Population synthétique

Dumont, Morgane; Carletti, Timoteo; Cornelis, Éric

Published in:
Vieillissement et entraide

Publication date:
2017

Document Version
Version revue par les pairs

Population synthétique: un outil pour une analyse spatiale fine des besoins futurs en soins de santé.

Dumont Morgane, Carletti Timoteo, Cornélis Eric

(NaXys, UNamur)

RÉSUMÉ

L'objectif de notre projet de recherche VBIH (Virtual Belgium In Health) est d'estimer les besoins en soins de santé des personnes âgées dans les années à venir (à l'horizon 2030) en Wallonie. Pour arriver à ce but, nous allons procéder par micro-simulation spatiale, c.-à-d. que nous allons modéliser chaque individu, avec ses caractéristiques propres, en le localisant au niveau de la commune.

Cette micro-simulation débute par la création d'une situation de départ : une population statique pour une année de référence. Ensuite, à l'aide de modèles, nous allons estimer la population, et donc la situation des besoins en soin de santé, d'une année (future) à partir de la population de l'année précédente et de données de transitions. Donc, une population initiale est indispensable pour initier ce processus. La protection de la vie privée et le manque d'accessibilité aux données individualisées existantes nous empêchent de considérer une population « réelle » de base et nous obligent à avoir recours à des techniques de population synthétique pour créer le point de départ de la simulation. Cette méthode consiste à prendre en considération toutes les données statistiques dont nous disposons sur la population et à recréer une population, dite 'synthétique', qui sera statistiquement semblable à la population belge. Cette population synthétique de départ sera caractérisée par des variables socio-économiques et des variables « santé » qui apparaissent pertinentes pour déterminer les besoins en soins de santé. Pour l'année de départ, 2011 a été choisi, car c'est l'année du dernier recensement de la population belge.

Ensuite, différents processus dynamiques doivent être modélisés pour simuler l'évolution temporelle de la population. Cela va du plus simple, qui est le vieillissement (chaque année, l'âge de tous les individus sera incrémenté d'un an),

aux plus complexes, comme l'entrée dans la vie active, les déménagements, les mariages, les divorces, les naissances, les décès,... Notons que ceci est un travail en progression et que tous les processus dynamiques n'ont pas encore été clairement définis.

Par conséquent l'essentiel de cet article est consacré à la manière dont la population synthétique de départ a été créée ; une part évoque cependant comment les processus évolutifs pourraient être mis en œuvre.

1 INTRODUCTION

Avec la croissance de l'espérance de vie, de nouveaux besoins de soins de santé apparaissent. En effet, confronté à une évolution démographique tendant à augmenter la part des personnes âgées dans la population, le système de soins de santé doit être adapté de manière à ce que l'offre continue à répondre à ces nouveaux besoins. Ainsi, la qualité de vie des aînés en Belgique pourra être maintenue. C'est dans cette perspective que s'inscrit notre projet de recherche VBIH (Virtual belgium In Health), financé par la Wallonie. L'objectif de celui-ci est de réaliser une micro-simulation spatiale de la population [8] et de la faire évoluer temporellement (jusqu'en 2030). Nous pourrons alors nous baser sur cet outil pour estimer prospectivement l'évolution, dans le temps et dans l'espace, des besoins en soins de santé des personnes âgées (maisons de repos, hôpitaux, ...).

Cette communication se centre sur les premières phases de notre recherche et décrit les méthodologies mises en œuvre pour mettre sur pieds la micro-simulation spatiale de la population. La section 2 sera ainsi dévolue à la « construction » de la population de base alors que la suivante esquissera les processus envisagés pour la phase ultérieure, consacrée à l'évolution temporelle de cette population. Enfin, une conclusion évoquera les perspectives futures.

2 POPULATION SYNTHÉTIQUE DE DÉPART

Dans une perspective de prospective, il faut d'abord bâtir une situation de base dont on pourra ensuite simuler l'évolution temporelle. Par conséquent, il est d'abord indispensable que nous modélisions une population initiale pour une année de référence (2011); 2011 a été choisie comme année de base, car c'est l'année la plus récente pour laquelle nous disposons d'un corpus de données relativement satisfaisant, notamment grâce au Census 2011 de la Direction générale

Statistique et Information économique du ministère des affaires économiques. Cette population initiale doit être caractérisée par les déterminants que nous estimons pertinents pour associer des besoins de soins de santé aux individus. Ces caractéristiques peuvent être liées tant aux individus (p.ex. l'âge) qu'aux ménages (p.ex. une personne isolée n'aura pas nécessairement les mêmes besoins qu'une personne vivant en couple).

Cependant, une base de données regroupant l'ensemble des informations dont nous souhaitons disposer pour chaque individu, chaque ménage n'existe pas. De plus, les données partielles qui pourraient être disponibles ne sont pas (ou pas facilement) accessibles à un niveau désagrégé pour des raisons de protection de la vie privée.

Dès lors, nous avons besoin de créer une population synthétique contenant des individus groupés en ménages. En effet, une population synthétique est construite de manière à ce que, statistiquement, les indicateurs agrégés calculés sur celle-ci soient les plus proches possibles de ceux disponibles pour la population réelle et ce, quel que soit le niveau d'agrégation spatiale avec lequel ces indicateurs sont fournis. Par ailleurs, cette population synthétique se caractérise par des individus et des ménages auxquels on associe l'ensemble de tous les attributs pertinents qui nous sont nécessaires.

Ensuite, comme notre objectif est de pouvoir localiser les zones qui auront plus de besoins en soins de santé que les autres, avec une granularité spatiale la plus fine possible mais, que par ailleurs, les statistiques les plus désagrégées disponibles ne le sont qu'au niveau communal, nous choisissons de construire, et donc localiser, notre population synthétique commune par commune.

Enfin, le processus de création de la population synthétique se passe en deux étapes. Dans un premier temps, on "fabrique" les individus qu'ensuite on regroupera en ménages. Les deux sous-sections suivantes vont successivement envisager ces deux aspects.

2.1.1 Caractéristiques de base

Les individus sont générés en utilisant des tables de contingence issues de données venant du Registre National [6] et du Census2011 [7]. Le Registre National donne des informations sur l'âge, le genre, la commune ainsi que sur la taille et le type du ménage et enfin sur le lien de l'individu vis à vis de son chef de ménage (conjoint, enfant, etc.). Par contre des caractéristiques nous apparaissant déterminantes dans les besoins de soins de santé ne sont pas présentes dans le Registre National; il s'agit du plus haut diplôme obtenu et du statut professionnel, que nous pouvons utiliser comme proxys des revenus du ménage. Dans le Census2011, ces informations

sont présentes, mais, pour des raisons de respect de la vie privée, il n'est possible d'accéder qu'à des tables contenant maximum quatre variables simultanément. Par conséquent, il n'est pas possible d'obtenir des totaux marginaux pour des catégories définies par le croisement de toutes les caractéristiques que nous souhaitons prendre en compte. Il faut donc combiner les données issues de ces deux sources pour obtenir des individus dotés de toutes les caractéristiques qui nous intéressent.

