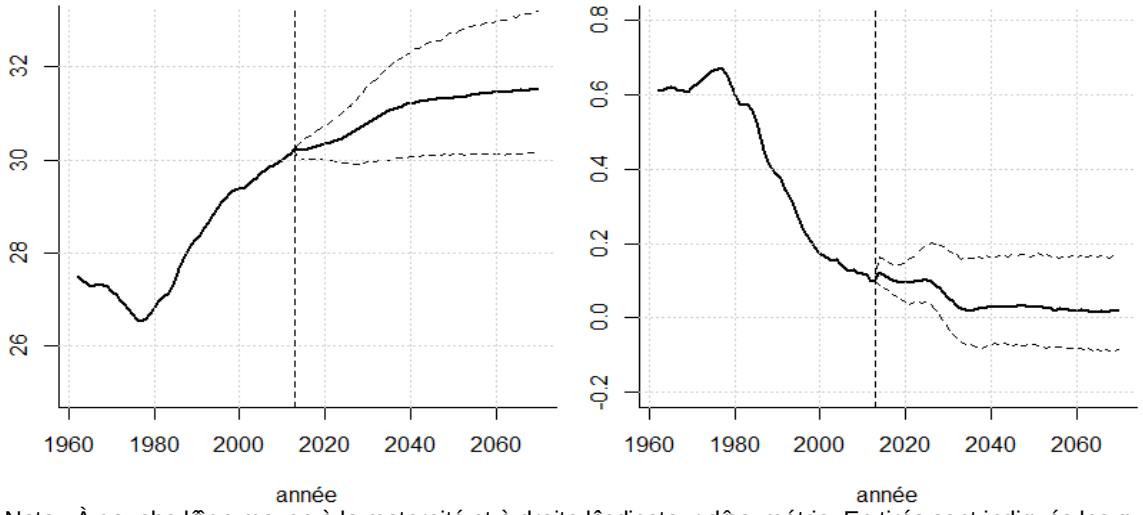
Les taux de fécondité par âge commencent à se stabiliser à partir de 2050 (figure 14). L'âge moyen à la maternité augmente mais à un rythme de moins en moins rapide, pour atteindre une valeur comprise entre 30,2 et 33,2 ans en 2070 (intervalle de confiance à 95 %). La distribution des taux de fécondité par âge se décale alors de plus en plus vers la droite et devient de plus en plus symétrique comme le montre l'évolution de l'indicateur d'asymétrie dont la médiane tend vers 0 (figure 15).

Figure 14 : Évolution des taux de fécondité estimés et projetés



Note : Sur la figure de gauche : taux de fécondité à 20 ans (en bas à droite) et à 35 ans (en haut à droite). Sur la figure de droite : taux de fécondité en 2070, pour tous les âges. En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee, calcul auteur

Figure 15 : Évolution de l'âge moyen à la maternité et de l'asymétrie de la distribution des taux de fécondité par âge

