

**Faculté des sciences économiques,
sociales, politiques et de communication**

Comprendre les comportements en matière de fécondité en Belgique :

Fusionner des données administratives et d'enquêtes

Auteur : Armel LEKEUFACK

Promotrices : Ester RIZZI, *UCLouvain*

Ingrid SCHOCKAERT, *Statistics Flanders*

Lecteur : Jean-Paul SANDERSON, *UCLouvain*

Année académique : 2024-2025

Master de spécialisation en méthodes quantitatives en sciences sociales

École des Sciences Politiques et Sociales (PSAD)

Comprendre les comportements en matière de fécondité en Belgique : fusionner
des données administratives et d'enquêtes

Mémoire rédigé en vue de l'obtention du
Master de Spécialisation en Méthodes Quantitatives en Sciences Sociales

Par :

Armel LEKEUFACK

Mat : 16 802 100

Promotrices :

Ester RIZZI, UCLouvain

Ingrid SCHOCKAERT, Statistics Flanders

Année académique 2023-2024

Déclaration de déontologie

« Je déclare sur l'honneur que ce mémoire a été écrit de ma plume, sans avoir sollicité d'aide extérieure illicite, qu'il n'est pas la reprise d'un travail présenté dans une autre institution pour évaluation, et qu'il n'a jamais été publié, en tout ou en partie. Toutes les informations (idées, phrases, graphes, cartes, tableaux...) empruntées ou faisant référence à des sources primaires ou secondaires sont référencées adéquatement selon la méthode universitaire en vigueur.

Je déclare avoir pris connaissance et adhère au **Code de déontologie pour les étudiants en matière d'emprunts, de citations et d'exploitation de sources diverses** et savoir que le plagiat constitue une faute grave. »



Lekeufack Armel

Remerciements

Je tiens à exprimer ma profonde gratitude à la professeure Ester Lucia Rizzi, promotrice de ce mémoire, sans qui ce travail n'aurait pas été possible. Sa disponibilité, sa patience et son ouverture d'esprit ont été essentielles à la réalisation de ce projet. De plus, la confiance qu'elle m'a accordée m'a permis de bénéficier d'un stage dont les apports ont été inestimables, tant sur le plan des compétences acquises que des échanges enrichissants auxquels j'ai eu l'honneur de participer ; pour cela, je lui témoigne toute ma reconnaissance.

Je remercie également Ingrid Schockaert, co-promotrice de ce mémoire. Son soutien, sa disponibilité et son empathie ont grandement facilité le déroulement de mon stage au sein de *STATISTICS FLANDERS*, mon intégration dans cet environnement qui m'était alors inconnu et de grandir tant sur le plan professionnel qu'humain. Je tiens aussi à remercier Jan, Jorre, Marije et Noëmi pour leur accompagnement, les orientations méthodologiques et leur disponibilité lorsque j'étais face à des difficultés. Vos questions ont grandement contribué à la construction des éléments présentés dans ce travail.

Je souhaite également remercier l'équipe enseignante et administrative du master de spécialisation en méthodes quantitatives en sciences sociales pour leur accompagnement, ainsi que pour m'avoir accordé une dispense et autorisé à réaliser un stage. Merci d'avoir rendu ce travail possible !

Je remercie également mes camarades de promotion avec qui nous nous sommes parfois demandé si nous y arriverons !

Je poursuis en remerciant mes « vraies » : Marie Claude, Guylène et Mélanie pour les moments de détente qui m'ont permis de surmonter les périodes de détresse et de stress dans cet environnement auquel nous n'étions pas habitués.

Enfin, je termine par des remerciements à ma famille, restée au Cameroun, mais qui suit de près mes « aventures » ! Ce travail est le fruit de nos efforts ; soyez-en remerciés.

Tables de matières

Déclaration de déontologie	i
Remerciements	ii
Tables de matières.....	iii
Listes des tableaux	v
Introduction générale	1
0.1. Contexte et justification	1
0.2. Problématique.....	2
0.3. Objectifs et hypothèses.....	3
0.4. Plan du travail.....	4
Chapitre I. Revue de la littérature : la fécondité et ses déterminants.....	5
I.1. Un mot sur les théories de la transition démographique.....	5
I.2. La fécondité et ses déterminants	5
I.2.1. Fécondité et structure démographique	5
I.2.2. Les déterminants de la fécondité.....	6
I.3. Cas particulier de la religion comme explicatif de fécondité.	8
I.3.1. Ce que l'on entend par religion et religiosité.	8
I.3.2. Relation entre religion et fécondité	9
I.3.3. Quel rang prendre en compte pour la naissance ?	10
I.4. Pour conclure : religion, religiosité et fécondité.....	10
Chapitre II. Données.....	12
II.1. <i>European Social Survey</i>	12
II.1.1. L'enquête et ses objectifs.....	12
II.1.2. Les Vagues 5 et 6 de la ESS de 2010 et 2012	13
II.1.3. Accès aux données.....	13
II.2. Du registre national à DEMOBEL	14
II.3. Pourquoi DEMOBEL et ESS ?.....	14
Chapitre III. Méthodes.....	16

III.1. Fusion statistique des données ou <i>Statistical Matching</i> : théorie.....	16
III.1.1. Généralités sur la fusion des données et importance.....	16
III.1.2. Grandes étapes de la fusion des données.....	18
III.1.3. Fusion statistique et informations auxiliaires	25
III.2. Fusion statistique des données ou <i>statistical matching</i> : méthode	26
III.2.1. Choix des variables spécifiques et mesures : religion et fécondité	26
III.2.2. Harmonisation des variables communes	28
III.2.3. Analyse de la distribution et choix des variables de fusion.....	28
III.2.4. Fusion statistique de la ESS et des données administratives.....	29
III.2.5. Évaluation de la qualité de la fusion des données	30
III.3. Analyse de la relation entre fécondité et religion.....	30
III.3.1. Pourquoi étudier l'arrivée d'un second enfant ?.....	30
III.3.2. Recodage et synthèse de nouvelles variables	31
III.3.3. Régressions logistiques	31
Chapitre III : Résultats et discussion	33
III.1. Résultats	33
III.1.1. Fusion en considérant ESS comme receveur et DA comme donneur	33
III.1.2. Relation entre fécondité et religion	38
III.2. Forces et limites de la méthode.....	44
III.2.1. Quelles fiabilités pour la fusion statistique ?.....	44
III.2.2. Quelle transformation pour quels résultats ?	44
Conclusion générale	46
Références bibliographiques	49
Annexes	54

Listes des tableaux

Tableau II : définition des notions de religion et de religiosité	9
Tableau I. approches possibles en fonction des objectifs de la fusion et du paramétrage.	24
Tableau III : liste des variables communes retenues	33
Tableau IV : caractéristiques des variables recodées.....	34
Tableau V : résumé de l'analyse des similarités entre les paires de variables.....	35
Tableau VI : anova de la régression pour le choix des meilleurs prédicteurs du degré de religiosité.....	36
Tableau VII : anova de la régression pour le choix des meilleurs prédicteurs des naissances	36
Tableau VIII : différences de distributions des variables spécifiques en fonction des donneurs et receveurs	37
Tableau IX : statistiques descriptives des variables utilisées.....	39
Tableau X : modèle explicatif de l'effet du degré de religiosité sur la variable dépendante...	41
Tableau XI : facteurs explicatifs de l'événement avoir un second enfant	42

0.1. Contexte et justification

La Belgique, comme l'ensemble des pays d'Europe, connaît depuis de nombreuses décennies un vieillissement progressif de sa population et une baisse des naissances (Vodisek, 2023). Cette situation, qu'on peut sans doute lier à l'évolution du comportement des femmes et des hommes face à la natalité (Goldscheider et *al.*, 2015), laisse entrevoir une baisse avérée de la fécondité et un recul de l'âge de la mère à la naissance de leur premier enfant. De plus, après le premier enfant, il est davantage plus courant de rencontrer des parents qui ne veulent pas en avoir un autre, que ce soit dans une perspective à court terme ou de manière définitive (Adsera, 2011 ; McDonald et *al.*, 2011). Selon Goldscheider et *al.* (2015), dans les pays industrialisés, le nombre d'adultes qui retardent ou évitent les liens formels d'engagement est croissant. Ces adultes entrent ainsi dans des relations de cohabitation ou vivent en dehors d'un partenariat ce qui a pour conséquence le recul du mariage (Doignon et *al.*, 2020). Les auteurs susmentionnés précisent que la procréation ou l'éducation des enfants sont de plus en plus séparées du mariage avec une forte augmentation des enfants nés hors mariage de mères célibataires ou dans le cadre d'une cohabitation d'individus non mariés ; et ceux dans un contexte de questionnement de la position de la femme dans la société et de l'augmentation de sa participation au marché du travail (Costa et *al.*, 2011).

Un constat s'impose : les comportements en matière de fécondité évoluent et ceux en fonction des mutations des sociétés. De nombreux facteurs peuvent être mis en cause dans ce changement parmi lesquels la religion. Il s'agit en effet d'un important facteur dans l'organisation des sociétés humaines (Krech, 2014). Du choix du style de vie, des aliments à consommer, des couleurs politiques à défendre ou du positionnement écologique, le rapport à la religion a guidé et guide encore les options de vie de millions d'individus sur la Terre. Qu'ils soient croyants ou non-croyants, la religion ou plus largement la religiosité, influe considérablement sur les options de vie notamment en ce qui concerne le mariage ou l'enfantement ; mais aussi sur l'ordre chronologique entre ces deux événements (Costa et *al.*, 2011). Ainsi, la sécularisation et la laïcisation des sociétés occidentales ont été identifiées comme étant en partie des facteurs explicatifs des changements observés en matière de fécondité en Belgique et dans le monde (Bailly et *al.*, 2011). Le transfert des biens du domaine du religieux vers le domaine public a contribué à redéfinir les rapports entre les genres (Woodhead, 2012) et donc de remettre en question la totalité ou une partie des

enseignements véhiculés par la religion y compris ceux en matière de sexualité et de maternité.

De ce qui précède, il est évident qu'il existe des liens inextricables entre les comportements en matière de fécondité et les croyances religieuses. Comprendre ces liens est d'une contribution significative à la définition des politiques démographiques efficaces surtout en contexte actuel où la religion tend à retrouver sa place dans les débats publics (Ghia, 2018). Il est cependant de plus en plus complexe d'en percevoir l'étendue (des dits liens) compte tenu de la multiplicité des bases de données existantes (Bounebach et *al.*, 2018). L'information joue en effet un rôle clé dans la compréhension des phénomènes (D'Orazio et *al.*, 2006a). De nombreuses bases de données existent pour comprendre la fécondité en Belgique. Cependant, la compréhension de ses liens avec d'autres variables dont la religion est peu certaine si l'on se limite à une seule source de données. Il est ainsi possible en combinant plusieurs sources de données d'obtenir des informations fiables sur la fécondité et les facteurs qui peuvent l'affecter (Brée et *al.*, 2016 ; Grimmeau et *al.*, 2012). Ceci passe par un processus rigoureux d'intégration des différentes sources de données.

La fusion des données, souvent remplacée dans la littérature par les termes d'appariement, de couplage ou de chainage en fonction des méthodes et des variables impliquées, est utilisée dans de nombreux domaines pour combiner soit des données de cohorte, soit des données administratives ou encore des données d'enquêtes (Bounebach et *al.*, 2018 ; Harron et *al.*, 2017 ; Lloyd et *al.*, 2022 ; Perlberg et *al.*, 2014 ; Wang & Ling, 2012). Dans certains cas, un identifiant unique permet de coupler deux bases de données. Il peut s'agir d'un numéro d'identification des individus ou des unités d'échantillonnage (ménages, entreprise, etc.) commun à système de collecte des données (Bounebach et *al.*, 2018). De telles variables restent cependant rares dans l'éventail des bases de données disponibles. La multiplication des bases de données et le volume d'informations contenues dans ces bases font de la fusion un réel défi et appellent à une réflexion constante sur l'automatisation des méthodes de fusion (D'Orazio et *al.*, 2006a ; Rodgers, 1984).

0.2. Problématique

La sécularisation entamée en Europe dès le début du XXe siècle a permis de renvoyer la pratique de la vie religieuse dans le domaine de l'intime et donc hors de la place publique à travers un processus de laïcisation de l'État et des institutions publiques (Honoré, 2021). Elle a contribué à réduire le pouvoir de l'église et de la religion dans la gestion des affaires de la sphère publique, mais aussi privée. Cette question que l'on croyait réglée réémerge (Ghia,

2018 ; Honoré, 2021) et suscite une attention particulière surtout en ce qui concerne la situation démographique préoccupante et la fécondité dans le contexte occidental auquel la Belgique n'échappe pas. La seconde transition démographique, amorcée dans les années 50, est le résultat d'une profonde mutation qui est toujours à l'œuvre (Lesthaeghe, 2011). Ainsi, depuis le début des années 70 (précisément entre 1971 et 1970)¹, la fécondité en Belgique est restée en deçà du seuil de renouvellement de la population qui est de 2,1 enfants par femme. Le rôle de la religion comme déterminant de cette situation est largement documenté (Adsera, 2011 ; Brée et al., 2016 ; Peri-Rotem, 2020). Bein et al. affirment ainsi qu'« *on a constaté que les gens qui sont plus religieux ont tendance à désirer — et à avoir — plus d'enfants que ceux qui sont moins religieux* » (2017, p. 185). La religiosité exercerait donc une influence sur les intentions de fécondité, mais aussi sur le nombre d'enfants effectif par femmes. Ainsi, ce travail se propose à partir du développement d'un processus de fusion statistique de deux bases de données de réponse à la question : quelle est l'influence de la religion sur l'arrivée d'un deuxième enfant pour les femmes qui en ont déjà un ?

0.3. Objectifs et hypothèses

Ce travail a pour objectif général d'analyser les comportements en matière de fécondité en Belgique en rapport avec la religion à partir des données administratives (DA) et de la *European Social Survey* (ESS). Spécifiquement, il est question :

- de réaliser une fusion statistique des DA de 2011 et des vagues 5 et 6 de la ESS ;
- d'analyser l'influence de la religiosité sur la probabilité d'avoir un deuxième enfant pour les femmes qui en ont déjà un.

L'hypothèse principale de cette recherche est que : la fusion des DA et de la ESS permet une meilleure compréhension des comportements de fécondité des femmes en Belgique en combinant leur précision territoriale aux données attitudinales de la ESS. Elle permet de formuler les deux hypothèses :

- La fusion des DA et de la ESS produit d'excellents résultats en raison d'une faible dissimilarité entre les distributions des variables dans les données d'origine et celles de la base de données produite par la fusion. Elle permet ainsi de mieux comprendre la relation entre la fécondité et la religion.

¹ Données disponibles sur www.statbel.fgov.be. Statbel est l'office belge de statistique qui collecte, produit et diffuse des chiffres fiables et pertinents sur l'économie, la société et le territoire belges. Il s'agit de l'organisme de références en matière statistique sur le Royaume de Belgique.

- Le degré de religiosité influence positivement sur la probabilité d'avoir un deuxième enfant chez les femmes qui en ont déjà un.

0.4. Plan du travail

La suite du travail se structure en 4 chapitres. Le premier chapitre situe le travail dans son contexte en faisant une revue de la littérature autour de la fécondité, des théories de la transition démographique, mais aussi de la religion et de la religiosité. Le chapitre 2 fait une présentation des données mobilisées en faisant un historique de la collecte, de la structuration et surtout de la justification des données utilisées dans le cadre du travail. Le chapitre 3 quant à lui présente les détails méthodologiques. Il présente dans un premier temps les généralités sur la fusion statistique avant de s'intéresser aux techniques utilisées dans le cadre du travail. Dans un second temps, il se poursuit par la présentation des techniques mobilisées dans la compréhension de la relation entre la fécondité et la religiosité. Le chapitre 4 présente et discute les résultats obtenus. Le travail s'achève enfin sur une conclusion générale.

Chapitre I. Revue de la littérature : la fécondité et ses déterminants

Ce premier chapitre situe le travail dans son cadre théorique en abordant la fécondité et ses déterminants. Il débute sur un bref rappel des théories de la transition démographique avant de revenir sur les généralités en ce qui concerne la fécondité et ses liens avec les notions de religion et la religiosité.

I.1. Un mot sur les théories de la transition démographique

Du 18^e au 19^e siècle, le monde surtout occidental a connu une transformation de sa structure démographique. Il a été en effet observé une diminution de la fécondité et aussi de la mortalité. C'est ce qui est connu comme étant la première transition démographique (Lesthaeghe, 2011). Cette première transition démographique fait suite à la révolution technique et au progrès scientifique majeur qui ont eu lieu durant cette période notamment dans le domaine des sciences et de la médecine. La seconde transition démographique émerge dans les années 1950 et se caractérise par de nombreux aspects, dont une fécondité désormais inférieure au seuil de remplacement de la population et la réduction du pouvoir de la religion dans la sphère sociale et politique (Lesthaeghe, 2011). Cette situation conduit dès lors au processus de diminution de la taille de la population dans de nombreux pays européens (Vodisek, 2023). Un autre fait important de cette seconde transition est le prolongement de durée de vie et le vieillissement de la population. C'est dans ce contexte de la seconde transition que se situe la situation démographique actuelle de la Belgique.