Sur base des données du Registre National, nous pouvons créer directement nos individus synthétiques caractérisés par les variables présentes dans celui-ci (âge, genre, commune, etc.). Ensuite, nous souhaitons « ajouter » à chaque individu un diplôme et un statut professionnel. Pour cela, nous avons choisi de nous baser sur trois tables du Census2011. Tout d'abord, nous avons considéré une table croisant l'âge, le genre, la commune et le diplôme, ainsi qu'une autre table croisant l'âge, le genre, la commune et le statut professionnel. Ces deux tables auraient été suffisantes si nous pouvions considérer que la corrélation entre diplôme et statut professionnel est entièrement déterminé via l'âge, le genre et la commune. Cette hypothèse étant douteuse, nous avons choisi de considérer une table supplémentaire comprenant simultanément le diplôme et le statut professionnel. En procédant à une analyse de corrélation, nous avons choisi de garder également le genre et l'âge dans les variables croisées dans cette table (et donc de « laisser tomber » la commune). Nous avons donc utilisé trois tables du Census2011 croisant les variables, comme repris dans la table 1.

Caractéristique	Table 1	Table 2	Table 3
Commune	×	×	
Classe d'âges	×	×	×
Genre	×	×	×
Diplôme	×		×
Statut professionnel		×	×

Table 1: Croisements présents dans les tables extraites du Census2011.

Avant d'aller plus loin, il convient de noter que le recensement Census2011 reprend également les données relatives aux demandeurs d'asile qui, par contre, ne sont pas présentes dans le Registre National. La première étape de notre démarche a donc consisté à remettre ces tables à l'échelle afin de correspondre aux marginales du Registre National.

Pour ajouter les deux variables servant d'indicateurs du revenu, nous avons recours à une technique largement répandue : l'IPF ("Iterative Proportional Fitting procedure" [4] et [1]). Celle-ci permet de construire, sur base des tables croisées et agrégées disponibles (appelées contraintes), un tableau croisé de toutes les caractéristiques prises en compte et, donc, de générer l'ensemble des individus avec les caractéristiques désirées. Cette méthode démarre avec une table complète de départ et l'adapte, itération après itération, pour respecter exactement une des contraintes, tout en gardant les proportions actuelles.

Nous avons pris comme table de départ celle qui n'est pas spatialement désagrégée (donc celle croisant âge, genre, diplôme et statut professionnel), que nous avons dupliquée pour chaque commune. Ensuite, le processus itère pour respecter au mieux les contraintes qui sont fournies par les deux autres tables. Nous obtenons ainsi une table finale comprenant les effectifs par âge, genre, commune, diplôme et statut professionnel. Cette table est utilisée comme distribution pour tirer aléatoirement (et sans remise) un diplôme et un statut professionnel à chaque individu présent dans le Registre National.

Nous obtenons donc une population synthétique comprenant toutes les personnes du Registre National avec les caractéristiques suivantes:

- Age (classes par tranche de 5 ans),
- Genre,
- Commune,
- Diplôme (plus haut niveau obtenu dans la classification CITE[1]),
- Statut professionnel (inactif, travailleur, chômeur),
- Taille du ménage (de 1 à '6 et plus'),
- Type de ménage (couples avec ou sans enfants, cohabitants avec ou sans enfants, familles monoparentales, ménages collectifs, personnes isolées, autres)
- Lien avec le chef de ménage (chef, conjoint, enfant, parent, ...).

La population synthétique ainsi obtenue est statistiquement semblable à la « réelle » population belge. Pour quantifier les différences, notre population dite 'synthétique' est agrégée afin de créer des tables du même type que les tables de contingence de départ. La comparaison entre ces tables produites et celles souhaitées, c.-à-d. celles issues du Census 2011 nous permet d'observer si le processus de construction a bien produit une population synthétique

[1] Classification Internationale Type de l'Education (allant de 1 : « Enseignement Primaire » à 8 : « Niveau doctorat ou équivalent »))

statistiquement semblable à la « vraie » population. Pour illustrer cela, nous employons des graphiques où chaque point correspond à une catégorie précise (combinaison possible des variables agrégées). L'abscisse d'un point est le nombre de personnes dans notre population synthétique qui appartiennent à cette catégorie et l'ordonnée est le nombre de personnes dans la table de départ (remise à l'échelle).

La figure 1 nous montre que notre simulation est très proche de la réalité pour toutes les catégories de diplôme, commune, âge et genre puisque chaque abscisse est égale à chaque ordonnée. La table 1 est donc bien respectée.

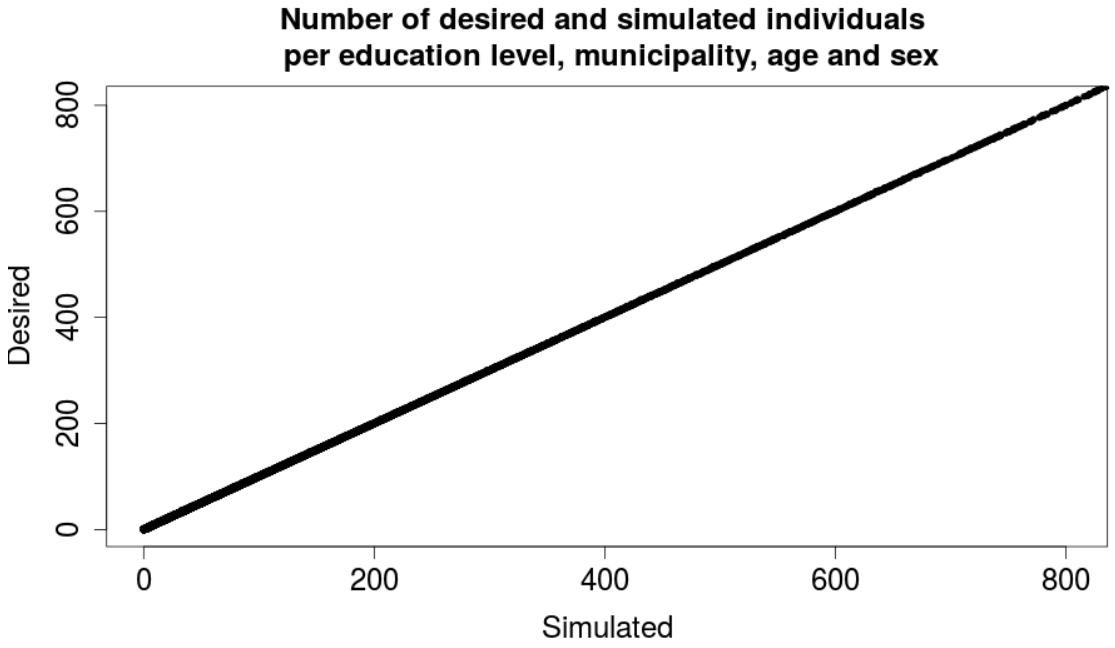

Figure 1 –Qualité de la simulation par rapport à la première table:
diplôme en fonction du genre, de l'âge et de la commune

Sur la figure 2, nous observons également une droite où les ordonnées et les abscisses sont identiques. Les effectifs des catégories de statuts, communes, âges et genres de notre simulation correspondent donc également à la réalité observée.

Notons que dans notre processus d'ajout du diplôme et du statut professionnel, nous avons d'abord essayé de considérer les trois tables comme contraintes fortes. Cependant, l'IPFP ne nous permettait jamais de coller parfaitement à ces trois tables simultanément. Nous avons alors choisi les deux premières tables comme contraintes fortes, puisque notre but est de nous rapprocher le plus possible de la population par commune. Il est donc très satisfaisant de respecter d'abord ces deux contraintes fortes au niveau communal.