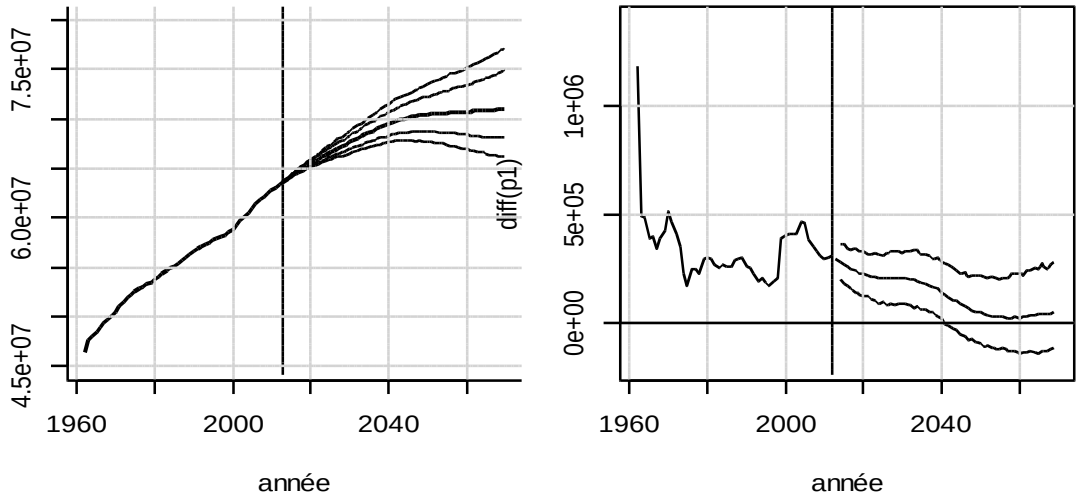


Note : À gauche l'âge moyen à la maternité et à droite l'indicateur d'asymétrie. En tirés sont indiqués les quantile d'ordre 2,5% et 97,5% et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

Projections de la population totale : une croissance vraisemblablement forte d'ici à 2040, et beaucoup plus faible ensuite

La population totale de la France métropolitaine va continuer à augmenter pour atteindre en 2070 un niveau compris entre 66,1 millions et 77,2 millions avec une probabilité de 95 % et entre 68,1 millions et 75,0 millions avec une probabilité de 80 % (figure 16). La projection médiane correspond à un niveau de 71,0 millions d'habitants en 2070. La population de France métropolitaine pourrait donc augmenter continûment tout au long des cinquante prochaines années ou bien augmenter et commencer à décliner vers 2050. Il y a, selon la modélisation retenue ici, une probabilité de 1 % que la population commence à baisser dès 2040 (c'est-à-dire que la population atteigne son maximum en 2040) et une probabilité de 19 % pour 2050. L'incertitude sur la taille de la population d'après la modélisation retenue est assez faible jusque vers 2040-2050, puis augmente plus fortement avec les années ensuite.

Figure 16 : Évolution de la taille totale de la population et de la croissance annuelle de la population, passée et projetée

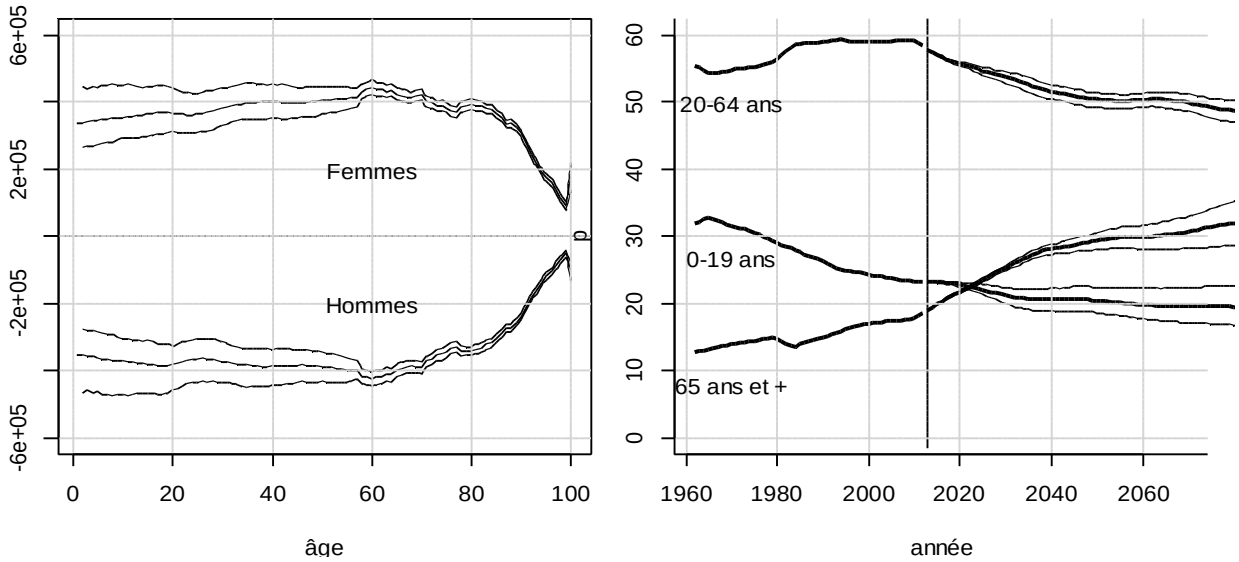


Note : À gauche la taille de la population et à droite la croissance annuelle de la population. En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 %, en pointillés ceux d'ordre 10 % et 90 % et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