I.2. La fécondité et ses déterminants

I.2.1. Fécondité et structure démographique

La population de plusieurs des pays de l'Union européenne (UE) dont la Belgique est vieillissante (Bourguignon et al., 2019 ; Vodisek, 2023). L'Italie se classe comme championne d'Europe avec une fécondité qui n'est plus allée au-dessus de 2 depuis 1977. Pire encore, l'accroissement naturel de sa population devient nul, voire négatif, depuis 2004. Les naissances ne sont donc plus suffisantes pour assurer le renouvellement de la population (Vodisek, 2023). La fécondité est comprise ici comme un indice qui indique « *le nombre moyen d'enfants que mettrait au monde au cours de sa vie féconde une génération qui aurait des taux par âge identiques à ceux observés l'année considérée* » (Vodisek, 2023, p. 6). Cet indice est rencontré dans la littérature sous l'appellation d'indice conjecturel (IWEPS, 2021 ; Volant et al., 2019). Il est la somme des taux de fécondité par âge ; le taux de fécondité par âge étant le rapport des naissances vivantes des femmes d'un âge donné à l'effectif moyen

des femmes de cet âge. Il est plus commun de le considérer comme étant le nombre d'enfants pour les femmes en âge de reproduction au sein de la population (Buelens, 2021).

En 2017, le nombre d'enfants par femme était estimé à 1,59 en moyenne dans l'UE avec un peu plus de 1,6 en Belgique. Il est passé de 1,61 en 2019 à 1,57 en 2021 (IWEPS, 2023); et à 1,53 en 2022 selon l'institut national des statistiques (Statbel)². Selon IWEPS (2023), ce taux est largement inférieur au pic de 1,85 observé en 2010 et reste inférieur au seuil de renouvellement de la population belge fixé à 2,1. En Belgique, les comportements en matière de fécondité changent. De plus en plus de personnes décident de ne pas faire d'enfants, ou de retarder le moment d'en faire. Dans un premier temps, le choix est volontaire. Dans un second temps, certaines personnes qui décident de retarder leur parentalité s'aperçoivent souvent trop tard que le moment est passé sans que les conditions soient réunies; il s'agit donc d'une décision involontaire (McDonald et *al.*, 2011). Statbel note ainsi une progression de l'âge moyen de la mère à la première naissance qui atteint 31,25 ans en 2022 contre 29,08 en 1998. La conséquence directe de cette transformation démographique est une modification de la pyramide des âges avec un élargissement du sommet et un rétrécissement de la base; synonyme d'un vieillissement de la population belge combiné à une baisse de la mortalité.

I.2.2. Les déterminants de la fécondité

La section plus haut a permis de se rendre compte des changements qui s'opèrent dans les comportements en matière de fécondité en Belgique. De nombreux démographes voient dans le phénomène actuel de baisse de la fécondité une réaction adaptative face aux crises sociales, économiques, politiques et plus récemment sanitaires et écologiques (Bourguignon et *al.*, 2019; Leridon, 2015; Makki, 2021). Ces grands groupes de paramètres regroupent des déterminants identifiés comme pouvant être explicatifs de la fécondité. La liste de ces déterminants est longue et variée et peut être associée négativement ou positivement à la fécondité. Des auteurs ont pu identifier des variables comme le niveau d'instruction, le revenu, le statut d'occupation, le lieu d'habitation, le nombre d'enfants déjà nés, le rapport à la religion et au genre, le sens de la famille, les politiques familiales et dans une certaine mesure l'immigration comme principaux déterminants de la fécondité (Bein et *al.*, 2017; Buelens, 2021; Dantis et *al.*, 2023; Kim, 2017; Peri-Rotem, 2020; Singh et *al.*, 2017).

² Données obtenues à partir du lien [Natalité et fécondité | Statbel \(fgov.be\)](https://statbel.fgov.be/fr/themes/population/natalite-fecundite)

À partir de la lecture de l'article produit par Buelens (2021) sur les contrastes interurbains de la fécondité bruxelloise, le coût qu'implique la parentalité est identifié comme un déterminant majeur de la fécondité (facteur socio-économique). Ce coût peut être par exemple évalué non seulement en valeur monétaire, mais aussi par le temps accordé dans l'éducation de l'enfant, la perte de salaire, une réduction des opportunités de carrière ou encore une privation des libertés individuelles. Les individus seraient donc de plus en plus enclin à sacrifier la parentalité au profit d'une option de carrière compte tenu de son coût. Ce coût se répercute tant sur le nombre d'enfants par femme que sur le calendrier des maternités. Dans le même ordre d'idée, le niveau de vie, couplé au lieu de vie, influence la fécondité. Ainsi, les femmes vivant dans les milieux défavorisés dans les centres urbains sont susceptibles d'avoir un taux de fécondité plus élevé que celui des autres femmes ; renvoyant certainement au fait qu'elles sont plus enclines à faire le « sacrifice de la parentalité » (Buelens, 2021). Il y a ainsi une relation négative entre le revenu et la fécondité (Doliger, 2008 ; Schoumaker, 1999).

D'autres facteurs culturels agissent comme déterminants de la fécondité dans le contexte européen. Costa et *al.* (2011) notent par exemple le cas de la concurrence entre parentalité et loisirs, la diffusion de la sécularisation ou encore la montée de l'individualisme et la transformation des valeurs par rapport à la famille. On peut ajouter à cette catégorie le rapport au genre qui influence fortement la fécondité tant dans les pays dits du Nord, que ceux du Sud. En effet, l'implication différenciée des hommes et des femmes dans la parentalité est largement débattue (Samuel & Attané, 2005). Des études montrent que lorsque la répartition des tâches ménagères au sein d'un couple est plus égalitaire, la tendance observée est à un comportement de la femme plus favorable à la fécondité. Pour le cas de la migration, Volant et *al.* (2019) concluaient que la contribution des femmes d'origine étrangère était non significative à la fécondité, mais significative sur le plan de la natalité. Ceci est lié à la faible proportion des femmes immigrées par rapport à la population totale de femmes en âge d'avoir un enfant. Pour que cette contribution soit significative, il faudrait que : (1) ces femmes représentent une fraction plus importante des mères et (2) surtout que leur fécondité soit très supérieure à la moyenne (Volant et *al.*, 2019, p. 2) ce qui est difficilement observable.

Concernant les politiques familiales, Rizzi & Rees (2023) dans une étude comparative des politiques familiales dans un groupe de pays d'Europe notent un accroissement des fonds pour le soutien des familles en Belgique. Ces soutiens servent au financement des garderies et des structures d'enseignement préprimaire, des dépenses publiques et des assistances aux

jeunes personnes et aux logements. Ces politiques régissent aussi le secteur du travail en permettant au mieux de garantir un équilibre entre la vie de famille et la vie professionnelle. Sont pris en compte, dans ce second cas, les congés de maternité tant pour la mère que pour le père (congés de paternité), ensuite des congés parentaux qui représentent un supplément accordé aux parents. Rizzi & Rees (2023) concluent comme d'autres auteurs (Peri-Rotem, 2020 ; Wood & Marynissen, 2019) que les mesures constitutives des politiques familiales peuvent avoir un effet significatif sur la fécondité. Erlandsson (2017), observe ainsi que les femmes qui bénéficient de l'allocation de garde d'enfant à domicile ont un risque plus élevé d'avoir un deuxième ou un troisième enfant que les femmes qui n'en bénéficient pas.

I.3. Cas particulier de la religion comme explicatif de fécondité.

La littérature est abondante sur le sujet de la religion, de la spiritualité et des relations avec la fécondité (Bein et *al.*, 2017 ; Dantis et *al.*, 2023 ; Krotki & Lapierre-Adamcyk, 1968 ; Lebugle Mojdehi, 2015 ; Leridon, 2015 ; Peri-Rotem, 2020 ; PERSEE, 1957 ; Samuel & Attané, 2005 ; Somers & Van Poppel, 2003). Avant de parler de cette relation, il est important de préciser ce qu'on entend par religion et religiosité.

I.3.1. Ce que l'on entend par religion et religiosité.

La religion et la religiosité se caractérisent par des faits religieux. Honoré (2021) précise en ce sens qu'il y a dans le fait religieux deux dimensions sous-jacentes, dont (1) la religion, qui renvoie à une structure sociale et un système de normes externes et (2) la religiosité qui renvoie à la personne. En ce sens, la religion est le fait d'appartenir à un groupe ou à une communauté et d'obéir à ses règles tandis que la religiosité relève plus du degré de croyance ou spiritualité propre à l'individu dans la pratique d'une religion. La religion a été et demeure toujours un facteur important dans les choix de vie des hommes. Elle est en effet perçue comme pouvant avoir un impact important sur les comportements et les habitudes (Honoré, 2021).

Dans le contexte de la mondialisation actuelle, de migrations et de transformations des sociétés, la religion fait son retour dans les débats publics notamment en ce qui concerne les naissances ou le mariage, le genre ou encore la famille (Bein et *al.*, 2017 ; Ghia, 2018). Ainsi, les rapports à la religion changent. Les individus, au lieu des rendez-vous dominicaux ou des activités en communautés, se tournent davantage vers des pratiques plus individuelles comme des prières individuelles ou en groupes. Dans ce contexte, il est possible d'observer

dans la littérature, un débat sur l'utilité ou non d'étudier de manière séparée les questions de religion et de spiritualité ou religiosité (Honoré, 2021).

Tableau I : définition des notions de religion et de religiosité (de Honoré, 2021)

Notions	Religion	Religiosité
Définitions	Un ensemble de croyances et de dogmes qui définissent le rapport de l'homme avec le sacré ainsi que les pratiques et les rites qui en résultent.	Un construit personnel à travers lequel chaque individu appréhende son rapport à la religion
Caractérisées par :	Un dispositif matériel : des institutions et des organisations, des lieux, des objets. Un dispositif normatif : des dogmes, des discours, des normes de comportements. Un dispositif théologique : des notions particulières de transcendance et de divinité, de vérité, d'idéal et de quête de cet idéal. Une communauté	Un mode de comportement pour soi : formé par la pratique religieuse intime ainsi que par l'influence de la posture et de l'engagement religieux de la personne sur son comportement. Un mode d'interaction : formé par la pratique religieuse, sociale et communautaire ainsi que par l'influence de la posture et de l'engagement religieux de la personne sur ses rapports aux autres. Un système de croyances propres à l'individu formé par l'articulation de ses convictions et de ses connaissances

I.3.2. Relation entre religion et fécondité

La compréhension et le contrôle des comportements en matière de fécondité ne sont pas une ambition nouvelle comme souligné par Leridon (2015). Selon lui, la réflexion sur le devenir des sociétés et surtout en ce qui concerne ses perspectives démographiques est restée pendant longtemps une réflexion qui ne pouvait s'appuyer sur aucune statistique fiable. Elle relevait plutôt de la philosophie, de la morale, des sciences politiques et surtout de la religion. Cette dernière a joué un grand rôle dans les politiques familiales qui se sont succédé au cours des siècles passés.

La littérature sur la relation liant la religion et la fécondité dans les pays occidentaux est abondante. Elle permet de mettre en évidence un lien réel qui existe entre ces 2 variables. La religion a donc un pouvoir explicatif élevé de la fécondité. Peri-Rotem (2020) montrait ainsi qu'une religiosité forte pouvait être associée à une fécondité élevée dans les cas de la

France et de la Grande-Bretagne. En outre, elle pouvait permettre d'atténuer les effets de la relation négative entre le niveau d'instruction et la taille des familles. L'autrice parvient aussi à mettre en évidence que ce n'est pas plus la religion qui influence la fécondité que le degré de religiosité. Une personne non croyante sera donc moins affectée qu'un croyant qui a son tour, sera affectée dépendamment du fait qu'il soit pratiquant ou pas. Bien entendu, il existe d'autres variables de contrôle qui peuvent et doivent être pris en compte pour l'étude de ces relations. En ce sens, le niveau d'éducation peut influencer à la fois la religion ainsi que la fécondité (Peri-Rotem, 2020). D'autres variables comme l'âge, le statut matrimonial ou le pays de naissance ou encore l'égalité entre les genres par exemple, peuvent être mobilisées (Bein et *al.*, 2017).

I.3.3. Quel rang prendre en compte pour la naissance ?

Le nombre d'enfants déjà mis au monde par la femme est une variable très importante à prendre en compte dans l'étude du lien entre la religion et la fécondité. Le choix est fait dans cette étude de partir du premier enfant déjà mis au monde pour étudier l'influence de la religion sur les chances d'en avoir d'autres. Dantis et *al.* (2023) font le choix d'analyser la fécondité (mesurée par l'intention d'avoir un enfant) par le second et le troisième enfant même si les détails sur ce choix restent difficiles à cerner. Bein et *al.* (2017) quant à eux se sont intéressés à l'influence de la religiosité sur l'arrivée d'un enfant indépendamment du fait que la femme en ait déjà ou pas. Ils ont ainsi considéré les intentions de fécondités chez les femmes sans enfant, avec 1, 2, 3 enfants ou plus. Dans les études qui s'intéressaient aux aspects économiques et non à la religiosité, avoir un enfant a été la variable d'intérêt la plus rencontrée (Dantis & Rizzi, 2020 ; Matias & Fontaine, 2017). En Finlande, Erlandsson (2017) utilisait comme indicateur de la fécondité, la transition au second et troisième enfant pour étudier l'impact de l'allocation de garde d'enfant à domicile sur la fécondité. En Italie, Rinesi et *al.* (2011) étudiaient la réalisation des intentions de fécondité en se basant sur les femmes avec un enfant, compte tenu du fait que la faible fécondité observée en Italie est due à l'absence d'intentions de seconde naissance. Les choix sont donc variés et tiennent compte des objectifs de la recherche.

I.4. Pour conclure : religion, religiosité et fécondité

La seconde transition démographique qui a lieu actuellement en Belgique invite à s'interroger sur ses causes profondes ainsi qu'à réfléchir sur les moyens à

mettre en œuvre pour sa compréhension. La fécondité est depuis quelques décennies en deçà du seuil de renouvellement de la population belge. Cette situation est la base d'un vieillissement de la population avec tout ce que cela implique comme mutations sur les plans démographiques, sanitaires ou encore de la planification (politique des retraites, de la santé, de l'emploi ou encore des finances publiques). La complexité dans la compréhension de ces causes est d'autant plus importante que les bouleversements qui s'enchainent (crises environnementales, économiques, sanitaires) s'accompagnent de leurs lots de changements. La maternité est davantage postposée, voire non envisagée. La religion ou la religiosité se positionne comme un facteur très important pour la compréhension de cette transition au regard de son pouvoir sur la scène publique. Ce chapitre, rédigé dans les limites d'un mémoire de master, fait un panorama des notions clés mobilisées dans le travail. Il a retracé notamment des généralités sur la fécondité et ses déterminants avec un point d'intérêt sur la religion. Il situe ainsi le lecteur dans le champ conceptuel et théorique dans lequel s'inscrit ce travail.

Les données mobilisées dans ce travail proviennent de deux sources principales. La *European social survey* (ESS) et les données administratives. Le chapitre est organisé en 3 principales sections : la première présente les données issues de la ESS ; la deuxième se concentre sur les données administratives et, enfin, la troisième section revient sur la justification des données mobilisées.

II.1. *European Social Survey*

II.1.1. L'enquête et ses objectifs

La ESS est une enquête européenne initiée en 2002 et qui a lieu tous les deux ans. Elle en est actuellement rendue à sa 11^e vague. La 12^e quant à elle, est en cours de préparation (en 2023). L'enquête couvre près de 38 pays européens, dont Israël. Il est cependant important de noter qu'une seule partie des pays a participé à toutes les vagues de l'enquête comme c'est le cas pour la Belgique. Ses trois objectifs sont de : (1) suivre et interpréter l'évolution des attitudes et des valeurs du public en Europe et d'étudier leur interaction avec les institutions européennes en mutation ; (2) faire progresser et consolider les méthodes améliorées de mesure des enquêtes transnationales en Europe et au-delà ; (3) développer une série d'indicateurs sociaux européens, y compris des indicateurs attitudinaux.

Depuis 2002, plus de 430 000 personnes des 38 pays ont été interviewées dans le cadre des enquêtes ESS avec un intérêt principal sur les attitudes, les comportements et les valeurs (Péri et *al.*, 2023). La ESS est une enquête de panel utilisée dans de nombreux travaux pour l'étude des relations entre variables diverses comme le terrorisme et ses différents effets (Péri et *al.*, 2023), la demande physique et les douleurs musculosquelettiques au niveau individuel et professionnel (Gulseren, 2022), le bien-être subjectif et les revenus (Bauknecht & Merkel, 2022), les rapports de genre et la consommation d'énergie (Bartiaux, 2022), la relation entre la perception du climat et le moral identitaire (Verschoor et *al.*, 2020 ; Welsch, 2021) ou encore la relation entre la religion et la dépression (Schnittker, 2020).

