DOCTORAT BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE NANTES

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE NANTES

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs interactions*

Par

Boyam Fabrice YAMEOGO

Méthodologie de calibration d'un modèle multimodal des déplacements pour l'évaluation des externalités environnementales à partir de données ouvertes (open data) : le cas de l'aire urbaine de Nantes

Thèse présentée et soutenue à Nantes, le 15 décembre 2021
Unité de recherche : Laboratoire Ease, Université Gustave Eiffel

Rapportrice et Rapporteur avant soutenance :

Juliette ROUCHIER	Directrice de Recherche CNRS, Université Paris Dauphine - Université Paris Sciences Lettres
Guillaume DEFFUANT	Directeur de Recherche INRAE, INRAE Clermont Ferrand

Composition du Jury :

Président :	Éric CORNELIS	Professeur des Universités, Université de Namur
Examinateurs :	Yannick AOUSTIN	Professeur des Universités, Université de Nantes
	Nicolas COULOMBEL	Ingénieur en chef du corps des Ponts Eaux et Forêts, École des Ponts - ParisTech
Dir. de thèse :	Pierre-Olivier VANDANJON	Chargé de Recherche dév. durable HDR, Université Gustave Eiffel, Nantes
Encadr. de thèse :	Pierre HANKACH	Chargé de Recherche dév. durable, Université Gustave Eiffel, Nantes
Encadr. de thèse :	Pascal GASTINEAU	Chargé de Recherche dév. durable, Université Gustave Eiffel, Nantes

Invités :

Jean CALIO	Ingénieur, SNCF Réseau, Saint Denis
Gabriel PLASSAT	Ingénieur, Ademe, Sophia-Antipolis

Remerciements

Mes premiers remerciements sont adressés à mon directeur de thèse Pierre-Olivier Vandanjon et mes deux encadrants Pascal Gastineau et Pierre Hankach. La thèse n'est pas qu'une aventure scientifique. L'aspect humain occupe une place primordiale. Vous avez fait preuve de bienveillance à mon égard et je vous en suis reconnaissant. Vous étiez toujours disponibles lorsque j'avais besoin d'aide tant sur le plan professionnel et personnel. Vos retours et apports m'ont permis de progresser durant ces trois années de thèse.

Je remercie chaleureusement Juliette Rouchier et Guillaume Deffuant, qui ont accepté d'être les rapporteurs de cette thèse. Je tiens également à remercier Eric Cornélis, Yannick Aoustin et Nicolas Coulombel pour leur participation à mon jury de thèse comme examinateurs.

Ma thèse a bénéficié du financement de l'ADEME et de la SNCF mobilités TER Pays de la Loire. Grâce à ces deux partenaires, j'ai pu mener mes recherches dans de bonnes conditions. Merci particulièrement à Gabriel Plassat de l'ADEME, Jean Calio et Josée Milcendeau de la SNCF qui ont suivi mes travaux d'un œil attentif durant ces trois années.

Merci à Arnaud Can pour sa participation à mon comité de suivi de thèse et sa constante disponibilité.

J'ai été très bien accueilli au laboratoire EASE. Je tiens particulièrement à remercier Véronique Cérezo, l'ancienne directrice du laboratoire et Minh-Tan Do, le directeur actuel que j'ai tant sollicité pour les différentes démarches administratives de la thèse. J'ai

également apprécié les moments de convivialité avec mes collègues du laboratoire. Une mention particulière à Denis et Bogdan pour les discussions enrichissantes et les débats interminables. A mon collègue de bureau Trinh, qui m'a tant de fois entendu râler (et qui a fini par faire de même), je te souhaite du courage pour ta dernière année de thèse. Je me dois également de citer Cyrille, l'assistante du laboratoire pour sa bienveillance naturelle et son éternel sourire.

Merci également à mes collègues de l'UMRAE, Bill et Pierre pour les heures de covoiturage, les mots d'encouragement et les conseils prodigués.