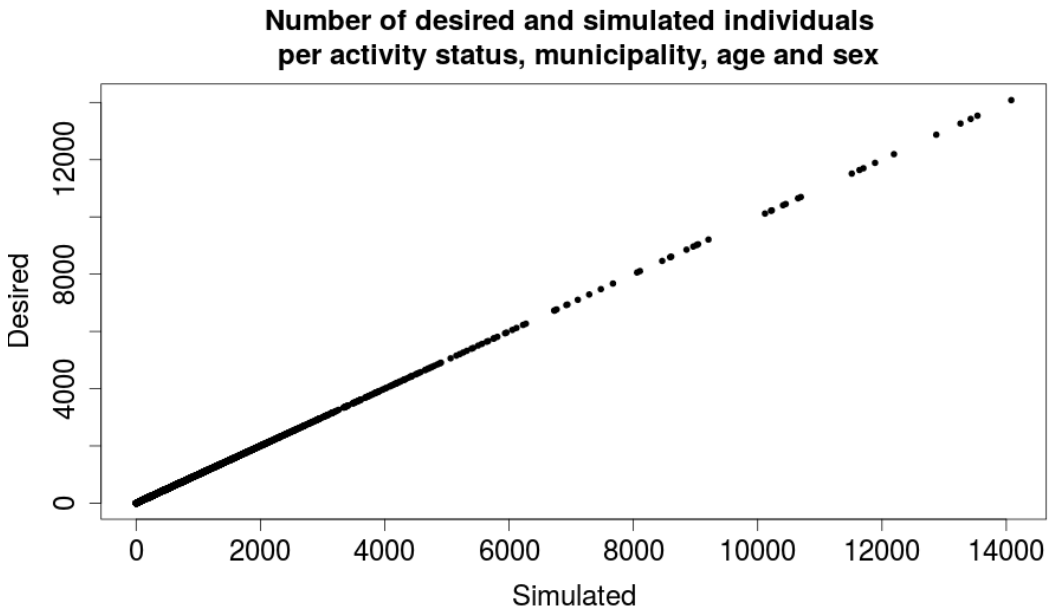

Figure 2 – Qualité de la simulation par rapport à la deuxième table : statut professionnel en fonction du genre, de l'âge et de la commune

La troisième table est nationale et nous permet simplement d'avoir une idée des corrélations entre le diplôme et le statut professionnel. Elle n'est utilisée que pour initialiser la matrice de poids (correspondant au croisement des quatre variables considérées : diplôme, statut professionnel, âge et genre) par commune. Il va donc s'avérer logique d'avoir un résultat un peu plus nuancé pour cette dernière table.

À la Figure 3, nous observons qu'en effet, les résultats sont légèrement moins bons pour le croisement des catégories d'âge, genre, diplôme et statut. Cependant, les points restent malgré tout dans une bande étroite autour de la droite « idéale » et nous n'avons aucun point complétement aberrant.

De plus, la table 3 n'était pas spatialement désagrégée, donc les légers écarts sont répartis sur l'ensemble des communes. Rappelons que nous avons essayé d'ajouter cette troisième table comme contrainte, mais que cela amenait à un moins bon respect des deux premières tables. Or, dans le cadre de notre recherche, nous sommes très intéressés par les différences communales ; donc, nous avons préféré nous éloigner légèrement plus de cette table nationale pour être très proche des tables communales.

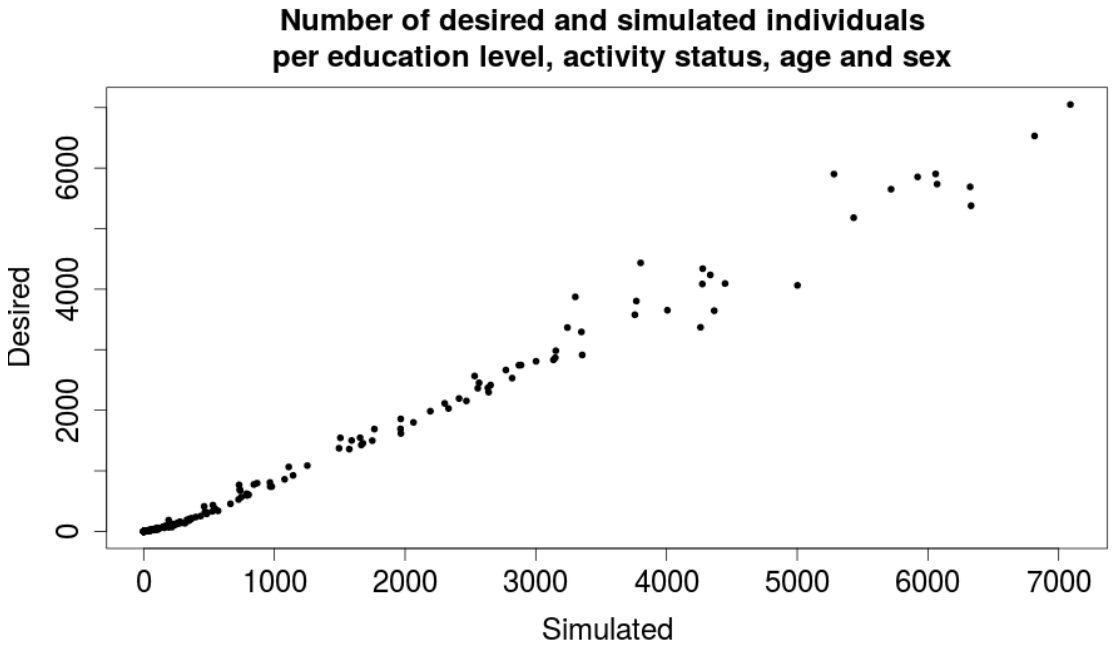

Figure 3 – Qualité de la simulation par rapport à la troisième table : diplôme, statut professionnel, genre et âge

2.1.2 Caractéristiques de santé

Notre projet étant de prédire les besoins en soins de santé des personnes âgées, nous avons également déterminé des caractéristiques supplémentaires qu'il serait pertinent et possible d'associer à nos individus. Pour cela, en collaboration avec l'OWS et le service gériatrique du CHU-Mont-Godinne, nous avons défini cinq groupes de maladies qui affectent les personnes âgées et qui peuvent être identifiées par des médicaments traceurs. Grâce à la base de données Pharmanet[2], nous avons pu ainsi produire des tables reprenant les maladies choisies par âge, genre et commune. Les maladies prises en compte sont:

- Parkinson,
- Ostéoporose,
- Diabète,
- Douleurs chroniques,

[2] « Pharmanet est une banque de données sur les prestations pharmaceutiques effectuées par les pharmacies publiques et remboursées par l'assurance soins de santé obligatoire. » Extrait du site de l'INAMI

- Maladies pulmonaires chroniques.

Connaissant le nombre de personnes ayant chaque maladie par classe d'âge, genre et commune, nous tirons aléatoirement les individus atteints au sein de ces classes pour leur associer cette caractéristique de maladie dans la population synthétique.

Par conséquent, à l'issue de cette phase du travail, nous disposons d'une population d'individus synthétiques pour chacune des communes belges ; ces individus étant caractérisés par

- leur genre ;
- leur âge (tranches d'âges de 5 ans) ;
- leur diplôme ;
- leur statut professionnel ;
- le type de leur ménage ;
- la taille de leur ménage ;
- leur lien par rapport à leur chef de ménage ;
- la présence ou l'absence de chacune des cinq maladies considérées.