L'enquête se compose de deux types de thèmes généraux : le premier se répète à chaque vague de l'enquête. Il comprend des thèmes comme les médias, les politiques ou les caractéristiques sociodémographiques. Le second est plus spécifique et varie entre les vagues de l'enquête. L'immigration, la santé et les soins, le bien-être et le travail, le vieillissement, le changement climatique ou encore la religion sont des thèmes abordés. Cette polyvalence fait de la ESS un modèle des enquêtes internationales (Péri et *al.*, 2023). L'enquête se concentre

sur les personnes âgées de 15 ans et plus et qui résident dans des ménages privés, sans distinction de nationalité, de citoyenneté, de langue ou de statut.

II.1.2. Les Vagues 5 et 6 de la ESS de 2010 et 2012³

Les vagues 5 (ESS5) et 6 (ESS6) de la ESS de 2010 et 2012 ont été utilisées dans le cadre de ce travail. En Belgique, les données de la ESS5 et de la ESS6 ont été collectées sur le même univers que pour toutes les enquêtes ESS. Les centres comme les monastères, les maisons de retraite ou des centres d'hébergement ne sont pas pris en compte dans la définition de la population. Elles couvrent les 10 provinces de la Belgique (considérées comme les unités primaires d'échantillonnage) en plus de la région Bruxelles-Capitale. La procédure d'échantillonnage est basée sur les données officielles du Registre national. Le plan d'échantillonnage pour ces enquêtes est intéressant dans le sens où il couvre l'étendue du territoire belge en incluant toutes les classes sociales. Il s'agit d'un plan stratifié à 2 niveaux : un premier pour les 11 unités primaires d'échantillonnage dans lesquelles sont choisis 363 points d'échantillonnages distribués dans les 221 communes de la Belgique en fonction du poids de la population. Les participants sont sélectionnés au hasard en raison de 9 participants par point d'échantillonnage, soit un total de 3 267 personnes sélectionnées par vague. Le taux de réponse pour la ESS5 était de 53,43 %, soit un total de 1 704 personnes enquêtées. Le taux de réponse de la ESS6 quant à lui s'élevait à 58,7 %, soit un total de 1869 personnes enquêtées (annexe 1). La base de données utilisée pour la suite du travail était constituée de femmes âgées de 18 ans et plus et résidant en Belgique soit, 1766 femmes.

II.1.3. Accès aux données

Les données sont accessibles gratuitement sous la condition d'une inscription au préalable. L'onglet Search European Social Survey (nsd.no) offre un panorama de toutes les vagues de l'enquête ainsi que les fichiers de données et la documentation qui accompagne chaque vague. Il est possible pour chaque vague d'avoir plus d'un millier de variables dans le

³ Les informations présentées dans cette sous-section proviennent des références suivantes :

- European Social Survey European Research Infrastructure (ESS ERIC). (2018). ESS5 Data Documentation. Sikt - Norwegian Agency for Shared Services in Education and Research.
- European Social Survey European Research Infrastructure (ESS ERIC). (2018). ESS6 - integrated file, edition 2.4 [Data set]. Sikt - Norwegian Agency for Shared Services in Education and Research.
- European Social Survey European Research Infrastructure (ESS ERIC). (2018). ESS5 - integrated file, edition 3.4 (Austria not included) [Data set]. Sikt - Norwegian Agency for Shared Services in Education and Research.
- European Social Survey European Research Infrastructure (ESS ERIC). (2018). ESS6 Data Documentation. Sikt — Norwegian Agency for Shared Services in Education and Research.

fichier de données incluant les répétitions. Des variables ne répondant pas au besoin de l'étude n'ont pas été prises en compte. Les données brutes téléchargées concernaient uniquement les vagues 5 et 6, pour la Belgique, soit 3 573 observations (hommes inclus).

II.2. Du registre national à DEMOBEL

Le Registre national (RN) ou registre des personnes physiques est un système centralisé de traitement de l'information relative à l'identification des personnes physiques c'est-à-dire les citoyens. Il a l'avantage de l'exhaustivité des détails territoriaux. Chaque individu est identifié avec une excellente précision. Statbel met les données du RN à la disposition de la communauté scientifique et aux bureaux statistiques à travers de son projet DEMOBEL. Les bases de données DEMOBEL sont divisées en données « stock » et « flow »⁴. Les données du stock se réfèrent à la population au 1^{er} janvier de chaque année tandis que les données du flow se réfèrent aux événements (naissances, décès, migrations, etc.) au cours de l'année. Ces données sont pseudonymisées par des clés de 16 caractères uniques à chaque personne et commun aux fichiers de données. Ceci est un moyen efficace pour faciliter l'intégration des sources données.

À *STATISTICS FLANDERS* (organe en charge des statistiques de la Région flamande), les données du stock de 2011 ont été utilisées. Il s'agit d'une base de données constituée de 33 variables pour 10 951 266 observations. Il résume pour l'ensemble de la population les caractéristiques des personnes, de ses parents et aussi du ménage. Compte tenu de la qualité de certaines variables (statut professionnel et plus haut niveau d'éducation), des variables de cette base ont été combinées avec d'autres provenant du Census de 2011 via la fonction *merge ()*. Une autre base de données qui retrace les naissances de 1998 à 2016 a été combinée via la même fonction *merge ()* aux bases de données précédentes en utilisant les identifiants communs. C'est cette base finale qui a été considérée comme « données administratives » (voir annexe 2).

II.3. Pourquoi DEMOBEL et ESS ?

D'abord, il n'y a actuellement aucune base de données qui permettent de relier la religiosité et le comportement de fécondité en Belgique. La fusion des deux bases de données se positionne comme solution à cette situation. Ensuite, les données de la ESS et celles mises à disposition par DEMOBEL sont collectées sur une même population. La ESS connaît

⁴ Pour en savoir plus sur le projet DEMOBEL : <https://statbel.fgov.be/fr/nouvelles/demobel-nouvelle-base-de-donnees-demographique>

comme toutes les enquêtes en face à face, une baisse continue du taux de réponse liée à une diminution de la participation volontaire. Le taux de réponse supérieur à 50 % observé en 2010 est passé à 39,19 % en 2020 en Belgique. Dans cette vague (2020), l'enquête n'a couvert que les 78 plus grandes municipalités, soit 48,8 % de la population belge d'où le choix de la vague 5 et 6 qui couvre l'ensemble du territoire et de la population avec des taux de réponse plus satisfaisants.

Chaque vague de ESS contenant un nombre d'observations très faible, la combinaison des vagues 5 et 6 avait pour objectif d'en accroître l'effectif. Ce choix se base sur l'hypothèse d'un effet non significatif du temps séparant la date de référence (2011 pour les données administratives) des dates de réalisations des vagues 5 (2010) et 6 (2012). Le choix de la période de référence de 2010-2012 tient aussi compte du fait que c'est à ce moment que la Belgique connaît un des pics les plus importants du nombre de naissance par femme (IWEPS, 2023). Plus généralement, le choix de cette période a eu pour effet de limiter non seulement les effets liés à la crise sanitaire de la COVID19, mais aussi ceux liés à l'écoanxiété de plus en plus présente, notamment chez les plus jeunes générations. Ces choix se tiennent compte des enjeux de la recherche et de sa dimension réduite à celle d'un mémoire de master. Toutefois, ces éléments pourraient être pris en compte dans des développements futurs.

Ce chapitre fait une revue sur la fusion statistique des données en contexte général avant de décrire l'approche telle que conçue pour les besoins de cette étude. Il se termine sur la présentation de la méthodologie utilisée pour l'étude du comportement en matière de fécondité à partir de la base de données produite.

III.1. Fusion statistique des données ou *Statistical Matching* : théorie

III.1.1. Généralités sur la fusion des données et importance

Les données d'enquêtes représentent l'une des principales sources de données dans les statistiques officielles (Rezaei Ghahroodi, 2023). Le secteur des enquêtes et des sondages d'opinion est confronté à une baisse de plus en plus importante des taux de réponse. Cette baisse s'explique notamment par la croissance des sollicitations, de la multiplication des modes de contacts utilisés. Il fait aussi face aux coûts élevés pour la mise en place d'une enquête, pour laquelle, à cause de la charge du répondant ou *respondent burden* en anglais, il n'est souvent pas possible de collecter des informations sur certaines variables d'intérêts (Rezaei Ghahroodi, 2023) du moins de manière conjointe. De plus, les réflexions citoyennes qui ont précédé l'entrée en vigueur le 25 mai 2018 du Règlement Général Européen sur la Protection des Données ont fait prendre conscience aux citoyens de la valeur des informations les concernant⁵. Dans ce contexte, les chercheurs et autres utilisateurs de données doivent innover en permanence dans la collecte des données, proposer des questionnaires plus simples et plus courts et des modes de collecte adaptés, et ce afin de limiter le fardeau de réponse. Ainsi, plutôt que de recueillir l'ensemble des informations souhaitées auprès d'un unique échantillon, ils sont de plus en plus contraints de collecter une partie des variables d'intérêt auprès d'un premier échantillon et l'autre partie auprès d'un second échantillon, disjoint du premier.

L'enjeu est dès lors de combiner ces sources d'information multiples de manière à obtenir une information complète : c'est ici qu'intervient l'intégration des données. Il existe différentes techniques d'intégration des données. D'Orazio et *al.* (2006a) les regroupent en trois catégories dont la première est celle du couplage des données ou *record linkage* (en anglais) qui se base sur la présence d'un identifiant unique commun deux bases comme

⁵ Règlement (UE) 2016/679 du Parlement Européen et Conseil — du 27 avril 2016 — relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données), 2016.

principe pour la fusion de celles-ci. La seconde est l'échelle statistique ascendante (descendante) ou *statistical Up (Down)* qui a été développée dans les domaines de la recherche pour adapter les données collectées à différents niveaux d'agrégation afin d'élargir ou de restreindre l'information. Enfin, la troisième méthode est celle de la fusion statistique des données ou *statistical matching* qui consiste à compléter un échantillon dit « receveur », où certaines informations ne sont pas présentes, à l'aide d'un échantillon dit « donneur » portant sur d'autres individus. Pour Lewaa et *al.* (2023), le couplage et la fusion statistique des données sont les principales méthodes mobilisées dans l'intégration des données.

La fusion statistique est un cas particulier du traitement de la non-réponse. C'est une approche basée sur un modèle pour produire des informations croisées sur des variables collectées via plusieurs sources, généralement des enquêtes, issues de la même population (Rezaei Ghahroodi, 2023). Les méthodes de fusion statistique de données visent à répondre à la problématique d'identifier des unités identiques au sein de différents fichiers. Apparue au milieu des années 60 (Okner, 1972), leur usage s'est grandement répandu dans les années 80-90 dans le domaine des études de marché et d'audience, car elles offrent une réponse économique et pragmatique à un besoin croissant d'informations économiques. Étant donné qu'un recueil complet d'informations peut s'avérer complexe et coûteux, le résultat de la fusion statistique des données est un fichier complet, simple d'utilisation et de traitement par des logiciels standard. C'est en ce sens que la fusion statistique devient de plus en plus, une des méthodes utilisées afin de rapprocher des données provenant d'enquêtes différentes.

La première publication sur la fusion statistique des données remonte à Okner (1972) (D'Orazio et *al.*, 2006a ; Rezaei Ghahroodi, 2023). Il produisit en 1972 dans son article intitulé « *constructing a new data base from existing microdata sets: the 1966 merge file* », un document nommé « 1996 MERGE file » issu de la fusion d'informations contenues dans l'enquête sur les opportunités économiques de 1967 et du fichier fiscal de 1966 des États-Unis (US). Il s'agit d'un travail précurseur dans le domaine de la fusion statistique des données. L'auteur part du constat de l'existence d'une multitude de bases de données traitant des revenus totaux aux US et d'une absence d'un fichier rapporté au niveau de ménages et pose ainsi la nécessité de la création d'une nouvelle base de données. Cette méthode connaîtra une grande diffusion à partir des années 80 (Barry, 1988) où elle est utilisée dans de nombreux domaines, notamment celui des sciences sociales (Donatiello et *al.*, 2022 ; Moretti & Shlomo, 2023) et économiques (Perchinunno et *al.*, 2020 ; Rodgers, 1984), celui de la

santé (Magnoni et *al.*, 2023), des politiques publiques (D’Alberto et *al.*, 2021), des sciences de la conservation notamment, l’évaluation des politiques de conservation de la biodiversité (Schleicher et *al.*, 2020) ou encore de l’éducation (Lecocq et *al.*, 2016). Trewin (2023) signalait son importance dans les études longitudinales par la combinaison de nombreuses bases de données pour suivre des événements comme la mortalité des individus en Australie et en Nouvelle-Zélande.

En somme, la fusion statistique des données offre un potentiel énorme dans l’exploitation et l’extraction d’informations des bases de données possédant un ensemble de variables communes, d’autres variables spécifiques (observés séparément pour chaque source considérée), et aussi des unités d’observations distinctes (D’Orazio et *al.*, 2006a). Elle est d’autant plus importante qu’elle mobilise de moyens financiers et humains plus réduits pour la production de données dont l’accessibilité sur le terrain est de plus en plus compromise. Elle souffre cependant de faiblesses dont celles d’ordre méthodologique, lié d’abord à la nature des données collectées, à la structure de base des données et enfin aux modèles utilisés pour le processus de fusion (Scanu, 2011).

III.1.2. Grandes étapes de la fusion des données

Pour la suite des articulations de cette section, prenons en compte cette formulation : 2 bases de données A et B possèdent des variables communes regroupées sous l’ensemble X. La base A possède la variable d’intérêt Y et la base B, la variable Z. Les variables Y et Z sont spécifiques à A et B et justifient le processus de la fusion statistique. L’enjeu est de combiner A et B pour obtenir un nouveau fichier de données C, contenant les variables Y, Z et X.

III.1.2.1. Considérations initiales

La fusion des données repose sur un principe bien établi : les 2 bases de données sont collectées indépendamment l’une de l’autre. Avant la réalisation de la fusion statistique des données, des considérations importantes sont à prendre. D’abord, il est nécessaire de faire l’hypothèse ou postulat d’une indépendance conditionnelle ou *conditional independent assumption* (CIA) en anglais. Ce postulat considère Y et Z comme indépendantes par rapport aux variables communes X (D’Orazio et *al.*, 2006b). Il s’agit de l’hypothèse à la base des différentes méthodes de fusion statistique (Lecocq et *al.*, 2016). Ensuite, il faut préciser l’approche envisagée, possible à partir des objectifs poursuivis. Ces approches peuvent être micro ou macro, paramétriques ou non paramétriques (D’Orazio, 2011 ; Scanu, 2011).

Dans l'approche micro, D'Orazio, et *al.* précisent que l'objectif est la construction d'un fichier complet dans le sens où, toutes les variables d'intérêts, bien que collectées séparément y sont contenues. Sa nature synthétique est liée à son mode production. En effet, il n'est pas le produit d'une observation directe d'un ensemble de la population, mais le fruit d'une exploitation d'informations contenues dans les fichiers sources de manière appropriée. L'approche macro quant à elle utilise des fichiers sources pour l'estimation directe de la fonction de distribution conjointe des variables d'intérêt qui n'ont pas été observées en commun (2006a, p. 2-3). L'approche micro est la plus utilisée principalement à cause de sa facilité d'usage dans les analyses, mais aussi de son potentiel pour satisfaire le besoin d'informations des universités, des centres de recherches, des institutions publiques (Conti et *al.*, 2016 ; Scholtus et *al.*, 2014 ; Van der Putten et *al.*, 2002). Elle est aussi considérée comme une approche prédictive. L'approche macro reste bien adaptée quand le but est de produire un tableau de contingence pour des variables non observées conjointement ou des coefficients de corrélations en rapport à une population entière (D'Orazio et *al.*, 2006b ; Moretti & Shlomo, 2023).

Enfin, il est tout aussi important de définir de quel type de paramétrage il est question notamment paramétrique ou non paramétrique. Les tests paramétriques sont connus comme celles reposant sur les principes d'une loi comme c'est le cas avec la loi de la distribution normale ou celle de poisson. Les tests non paramétriques quant à eux ne se reposent pas sur les principes d'une loi. Ils sont particulièrement utiles quand le nombre d'observations est très faible ($n < 30$) (D'Orazio et *al.*, 2006a).