La structure de la population va également être modifiée, comme en témoigne la pyramide des âges en 2070 dont la base est bien plus droite et fine que la pyramide des âges actuelle. Certains groupes d'âges vont ainsi baisser en proportion, notamment les plus jeunes (figure 17) : la part des 0-19 ans va continuer à diminuer lentement pour atteindre en 2070 un niveau médian de 19 %, celles des 20-64 ans suit le même profil, avec en 2070 un niveau médian de 50 %. Au contraire, la part des 65 ans et plus dans la population va probablement continuer d'augmenter pour en 2070 être plus importante que la part des moins de 20 ans. Elle est passée de 13% en 1962 à 19% en 2013 et pourrait atteindre, avec une probabilité de 95 %, entre 28 % et 33 % de la population en 2070.

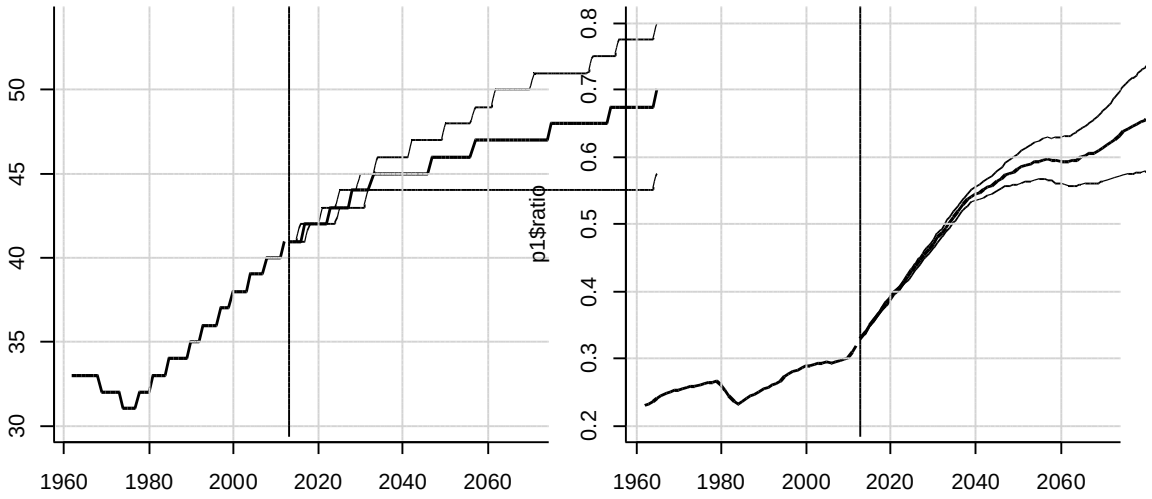
La population va donc continuer à vieillir. L'âge médian de la population, qui est en 2013 de 41 ans, pourrait être, avec une probabilité de 95 %, compris entre 44 et 50 ans en 2070. En conséquence, le ratio entre les personnes de 65 ans et plus et les personnes de 20 à 64 ans risque de fortement augmenter dans les années à venir. L'augmentation rapide et linéaire de ce ratio d'aujourd'hui au début des années 2040 est principalement due au vieillissement des générations nombreuses nées pendant le baby-boom. En effet, les personnes nées en 1946, début du baby-boom, ont eu 65 ans en 2011 et celles nées à la fin du baby-boom en 1975 auront 65 ans en 2040. Le ratio des 65 ans ou plus sur les 20-64 ans, aujourd'hui de 0,33, atteindrait selon les modèles utilisés, une valeur comprise entre 0,56 et 0,67 avec une probabilité de 95 % en 2070 (figure 18).