2.2 GROUPEMENT DES INDIVIDUS EN MÉNAGES

L'étape suivante consiste à grouper ces individus pour former les ménages. Cela constitue en fait un énorme problème d'optimisation combinatoire, puisqu'il s'agit de regrouper l'ensemble des individus dont nous disposons en des sous-groupes (les ménages) en veillant à nous rapprocher le plus possible des statistiques sur les ménages dont nous disposons. Nous utilisons, par exemple, pour les couples des statistiques indiquant les différences d'âges entre conjoints. Pour les enfants, nous disposons de statistiques semblables mais où on reprend l'âge des enfants en fonction de l'âge du chef de ménage. Ces statistiques sont disponibles en fonction du type de ménage. Pour illustrer le problème, nous allons expliquer plus en détails ce qui se passe pour les couples mariés avec enfants (figure 4). Afin de respecter au mieux les données que l'on connait déjà concernant le ménage de chaque individu, nous allons commencer par considérer chaque type de ménages séparément, car ces variables sont exclusives. En effet, quelqu'un qui vit en couple ne peut pas aussi vivre avec quelqu'un qui vit dans un ménage collectif. Ensuite, nous séparons encore, dans chacun des types, par taille de ménages. Enfin, au sein de chaque taille de ménages, nous considérons les types de lien avec le chef. Ainsi, une fois dans les bons type et taille de ménages, pour créer un ménage, il reste à choisir les individus pour que les liens soient

respectés. Par exemple, pour les couples mariés avec enfants, pour la taille 4, nous devons choisir un chef, un conjoint et 2 enfants.

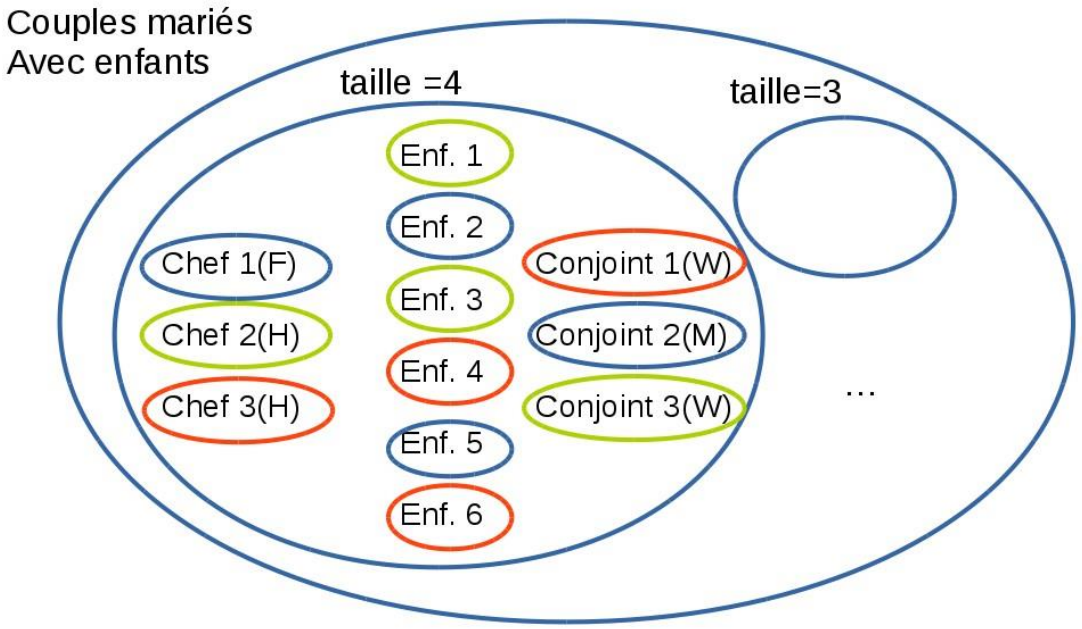

Figure 4 – Illustration des groupements pour les couples mariés avec enfants

Pour cela, nous devons commencer par déterminer un critère d'optimalité pour quantifier la qualité de l'assemblage. Par exemple, ayant accès à la distribution des différences d'âges au sein des couples mariés, minimiser l'écart entre cette distribution et celle observée dans notre groupement est un tel critère. Mais, dans la mise en œuvre de cette phase, nous pouvons prendre en compte simultanément plusieurs critères de ce type. L'ensemble des critères considérés dépend essentiellement des informations statistiques, des corrélations pertinentes et disponibles. Pour grouper les ménages, plusieurs méthodes sont possibles.

Nous avons dans cette recherche commencé par tester une méthode appelée "Recuit Simulé"[3] qui est une technique d'optimisation combinatoire. Nous avons préféré celle-ci à un algorithme génétique, qui nécessite à chaque itération un grand nombre de duplications de la population totale, contrairement au Recuit Simulé qui ne prend en compte qu'une seule fois la population et la modifie d'étape en étape. Un Recuit Simulé a été testé, mais prend énormément de temps pour arriver à une solution moins proche de la réalité que celle obtenue avec une méthode plus traditionnel. Néanmoins, cette méthode nécessiterait plus d'investigations et est plus facilement réalisable qu'une méthode traditionnelle pour les problèmes multi-objectifs. Les bases du Recuit Simulé sont expliquées dans la section ci-dessous. Une seconde méthode, plus classique, a été utilisée ; elle consiste à prendre en compte les distributions telles quelles et à procéder à des

tirages aléatoires au sein de ces lois de probabilité. Cette méthode est développée dans la seconde sous-section ci-dessous.

2.2.1 Création des ménages à l'aide d'un Recuit Simulé

Le "Recuit Simulé" est divisé en deux étapes consécutives. Tout d'abord, nous générons un assemblage "aléatoire". Ce groupement est cependant réalisé en respectant les informations déjà disponibles (au travers des individus et plus spécifiquement des données utilisées issues du Registre National) à savoir les tailles et types de ménages, ainsi que les liens 'conjoint' et 'enfant'. Le respect des liens connus via les données est assuré et l'aspect de tirage aléatoire n'est présent que dans le choix des personnes (présentant les bonnes caractéristiques notamment de lien).

Une fois un premier groupement purement aléatoire (mais respectant les informations issues des données) effectué, il faut alors passer au "recuit simulé" proprement dit. Nos critères sont essentiellement axés sur les différences d'âges, entre conjoints, entre parents et enfants et/ou entre enfants. Ceux-ci nous permettent de définir une fonction objectif qui sera une jauge permettant de mesurer la "qualité" du groupement; elle représente en quelque sorte la proximité par rapport à un groupement en ménages optimal, le plus statistiquement proche de la réalité telle qu'elle peut être lue au travers des données disponibles.

Le processus du "recuit simulé" consiste à permuter, au sein d'une même catégorie (définie par le type et la taille du ménage), deux individus présentant le même lien (p.ex. deux enfants) entre deux ménages créés et à évaluer si le nouveau groupement issu de cette permutation est plus proche de la configuration optimale que l'ancien (en termes mathématiques : maximise davantage la fonction objectif). Si c'est le cas, on le conserve, sinon, cette mauvaise combinaison peut néanmoins être gardée dans l'ensemble des ménages mais avec une probabilité décroissante en fonction du nombre d'itérations déjà réalisées et de l'écart par rapport à la mesure d'optimalité calculée au pas précédent. L'idée étant que des "mauvais" choix initiaux peuvent permettre de mieux explorer l'espace de solutions et donc atteindre un meilleur résultat final. Ensuite, on tente une autre permutation. Ce processus est répété jusqu'à un critère d'arrêt: soit la population stagne et on ne trouve plus de permutations pertinentes, soit l'algorithme a atteint un nombre maximum d'itérations défini au départ.