III.1.2.2. Harmonisation des bases données

De nombreuses incohérences peuvent exister dans les bases de données existantes compte tenu des méthodes impliquées dans leur production. L'harmonisation est une étape très importante qui permet une homogénéisation des deux bases de données et d'affiner ainsi les résultats de la fusion. Van der Laan cité par D'Orazio et *al.* (2006a) distingue huit principales actions dans l'harmonisation de 2 bases de données. Les 5 premières actions sont typiques de la phase d'harmonisation. Les 2 actions qui suivent ne sont utilisées que lorsque les bases de données sont affectées par des erreurs (données manquantes, erreurs d'échantillonnage). La dernière étape permet la synthèse de nouvelles variables à partir de celles présentes dans les bases de données. Ces actions sont présentées ci-contre.

1. L'harmonisation des unités de définition : ces unités peuvent être des ménages, des individus, etc.

2. L'harmonisation des périodes de références qui peut se résumer à l'uniformisation des périodes de collecte des données.
3. La définition des populations : les caractéristiques qui définissent une population peuvent varier d'une enquête à une autre. Il peut par exemple être difficile, voire impossible de fusionner une enquête sur une population d'hommes uniquement avec celle sur une population de femmes uniquement.
4. L'harmonisation des variables : il est question ici de gérer les problèmes posés par les majuscules, les minuscules, les codifications ou encore les problèmes liés à la définition des variables elles-mêmes.
5. L'harmonisation des classifications.
6. L'ajustement pour la mesure des erreurs.
7. L'ajustement pour les données manquantes.
8. La synthèse de variables au cours de laquelle certaines variables peuvent être recodées ; et d'autres, créées.

Les choix des actions reposent aussi sur le fait que les deux sources de données soient biaisées ou incohérentes, comme précisé dans l'ouvrage « *Statistical matching : theory and practice* » de D'Orazio et al. (2006a). Lorsque les périodes de références des sources de données sont différentes. Dans ce cas, les 3 premières actions de la liste plus haut sont les mieux indiquées. Ensuite, des données peuvent partager la même période de référence, mais pour des populations différentes ; dans ce cas, se référer aux actions 4 et 5. Pour faire face à ce problème, des hypothèses de départ peuvent être formulées. Les incohérences dans les bases de données sont aussi liées à la définition des variables et à leurs catégorisations. Elles ne peuvent pas être résolues par des hypothèses. Un travail important doit être réalisé pour l'harmonisation des variables d'intérêts pour la fusion statistique dans les bases de données. Cependant, certaines variables ne peuvent pas être harmonisées et doivent donc être exclues de la fusion. D'autres par contre peuvent être substituées par la transformation d'informations disponibles. Plus évidents encore, d'autres peuvent juste être recodés et recatégorisés.

III.1.2.3. Sélection des variables de fusion

Chaque enquête permet de collecter des informations sur des variables variées regroupées dans des bases de données. Ainsi, plusieurs bases de données peuvent partager tout un ensemble de variables communes qui peuvent être utilisées dans le processus de fusion. On peut par exemple citer l'âge, le genre, le lieu de résidence, le statut professionnel ou encore le niveau d'éducation. Il est généralement d'usage de se baser sur la littérature et

aussi sur des méthodes statistiques utilisées qui elles-mêmes, reposent très souvent sur les caractéristiques de ces variables pour la sélection de ces variables communes (D’Orazio, 2017). La littérature permet en effet d’identifier des variables qui peuvent être intégrées au processus. Les méthodes statistiques permettent cependant une sélection de celles qui seront effectivement utilisées pour la fusion. Ainsi, toutes les variables communes ne sont pas utilisées. Il est parfois nécessaire ou utile de les regrouper en 2 groupes dont celui des variables de fusion X_f ou *matching variables* et celui des variables non directement utilisés dans la procédure de fusion (D’Orazio, 2011). La sélection des variables de fusion X_f pour chaque base de données A et B repose sur l’évaluation des relations entre X et Y dans A ; et aussi sur l’évaluation des relations entre X et Z dans B. Les variables de fusion X_f retenues sont l’union des deux sous-ensembles des variables communes identifiées.

En pratique, D’Orazio (2011) précise qu’une analyse multivariée peut être une solution acceptable pour la sélection des variables de fusion. Elle permet de calculer dans un premier temps les associations ou corrélations entre chaque variable commune et la variable Y dans la base A ; et dans un second temps, les associations ou corrélations entre chaque variable commune et la variable Z dans la base B. Un test de chi-carré de Pearson peut être utilisé pour tester l’association entre des variables catégorielles. Dans le cas de variables ordinal (variables catégorielles dont les modalités sont ordonnées, représentant un niveau de gradation), c’est celui de Kendall ou Somer qui peut être utilisé par exemple. Plus généralement, le test de corrélation est très souvent utilisé pour tester l’hypothèse d’indépendance linéaire entre les variables continues (D’Orazio et al., 2006a). Le type de variable influence fortement les méthodes qui doivent être utilisées pour la sélection des variables de fusion. De nombreuses autres corrélations ou associations peuvent donc ainsi être observées à partir d’indicateurs comme le coefficient de corrélation de Pearson en présence d’échelles d’intervalles et/ou ordinales ou encore le coefficient de corrélation de rang de Spearman pour des échelles d’intervalles (D’Orazio, 2011).

III.1.2.4. La fusion statistique des données

Différentes techniques sont utilisées pour la fusion des données. Dans le cas de cette sous-section, seules celles qui sont les plus courantes seront présentées. Ces techniques sont en constantes évolutions compte tenu des progrès fulgurants que connaît le domaine de l’informatique surtout en ce qui concerne la programmation et l’automatisation des processus de calcul. Sous le postulat d’indépendance conditionnelle, D’Orazio et al. (2006a) propose d’utiliser différentes techniques d’imputation, dont celle dite de *hot deck* ; qui est la méthode

propre aux approches micro non paramétriques. Elle permet la production d'un jeu de données synthétiques complètes à partir du remplacement aléatoire ou imputation des variables manquantes d'un fichier receveur à partir d'un autre considéré comme donneur. Les auteurs précisent qu'il est d'usage de considérer le plus petit fichier de donnée comme receveur, et le plus grand comme donneur. Ceci se justifie par le fait qu'il sera possible pour le donneur de donner plus d'une observation au receveur. Ils classent cette famille de méthode en 3 catégories dont le *random hot deck*, le *rank hot desk* et le *distance hot desk*⁶

La *distance hot desk* est la plus utilisée et se base sur la distance du plus proche voisin pour réaliser les imputations à partir de l'estimation d'une compatibilité entre les variables de fusion. Il s'agit de la méthode de fusion la plus utilisée notamment par Okner (1972). Chaque observation du fichier peut être utilisée plus d'une fois. On parle d'une méthode sans contrainte. Dans le cas où l'approche est conditionnée au fait que chaque observation ne soit utilisée qu'une seule fois, la méthode est dite sous contrainte (Rodgers, 1984). Selon Lecocq et al., « *l'appariement sur le plus proche voisin n'est toutefois pas toujours performant, notamment lorsque le voisin le plus proche se trouve à une distance éloignée de l'individu traité à appairer. Des individus très différents peuvent alors être jumelés. D'un point de vue pratique, il convient d'ordonner aléatoirement les données avant d'appairer, puisque le pairing s'effectue par ordre* » (2016, p. 81). L'ordination des individus avant le pairing est donc très importante ; auquel cas, elle constituerait une source de biais. Il s'agit ici d'une limite majeure à prendre en compte lors de l'utilisation de la méthode.

Le *random hot deck* procède aléatoirement pour sélectionner une observation dans le fichier donneur pour le faire correspondre à une observation dans le fichier receveur (D'Orazio et al., 2006b). Il est possible à ce niveau de procéder par des stratifications qui permettent d'améliorer la robustesse de l'algorithme. Cette stratification peut se faire par une ou plusieurs variables comme le lieu de résidence (ne faire correspondre que les individus d'une même aire géographique) uniquement ou le lieu de résidence et le statut professionnel (les individus d'une même aire et exerçant le même emploi ou dans le même secteur d'activité) par exemple. Elle permet d'éviter le jumelage d'individus trop distants par la création de strates au sein desquelles les individus seront appariés (Lecocq et al., 2016). L'une des limites de cette méthode est que les individus peuvent ne pas être appariés alors que la distance les séparant est faible, en raison du fait qu'ils sont situés dans des strates

⁶ Les termes sont mis en anglais pour conserver tout leur sens, car leur équivalent n'a réellement pas été trouvé en français.

différentes. Il peut en résulter une diminution de la puissance statistique et un accroissement du nombre d'individus non appariés qui est très souvent supérieur à celui obtenu avec la méthode du plus proche voisin (Lecocq et *al.*, 2016). Elle offre cependant de meilleurs résultats en raison d'une plus forte ressemblance entre les individus jumelés.

Les approches micro sous réglages paramétriques regroupent 2 catégories de méthodes comme décrites par D'Orazio et *al.* (2006a). La première se base sur la correspondance moyenne conditionnelle. Il s'agit de remplacer chaque élément manquant par l'espérance de la variable manquante compte tenu des variables observées. Cela peut être fait de manière simple lorsque les variables dans Y et Z sont continues. La deuxième catégorie de méthode concerne le tirage basé sur des distributions prédictives conditionnelles. Ces approches comme l'essentiel des approches paramétriques font appel aux régressions pour l'imputation de valeur de manière aléatoire. Les approches macro se basant sur des réglages paramétriques reposent aussi sur les statistiques de distribution. Elles reposent sur trois approches dont (1) celle sur la distribution univariée normale, (2) sur la distribution multinormale qui est une extension de première appliquée à un cas général multivarié et enfin, (3) sur une distribution multinomiale. L'approche macro ne faisant pas partie de ce travail, ces méthodes ne seront pas davantage développées.

D'Orazio et *al.* (2006a) notent enfin que de nombreuses recherches font appel à des approches mixtes (paramétrique et non paramétrique) dans la réalisation des fusions statistiques des données. Le mélange des approches consiste d'abord à adopter un modèle paramétrique, et ensuite obtenir le jeu de données synthétiques complet par le biais d'une approche micro non paramétrique. L'idée est d'exploiter deux propriétés propres à chacune des approches : le modèle paramétrique est plus parcimonieux que celui non paramétrique ; le modèle non paramétrique est cependant plus robuste à la mauvaise ou non spécification du modèle. D'une manière générale, les méthodes mixtes peuvent être considérées comme une procédure en deux étapes : une première étape qui consiste à estimer les paramètres du modèle paramétrique et une seconde qui consiste à utiliser l'une des techniques *hot deck* fonction des paramètres estimés lors de la première étape avec 2 fichiers, dont un donneur et l'autre receveur (D'Orazio et *al.*, 2006a, p. 47).

Tableau II. Approches possibles en fonction des objectifs de la fusion et du paramétrage.
(Source : auteur à partir de D’Orazio et *al.*, 2006a ; Scanu, 2011).

	macro	micro	Mixte
paramétrique	Distributions normales, multinormales et multinomiales	Distributions normales, multinormales et multinomiales	Combine le paramétrique et non paramétrique (uniquement micro)
Non paramétrique	Autres	<i>Hot-deck</i>	

L’ensemble des techniques de fusion statistique des données développées sous la CIA se base sur les 4 paramétrages de base à savoir macro paramétrique, macro non paramétrique, micro paramétrique ou micro non paramétrique (tableau I). La méthode basée sur le score de propension très rencontrée dans la littérature est réalisée en suivant les principes de la méthode *hot deck* (Lecocq et *al.*, 2016 ; Magnoni et *al.*, 2023).

III.1.2.5. Évaluation de la qualité de la procédure de fusion

Cette évaluation doit tenir compte de l’hypothèse faite quant aux données de départ, du modèle, de l’approche choisie (macro ou micro) et des techniques qui seront utilisées pour l’analyse de la base de données produite (D’Orazio et *al.*, 2006a). Un consensus n’est pas disponible entre toutes les propositions de techniques d’évaluation de la qualité de la fusion existantes. Selon D’Orazio (2011), l’erreur des moyennes carrées (EMC) est utilisée en statistique pour évaluer la précision des paramètres d’estimations. Cependant, des processus qui font intervenir une phase de fusion statistique des données rendent très difficile son estimation. Dans le cas des approches macro, l’auteur précise que les procédures alternatives consistent à évaluer la fiabilité des estimations obtenues à la fin de la fusion soit (1) en vérifiant si elles sont cohérentes avec le cadre de la fusion ou (2) en les comparant avec des estimations provenant de sources externes à condition qu’elles n’aient pas été utilisées dans la procédure de fusion.

Dans le cas des approches micro, Rässler (2002) citée par D’Orazio et *al.* (2006a) propose un cadre d’évaluation de la validité de la procédure de fusion qui repose sur 4 niveaux d’évaluation dont : (1) la reproduction des variables inconnues dans le fichier destinataire ; (2) la façon dont la distribution est reflétée dans l’ensemble des données fusionnées ; (3) la préservation de la structure de corrélation pour les distributions conjointes et marginales qui ne peut être évaluée qu’à partir de simulations (se référer à Paas, 1986 cité par D’Orazio (2011)) ; (4) la préservation de la distribution marginale dans le fichier de

données fusionnées à partir de tests statistiques comme le test chi-carrée ou des mesures de similarité/dissimilarité si les variables sont catégorielles. Pour des variables continues, on peut utiliser des tests comme celui de Kolmogorov-Smirnov, Mann-Whitney ou Wald-Wolfowitz par exemple. Des auteurs (Lecocq et *al.*, 2016 ; Tran & Osier, 2023) proposent comme approche pour l'évaluation du processus de fusion de comparer les individus selon les groupes, avant et après appariement, sur plusieurs variables. En effet, une fusion de qualité permettra d'observer peu ou pas de différences entre les variables. Dans le cas contraire, la fusion doit être reprise jusqu'à l'obtention d'un résultat satisfaisant.

III.1.3. Fusion statistique et informations auxiliaires

L'un des problèmes communs de la fusion statistique des données est lié à la CIA entre les variables observées dans les différentes sources de données (Moretti & Shlomo, 2023). Cette hypothèse est particulièrement forte et se vérifie rarement dans la pratique. Dans le cas où elle ne se vérifie pas, les résultats de la fusion dérivés de cette hypothèse ne seront pas valables (D'Orazio, 2022). C'est pour pallier cela que des informations dites auxiliaires peuvent être utilisées pour améliorer le processus (D'Orazio et *al.*, 2006a ; Moretti & Shlomo, 2023). Elles sont utilisées quand la CIA ne tient pas. Il s'agit dans certains cas de fichiers de données dans lesquels les variables d'intérêt Y et Z et, dans certains cas X, ont été observées conjointement (D'Orazio et *al.*, 2006b). Leur intégration dans le processus de fusion permet d'assurer le contrôle dudit processus. Ces informations auxiliaires peuvent donc être utilisées pour valider la relation entre les variables d'intérêts (Y et Z) dans le nouveau fichier synthétique de données obtenu et donc de confirmer ou pas la CIA (Tran & Osier, 2023). Dans le cas où cette CIA est justifiée, son utilisation n'est plus nécessaire.

Pour généraliser, la fusion des données peut soit se faire sous la CIA ou en utilisant des informations auxiliaires (Rezaei Ghahroodi, 2023). En utilisant la CIA, on fait l'hypothèse non pas d'une indépendance conditionnelle, mais celle du monde le proche pour reprendre l'expression « *the closest world* » de D'Orazio et *al.* (2006b). Dans le cas des approches mixtes, elles sont aussi utilisées comme c'est cas avec la méthode mixte multivariée pour la fusion statistique proposée par Claramunt González et *al.* (2023). Les auteurs ont en effet utilisé un jeu de données supplémentaires pour améliorer le calcul des valeurs intermédiaires, et des contraintes pour améliorer la sélection des donneurs. L'une des faiblesses de l'utilisation des informations auxiliaires et aussi un facteur entravant son déploiement repose sur le fait que de telles bases données (complètes) ne sont pas toujours disponibles, les données disponibles étant de plus en plus partielles (D'Orazio et *al.*, 2006b).

III.2. Fusion statistique des données ou *statistical matching* : méthode

Après une partie théorique présentée plus haut, cette seconde section se concentre sur l'approche conçue et mise en œuvre dans le cadre du travail. Dans cette étude, le choix a été fait de procéder par la méthode de D'Orazio et *al.* (2006a) décrite dans l'œuvre « *statistical matching : theory and practice* » et l'utilisation du package StatMatch dans sa dernière version 1.4.1 (D'Orazio, 2022) sur R. Ce choix est lié à sa facilité de manipulation et à sa capacité à implémenter les techniques les plus communes de la fusion des données (D'Orazio, 2013). La fusion des données est réalisée sous l'hypothèse de l'indépendance conditionnelle. Le besoin d'un fichier synthétique de données l'ancre dans une approche micro non paramétrique.