Figure 17 : Pyramide des âges en 2070 et évolution de la proportion de certains groupes d'âge.



En tirés sont indiqués les quantile d'ordre 2,5% et 97,5% et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

Figure 18 : Évolution de l'âge médian de la population et du ratio des personnes de 65 ans et plus sur les personnes de 20-64 ans.



Note :
À gauche, l'évolution de l'âge médian (en année) et à droite l'évolution du ratio des personnes de 65 ans et plus sur les personnes de 20 à 64 ans. En tirés sont indiqués les quantiles d'ordre 2,5 % et 97,5 % et en trait plein la médiane des distributions a posteriori.
Champ : France métropolitaine.
Source : Insee estimations de population et statistiques de l'état-civil (1962-2013), calculs auteur (2013-2070).

On peut comparer ces projections probabilistes aux projections déterministes réalisées par l'Insee. Les projections déterministes concernant la France métropolitaine ne concernent que la période 2013-2050 (https://www.insee.fr/fr/statistiques/2859843). Le scénario central retenu conduit à une population légèrement plus élevée que la médiane de nos projections probabilistes : en 2050 la population de France métropolitaine atteindrait 71,7 millions d'habitants selon la première projection et 70,5 millions selon la deuxième. Par ailleurs, l'intervalle de confiance estimé par les projections probabilistes est bien plus faible que l'intervalle entre le scénario de population haute et le scénario de

population basse, qui sont les scénarios extrêmes des projections déterministes. L'écart entre les deux scénarios déterministes extrêmes est de 11,1 millions d'habitants en 2050, alors que l'intervalle de confiance des projections déterministes cette même année est de 5,7 millions pour l'intervalle de confiance à 95% et de 3,6 millions pour celui à 80%

7. Discussion

Les projections de population probabilistes offrent un nouvel éclairage sur l'évolution possible de la population française. Elles permettent, sous certaines hypothèses de modélisation, de quantifier l'incertitude sur l'évolution à venir des indicateurs démographiques, et notamment sur l'évolution de la taille totale de la population. Elles présentent donc un avantage certain sur les projections déterministes basées sur des scénarios dont la probabilité d'occurrence n'est pas quantifiée. N'importe quel indicateur démographique, comme l'espérance de vie, l'âge moyen à la maternité ou la part des 65 ans et plus, peut être déterminé avec un certain degré de probabilité. Une des difficultés potentielles de l'interprétation des résultats provient du fait qu'il ne faut pas penser en termes ponctuels, mais plutôt en termes de distribution de probabilité, tout comme un dé, même pipé, ne peut se résumer à une seule de ces six faces. C'est plutôt en donnant la probabilité d'apparition de chaque face qu'on aura une bonne description du dé et de ce qu'on peut en attendre lorsqu'il sera lancé. Une fois cette difficulté surmontée, l'interprétation et l'utilisation de projections de population probabilistes offrent une grande liberté et une grande souplesse. A contrario, les résultats des projections déterministes deviennent compliqués à utiliser et à communiquer quand le nombre de scénarios envisagés est démultiplié sous l'effet du croisement de plusieurs hypothèses.