Nous avons testé cette méthode uniquement pour la création des couples pour la commune de Namur vu l'important temps de calcul nécessaire. Pour établir notre critère d'optimalité, nous utilisons exclusivement une table issue du Registre National reprenant l'âge de l'épouse par rapport à l'âge de son conjoint. Comme expliqué précédemment, nous partons d'un groupement aléatoire et faisons des échanges. Après chaque échange, nous devons savoir si cette permutation

a permis de se rapprocher de la table théorique, ou pas. La fonction dite 'fitness' quantifiant la qualité de la population simulée doit être entre 0 et 1 avec 1 exprimant un parfait accord entre données réelles et simulées. Nous avons donc choisi la fonction :

$$Fitness = 1 - \frac{Erreur\_totale\_absolue}{Erreur\_maximale\_possible} = 1 - \frac{\sum_{i=1}^{n} | Obs_i - Reel_i |}{2 * nombre\_menages}$$

où *Reel* est la table issue du Registre National, *Obs* la table correspondant à la simulation en cours (agrégée pour être comparable à *Reel*) et *n* le nombre de cellule dans chacune de ces tables. Nous voyons que, pour ramener l'erreur entre 0 et 1, nous divisons par deux fois le nombre de ménages (de type couple) à générer. Cela est dû au fait que, s'il manque un couple dans une combinaison âge homme et âge femme, cela signifie qu'un autre groupe aura une combinaison de trop. Par conséquent, chaque mauvaise combinaison est répercutée deux fois. Le principe de la méthode est de toujours garder un échange qui a amélioré la population, mais également, de parfois garder un échange qui l'a « faiblement » détériorée. Cela permet de ne pas 'stagner' dans des maxima locaux. La probabilité de garder un mauvais échange est donnée par:

$$P = e^{\frac{-erreur}{temperature}}$$

où la température est une fonction décroissante du nombre d'itérations du processus de construction de la population. De cette manière, plus on avance dans les itérations, moins on a de chances de garder une moins bonne population. La fonction choisie, basée sur les explications de [3] et adaptée en fonction d'essais, est:

$$Temperature = \alpha^{\frac{iteration}{500}} * T0$$

Avec $T0$ la température initiale, fixée à 0.1, et α fixé à 0.95. Remarquez que, pour passer d'une population à la suivante, on peut décider combien d'échanges sont faits. Pour pouvoir prendre une décision à ce sujet, nous avons testé différentes simulations reprise à la figure 5.

Remarquez que pour ce test, nous n'avons pas encore inséré la probabilité de garder un moins bon choix.

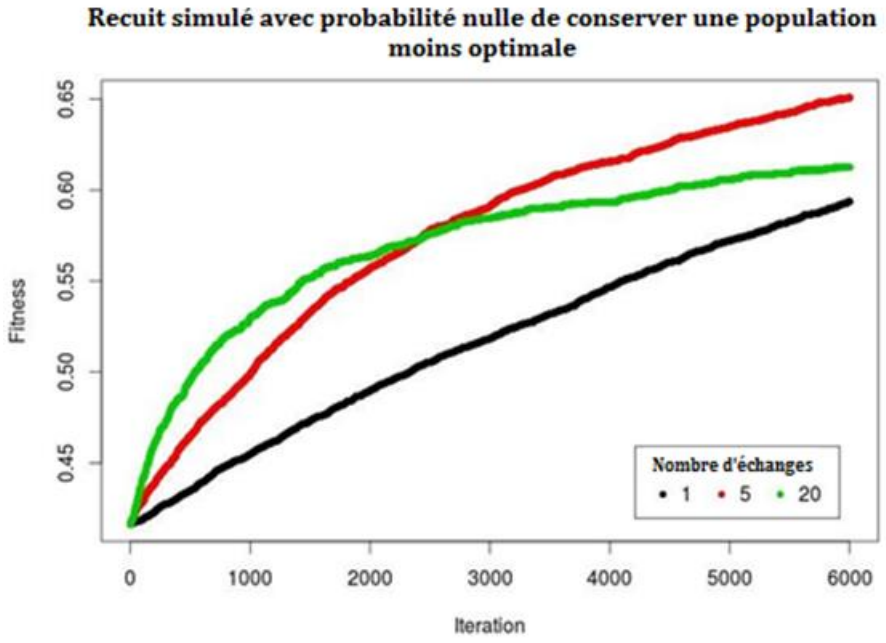

Figure 5 – Évolution de la fitness en fonction du nombre d'échanges par itération.

Nous pouvons voir que faire 20 échanges permet d'améliorer très vite la population au départ, mais ralenti l'amélioration par la suite. Au départ, le moins bon est de ne faire qu'un seul échange. Suite à cette observation, nous avons décidé de faire varier le nombre d'échanges par itération.

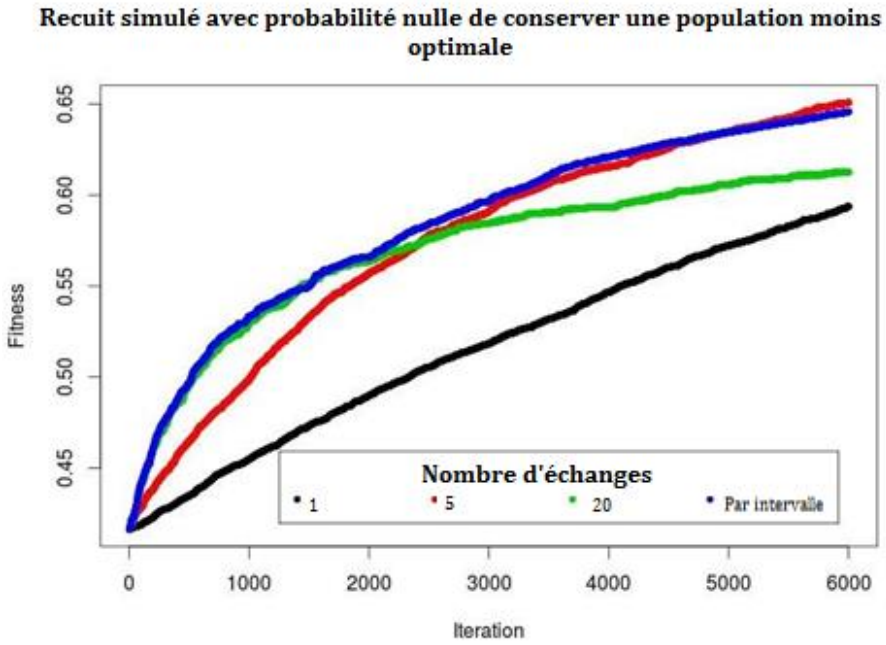

Figure 6 – Évolution de la fitness en fonction du nombre d'échanges par itération avec variation du nombre d'échanges par intervalles d'itérations

Nous procédons à 20 échanges pour les 2000 premières itérations, ensuite à 5 échanges jusqu'à 4000 itérations puis à 1 seul échange pour le reste. Nous constatons, à la figure 6, que cela permet d'augmenter la fitness rapidement au départ, tout en restant dans une amélioration satisfaisante par la suite. Notez que nous avons essayé des fonctions décroissantes plus évoluées, mais rien n'a été plus concluant que cela.

Pour Namur, la création des couples avec les choix décrits ci-dessus est illustrée à la figure 7. On peut voir que l'évolution est bien présente et que l'on passe d'une fitness de 0.44 à une fitness de 0.90. Cependant, cela prend quand même 125 000 itérations.