III.2.1. Choix des variables spécifiques et mesures : religion et fécondité

III.2.1.1. La religion dans la ESS

La mesure de la religion dépend des enquêtes, du contexte du pays dans lequel elles se déroulent ou du système de nomenclature. Il est donc possible d'observer dans des enquêtes, différentes variables qui renvoient toutes à la religion ou plus largement à la religiosité. Ainsi, une question qui serait : « appartenez-vous à une religion ou à une obédience religieuse particulière ? » et dont la réponse laisserait le choix à « oui » ou à « non » permettrait déjà d'avoir des informations sur le fait que l'individu soit croyant ou pas. Dans l'affirmative, d'autres variables peuvent permettre d'identifier l'appartenance ainsi que le niveau d'implication. Les variables mesurant la religion sont dans de nombreuses études, le résultat d'une agrégation à partir de celles qui sont disponibles (Peri-Rotem, 2020 ; Dantis et *al.*, 2023). Un indicateur très utilisé de la religiosité est la fréquence de participation aux services religieux (Bein et *al.*, 2017, Dantis et *al.*, 2023).

Dans la ESS, plusieurs variables traitent de la religion. Selon Honoré (2021), de plus en plus de croyants font le choix de se sortir du groupe pour pratiquer leur religion. Cela inclut des prières et autres pratiques religieuses de manière individuelle ou en groupe très restreint. Il est possible d'observer des croyants qui se disent n'appartenir à aucune religion ou obédience particulière. Ainsi, en se basant sur le seul fait de l'appartenance religieuse, on court le risque de les considérer comme non croyants. Dans cette étude, le degré de religiosité a été retenu comme indicateur pour la religion. Cette variable permet de mesurer, indépendamment de l'obédience religieuse, l'importance de la religion ou simplement de Dieu dans la vie des individus. Cette question de l'importance de Dieu comme indicateur a

d'ailleurs été utilisée par Rizzi (2013) dans le contexte spécifique de l'Italie. La variable degrés de religiosité est une variable catégorielle ordinale de 11 modalités codées de 0 pour non-croyant à 10 pour très croyant et concerne 1 766 femmes.

III.2.1.2. La fécondité dans les données administratives

La fécondité est considérée dans de nombreuses études dans une dimension prospective. Ainsi, des auteurs vont utiliser les intentions de fécondité pour la mesurer. L'indicateur est souvent une variable dichotomique (oui ou non) sur laquelle sont appliquées des régressions logistiques (Bein et *al.*, 2017 ; Dantis et *al.*, 2023). Il peut aussi être une variable catégorielle qui est une projection des intentions sur les années futures (ne veut pas d'enfants, en veut dans l'année, dans 2 ans ou plus) comme utilisées par Kim (2017) qui le considère comme étant un bon prédicteur des comportements en matière de fécondité. D'autres études par contre, mesurent la fécondité par la variable classique du nombre d'enfants par femme (et par génération) (Peri-Rotem, 2020). Il s'agit d'une variable qui n'est pas très souvent disponible au moment de l'enquête, mais qui est synthétisée à travers d'autres variables comme celui du nombre d'enfants déjà mis au monde par chaque femme et l'âge. Définir un indicateur pour l'évaluation du comportement de fécondité dans le contexte de cette étude s'est révélé être un exercice assez contraignant. Deux indicateurs ont ainsi été créés dans un premier temps et uniquement pour le processus de fusion. Il s'agit du nombre de naissances par femme d'au moins 18 ans en 2011 et du nombre d'enfants pour la période de 2012 à 2016 (5 ans) pour 5 598 920 femmes. Dans un second temps, le nombre d'enfants par femme est utile pour sélectionner les femmes avec un enfant en 2011 et voir qui a eu un deuxième enfant dans les cinq ans qui suivent entre 2012 et 2016.

Tableau III : caractéristiques de la population des deux échantillons utilisés dans la fusion.

Critères	Bases de données	
	ESS 2010 et 2012	DA 2011
Genre	femmes	femmes
Lieu de résidence	Belgique	Belgique
Âge	18 ans +	18 ans +
Effectif	1766 femmes	5 598 920 femmes
Variable spécifique	Religiosité	Nombre de naissances

III.2.2. Harmonisation des variables communes

L'harmonisation a été une étape majeure dans le processus de fusion des données et aussi la plus chronophage. Elle a été réalisée suivant les étapes décrites ci-contre.

III.2.2.1. Identification des variables communes

Le choix des variables communes n'est pas le fait d'un simple hasard. D'Orazio et *al.* (2006a) précisent que ce choix doit être fonction de l'intérêt de l'étude, de la structure des données, mais aussi de la littérature et de la consultation d'experts. L'identification des variables communes a été fonction de leurs disponibilités dans les deux bases données par la confrontation desdites listes. Ainsi, les variables communes suivantes ont été identifiées : l'âge du répondant, son pays de naissance et sa nationalité, le statut marital, la situation professionnelle, le plus haut niveau d'éducation, la province de résidence et des informations sur les parents (pères et mères) telles que le pays de naissance, la nationalité.

III.2.2.2. Harmonisation des variables

L'harmonisation des variables retenues consistait à analyser la définition et la qualité des variables et à procéder à leurs uniformisations. Cette étape a permis d'exclure des variables de faible qualité et celles dont l'harmonisation des définitions était impossible. À cette étape, les données sur les parents ont été écartées. La variable pays de naissance a été recatégorisée en 2 modalités (Belgique et hors Belgique), de même que la variable nationalité (belge et non belge) et le statut d'occupation (travaille ou ne travaille pas). Le niveau d'éducation contient 4 catégories (aucun diplôme, primaire, secondaire et supérieur). Le statut matrimonial légal quant à lui contient 3 modalités (femmes en union légale, femmes qui ne sont pas engagées dans une union, femmes qui n'ont jamais été engagées dans une union légalement reconnue). Le lieu de résidence correspond à la province avec 11 modalités. Ces variables sont de bonnes qualités avec une très faible proportion de données manquantes. Il s'agit d'un problème qui entrave cependant le fonctionnement de l'algorithme de fusion. Le problème a été géré par imputation via le package MICE sur la base d'une imputation par correspondance prédictive idéale pour les variables catégorielles (Tran & Osier, 2023).

III.2.3. Analyse de la distribution et choix des variables de fusion

Les variables communes utilisées dans le processus de fusion doivent respecter le double critère d'une distribution conjointe plus ou moins identique et de la meilleure prédiction possible avec la variable spécifique. Toutes les variables communes n'étant pas utilisables dans le processus de fusion (D'Orazio et *al.*, 2006a), il est dès lors nécessaire d'en

sélectionner celles qui s’y prêtent le mieux. L’analyse des distributions a reposé sur le calcul du coefficient de Bhattacharya (D’Orazio et *al.*, 2017). Il s’agit d’un coefficient qui permet de mesurer la similarité entre deux distributions de probabilité. Plus il est proche de 1, plus les distributions sont similaires. À ce coefficient a été associé le calcul des distances totales de variation qui mesurent à l’inverse la différence entre deux distributions de probabilité. Plus elle est faible, plus les distributions sont similaires. La fonction *comp.prop* de Statmatch a été utilisée. La capacité de prédiction a été testée au moyen de régressions linéaires multiples. En effet, la fécondité est une variable continue. Le degré de religiosité quant à lui, bien qu’il soit une variable catégorielle, le nombre de ses modalités (11) et leur classement (de non croyant à très croyant) a justifié l’utilisation d’une régression linéaire multiple. Les variables indépendantes étant catégorielles, une fonction *step ()* a été appliquée au modèle. L’application d’une *anova* sur la fonction *step ()* a permis d’identifier les meilleurs prédicteurs qui ont été utilisés pour la fusion.

III.2.4. Fusion statistique de la ESS et des données administratives

Les formulations de « donneur » pour les DA et de « receveur » pour la ESS ont été initialement retenues. D’Orazio (2017), recommande en effet de considérer la base avec le plus petit effectif comme receveur. À l’inverse, une autre fusion en considérant les DA comme receveur et ESS comme donneur a été testée pour une comparaison des résultats. La faisabilité de cette dernière option contribuerait d’abord à accroître la taille de la population qui surpasserait les limites imposées par celle de la ESS (environ 1 700 personnes par vague). Ceci serait d’un apport substantiel pour bénéficier de la richesse des données administratives, surtout en ce qui concerne leur caractère territorial, pour des analyses les plus fines sur un ou des ensembles de la population belge. Des auteurs (Rezaei Ghahroodi, 2023 ; Wilkens, 2018) recommandent de faire le choix du receveur ou donneur en fonction des caractéristiques des bases de données et aussi des objectifs de la recherche.

La fusion statistique s’est faite sur la base de distance du plus proche voisin ou *distance hot deck* (*NND.hotdeck* sur StatMatch). Le principe de fonctionnement est le suivant : pour une observation dans le receveur, l’algorithme cherche dans le donneur celle qui lui est la plus proche selon la fonction de distance choisie. Les variables de fusion étant catégorielles, la fonction *gower.dist* du package *proxy* (il s’installe automatiquement avec StatMatch) a été utilisée (D’Orazio, 2022). Une stratification en fonction de la province de résidence a été ajoutée au modèle par ailleurs non contraint (une observation du donneur pouvait être utilisée *n* fois dans le receveur). La fonction *NND.hotdeck ()* ne crée pas

directement un fichier de données synthétique, mais une correspondance entre les observations du receveur et du donneur en fonction des classes d'imputation (province). Pour obtenir le fichier de données, la fonction *create.fused()* a été utilisée (D'Orazio, 2013). Cette dernière étape a permis d'obtenir le résultat final qui est un fichier synthétique de donnée contenant en même temps les variables spécifiques de départ et les variables communes.

III.2.5. Évaluation de la qualité de la fusion des données

Évaluer le résultat d'une fusion statistique n'est pas simple compte tenu de nombreux facteurs dont : le fait que le phénomène étudié concerne des paramètres non observés ensemble ; une différence dans les données et enfin dans les moyens de collectes (D'Orazio et al., 2006a). Toutefois, cette évaluation a été réalisée en se basant sur l'analyse des distributions marginales des variables dans le donneur et celles dans le fichier synthétique final selon Rässler (2002) citée par (D'Orazio et al., 2006a). En somme, les distributions marginales des variables des DA ont été comparées à celles présentes dans les fichiers synthétiques de données obtenues en calculant la distance totale de variation (TVD).

III.3. Analyse de la relation entre fécondité et religion

Le but est ici de partir d'une situation de base en 2011 pour analyser le comportement en matière de fécondité des femmes en Belgique qui ont déjà un enfant et ceux dans les 5 années suivantes (2012-2016). La fusion de données est essentielle dans ce processus en ce sens qu'elle représente la seule manière d'étudier le comportement (actions observables des individus) des femmes contrairement aux attitudes (ressenti et prédispositions à agir de certaines manières). Le fichier de données créé a été ainsi utilisé pour étudier ce comportement avec la religion comme principale variable explicative suivant les étapes décrites ci-contre.

III.3.1. Pourquoi étudier l'arrivée d'un second enfant ?

La situation démographique en Europe et en Belgique nécessite une attention particulière en ce sens qu'il y a un recul de l'âge à la première naissance et aussi une tendance à ne pas vouloir d'enfants. Au lieu de s'intéresser à la troisième naissance comme habituellement observée dans la littérature, cette étude questionne la religion comme déterminant de l'arrivée d'un second enfant. Son importance tient du fait que dans la population des femmes habitant en Belgique et qui ont fait le choix de la maternité (ont déjà un enfant), elle essaye de comprendre quels sont les éléments qui influencent ou conditionnent l'arrivée d'un second.

Le but est donc de fournir des éléments d'explications du faible taux de fécondité. Plus simplement, ce choix répond à la préoccupation de comprendre les mécanismes qui maintiennent les naissances en deçà du seul de renouvellement en Belgique. Il s'inscrit en droite ligne du choix fait par Rinesi et *al.* (2011). En effet, en étudiant la réalisation des intentions de fécondité, le choix de travailler sur la deuxième naissance est d'avantage lié à une tendance croissante des femmes à avoir un enfant et ne pas en vouloir un second et ceci pour le cas de l'Italie, comme illustré par les auteurs.

III.3.2. Recodage et synthèse de nouvelles variables

Dans le fichier synthétique obtenu après la fusion des données, la fécondité est représentée par deux variables qui sont les naissances de 1998 à 2011 et les naissances de 2012 à 2016. À partir de ces deux variables, une nouvelle a été créée. Il s'agit d'une variable dichotomique codée 1 et 0 qui représente le fait d'avoir eu ou non un autre enfant pour les femmes qui en avait déjà un : il s'agit de la variable dépendante. Dans cette variable dépendante 0 correspond aux femmes qui ont un enfant entre 1998 et 2011 et qui n'en ont pas eu entre 2012 et 2016. Quant au code 1, il correspond aux femmes qui ont eu un enfant entre 1998 et 2011 et qui en ont eu un autre ou plus dans les 5 années qui ont suivi. Le degré de religiosité a aussi été recodé de manière à avoir non croyant pour la modalité 0 ; peu croyant qui correspond aux modalités allant de 1 à 5 ; croyant qui regroupent les modalités 6 et 7 et enfin, très croyant pour les modalités de 8 à 10. L'âge de fécondité de la femme telle que considérée dans cette étude est compris entre 18-49 ans. Cette limite de 49 est généralement celle observée dans la littérature (Dantis & Rizzi, 2020). Aussi, le choix est lié au fait que passé cet âge, les chances d'avoir un (autre) enfant deviennent quasi nulles.

III.3.3. Régressions logistiques

Un modèle de régression logistique simple a d'abord été créé entre la variable dépendante et le degré de religiosité compte tenu du fait que la variable dépendante est binaire. Un second modèle, cette fois-ci multiple, intégrant en plus de la religion, l'âge, le pays de naissance, le plus haut diplôme, le statut matrimonial, l'occupation et la province de résidence a été ajusté. Ces variables ont servi de variables de contrôle pour l'ajustement du modèle. Ceci a permis de mieux comprendre l'effet isolé de la religion sur l'événement « avoir un autre enfant chez les femmes qui en avait déjà un », en tenant compte des influences potentielles des autres facteurs démographiques et socio-économiques. Il s'agit de pratiques bien répandues bien qu'elles concernent très souvent les facteurs attitudinaux comme les intentions de fécondités (Dantis et *al.*, 2023 ; Kim, 2017).

L'interprétation des résultats du modèle a été faite sur la base des *odd ratio* (rapport de cote) et non des coefficients, parce que plus utilisés en sciences sociales (Rizzi, 2013). Selon l'autrice, ils ont l'avantage d'une relative facilité d'interprétation et : « *leur effet sur la variable dépendante s'exprime [...] en termes d'augmentation [si l'OR est supérieur à 1] ou de diminution [si l'OR est inférieur à 1] de l'odds de réalisation de la variable dépendante* » (Rizzi, 2013, p. 16) qui est « avoir un second enfant ».

III.1. Résultats

III.1.1. Fusion en considérant ESS comme receveur et DA comme donneur

Cette partie présente les différents résultats à chaque étape de la fusion des données.

III.1.1.1. Caractéristiques des variables communes et spécifiques

Nous avons commencé cette partie en identifiant toutes les variables qui sont communes aux bases de DA et ESS. Elles ont été identifiées en comparant une à une les variables des DA à ceux présents dans la ESS. Ce choix est lié au fait que la liste des variables est plus réduite dans la première. Cette étape a permis d'identifier 11 variables communes. Notons que des variables pouvaient apparaître en double dans la ESS, nous avons donc fait le choix de celles ayant la meilleure qualité (absence ou faible proportion de données manquantes) et de définition la plus proche possible. Se basant sur le même critère de la qualité, 4 autres variables ont été exclues. Il s'agit des variables concernant les pays de naissances et les nationalités des parents. Les 7 variables communes retenues dans l'étude sont présentées dans le tableau ci-contre.

Tableau IV : Liste des variables communes retenues

Variables		Caractéristiques
1. Âge		Variable continue
2. Pays de naissance		Variable catégorielle
3. Nationalité		Variable catégorielle
4. Occupation		Variable catégorielle
5. Plus haut niveau d'éducation		Variable catégorielle
6. Statut marital		Variable catégorielle
7. Province de résidence		Variable catégorielle
Variables spécifiques		
	ESS	DA
8. Religion	Variable catégorielle	-
9. Naissance	-	Variable continue
10. Pondérations	Présentes	-

*le signe « — » fait référence à l'absence de la variable dans la base de données correspondante.

Le tableau suivant présente les modalités des variables après uniformisation. À cette étape, les univers des deux bases de données étaient déjà uniformisés. La faible proportion des effectifs des certaines modalités a justifié leurs regroupements en des modalités de niveau supérieur. Les données des pays de naissance ont par exemple été regroupés en modalités « Belgique » et « hors Belgique » compte tenu du fait qu'une redéfinition par pays aurait conduit à créer des modalités à effectif nul. Un des principes de la fusion statistique des données est d'éviter d'avoir des modalités vides. Ainsi, nous avons fait le choix de recoder les variables de manière à ce qu'elles aient des modalités aux effectifs suffisants. Ceci s'est fait en procédant par des recodages successifs jusqu'à l'obtention d'un effectif jugé satisfaisant (effectifs minimums par modalité égale à 5)⁷.