Mes amies et amis, Carine, Nathalie, Carolle, Wiyao, Abdou merci pour tout.

Je souligne avec toute ma gratitude, le rôle primordial de ma famille dans l'aboutissement de ce travail. Merci maman, papa, petite soeur et mamie pour le soutien affectif en particulier durant certains moments difficiles que j'ai connus. Merci de m'avoir donné la force de continuer et d'avancer.

Enfin, Michèle Carine, mon amie, ma confidente et ma partenaire de tous les jours. Tu as su apaiser mes craintes, mon stress et me réconforter durant toutes ces années. Merci d'être là et de toujours me motiver.

Préambule

Cette thèse a été cofinancée par l'Agence de la transition écologique (ADEME) et la Société nationale des chemins de fer français (SNCF mobilités). Elle s'est déroulée au sein du laboratoire Ease (Environnement, Aménagement, Sécurité et Eco-conception) de l'université Gustave Eiffel. Un partenariat de recherche SNCF/Université Gustave Eiffel accompagne cette thèse sous la responsabilité scientifique de Jean Calio, expert« Data, Mobilité et Territoires et Développement Durable » et de Pierre-Olivier Vandanjon, chargé de recherche à l'université Gustave Eiffel. Cette thèse contribue à un projet plus global de la SNCF sur une nouvelle offre de mobilité innovante entre le quartier Doulon de Nantes et Carquefou dont la cheffe de projet est Josée Milcendeau. Coté ADEME, cette thèse a été suivie par l'ingénieur référent Gabriel Plassat.

Résumé

Ce travail s'articule autour de la thématique de la population synthétique, pierre angulaire de tout modèle multi-agents basé sur les activités. L'objectif de la thèse est de proposer des solutions permettant de pallier certaines insuffisances relevées dans la littérature sur ce sujet. Dans un premier temps, les principales méthodes de génération de population synthétique à deux niveaux (individus et ménages) sont présentées de façon détaillée. La recension de ces méthodes s'achève sur la proposition d'un arbre de décision qui permet de faire un choix raisonné entre les méthodes présentées. Une population synthétique d'individus répartis dans des ménages est générée dans un second temps en comparant différents algorithmes. Après génération de la population synthétique, une méthodologie d'attribution de caractéristiques supplémentaires à partir de données agrégées est développée. L'ajout d'une nouvelle caractéristique est formulé comme une maximisation de l'entropie en associant les attributs disponibles dans la population synthétique et les données agrégées. La validation de cette méthodologie a consisté à l'affectation d'un revenu à un nombre conséquent de ménages synthétiques. Enfin, une approche innovante de spatialisation d'une population synthétique à une échelle géographique plus fine est proposée. Le processus d'affectation est modélisé comme un problème de programmation quadratique mixte en nombres entiers. Bien qu'appliqués à des données françaises librement accessibles, la totalité des algorithmes implémentés restent génériques et transposables à d'autres sources de données.

Abstract

This work focuses on the topic of synthetic population, which serves as the cornerstone of any activity-based model. The objective of this thesis is to propose solutions that overcome some of the shortcomings found in the literature on this subject. First of all, the main methods for generating a two-layered synthetic population (individuals and households) are presented in detail. The review of these methods is completed by proposing a decision tree that allows for a reasoned choice between the various methods presented. A synthetic population of individuals distributed in households is generated during a second step by comparing the different algorithms. Afterwards, a methodology is developed for assigning additional characteristics from the aggregated data. The addition of a new feature is formulated as an entropy maximization problem by combining the attributes available in the synthetic population with the aggregated data. The methodology validation step consists of assigning an income to a large number of synthetic households. Lastly, an innovative approach to spatialize a synthetic population at a finer geographical scale is proposed. This assignment process is modeled as a mixed-integer quadratic programming problem. Although applied here to freely available French data, all the algorithms implemented remain generic and transposable to other data sources.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Contraintes associées aux transports dans les villes