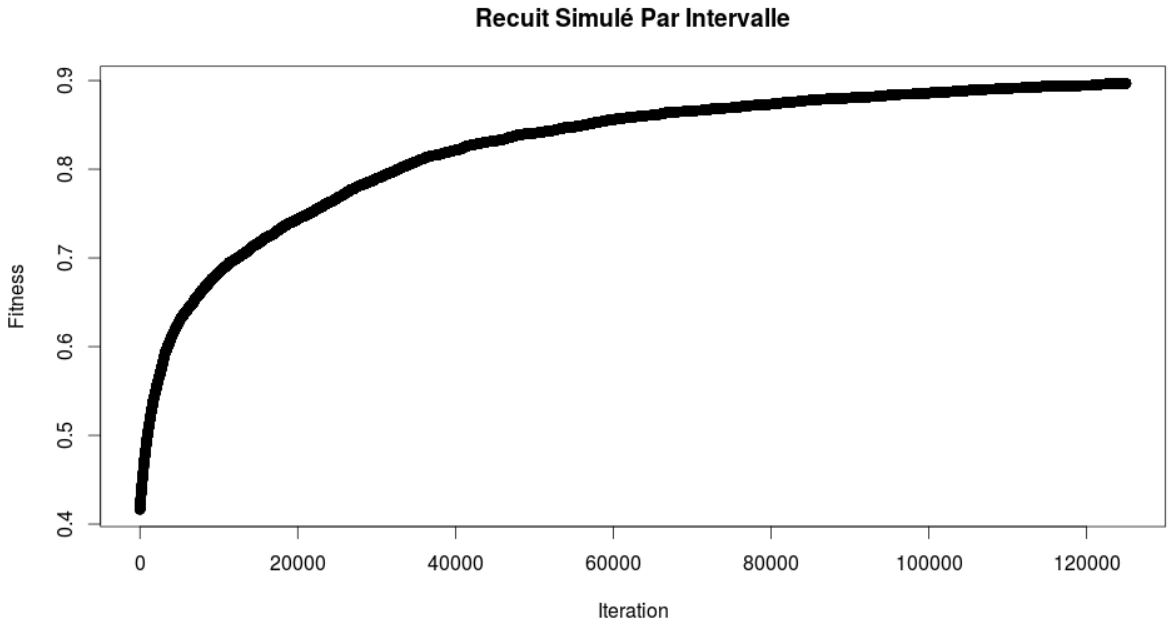

Figure 7 – Résultat du Recuit Simulé pour les couples de Namur

2.2.2 Création des ménages par les probabilités

La seconde méthode utilisée, plus classique, prend en compte les distributions connues telles quelles et procède à des tirages aléatoires au sein de ces lois de probabilité. Cependant, les tables de différences d'âge entre conjoints ne reprenant pas les tailles de ménages, cette méthode ne nous permet pas nécessairement de respecter pour tous les individus exactement les tailles de ménage qui leur sont associées tout en respectant exactement les tables de différences d'âge. Pour cette raison, nous avons choisi de rester dans les mêmes tailles de ménages tant que possible, mais de permettre des assemblages d'individus associés à des tailles de ménages différentes pour respecter les différences d'âges. Ce choix a été fait pour essayer d'assigner un maximum de personnes à des ménages, sans créer de couples improbables. Il nous paraissait moins important de respecter les tailles que les écarts d'âges.

En appliquant cette méthode commune par commune, nous obtenons un groupement qui a une fitness telle que définie par le Recuit Simulé de 1. Nous pouvons analyser les résultats pour chaque statistique que nous avons essayé de respecter. Toutes les statistiques utilisées ici proviennent du Registre National. Tout d'abord, nous avions une table comprenant les couples mariés (avec ou sans enfants) reprenant par commune le nombre de ménages par âge du conjoint, âge du chef de ménage, genre du conjoint et genre du chef. Ensuite, nous avions une table reprenant par commune et par type de ménages, le nombre de combinaisons entre âge d'un enfant et âge du chef de ménage. Ces résultats sont représentés respectivement aux figures 8 et 9.

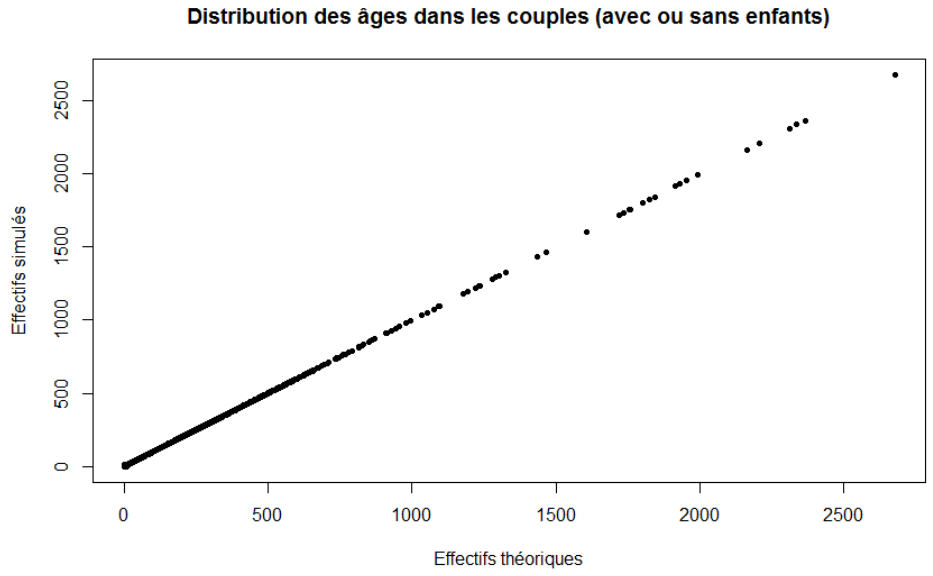

Figure 8 – Résultat pour les âges au sein des couples avec ou sans enfants

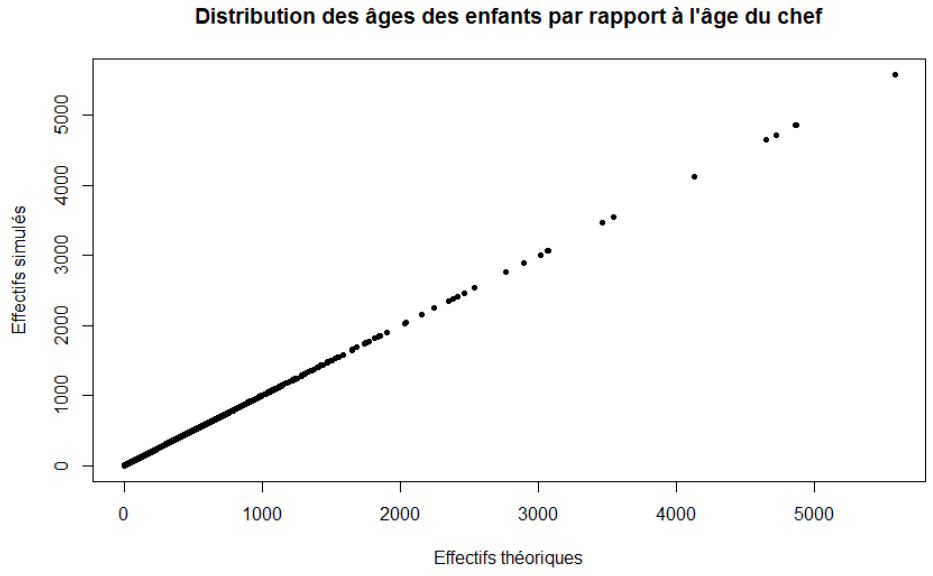

Figure 9 – Résultat pour les âges des enfants et chef de ménage

Nous pouvons voir que, dans les deux cas, les statistiques sur les différences d'âges sont très bien respectées. Ceci est logique, puisque, comme expliqué précédemment, nous avons mis la priorité sur le respect des écarts d'âges au détriment des tailles de ménages.