Tableau V : caractéristiques des variables recodées

Variables	Modalités de la variable
1. Âge⁸	1. 18 à 25 ans, 2. 26 à 35 ans, 3. 36 à 45 ans, 4. 46 à 55 ans, 5. 56 à 65 ans, 6. 66 ans et plus.
2. Pays de naissance	1. Belgique, 2. hors Belgique.
3. Nationalité	1. Belge, 2. Non belge.
4. Occupation	1. travaille, 2. ne travaille pas.
5. Niveau d'éducation	1. aucun, 2. primaire, 3. secondaire, 4. supérieur.
6. Statut marital	1. femmes en union légale, 2. femmes qui ne sont pas engagées dans une union, 3. femmes qui n'ont jamais été engagées dans une union légalement reconnue.
7. Province de résidence	1. Flandre occidentale, 2. Flandre orientale, 3. Anvers, 4. Limbourg, 5. Brabant Flamand, 6. Brabant Wallon, 7. Hainaut, 8. Namur, 9. Liège, 10. Luxembourg, 11. Bruxelles-Capitale.

III.1.1.2. Variables communes versus variables de fusion

Similarité des variables communes

Le tableau V montre les résultats des analyses de similarité entre les paires de variables communes. Ces résultats présentent les coefficients de similitudes de Bhattacharya et les distances totales de variation (TVD) pour les variables. Les coefficients de

⁷ La limite est liée à la littérature sur la question. Voir D'Orazio et *al.* (2006a).

⁸ L'âge a été recodé en une variable catégorielle pour les besoins de l'analyse. Voir Rezaei Ghahroodi, 2023.

Bhattacharya sont élevés pour toutes les variables ($> 0,95$), ce qui indique une forte similarité entre les distributions des variables pour chaque paire comparée. Pour ce qui est de la TVD, elle est inférieure ou égale à 0,05 pour 4 paires de variables que sont : le pays de naissance (0,03); la nationalité (0,04); la province de résidence (0,04) et l'âge (0,05). Elle est cependant plus élevée pour la variable « Occupation » (0,27). Ceci suggère une différence plus élevée pour cette variable dans les deux bases de données.

Tableau VI : résumé de l'analyse des similarités entre les paires de variables

Variables	Coef. de Bhatt.	TVD
Groupe d'âge	0,999	0,05
Pays de naissance	0,999	0,03
Nationalité	0,998	0,04
Occupation	0,950	0,27
Plus haut niveau d'éducation	0,992	0,12
Statut marital	0,990	0,12
Province de résidence	0,998	0,04

Sur la base du critère de la forte similarité et de la faible dissimilarité, l'occupation a été écartée de la liste des variables de fusion puisque c'est celle qui s'en écartait le plus. Bien que les TVD du plus haut niveau d'éducation et du statut marital soient aussi élevés, celles-ci ont été aussi incluses dans la liste compte tenu de la valeur du coefficient de Bhattacharya.

Capacités prédictives et choix des variables de fusion

La capacité prédictive des 6 variables communes retenues testée au moyen des régressions linéaires multiples. Dans le cas de la corrélation entre le degré de religiosité et les variables communes dans la ESS, les résultats montrent que certaines de ces variables sont de meilleurs prédicteurs que d'autres. Le tableau VI présente les meilleurs prédicteurs du degré de religiosité. Il s'agit du sous-ensemble de variables explicatives qui fournissent le meilleur ajustement du modèle de régression initiale. Le modèle de régression linéaire multiple permet d'analyser les corrélations entre les variables explicatives et le degré de religiosité. La fonction *step* permet cependant de choisir les meilleurs prédicteurs en répétant par itérations successives par ajout ou un retrait de variables à partir du modèle initial, la combinaison des variables qui prédisent au mieux la variable dépendante. Elle a ainsi permis de ne retenir que les variables dont la capacité de prédiction était significative au seuil de 0,05 %. Il s'agit dans

l'ordre de l'âge ($p < 0,001$), du pays de naissance ($p < 0,001$), du statut marital ($p < 0,01$) et enfin du plus haut niveau d'éducation ($p < 0,05$).

Tableau VII : anova de la régression pour le choix des meilleurs prédicteurs du degré de religiosité.

Variable dépendante : <i>degré de religiosité</i>					
	<i>Df</i>	<i>Sum sq</i>	<i>Mean sq</i>	<i>fvalue</i>	<i>pvalue</i>
Âge	5	1393	278,61	13,965 5	2,070e-13 ***
Niveau d'éducation	3	171	57,07	2,860 9	0,035 683 *
Statut marital	2	235	117,53	5,891 2	0,002 819 **
Pays de naissance	1	596	596,11	29,880 7	5,254e-08 ***
Résidus	1754	34 992	19,95	-	-

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; non significatif $p > 0,10$

Les résultats différents ont été obtenus en analysant la corrélation entre le nombre de naissances et les variables commune dans le donneur (DA). En effet, en plus des 4 variables mentionnées plus haut, la province s'est révélée être un bon prédicteur des naissances ($p < 0,05$ %). L'âge, le niveau d'éducation, le statut marital, la province de résidence et le pays de naissance influencent significativement sur les naissances ($p < 2,2 \cdot 10^{-16}$). L'analyse de la somme des carrées (Sum Sq) montre qu'elles sont mieux corrélées à l'âge et statut marital ; viennent ensuite le niveau d'éducation, la province de résidence et le pays de naissance. La suite du processus de fusion a donc été réalisée en se basant sur l'âge, le pays de naissance, le statut marital et le niveau d'éducation comme variables de fusion.

Tableau VIII : anova de la régression pour le choix des meilleurs prédicteurs des naissances

Variable dépendante : <i>nombre de naissances</i>					
	<i>Df</i>	<i>Sum sq</i>	<i>Mean sq</i>	<i>fvalue</i>	<i>pvalue</i>
Âge	5	77 506 464	15 501 293	373 375,9	$< 2,2 \cdot 10^{-16}$ ***
Niveau d'éducation	3	2 174 643	724 881	17 460,0	$< 2,2 \cdot 10^{-16}$ ***
Statut marital	2	41 436 155	20 718 078	499 031,4	$< 2,2 \cdot 10^{-16}$ ***
Province de résidence	10	379 046	37 905	913,0	$< 2,2 \cdot 10^{-16}$ ***
Pays de naissance	1	375 936	375 936	9055,1	$< 2,2 \cdot 10^{-16}$ ***
Résidus	5 598 898	232 447 111	42	-	-

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; non significatif $p > 0,10$

III.1.1.3. Résultats de la fusion et validité

Distribution marginale

Le processus de fusion a été réalisé avec ESS comme receveur et ensuite comme donneur. Le résultat de l'analyse des distributions marginales est présenté dans le tableau VIII ci-contre.

Tableau IX : différences de distributions des variables spécifiques en fonction des donneurs et receveurs

Processus 1			Processus 2		
Donneur : ESS — receveur : DA			Donneur : DA — receveur : ESS		
Degré de religiosité	dist_don ⁹	dist_final	Nombre de naissances	dist_don	dist_final
0	0	0,03	0	0	-6,21
1	0	0,80	1	0	3,01
2	0	-1,91	2	0	2,09
3	0	-0,05	3	0	0,83
4	0	1,19	4	0	0,19
5	0	1,20	5	0	0,06
6	0	-1,89	6	0	0,04
7	0	-1,63	7	0	76,67
8	0	1,59	8	0	11,83
9	0	-0,01	9	0	8,32
10	0	0,58	10	0	2,49
-	-	-	11	0	0,51

Dans le tableau, dist_don et dist_final sont des tableaux de contingence normalisés (tableaux de fréquences relatives) pour la variable naissance dans le processus 1 et le degré de religiosité dans le processus 2. Le nombre de naissances va 0 à 11 enfants par femme dans la base de DA constituée. Dist_final montre des écarts mineurs entre les distributions de la variable naissance dans DA (donneur) et le fichier synthétique de données obtenu. Il représente l'écart des modalités de la variable observée dans le fichier synthétique en fonction de sa version standardisée dans le donneur. Le degré de religiosité va de 0 pour non croyant à 10 pour très croyant. L'analyse de dist_don et dist_final montre un très grand écart principalement concentré aux modalités 7 (76,67), 8 (11,83) et 9 (8,32). Les valeurs positives

⁹ Il s'agit d'une donnée standardisée de la distribution marginale de la variable d'intérêt dans la base donneuse. dist-final est l'écart par rapport à cette distribution initiale standardisée.

indiquent une fréquence plus élevée dans le fichier synthétique des données par rapport à celui dans le donneur ; tandis que les valeurs négatives indiquent une fréquence plus faible.

Distance totale de variation et qualité de la fusion

Le calcul de la distance totale de variation entre les distributions (nombre de naissance dans le processus 1 et degrés de religiosité dans le processus 2) montre une dissimilarité globale très forte dans le processus 2 (ESS comme receveur et DA comme donneur). Cette dissimilarité est en effet de 112,26 % par rapport à une distribution de référence dans le donneur (DA) dans le processus 2, soulignant une différence majeure entre les distributions observées et attendues. Ce grand écart signale des changements significatifs produits lors du processus d'appariement. En ce qui concerne le processus 1, cette dissimilarité est de 11 %. Cette valeur montre une dissimilarité modérée prouvant ainsi que les deux distributions sont proches, bien qu'il existe des différences notables. La comparaison des distributions marginales et des dissimilarités montre que le processus qui considère les données administratives comme receveur et ESS comme donneur produit de meilleurs résultats de celle qui privilégie l'inverse. Le fichier synthétique de données obtenu à partir du processus 1 est celui qui a donc été utilisé dans l'étape suivante pour l'étude de cas à savoir, l'impact du degré de religiosité sur la probabilité d'avoir ou pas un autre enfant pour les femmes qui en avait déjà un.

III.1.2. Relation entre fécondité et religion

III.1.2.1. Analyses descriptives

La base de données est un fichier synthétique de données composé de femmes résidant en Belgique, ayant un enfant et en âge de procréer (18-49 ans) en 2011. Elle compte 9 variables : le degré de religiosité, a eu un deuxième enfant entre 2012 et 2016, l'âge, le statut marital, le plus haut niveau d'éducation, le pays de naissance, la province de résidence, l'occupation et la nationalité. Le tableau II présente une analyse descriptive de ces différentes variables. Il montre que parmi les femmes qui avaient déjà un enfant, 26,3 % en ont eu un autre contre 73,7 % qui n'en ont pas eu et ceux sur la période de 2012 à 2016. En ce qui concerne le degré de religiosité, 19,2 % des femmes sont non croyantes ; 37,7 % sont peu croyantes contre 21,7 % qui sont croyants. Enfin, 21,4 % de la population est très croyante.

La répartition des individus par classe d'âge montre que 11,3 % des femmes ont entre 18 et 25 ans, 38,7 % entre 26 et 35 ans, 42,8 % entre 36 et 45 ans, et 7,16 % ont 46 ans et plus. On note un effectif élevé de femmes ayant déjà un enfant chez les 36-45 ans. En ce qui

concerne le plus haut niveau d'éducation, une majorité significative a atteint le niveau secondaire (55,2 %), suivie par celles ayant atteint le niveau supérieur (37,5 %). Une petite proportion n'a aucun diplôme (1,06 %) ou seulement un niveau primaire (6,20 %).

Tableau X : statistiques descriptives des variables utilisées

Variables	Pourcentages	Effectifs
A eu un autre enfant entre 2012-2016		
0 : n'a pas eu un autre enfant	73,7	356 503
1 : a eu un autre enfant	26,3	127 042
Degré de religiosité		
1.Non croyant	19,2	92 752
2.Peu croyant	37,7	182 246
3. Croyant	21,7	105 036
4.Très croyant	21,4	103 315
Classes d'âge		
18—25 ans	11,3	54 866
26—35 ans	38,7	187 309
36—45 ans	42,8	206 741
46—55 ans	7,16	34 629
Plus haut niveau d'éducation		
Aucun diplôme	1,06	5132
Primaire	6,20	29 968
Secondaire	55,2	266 879
Supérieur	37,5	181 566
Occupation		
Ne travaille pas	22,4	108 467
Travaille	77,6	375 078
Statut marital		
En union légale	61,5	297 322
Pas en union	11,9	57 612
Jamais été en union légale	26,6	128 611
Pays de naissance		
Belgique	78,7	380 529
Hors Belgique	21,3	103 016
Nationalité		
Belge	85,7	414 415
Non belge	14,3	69 130

Province de résidence

Anvers	14,7	71 265
Brabant flamand	9,31	45 010
Brabant wallon	3,55	17 147
Bruxelles-Capitale	11,2	54 089
Flandre occidentale	9,46	45 759
Flandre orientale	13,0	62 684
Hainaut	13,8	66 719
Liège	10,5	50 704
Limbourg	7,35	35 552
Luxembourg	2,54	12 268
Namur	4,62	22 348

Pour ce qui est de l'occupation, 22,4 % des femmes n'ont pas d'emploi contre 77,6 % qui sont en emploi. Pour le statut marital, 61,5 % des femmes sont en union légale contre 11,9 % qui ne le sont pas. Il s'agit pour ce dernier cas de veuves et des divorcées. Le reste (26,6 %) des femmes n'ont jamais été engagées dans une union légale montrant ainsi une proportion importante de femmes qui enfantent sans être impliquées dans les liens du mariage ou de tout autre engagement légal.

La majorité des femmes est née en Belgique, soit 78,7 % contre 21,3 % qui est née hors Belgique. Une proportion plus importante a une nationalité belge, soit 85,7 % contre 14,3 % qui sont non belges.

Enfin, en ce qui concerne la répartition géographique, les provinces d'Anvers, du Hainaut et de la Flandre orientale occupent la tête du classement avec respectivement 14,7 %, 13,8 %, et 13 % de la population. D'une manière générale, les résultats montrent une répartition des proportions des femmes de l'ordre de 53,82 % pour la Région flamande, 35 % pour la Région wallonne et 11,2 % pour la région Bruxelles-Capitale. Ces proportions s'adaptent bien aux proportions de la population totale de la Belgique de 2011 qui était de 57,6 % pour la Région flamande, 32,2 % pour la Région wallonne et 10,2 % pour région Bruxelles-Capitale¹⁰.

¹⁰ Cf. [Population | Statbel](#) pour plus de détails sur l'état de la population en 2011.

III.1.2.2. Résultats de la régression logistique

Régression logistique simple

Le modèle de régression simple permet de montrer que la religion est en effet un facteur déterminant de la fécondité ou plus spécifiquement de l'événement avoir un second enfant pour les femmes qui en ont déjà un ($p < 0,05$). Ainsi, l'*odds* ou la cote des femmes très croyantes est de 1,15 fois celui des femmes non croyantes ($OR = 1,15$; 95 % IC = 1,13 - 1,17). Les femmes qui sont très croyantes ont 1,15 fois plus de chance d'avoir un autre enfant que celles qui ne le sont pas. De même, les femmes croyantes ont 1,13 fois plus de chance d'avoir un autre enfant que celles non croyantes ($OR = 1,11$; 95 % IC = 1,09 - 1,13). Chez les femmes peu croyantes, cette chance est de 70 % par rapport aux femmes non croyantes, donc moins élevée.

Tableau XI : modèle explicatif de l'effet du degré de religiosité en fonction la variable dépendante

Variables	Coef	OR	[95 % IC]	p_values
(Intercept)	-0,97	0,38	0,37 - 0,39	0***
Peu croyant	-0,33	0,72	0,70 - 0,73	7,23 e-280 ***
Croyant	0,10	1,11	1,09 - 1,13	5,73 e25 ***
Très croyant	0,14	1,15	1,13 - 1,17	1,78 e-45 ***
Réf. = Non croyant		1,00		

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; non significatif $p > 0,10$

Régression logistique multiple

L'ajout des variables de contrôle dans le modèle précédemment analysé montre un rapprochement des OR des différentes modalités du degré de religiosité vers la modalité de référence. Bien qu'elles soient toujours significatives ($P < 0,05$), les chances d'avoir un autre enfant sont désormais de 1,1 fois plus élevées pour les femmes très croyantes ($OR = 1,11$; 95 % IC = 1,07 - 1,13), 1,05 fois pour les femmes croyantes ($OR = 1,05$; 95 % IC = 1,02 - 1,08) et 84 % pour celles qui sont peu croyantes ($OR = 0,84$; 95 % IC = 0,83 - 0,86) par rapport à celles qui ne le sont pas.

Les OR des femmes entre 36-45 ans et 45-49 ans se rapprochent de 0. Ceci est souvent inhabituel, mais traduit dans le cas de cette étude que les femmes qui situent dans cette tranche d'âge et qui ont déjà un enfant en ont rarement un autre. Très peu d'entre elles ont donc eu un autre enfant montrant ainsi une tendance à l'enfant unique. Cependant, les femmes âgées de 26 à 35 ans ont 30 % de chances de connaître l'événement contrairement à

celles âgées de 18 à 25 ans. Ceci suppose que ces dernières ont des chances multipliées par 2,78 d'avoir un autre enfant par rapport aux femmes qui ont entre 26 et 35 ans. Nous avons fait le choix d'exclure la variable nationalité de l'analyse compte tenu de la présence de la variable correspondant au pays de naissance qui rend mieux compte l'importance du facteur culturel associé au lieu d'origine.