D'après les modélisations décrites dans cet article et les simulations effectuées, la population de France métropolitaine devrait continuer à augmenter dans les décennies à venir. La population de France métropolitaine pourrait pourtant commencer à décroître avant 2070, avec une probabilité non négligeable, même si cette évolution est moins probable qu'une hausse ou une stabilisation. La structure de cette population sera également probablement modifiée : on s'attend à un vieillissement général de la population dû à l'augmentation de l'espérance de vie, à la stagnation de l'indice conjoncturel de fécondité et à l'arrivée continue des baby-boomers aux âges de la retraite. Le modèle utilisé pour projeter le solde migratoire est le plus simple des trois modélisations utilisées. L'absence de données détaillée par sexe et âge sur les entrées et les sorties empêchent d'utiliser une modélisation de Poisson faisant apparaître des taux, comme on l'a fait pour le nombre de décès et le nombre de naissances. De manière générale, les modèles de projection du solde migratoire sont moins sophistiqués et ont fait l'objet de moins d'effort de recherche que ceux concernant la mortalité et la fécondité, les données disponibles étant moins riches. On peut noter toutefois, que certains pays à registre, et notamment la Nouvelle Zélande, ayant des données détaillées d'entrées et de sorties, commencent à proposer des modélisations avancées des phénomènes migratoires, prenant en compte de nombreux paramètres, comme par exemple le niveau d'éducation de la population (Bryant & Zhang, 2014). Notre modélisation étant assez simple, il en résulte que la plupart des évolutions passées du solde migratoire sont considérées comme du bruit. Ce bruit étant ensuite propagé dans le futur, les intervalles de confiance du solde migratoire projeté sont très larges et reflètent de ce fait notre niveau d'incertitude sur l'évolution à venir des migrations. C'est pour cette raison que nous avons restreint l'estimation des paramètres du modèle (et donc en particulier de la variance du terme d'erreur) à la période 1995-2013, afin de ne pas prendre en compte les larges fluctuations du solde migratoire trop anciennes. Estimer le modèle sur une plus large période aurait conduit à une incertitude encore plus grande sur l'évolution future du solde migratoire.

Au contraire du solde migratoire, les évolutions de la mortalité sont très stables, et le modèle utilisé peut rendre compte de ces évolutions sans les considérer comme étant majoritairement du bruit. Il en résulte que les intervalles de confiances des taux de mortalité projetés et de l'espérance de vie projetée sont très faibles. Ceci peut apparaître trompeur, car on pourrait croire qu'on est presque

certain de ce qui va arriver. En réalité, il ne faut pas oublier que ces intervalles de confiance sur les niveaux futurs de la mortalité sont déterminés conditionnellement au fait que le modèle offre une bonne approche du réel. C'est bien en supposant que les tendances observées vont se poursuivre qu'on peut arriver à de tels niveaux de confiance concernant l'avenir de la mortalité. Malgré cela, le modèle utilisé ne prend pas en compte certaines spécificités de la mortalité en France. En premier lieu, il ne permet pas de projeter des évolutions du logarithme du quotient de mortalité à un âge donné différentes selon le sexe. De plus, il apparaît que les générations nées après la seconde guerre mondiale ont très peu de gain au niveau de la mortalité à âge donné, par rapport aux générations précédentes, et ce quel que soit l'âge (Blanpain & Buisson, 2016a). La modélisation retenue ne permet pas de prendre en compte de tels effets de générations : les écarts à la tendance générale sont alors traités comme du bruit, mis dans les termes d'erreur, plutôt que comme un effet bien identifié. Les espérances de vie projetées auxquelles on a abouti sont donc un peu plus faibles que celles obtenues par les projections de Blanpain et Buisson (2016a).

La modélisation de la fécondité est différente de celle du solde migratoire et de la mortalité. En effet, contrairement aux taux de mortalité, les taux de fécondité n'ont pas une évolution régulière au cours du temps. Ils peuvent augmenter puis diminuer, ou faire l'inverse, et de ce fait se croiser. Prolonger les taux de fécondité selon des tendances linéaires amènent ainsi à des situations qui apparaissent invraisemblables au vu d'autres indicateurs de la fécondité, comme l'indice conjoncturel de fécondité ou le maximum de fécondité atteint dans l'année, qui sont plus ou moins stables depuis 1975. L'idée a alors été de prolonger dans un premier temps l'ICF, qui est un indicateur qui reflète le niveau de la fécondité, selon la même méthode qu'on a projeté le solde migratoire. On a ensuite prolongé les taux de fécondité par âge selon la méthode de Wisniowski et al. (2015), et on a modifié ces taux pour retomber sur l'ICF projeté en premier lieu. On dispose ainsi d'une évolution assez réaliste des taux de fécondité par âge, dont la distribution se décale vers les âges plus élevés tout en devenant plus symétrique. Cette façon de procéder (projection d'un indicateur agrégé, puis ventilation par catégories détaillées) n'est pas nouvelle en soi et c'est aussi le schéma retenu par l'ONU. L'inconvénient est qu'il faut ici fixer un ICF de long terme et le niveau choisi joue bien évidemment sur les résultats.