Par conséquent, il est important de quantifier le nombre de ménages 'mal formés' dans le sens où des individus devant vivre dans des ménages de tailles différentes se retrouvent ensemble.

Dans ce but, nous considérons que la taille de ménage théorique est celle du chef et que les personnes auxquelles est associée une taille de ménage différente de celle du chef de leur ménage sont 'mal placées'. Pour cela, nous nous sommes restreints aux catégories de tailles inférieures à '6 et plus'. La figure 10 indique que, pour la meilleure commune, nous avons 0.7% d'individus dans ce cas. La pire commune, Aubange, a 1% d'individus vivant avec un mauvais nombre de cohabitant. Cette erreur est également tout à fait acceptable.

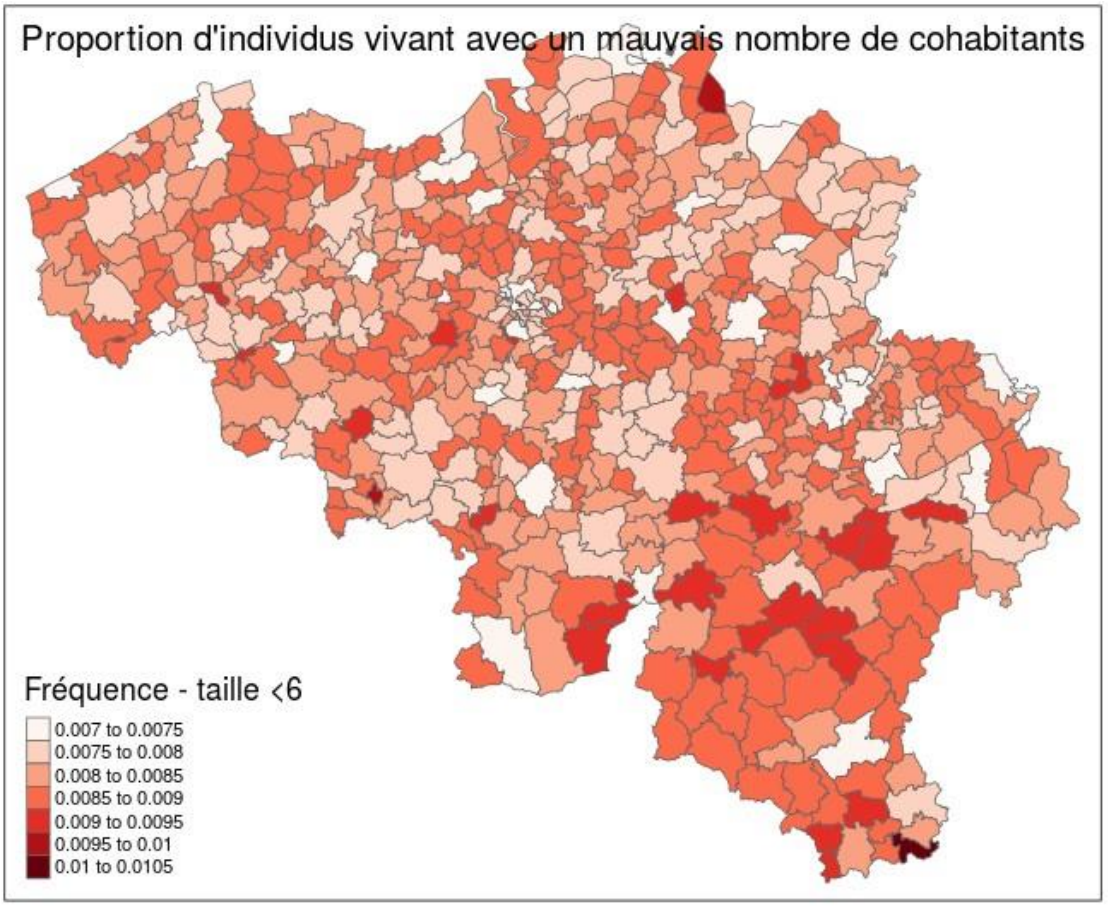

Figure 10 – Personnes vivant dans un « mauvais » ménage (en termes de taille du ménage)

Ayant mis la priorité absolue sur les écarts d'âges, il est possible que des individus restent non assignés à un ménage si plus aucune combinaison n'est possible en respectant les distributions. Il est donc intéressant d'observer la proportion de personnes qui ne sont pas assignées à un ménage dans chaque commune. Pour cela, nous divisons le nombre de personnes non assignées par le nombre total d'habitants de la commune et nous créons une carte reprenant cet indicateur. La figure 11 indique que la vaste majorité des communes a entre 0 et 0.1% d'individus restés seuls alors qu'ils auraient dû être dans un ménage. Les deux pires communes sont Baarle-Hertog et Martelange avec, respectivement 0.41% et 0.26% de personnes non assignées. Toutes les autres communes ont des indices inférieurs à 0.2%. Les résultats présents sur cette carte sont donc très satisfaisants.