Tableau XII : facteurs explicatifs de l'événement avoir un autre enfant

Variables	Coef	OR	[95 % IC]	p_values
(Intercept)	0,60	1,82	1,64 - 2,02	1,41 e-29 ***
1. Degré de religiosité				
Peu croyant	-0,17	0,84	0,83 - 0,86	1,13 e-47 ***
Croyant	0,05	1,05	1,02 - 1,08	8,86 e-05 ***
Très croyant	0,09	1,10	1,07 - 1,13	1,79 e-12 ***
Réf. = Non croyant		1,00		
2. Classe d'âge				
26—35 ans	-1,01	0,36	0,35 - 0,37	0,00 e+00 ***
36—45 ans	-4,09	0,02	0,02 - 0,02	0,00 e+00 ***
46—49 ans	-8,02	0,00	0,00 - 0,00	0,00 e+00 ***
Réf. = 18-25 ans		1,00		
3. Plus haut niveau d'éducation				
Primaire	0,17	1,18	1,07 - 1,31	1,33 e-03**
Secondaire	0,07	1,08	0,98 - 1,19	0,137
Supérieur	0,84	2,32	2,10 - 2,56	7,41 e-62 ***
Réf. = aucun niveau		1,00		
4. Statut marital				
Pas en union	-0,62	0,54	0,52 - 0,55	4,56 e-304***
Jamais été en union légale	-0,21	0,81	0,80 - 0,83	2,55 e-113***
Réf. = en union légale		1,00		
5. Occupation				
Travaille	0,01	1,01	0,99 - 1,03	0,276
Réf.= Ne travaille pas		1,00		
6. Pays de naissance				
Hors Belgique	0,05	1,05	1,03 - 1,08	2,79 e-05 ***

Réf. = Belgique		1,00		
7. Province de résidence				
Brabant flamand	-0,01	0,99	0,96 - 1,03	0,717
Brabant wallon	-0,03	0,97	0,93 - 1,02	0,304
Bruxelles-Capitale	0,22	1,25	1,21 - 1,29	3,19 e-45 ***
Flandre occidentale	-0,12	0,88	0,86 - 0,91	9,33 e-14 ***
Flandre orientale	-0,18	0,84	0,81 - 0,86	8,21 e-32 ***
Hainaut	-0,26	0,77	0,75 - 0,79	2,74 e-67 ***
Liège	-0,14	0,87	0,84 - 0,90	3,11 e-18 ***
Limbourg	-0,18	0,83	0,80 - 0,86	2,99 e-24***
Luxembourg	0,02	1,02	0,96 - 1,07	0,523
Namur	-0,07	0,93	0,89 - 0,97	9,35 e-04 ***
Réf. = Anvers		1,00		

*** p< 0.01 ; ** p<0.05 ; * p<0.10 ; non significatif p>0,10

En ce qui concerne le statut marital, l'OR de 0,54 (IC = 0,52 - 0,55) des femmes qui ne sont pas en union traduit ici le fait que les femmes mariées et ayant déjà un enfant voient leurs chances d'en avoir un autre multiplié par 1,85 par rapport aux femmes qui ne le sont pas (plus). De la même manière, celles qui n'ont jamais été en union ont des chances multipliées par 1,23. Les femmes qui n'ont jamais été en union ont donc plus de chance d'avoir un autre enfant que celles qui l'ont déjà été, mais qui ne le sont plus.

Concernant le plus haut niveau d'éducation, seuls les niveaux primaires et supérieurs produisent des résultats significatifs au seuil de 95 % (p< 0,05). Ces résultats montrent que les femmes avec un niveau primaire ont 1,18 fois plus de chance d'avoir un autre enfant par rapport aux femmes sans instruction (OR = 1,18, 95 % IC = 1,07 - 1,31). Celles du supérieur voient leurs chances multipliées par 2,32 (OR = 2,32, 95 % IC = 2,10 - 2,56) par rapport aux femmes sans instruction. Le pays de naissance produit aussi un effet. Les chances d'avoir un autre enfant sont multipliées par 1,05 pour les femmes nées en Belgique par rapport à celles nées hors Belgique (OR = 1,05, 95 % IC = 1,03 - 1,08). En considérant la province d'Anvers comme référence, on constate que les femmes qui résident dans la province de Bruxelles-Capitale ont 1,25 fois plus de chance de mettre au monde un autre enfant. Bruxelles-Capitale est la zone en Belgique où les femmes qui ont déjà un enfant sont susceptibles d'en avoir une autre et donc un nombre d'enfants plus élevés.

III.2. Forces et limites de la méthode

III.2.1. Quelles fiabilités pour la fusion statistique ?

L'intégration des données ouvre un champ énorme de possibilités en ce sens qu'elle permet de combiner des données de sources différentes pour la compréhension de phénomènes qui ne sont pas observés ensemble (Bounebach et *al.*, 2018). Quelle que soit l'option retenue, l'enjeu repose toutefois sur l'obtention de résultats qui se rapprocheraient au mieux d'une réalité qui serait dans l'idéal celle où ces phénomènes sont observés ensemble. Dans le cas de cette étude, le choix d'utiliser l'approche décrite par D'Orazio et *al.* (2006a) offre des résultats qui sont jugés encourageants et démontre de la possibilité de combiner des données d'enquête à celles qui sont administratives en considérant ces dernières comme receveur. Les similitudes plus élevées et les dissimilitudes plus faibles, ainsi que la distance totale de variation entre les variables quand DA est receveur conforte ce résultat.

Des précautions sont cependant à prendre. En effet, la nature différente des bases de données et leur taille très différente sont à prendre en compte pour une meilleure confirmation de ce résultat. En effet, D'Orazio et (2006a) conseille d'utiliser la plus petite base de données comme receveur. Il est cependant possible de faire autrement en fonction des objectifs de la recherche comme précisés par Rezaei Ghahroodi (2023) ou Wilkens (2018). Plusieurs options sont alors disponibles comme c'est le cas avec l'utilisation des contraintes. Dans le processus mobilisé dans cette étude, la fusion a été réalisée de sorte que chaque observation contenue dans ESS puisse être utilisée n fois ($k = n$) dans le processus de fusion. On pourrait penser à inclure une contrainte ($k = 1, 2, 5$, etc.) et comparer les résultats obtenus à ceux de cette étude.

III.2.2. Quelle transformation pour quels résultats ?

Durant la phase d'harmonisation des données, cette question revenait sans cesse : jusqu'à quel niveau la transformation des données n'altère-t-elle pas la qualité des données ? Il n'existe pas de consensus sur cette question dans la littérature. D'Orazio et *al.* (2006a) conseille de conserver les données originelles et d'en limiter au maximum la transformation. Aussi, les transformations doivent être fonction des objectifs de la recherche et de la qualité des données utilisées. Ainsi, les transformations devraient être réalisées uniquement lorsque cela est nécessaire comme c'est le cas lorsque les modalités et les effectifs des variables diffèrent. Dans cette étude, le faible effectif dans certaines modalités des variables a conduit à les regrouper en modalité d'un niveau plus élevé. C'est le cas par exemple du lieu de

résidence dans ESS détaillé jusqu'au niveau des communes, mais recodé au niveau des provinces compte tenu des effectifs très réduits dans les communes.

De plus, il est possible de procéder à la création de toutes les variables avant la fusion et non après comme réalisée dans le cadre de ce travail. Une fois de plus, le nombre d'observations limitées dans ESS fait qu'un durcissement de critères de sélection à l'entame conduirait à une réduction drastique de sa taille et la rendrait donc peu fiable et difficilement utilisable dans un processus de fusion. Un exemple est que, si dès le départ, l'étude avait fait le choix de ne retenir dans ESS que les femmes, en Belgique, âgé de 18 à 49 et ayant déjà un enfant, ce critère aurait limité la base de données à moins de 200 observations. Cette base de données a été jugée trop faible en effectif pour être utilisée (D'Orazio et *al.*, 2006a). Ainsi, les changements apportés dans le fichier synthétique de données, bien qu'elles soient les moins recommandées, ont permis de passer outre cette barrière. On pourrait voir ici l'opportunité du développement des techniques de création de données fictives comme moyen d'accroître les effectifs de la ESS et comme option pour contourner cette situation.

Conclusion générale

Pour conclure, ce travail est un essai méthodologique qui avait pour objectif de partir du développement d'un processus de fusion statistique de deux bases de données pour répondre à la question de l'influence de la religion sur l'arrivée d'un deuxième enfant pour les femmes vivant en Belgique et qui en avaient déjà un. Ceci a été fait en deux étapes : une première sur le développement d'une technique d'intégration des bases de données et une seconde sur l'étude du comportement des femmes en matière de fécondité à partir de la base de données obtenues. Les résultats montrent un potentiel significatif des opérations d'intégration des données de registre et d'enquêtes pour la compréhension de phénomènes non observés ensemble. La base de données obtenue n'est pas la réalité au sens strict du terme, mais un monde qui s'en rapproche ou « *closest world* » (D'Orazio et al., 2006b) qui est produite par la littérature.

Commençons ainsi cette section par la religion qui est le principal facteur explicatif mobilisé pour expliquer les comportements en matière de fécondité. Les résultats obtenus à partir du fichier synthétique montrent des chances accrues d'avoir un autre enfant lorsque les femmes sont croyantes et encore plus lorsqu'elles sont très croyantes. Les femmes croyantes ont ainsi tendance à avoir plus d'enfants que celles qui ne le sont pas comme observé par Bein et al. (2017) ou Dantis et al. (2023). Bien qu'ils se soient intéressés à la fréquence de participation aux services religieux et non au degré de religiosité, ils montrent tout de même un lien entre la religion et l'intention de fécondité. Les femmes croyantes ont donc une volonté accrue d'avoir plus d'enfants que celles qui ne le sont pas. Aussi, la naissance peut être un facteur de réflexion existentialiste, qui peut pousser les parents qui veulent exposer leur enfant à des valeurs religieuses positives à s'exposer eux-mêmes à ces enseignements. En plus de la religion, la communauté dans laquelle se développe la femme croyante se situe comme un facteur important à prendre en compte pour une approche holistique dans la compréhension des comportements en matière de fécondité ; le « familiarisme » (Dantis et al., 2023) offrant un soutien significatif dans la charge liée à la maternité.

Ensuite, il a été montré un lien positif entre le niveau d'éducation et la religion. Les femmes croyantes sont ainsi susceptibles de correspondre aux caractéristiques d'un niveau d'éducation élevé et d'un nombre d'enfants relativement plus important. Dantis et al. (2023) observait aussi au niveau européen, un lien positif entre le niveau d'éducation et l'intention d'avoir un second ou un troisième enfant. Dans le cas du présent travail, il a été observé que le niveau d'éducation pour les femmes ayant déjà un enfant impactait positivement les

chances d'en avoir un autre. Sans exclure le fait que d'autres facteurs puissent être à l'œuvre dans cette relation (les revenus, le cadre familial, le partage des tâches ménagères, etc.) on peut tout de même faire l'hypothèse que la vie religieuse et la vie en communauté qui en résulte s'inscrivent comme facteurs importants à prendre en compte pour comprendre cette situation. Ce résultat corrobore ceux de Peri-rotem (2020) qui établissait un lien positif entre le niveau d'éducation, la religion, et la fécondité.

L'âge pris ici comme facteur et inclus dans le modèle de régression permet d'observer que les chances d'avoir un autre enfant diminuent avec l'âge. L'étude montre que les femmes qui ont des enfants plus jeunes sont plus susceptibles de connaître une autre maternité contrairement aux autres. En effet, avec l'âge, les femmes voient leur fécondabilité décroître jusqu'à être nulle (Rizzi & Rosina, 2006). Avec le recul de l'âge au premier enfant en Belgique, le nombre de femmes avec plus d'un enfant connaîtra ainsi une décroissance. L'arrivée tardive de la première naissance peut ainsi être associée à la baisse du nombre de naissances et donc du nombre d'enfants par femmes. Pour compléter, Rizzi et Rosina (2006) concluent dans leur étude qu'un report progressif de la formation d'une famille dans les pays occidentaux peut avoir un effet négatif sur la capacité à concevoir, lié au fait que la composante biologique diminue avec l'âge, mais aussi au fait que la composante comportementale (activité sexuelle) diminue aussi de manière significative. L'activité sexuelle s'inscrit donc ici comme un élément qui doit être pris en compte pour analyser ce comportement.

Un autre point est la proportion élevée de femmes qui ont un enfant hors de tout lien d'engagement légalement reconnu¹¹. L'étude de la diffusion des cohabitations non maritales menée par Doignon et *al.* (2020) montre que ce comportement (cohabitation hors mariage) faible dans les années 1991 (moins de 10 % des cohabitations) atteint désormais les 50 % en 2015. De plus en plus de Belges font le choix d'une cohabitation sans engagement légal. Ce comportement est plus observé dans la Région wallonne que dans celle flamande (quelle que soit la période considérée par l'étude de Doignon et *al.* (2020)). Cette situation confirme ainsi un changement notable dans les comportements familiaux et les attitudes envers les structures traditionnelles du mariage qu'il serait intéressant d'étudier. Les liens avec la religion et d'autres paramètres sociaux devront davantage être investigués. La fusion statistique des données trouve ainsi toute son utilité.

¹¹ Notons qu'il s'agit des femmes qui n'ont jamais été mariées ou engagées dans une union légalement reconnue ; ces deux catégories étant celles reconnues dans le cadre de ce travail comme en union légale.

La mise en discussion de ces quelques points en rapport avec la littérature permet de montrer que la fusion des données administratives et de la ESS offre d'excellents résultats pour comprendre des comportements en matière de fécondité en Belgique ; ce qui confirme notre hypothèse spécifique 1. Aussi, il a été observé un lien positif entre la fécondité et la religiosité à partir du fichier synthétique de données obtenues confirmant ainsi l'hypothèse spécifique 2. La fusion des données administratives et d'enquêtes offre ainsi un potentiel énorme pour la compréhension des comportements sociaux. Son intérêt réside aussi dans sa capacité à combiner la précision territoriale des données administratives à celles des données attitudinales de l'ESS confirmant ainsi l'hypothèse générale énoncée à l'entame du travail.

En somme, cette étude illustre l'importance de l'intégration des données de sources diverses pour la compréhension de phénomènes démographiques de moins en moins observés ensemble lors de la collecte des données. Les résultats obtenus renforcent la validité de l'approche et ouvre la voie à des applications tant dans le domaine de la recherche que de l'aide à la prise de décision et de l'action publique. L'adapter au contexte actuel de crises multiples (pandémies, crises environnementales, écoanxiété, etc.) permettra sans nul doute d'apporter des explications et des pistes de solutions.

Références bibliographiques

- Adsera, A. (2011). Where Are the Babies? Labor Market Conditions and Fertility in Europe. *European Journal of Population / Revue Européenne de Démographie*, 27(1), 1-32.
- Bailly, N., Roussiau, N., & Fleury-Bahi, G. (2011). Étude des liens entre les croyances religieuses et spirituelles, la santé et l'âge : *Bulletin de psychologie*, 512(2), 149-154.
- Barry, J. T. (1988). An investigation of statistical matching. *Journal of Applied Statistics*, 15(3), 275-283.
- Bartiaux, F. (2022). Gender roles and domestic power in energy-saving home improvements. *Buildings and Cities*, 3(1), Article 1. <https://doi.org/10.5334/bc.232>
- Bauknecht, J., & Merkel, S. (2022). Differences in self-reported health between low- and high-income groups in pre-retirement age and retirement age. A cohort study based on the European Social Survey. *Health Policy OPEN*, 3, 100070. 6 p.
- Bein, C., Gauthier, A., & Mynarska, M. (2017). Une étude comparative de l'impact de la religiosité et de l'égalité des genres sur les intentions en matière de fécondité et leurs réalisations. *Cahiers de recherche sociologique*, 63, 185-220.
- Bounebach, S., Quantin, C., Benzenine, É., Obozinski, G., & Rey, G. (2018). Revue Bibliographique des Méthodes de Couplage des Bases de Données : Applications et Perspectives dans le cas des Données de Santé Publique. *Journal de la Société Française de Statistique*, 159(3), 79-123.
- Bourguignon, M., Brée, S., Eggerickx, T., & Sanderson, J.-P. (2019). Fécondité différentielle au début du XXe siècle selon la classe sociale des femmes. Les cas de la Belgique et de la France. *Annales de démographie historique*, 138(2), 179-211.
- Brée, S., Eggerickx, T., Sanderson, J.-P., & Costa, R. (2016). Comparaison des données rétrospectives de fécondité dans les recensements en Belgique et les enquêtes Famille en France. *Population*, 71(1), 85-120.
- Buelens, M. (2021). Contrastes intra-urbains de la fécondité bruxelloise. *The Journal of Research on Brussels*. <https://doi.org/10.4000/brussels.5277>
- Claramunt González, J., van Delden, A., & de Waal, T. (2023). Assessment of the effect of constraints in a new multivariate mixed method for statistical matching. *Computational Statistics & Data Analysis*, 177 (107569).
- Conti, P. L., Marella, D., & Scanu, M. (2016). Statistical Matching Analysis for Complex Survey Data With Applications. *Journal of the American Statistical Association*, 111(516), 1715-1725.
- Costa, R., Eggerickx, T., & Sanderson, J.-P. (2011). Les territoires de la fécondité en Belgique au 20e siècle. *Espace populations, sociétés*, 2, 353-375.
- D'Alberto, R., Zavalloni, M., Raggi, M., & Viaggi, D. (2021). A Statistical Matching approach to reproduce the heterogeneity of willingness to pay in benefit transfer. *Socio-Economic Planning Sciences*, 74 (100935). 19 p.
- Dantis, C., & Rizzi, E. L. (2020). Transition to first birth during the Great Recession : The case of Greece. *Genus*, 76(1), 21 p.