Afin d'améliorer les méthodes utilisées dans cet article et par conséquent les résultats, plusieurs pistes sont envisageables. Il s'agit en premier lieu de mieux comprendre les phénomènes migratoires et pour cela d'analyser les entrées de façon détaillée. Il serait aussi intéressant de se pencher sur des estimations des flux de sorties actuels et passés, qui est un travail assez récent en France, compte tenu des données disponibles. Pour la projection de la mortalité il serait utile d'intégrer un effet de génération et d'autoriser une évolution différente des taux de mortalité pour les femmes et les hommes. Plusieurs modèles sont pour cela envisageables, mais la difficulté reste que s'il y a trop de paramètres il y a un fort risque de non-identification du modèle ou de mauvaise convergence des chaînes de Markov servant à estimer les distributions *a posteriori*. Pour améliorer la projection des taux de fécondité par âge, on pourrait, comme cela a déjà été fait dans plusieurs travaux, trouver une modélisation paramétrique de la distribution des taux selon l'âge. Il suffirait alors, et ce n'est pas forcément une chose facile, de prolonger ces paramètres comme dans le cas de la modélisation de Lee-Carter, en détectant les régularités et les tendances dans l'évolution de ces paramètres. La distribution de la loi Bêta est une modélisation possible, mais sa forme arrondie ne représenterait pas bien les données. La loi Gamma à l'avantage de mieux refléter la distribution des taux de fécondité, mais elle est définie sur un support ouvert à droite. Il faudrait alors la tronquer pour ne pas avoir de résultats irréalistes. La fonction de Hadwiger est une troisième piste, car elle semble mieux adaptée pour modéliser la distribution de la fécondité. L'inconvénient est qu'estimer ses paramètres peut prendre du temps et que leur interprétation n'est pas forcément évidente. Enfin, pourquoi pas proposer une fonction *ad hoc*, qui reflète fidèlement les données observées ? On peut alors être tenté d'estimer de façon non-paramétrique la distribution des taux de fécondité, c'est-à-dire en réalité en utilisant un très grand nombre de paramètres. La difficulté réside alors en la projection de ces très

nombreux paramètres. On peut aussi penser, pour les trois composantes d'évolution de la population, à développer des modèles structurels, permettant d'expliquer l'évolution passée selon des mécanismes plus détaillés et reposant sur des variables externes, mais cela nécessite aussi d'avoir suffisamment d'éléments pour ensuite projeter l'évolution de ces variables. Par ailleurs, il serait très instructif de mener des études de sensibilité, qui permettraient de tester comment varie les résultats lorsqu'on modifie légèrement certaines hypothèses des modèles. Cela aiderait à mieux comprendre et quantifier quel est le rôle précis de chaque composante dans l'évolution de la population.

On le voit, de nombreuses améliorations sont sans doute possibles, et cela demandera des investissements importants de recherche dans la compréhension et la modélisation des migrations, de la mortalité et de la fécondité. Cela ne pourra qu'être bénéfique aux projections probabilistes de population, dont le degré d'incertitude dépend avant tout de nos connaissances (ou de nos ignorances) sur ces sujets. Enfin, il ne faudrait pas opposer les projections probabilistes de la population aux projections déterministes de la population. Ces dernières restent très utiles et permettent de tester ce qui se passerait à l'avenir, dans tel ou tel scénario. Les conclusions générales auxquelles on aboutit sont d'ailleurs très cohérentes avec celles des projections déterministes quant à l'évolution de la taille de la population et de sa structure par âge. Mais c'est avant tout aux utilisateurs des projections de population de choisir l'approche qui leur convient le mieux, selon leurs usages. Les projections probabilistes et déterministes sont deux manières différentes d'aborder l'incertain et d'essayer d'éclairer l'avenir.