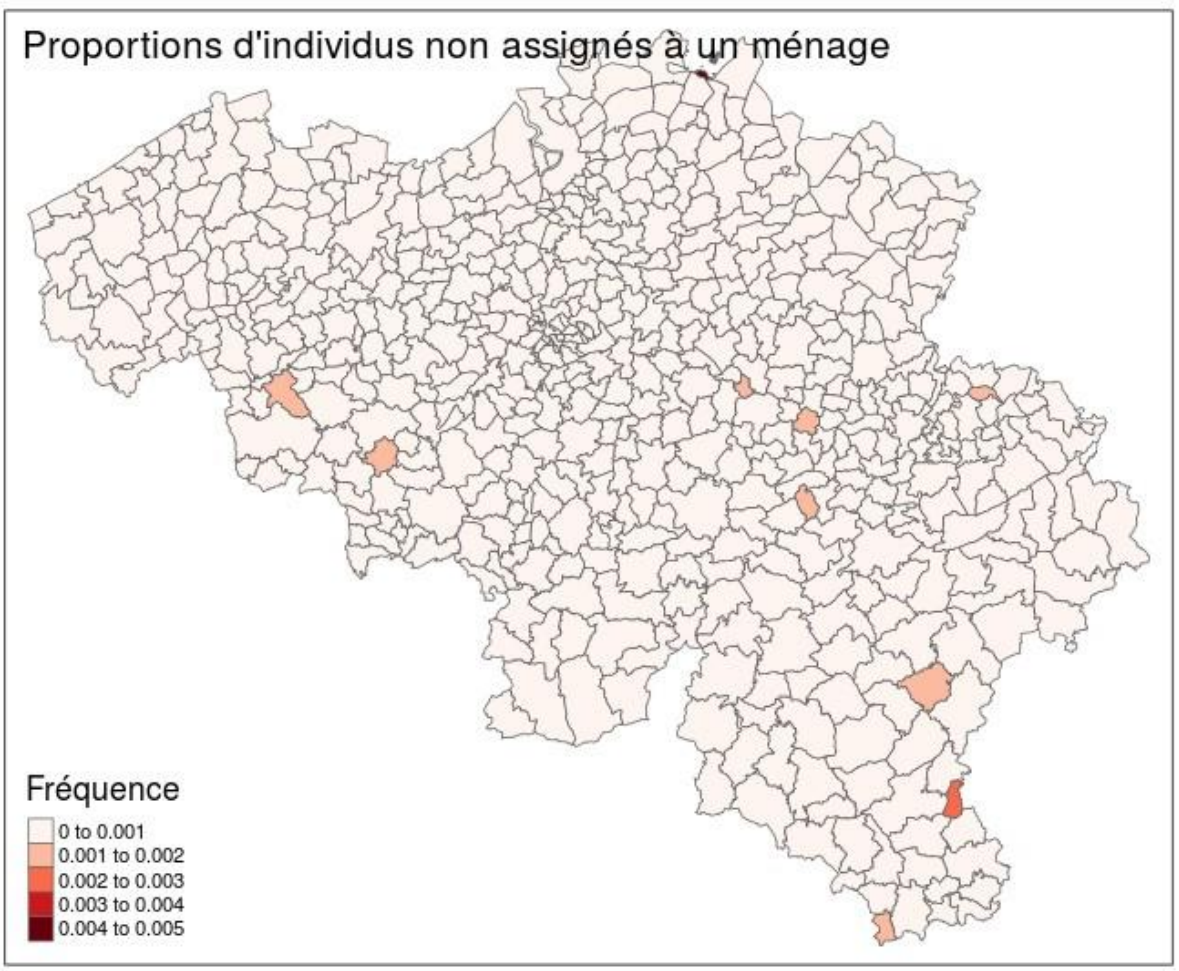

Figure 11 – Personnes non assignées à un ménage

3 ÉVOLUTION TEMPORELLE

La création de la population synthétique pour l'année de base n'est qu'un premier pas. Il nous faut ensuite modéliser les processus d'évolution temporelle de cette population [2]. Premièrement, d'une année à l'autre, nous vieillissons tous d'un an. Cependant, certains vont mourir et d'autres vont naître. Ceci ajoute déjà deux dynamiques à modéliser. Ensuite, chaque individu va évoluer, obtenir un nouveau diplôme, changer de statut professionnel, déménager, se marier, divorcer... Tous ces phénomènes vont générer des modélisations plus ou moins complexes. Par exemple, pour les mariages, nous devons procéder en plusieurs étapes. Tout d'abord, pour chaque individu, nous allons décider s'il désire se marier cette année ou non. Dans un second temps, il faudra créer les couples parmi ceux qui désirent s'engager. Ensuite, nous devons aussi déterminer quel(s) partenaire(s) va (vont) déménager et vers quelle commune.

Dans la plupart des exercices de projections démographiques de ce type, des taux sont utilisés. Par exemple, chacun a une probabilité fixée de se marier (suivant certains facteurs tels que l'âge et le genre) et un simple tirage aléatoire suivant cette loi est effectué pour déterminer si cette personne va s'engager cette année ou non. Notre modélisation désire plutôt considérer ces problèmes d'évolution en prenant davantage en compte les causalités de ces processus de choix (se marier ou pas, déménager ou non, ... [5]). Par exemple, le choix de se marier est le résultat d'un processus beaucoup plus complexe qu'une chance d'engagement. On peut imaginer que cette décision dépend du nombre de célibataires dans l'entourage de la personne, de ces relations passées, ou de bien d'autres paramètres. Un modèle de choix discret peut donc être développé, afin de déterminer quelles caractéristiques influencent ce choix et dans quelle mesure. De cette manière, au lieu d'appliquer à chaque personne une probabilité de se marier, nous allons observer les paramètres décisifs de cet individu et appliquer le modèle. Nous aurons alors, l'utilité (ou le "bonheur" fourni) de chacun des choix en fonction des variables observées. De là, l'engagement sera peut-être choisi mais basé sur des faits raisonnés et des processus explicatifs déduits de la réalité observable.

D'autres mécanismes (p.ex. les naissances ou les divorces) peuvent aussi être modélisés de la même façon. Le recours à ces techniques de choix discrets dépendra aussi des données qui seront disponibles pour calibrer ces modèles, ce qui est une étape incontournable avant leur utilisation.

D'un point de vue pragmatique, les dynamiques temporelles seront d'abord modélisées par des procédés statistiques consistant à simuler la survenance d'événements (naissance, mariage, déménagement, etc.) en fonction de distributions de probabilités établies en tenant compte de

caractéristiques associées aux individus ou aux ménages. Ensuite, en fonction de la disponibilité de données, des processus plus élaborés recourant aux choix discrets seront envisagés.

4 CONCLUSION

En l'état actuel de notre projet de recherche VBIH, nous disposons maintenant d'une population synthétique composée d'individus regroupés en ménages et localisés au niveau de la commune. De plus, des indicateurs « santé » ont également été associés à ces individus synthétiques.

Les mécanismes à mettre en œuvre pour simuler les dynamiques de l'évolution temporelle de cette population synthétique sont maintenant établis et vont pouvoir être mis en application.

Cette modélisation de l'évolution dynamique de la population permettra d'obtenir des portraits de la population future aux différentes années considérées dans notre horizon prospectif (2030). Sur base des corrélations mesurées entre les caractéristiques des individus et des ménages prises en compte et les besoins de soins de santé, il sera alors possible d'évaluer de manière prospective l'évolution, tant dans le temps que dans l'espace, de ces besoins. Cette connaissance des tendances de la demande de soins de santé pourra alors être utilisée pour une planification spatio-temporelle de l'offre.

5 REMERCIEMENTS

Nous remercions:

- la Wallonie, DGO 6, qui finance cette recherche dans le cadre du programme WB-Health;
- le groupe DEMO, IACCHOS, UCL, avec qui nous collaborons pour ce projet et qui nous a fourni des données;
- l'Observatoire Wallon de la Santé (OWS), DGO 5, qui est le parrain de cette recherche et a procédé à la recherche des données de santé;
- et la 'Plateforme Technologique Calcul Intensif (PTCI)' de l'Université de Namur, dont les moyens sont nécessaires aux modélisations informatiques développées.

6 RÉFÉRENCES

[1] M. Lenormand, G. Deffuant (2015), "Generating a synthetic population of individuals in households : Sample-free vs sample-based methods", Journal of Artificial Societies and Social Simulation, 16(4) :12. Disponible sur http://arxiv.org/pdf/1208.6403v2.pdf

[2] J. Barthélemy, PhD Thesis : "A parallelized micro-simulation platform for population and mobility behaviour - Application to Belgium", Unamur, 2014

[3] Holger H. Hoos, Thomas Stützle (2005) "Stochastic local search : foundations and applications", Morgan Kaufmann publishers, Elsevier.

[4] J. Barthelemy, P. Toint (2013), "Synthetic Population Generation Without a Sample". Transportation Science 47(2) :266-279. http://dx.doi.org/10.1287/trsc.1120.0408

[5] E. Cornelis, J. Barthelemy, X. Pauly, F. Walle (2012), « Modélisation de la mobilité résidentielle en vue d'une micro-simulation des évolutions de population », Les Cahiers Scientifiques du Transport, 62, pp. 65-84

[6] Tables de contingences fournies par le groupe DEMO, IACCHOS, UCL.

[7] Tables de données créées par la SPF Economie, disponibles sur http://census2011.be/censusselection/selectionFR.html.

[8] D. Dallas, G. Clarke, "Modelling the local impacts of national social policies: A Microsimulation Approach", paper presented at the 11th European colloquium on Theoretical and Quantitative Geography, England, 3d-7th September 1999