- Dantis, C., Rizzi, E. L., & Baudin, T. (2023). The Association between Religiosity and Fertility Intentions Via Grandparenting : Evidence from GGS Data. *European Journal of Population*, 39(1), 24 p.
- Doignon, Y., Eggerickx, T., & Rizzi, E. L. (2020). The spatial diffusion of nonmarital cohabitation in Belgium over 25 years : Geographic proximity and urban hierarchy. *Demographic Research*, 43, 1413-1428.
- Doliger, C. (2008). La fécondité et ses déterminants économiques : Becker vs Easterlin. *Revue économique*, 59(5), 955-971.
- Donatiello, G., D'Orazio, M., Frattarola, D., & Spaziani, M. (2022). The joint distribution of income and consumption in Italy : An in-depth analysis on statistical matching. *Review of official statistics*, 3, 75-109.
- D'Orazio, M. (2011). *Statistical matching : Accuracy of results choice of the matching variables*. ESSnet on Data Integration : Course on Statistical Matching, Rome.
- D'Orazio, M. (2013). *Statistical matching : Metodological issues and practice with R-StatMatch*. EUSTAT 55th international statistical seminar.
- D'Orazio, M. (2017). Statistical matching and imputation of survey data with StatMatch. *Italian National Institute of Statistics: Rome, Italy*.
- D'Orazio, M. (2022). *Statistical Matching or Data Fusion. Package « StatMatch »*. Version 1.4.1. <https://github.com/marcellodo/StatMatch>
- D'Orazio, M., Di Zio, M., & Scanu, M. (2006a). *Statistical matching : Theory and practice*. Wiley.
- D'Orazio, M., Di Zio, M., & Scanu, M. (2006b). Statistical Matching for Categorical Data : Displaying Uncertainty and Using Logical Constraints. *Journal of Official Statistics*, 22(1), 137-157.
- D'Orazio, M., Di Zio, M., & Scanu, M. (2017). The use of uncertainty to choose matching variables in statistical matching. *International Journal of Approximate Reasoning*, 90, 433-440.
- Erlandsson, A. (2017). Child Home Care Allowance and the Transition to Second- and Third-Order Births in Finland. *Population Research and Policy Review*, 36(4), 607-630.
- Ghia, F. (2018). Religion civile, culte public et religiosité privée. Réflexions à partir de Max Weber. *Société, droit et religion*, Numéro 8(1), 39-53.
- Goldscheider, F., Bernhardt, E., & Lappegård, T. (2015). The Gender Revolution : A Framework for Understanding Changing Family and Demographic Behavior. *Population and Development Review*, 41(2), 207-239.
- Grimmeau, J.-P., Decroly, J.-M., & Wertz, I. (2012). La démographie des communes belges de 1980 à 2010. *Courrier hebdomadaire du CRISP*, 2162-2163(37-38), 1-90.
- Gulseren, D. (2022). Physical hazards, musculoskeletal pain, and job control : Multilevel results from the European Social Survey. *Safety Science*, 156, 105901.
- Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimae, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2), 1-12.

- Honoré, L. (2021). Les enjeux de la religion et de la religiosité au travail. *@GRH*, 41(4), 167-198.
- IWEPS. (2021). *Fécondité et natalité, indice conjoncturel*. IWEPS. www.iweps.be.
- IWEPS. (2023). *Fécondité et natalité, indice conjoncturel*. IWEPS. www.iweps.be.
- Kim, E. H. — W. (2017). Division of domestic labour and lowest-low fertility in South Korea. *Demographic Research*, 37, 743-768.
- Krech, V. (2014). La religiosité comme seuil. *Archives de sciences sociales des religions*, 167, Article 167.
- Krotki, K. J., & Lapierre-Adamcyk, E. (1968). La fécondité au Canada selon la religion, l'origine ethnique et l'état matrimonial. *Population*, 23(5), 815-834.
- Lebugle Mojdehi, A. (2015). Baisse de la fécondité dans un contexte d'amélioration du statut des femmes ? Le cas du milieu rural iranien. *Autrepart*, 74-75(2-3), 67-84.
- Lecocq, A., Ammi, M., & Bellarbre, É. (2016). Le score de propension : Un guide méthodologique pour les recherches expérimentales et quasi expérimentales en éducation. *Mesure et évaluation en éducation*, 37(2), 69-100.
- Leridon, H. (2015). Théories de la fécondité : Des démographes sous influence ? *Population*, 70(2), 331-373.
- Lesthaeghe, R. (2011). The “Second Demographic Transition” : A Conceptual Map for the Understanding of Late Modern Demographic Developments in Fertility and Family Formation. *Historical Social Research*, 36(2), 179-218.
- Lewaa, I., Hafez, M. S., & Ismail, M. A. (2023). Mixed Statistical Matching Approaches Using a Latent Class Model : Simulation Studies. *Journal of Statistics Applications & Probability*, 12(1), 247-265.
- Lloyd, M., Dobranowski, K., Cheng, S. Y., Temple, V. A., Foley, J. T., Lunskey, Y., Lin, E., & Balogh, R. (2022). Matching Special Olympics registration data with administrative health databases : Feasibility and health status differences in children and youth with IDD. *Disability and Health Journal*, 15(101319), doi: 10.1016/j.dhjo.2022.101319.
- Magnoni, F., Corso, G., Maisonneuve, P., Massari, G., Alberti, L., Castelnovo, G., Leonardi, M. C., Sacchini, V., Galimberti, V., & Veronesi, P. (2023). A propensity score–matched analysis of breast-conserving surgery plus whole-breast irradiation versus mastectomy in breast cancer. *Journal of Cancer Research and Clinical Oncology*, 149(3), 1085-1093.
- Makki, R. (2021). La crainte de l'effondrement et le sentiment de la religiosité. *Figures de la psychanalyse*, 41(1), 251-258.
- Matias, M., & Fontaine, A. M. (2017). Intentions to Have a Child : A Couple-Based Process. *Family Relations*, 66(2), 231-243.
- McDonald, J. W., Rosina, A., Rizzi, E., & Colombo, B. (2011). Age and fertility : Can women wait until their early thirties to try for a first birth? *Journal of Biosocial Science*, 43(6), 685-700.
- Moretti, A., & Shlomo, N. (2023). Improving Statistical Matching when Auxiliary Information is Available. *Journal of Survey Statistics and Methodology*, 11, 619–642

- Okner, B. (1972). Constructing a New Data Base from Existing Microdata Sets : The 1966 Merge File. *Annals of Economic and Social Measurement*, 1(3), 325-362.
- Perchinunno, P., Mongelli, L., & d'Ovidio, F. D. (2020). Statistical matching techniques in order to plan interventions on socioeconomic weakness : An Italian case. *Socio-Economic Planning Sciences*, 71 (100 836), 9 p.
- Peri, G., Rees, D. I., & Smith, B. (2023). Terrorism and political attitudes : Evidence from European social surveys. *Regional Science and Urban Economics*, 99 (103864), 24 p.
- Peri-Rotem, N. (2020). Écarts de fécondité en fonction du niveau d'instruction : Le rôle de la religion en Grande-Bretagne et en France. *Population*, 75(1), 9-38.
- Perlberg, J., Allonier, C., Boisaubault, P., Daniel, F., Le Fur, P., Szidon, P., & Bourgueil, Y. (2014). Faisabilité et intérêt de l'appariement de données individuelles en médecine générale et de données de remboursement appliqué au diabète et à l'hypertension artérielle : *Santé Publique*, 26(3), 355-363.
- PERSEE. (1957). La fécondité des mariages à Amsterdam selon l'appartenance sociale et religieuse. *Population*, 12(2), 289-318.
- Rezaei Ghahroodi, Z. (2023). Statistical matching of sample survey data : Application to integrate Iranian time use and labour force surveys. *Statistical Methods & Applications*. <https://doi.org/10.1007/s10260-023-00693-2>
- Rinesi, F., Pinnelli, A., Prati, S., Castagnaro, C., & Iaccarino, C. (2011). The Transition to Second Child in Italy : Expectations and Realization. *Population (English Edition, 2002-)*, 66(2), 391-405.
- Rizzi, E. L. (2013). La régression logistique. In G. Masuy-Stroobant & R. Costa, *Analyser les données en sciences sociales. De la préparation des données à l'analyse multivariée*. (P.I.E. Peter Lang).
- Rizzi, E., & Rees, A. (2023). Belgian family policy from a comparative perspective : Does it support fertility and gender equity? *Quetelet Journal*, 9(1), 65-99.
- Rizzi, E., & Rosina, A. (2006). Does sex matter? Ageing and ability to conceive. *Vienna Yearbook of Population Research*, 4, 77-90.
- Rodgers, W. L. (1984). An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics*, 2(1), 91-102.
- Samuel, O., & Attané, I. (2005). Femmes, famille, fécondité. De la baisse de la fécondité à l'évolution du statut des femmes. *Revue Tiers Monde*, 182(2), 247-254.
- Scanu, M. (2011, septembre 28). *Course on data integration : Statistical matching. Introduction and motivation* [ESSnet DI course on data integration]. ESSnet on Data Integration : Course on Statistical Matching, Rome.
- Schleicher, J., Eklund, J., D. Barnes, M., Geldmann, J., Oldekop, J. A., & Jones, J. P. G. (2020). Statistical matching for conservation science. *Conservation Biology*, 34(3), 538-549.
- Schnittker, J. (2020). Religion, social integration, and depression in Europe : Evidence from the European Social Survey. *Social Science & Medicine*, 267 (112376), 8 p.

- Scholtus, S., van de Laar, R., & Willenborg, L. (2014). *Memobust Handbook on Methodology of Modern Business Statistics. Method: Statistical Matching Methods*. Statistics Netherlands.
- Schoumaker, B. (1999). Indicateurs de niveau de vie et mesure de la relation entre pauvreté et fécondité : L'exemple de l'Afrique du Sud. *Population*, 54(6), 963-992.
- Singh, A., Kumar, K., Kumar Pathak, P., Kumar Chauhan, R., & Banerjee, A. (2017). La structure spatiale de la fécondité indienne et ses déterminants. *Population*, 72(3), 525-549.
- Somers, A., & Van Poppel, F. (2003). Catholic priests and the fertility transition among Dutch Catholics, 1935-1970. *Annales de démographie historique*, 106(2), 57-88.
- Tran, T. T. H., & Osier, G. (2023). Identifying the disadvantaged in Luxembourg. Measuring multidimensional poverty by statistical matching. *Working Papers du STATEC*, 133, 26.
- Trewin, D. (2023). *Development of Sample Surveys in Australia and New Zealand over the Last 50 Years* n.d.
- van der Putten, P., Kok, J. N., & Gupta, A. (2002). *Data Fusion Through Statistical Matching* (SSRN Scholarly Paper 297501). <https://doi.org/10.2139/ssrn.297501>
- Verschoor, M., Albers, C., Poortinga, W., Böhm, G., & Steg, L. (2020). Exploring relationships between climate change beliefs and energy preferences: A network analysis of the European Social Survey. *Journal of Environmental Psychology*, 70 (101435), 11 p.
- Vodisek, D. (2023). Les enjeux du déclin démographique de la population italienne. L'Italie, laboratoire du vieillissement européen ? *Les Analyses de Population & Avenir*, 43(1), 1-26.
- Volant, S., Pison, G., & Héran, F. (2019). La France a la plus forte fécondité d'Europe. Est-ce dû aux immigrées ? *Population & Sociétés*, 568(7), 1-4.
- Wang, X., & Ling, J. (2012). Multiple valued logic approach for matching patient records in multiple databases. *Journal of Biomedical Informatics*, 45, 224-230.
- Welsch, H. (2021). How climate-friendly behavior relates to moral identity and identity-protective cognition: Evidence from the European social surveys. *Ecological Economics*, 185 (107026), 11 p.
- Wilkens, M. (2018). *Statistical Matching of the EWCS and EQLS* [Working Paper]. Accessible au lien [Working Paper Statistical Matching.docx \(live.com\)](#) consulté le 13 juillet 2023.
- Wood, J., & Marynissen, L. (2019). Qui reste à la maison ? Organisation du travail rémunéré et des congés après une naissance au sein des couples bi-actifs en Belgique. *Population*, 74(3), 323-354.
- Woodhead, L. (2012). Les différences de genre dans la pratique et la signification de la religion. *Travail, genre et sociétés*, 27(1), 33-54.

Annexe 1. Liste des variables dans les données administratives

Données			Labels
Census 2011	Stock 2011	Naissances	
ID_DEMO_C	ID_DEMO_C	ID_DEMO_C	Identifiant unique
-	MS_AGE		Âge du répondant
-	CD_CNTRY_BTH		Pays de naissance
-	CD_NATLTY		Nationalité
IND	-		Occupation
SIE	-		Statut professionnel
EDU	-		Plus haut niveau d'éducation
LMS	-		Statut marital légal
HAR	-		Condition de logement
TLQ	-		Types de logement
GEO	-		Lieu de résidence
-	-	Nai 9811	Nbre de naissance de 1998 à 2011 (créée)
-	-	Nai 1216	Nbre de naissance de 2012 à 2016 (créée)
	CD_NATLTY_PAR_1		Nationalité du parent 1
	CD_NATLTY_PAR_2		Nationalité du parent 2

Annexe 2. Liste des variables de la ESS

Variables	Labels
rlgdgr	Degrés de religiosité du répondant
agea	âge du répondant
cntbrth	Pays de naissance du répondant
ctzship	Nationalité
ctzcnr	A la nationalité du pays (d'enquête)
facntn	continent de naissance du père
facntr	Le père est né dans le pays (d'enquête)
fbrbcnt	Pays de naissance du père
mocntn	continent de naissance de la mère
mocntr	la mère est née dans le pays (d'enquête)
mbrncnt	Pays de naissance de la mère
tporgwk	Type d'organisation dans laquelle le répondant travaille ou a travaillé.
Isco08	occupation
edlvbe	Plus haut niveau d'éducation-Belgique
marsts	Statut marital légal
pweight	Pondération de la taille de la population
anweight	Pondération d'analyse

Résumé

La Belgique connaît depuis de nombreuses décennies un vieillissement progressif de sa population et une baisse des naissances. Cette situation est le résultat d'un changement de comportement en matière de fécondité lié à de nombreux facteurs, dont la religion. Dans un contexte où les bases de données sont de plus en plus partielles, ce travail analyse les comportements en matière de fécondité en Belgique en rapport avec la religion à partir de la fusion des données administratives belge et de la *European Social Survey* (ESS). Pour atteindre ce résultat, l'étude a analysé les chances pour une femme vivant en Belgique et ayant un enfant en 2011 d'en avoir un autre dans les 5 années suivantes. Le processus de fusion a reposé sur l'approche développée par D'Orazio et *al.* (2006) et l'utilisation du package Statmatch sur R. L'analyse des comportements a reposé sur l'usage d'une régression logistique avec comme variable cible : a eu un second enfant ou pas.

Les résultats montrent que la fusion des données administratives et ceux de la *European social survey* produisent des données viables pour l'étude des comportements sociaux en Belgique. Cette viabilité a été évaluée sur la base d'une faible dissimilarité entre les données d'origines et le fichier synthétiques de données obtenues. L'analyse a permis de montrer que la religion avait un impact positif sur la chance d'avoir un second enfant. Cette analyse permet aussi de faire un lien entre l'âge de la mère à l'arrivée du premier enfant et la situation démographique du pays notamment : la baisse du nombre d'enfants par femmes. Cette étude ouvre à des possibilités immenses de bénéficier de l'excellente qualité des données de registre pour la compréhension des comportements non observés à travers une seule source des données au niveau belge.

Mots clés : fécondité, religiosité, fusion statistique, Statmatch, comportement, Belgique.