Les villes sont confrontées à de nombreux enjeux associés aux systèmes de transport actuels : congestion des infrastructures de transport, retards constatés, utilisation de nouveaux services de mobilité. Sur le plan environnemental, ces enjeux ont déjà des conséquences importantes. Le changement climatique nous oblige à penser autrement les transports, secteur qui contribue fortement à l'émission de gaz à effet de serre. L'exposition aux nuisances sonores ou aux polluants atmosphériques générés par les moyens de transport constitue des facteurs de risque sanitaire pour les personnes concernées. En France par exemple, 48 000 décès par an pourraient être attribués à l'exposition chronique aux particules fines ($PM_{2.5}$) dont au moins la moitié dans les agglomérations de plus de 100 000 habitants (Pascal et al., 2016). Le bruit associé aux modes de transport est une des sources de stress et d'inconfort les plus répandues dans les zones urbaines (Dale et al., 2015). Au sein de la zone dense francilienne par exemple, le bruit des transports serait chaque année responsable de la perte de 107 766 années de vie en bonne santé, ce qui représente une perte de 10,7 mois par habitant en moyenne au cours d'une vie (BruitParif, 2019).

Agir sur les transports

Agir sur les transports et plus globalement sur les déplacements (aussi bien sur leur réalisation que sur le mode avec lequel ils sont effectués) apparaît donc comme un moyen de réduire à la fois les émissions de gaz à effet de serre, de polluants atmosphériques et les

nuisances sonores associées aux transports. Fort de ce constat, il devient donc indispensable de réinterroger les habitudes de déplacements et d'inventer de nouveaux modèles de mobilité, d'autant plus que l'écosystème du transport urbain est en pleine mutation. D'une part, de nouvelles offres émergent : trottinettes, autopartage, covoiturage, navette autonome ; d'autre part, les offres classiques se transforment : modernisation et extension des réseaux de transport en commun, service de transport à la demande, renouvellement des réseaux des équipements cyclables.

Evaluation des externalités environnementales

L'évaluation de ces offres est souvent très localisée avec des externalités parfois mal prises en compte (Emery et al., 2014 ; Kickhöfer and Nagel, 2016). A titre illustratif, l'extension d'une ligne de bus peut réduire certaines externalités comme la pollution et le bruit à des endroits localisés mais peut engendrer une augmentation de ces mêmes externalités à d'autres endroits. Une modélisation globale des déplacements urbains qui serait le support à ces évaluations systémiques et interdisciplinaires est donc indispensable.

L'approche actuelle de mesure des externalités environnementales consiste le plus souvent à estimer les niveaux d'exposition selon une modélisation statique qui repose sur un croisement entre la répartition spatiale des nuisances et les lieux de résidence. Le voisinage est alors utilisé dans de nombreuses situations comme aire d'exposition individuelle. Cette approche a toutefois quelques limites :

— elle ne tient pas compte des dynamiques temporelles de la mobilité [1] des individus et des niveaux de pollution des lieux traversés par ces derniers. Cela peut entraîner des erreurs dans le calcul des expositions moyennes (Park and Kwan, 2017) ;

— les expositions calculées sont statiques et s'appuient sur des indicateurs agrégés qui ne permettent pas de rendre compte de l'impact potentiel d'exposition brèves à des niveaux élevés.

Une modélisation adéquate de l'exposition au bruit ou à la pollution implique de considérer deux aspects essentiels à savoir la dynamique de mobilité des individus et le

1. La mobilité d'une personne correspond au nombre de déplacements réalisés par cette personne au cours d'une journée (Bonnel, 2002).

niveau d'agrégation retenu pour mesurer leurs comportements en matière de mobilité. Deux approches concurrentes de modélisation de la mobilité existent (Lemp et al., 2007) : la modélisation à quatre étapes et la modélisation multi-agents.