Bibliographie

[1] Alkema, L, Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F. & Buettner, T. (2010). Probabilistic projections of the total fertility rate for all countries. Center for statistics and the social sciences, University of Washington. Working paper n°97.

[2] Bellamy, V. & Beaumel, C. (2016). Bilan démographique 2015, le nombre de décès au plus haut depuis l'après-guerre. Insee Première n°1581.

[3] Bijak, J., Alberts, I., Alho, J., Bryant, J., Buettner, T., Falkingham, J., Forster, J. J., Gerland, G., King, T., Onorante, L., Keilman, N., O'Hagan, A., Owens, D., Raftery, A. & Ševčíková, H. (2015). Probabilistic population forecasts for informed decision making. Journal of Official Statistics, 31(4), 537-544.

[4] Bijak, J. & Bryant, J. (2016). Bayesian demography 250 years after Bayes. Population Studies, 70(1), 1-19.

[5] Blanpain, N. (2016). Les hommes cadres vivent toujours 6 ans de plus que les hommes ouvriers. Insee Première n°1584.

[6] Blanpain, N. & Buisson, G. (2016a). Projections de population 2013-2070 pour la France : méthode et principaux résultats. Direction des Statistiques Démographiques et Sociales, Insee. Document de travail n°F1606.

[7] Blanpain, N. & Buisson, G. (2016b). Projections de population à l'horizon 2070. Insee Première n°1619.

[8] Booth, H. (2006). Demographic forecasting : 1980 to 2005 in review. Demography and sociology program, the Australian National University. Working papers in demography n°100.

[9] Brutel, C. (2014). Estimer les flux d'entrées sur le territoire à partir des enquêtes annuelles de recensement. Direction des Statistiques Démographiques et Sociales, Insee. Document de travail n°F1403.

[10] Bryant, J. & Zhang, J. L. (2014). Bayesian forecasting of demographics rates for small areas : emigration rates by age, sex and region in New Zeland. Statistica Sinica Preprint n° SS-14-200tR3.

[11] Conseil d'orientation des retraites (2017). Évolutions et perspectives des retraites en France. Rapport annuel du COR.

[12] Costemalle, V. (2015). Projections de populations : l'ONU adopte une méthode bayésienne. Statistique et Société, 3(3), 9-14.

[13] Dunstan, K. & Ball, C. (2016). Demographic projections : user and producer experiences of adopting a stochastic approach. Journal of Officials Statistics, 32(4), 947-962.

[14] Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G. K. & Wilmoth, J. (2014). World population stabilization unlikely this century. Science, 346(6206), 234-237.

[15] Hyndman, R. J. & Booth, H. (2006). Stochastic population forecasts using functional data models for mortality, fertility and migration. Demography and sociology program, the Australian National University. Working papers n°99.

[16] Hyndman, R. J. & Ullah, S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis, 51(10), 4942-4956.

[17] Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K. & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries : projections with a Bayesian model ensemble. The Lancet.

[18] Lee, R. D. & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. Journal of the American Statistical Association, 87(419), 659-671.

[19] Lutz, W., Sanderson, W. & Scherbov, S. (2001). The end of world population growth. Nature, 412, 543-545.

[20] MacPherson, L. (2016). National population projections : 2016(base)-2068. Statistics New Zeland.

[21] Organisation des Nations Unies, Département des affaires économiques et sociales, Division Population (2017). World Population Prospects : The 2017 Revision, key findings and advance tables. Working Paper n° ESA/P/WP/248.

[22] Raftery, A. E., Chunn, J. L., Gerland, P. & Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. Demography, 50, 777-801.

[23] Raftery, A. E. (2014). Use and communication of probabilistic forecasts. University of Washington. arXiv:1408.4812v1.

[24] Régnier-Loilier, A. & Vignoli, D. (2011). Intentions de fécondité et obstables à leur réalisation en France et en Italie. Population-F, 66(2), 401-432.

[25] Wiśniowski, A., Smith, P. W. F., Bijak, J., Raymer, J. & Forster, J. J. (2015). Bayesian population forecasting : extending the Lee-Carter method. Demography, 52, 1035-1059.