Première approche : la modélisation à quatre étapes

Les premiers modèles à quatre étapes ont été développés aux Etats-Unis à partir des années 1950 avec pour objectif de faire face aux problèmes liés à la modélisation intra urbaine de la congestion. Ces modèles servent à prédire le nombre de déplacements effectués dans une aire urbaine donnée par motif, en fonction de l'origine et de la destination, du mode de transport utilisé et des itinéraires empruntés sur le réseau (Masson, 1998) et à anticiper la construction d'infrastructures routières (Antoni, 2016). Dans un modèle à quatre étapes, un individu doit décider de son déplacement en quatre questions (Cerema, 2015) :

1. Faut-il effectuer le déplacement ?
2. Vers quelle destination ?
3. Par quel mode de transport ?
4. Par quel itinéraire ?

Chaque question correspond à une des quatre étapes du modèle. Ces étapes sont respectivement la génération, la distribution, la répartition modale et l'affectation.

L'étape de génération consiste à déterminer les volumes de déplacements émis et reçus au sein de chaque entité géographique retenue. Le calcul se fait en référence aux caractéristiques socioéconomiques des zones retenues et de la mobilité des habitants (CERTU, 2003). A la fin de cette étape, le nombre de déplacements est censé être estimé depuis et vers chaque zone de la région étudiée (Antoni, 2016). La sortie d'un modèle de génération fournit le nombre d'émissions (déplacements sortants) et d'attractions (déplacements entrants) dans chaque zone.

A l'issue de la génération, l'origine et la destination des déplacements ne sont pas reliées. L'étape de distribution permet de les relier pour produire la matrice origines-destinations des déplacements sur l'aire d'étude (Bonnel, 2002). La deuxième étape dis-

tribue donc spatialement les déplacements. Cette distribution spatiale peut se faire selon deux méthodes différentes (de Dios Ortuzar and Willumsen, 2011). La première méthode (la plus utilisée) consiste à calculer une matrice origines-destinations soit à partir d'une approche agrégée (on considère un individu moyen qui représente les voyageurs d'une zone) ou soit à partir d'une approche désagrégée dans laquelle on considère des individus moyens. Chaque individu moyen représentant un groupe d'individus selon des critères établis comme le revenu, le motif du déplacement ou même l'activité entreprise à la fin du déplacement. La deuxième méthode consiste à considérer les motifs qui génèrent et attirent les déplacements en calculant une matrice émissions-attractions.

Une fois le motif fixé et le déplacement décidé, l'individu peut dans de nombreux cas choisir le mode de transport qu'il souhaite utiliser (Audard, 2006). Le choix du mode de transport est probablement l'une des étapes les plus importantes du modèle car il affecte l'efficacité générale du déplacement, surtout dans le milieu urbain (de Dios Ortuzar and Willumsen, 2011). La répartition modale (troisième étape) consiste à éclater la(les) matrice(s) origines-destinations en autant de matrices origines-destinations pour chaque mode pris en compte (Bonnel, 2002).

L'affectation, quatrième et dernière étape, procède à la répartition du trafic entre les différents modes de transport sur les réseaux essentiellement viaires et de transport en commun à partir des matrices origines-destinations par mode obtenues dans l'étape précédente. Les sorties de cette étape sont les réseaux chargés avec pour chaque arc le volume de véhicules ou de voyageurs, la vitesse et le temps de parcours (Cerema, 2015) et permettent de calculer les matrices origines-destinations de coûts généralisés [2] par mode.

Seconde approche : la modélisation multi-agents

La théorie qui sous-tend les modèles multi-agents également appelés systèmes multi-agents, est basée sur Hägerstrand et son concept d'espace-temps (Zhong et al., 2015).

2. Le coût généralisé essaye d'intégrer l'ensemble des coûts réellement supportés par l'utilisateur d'un mode de transport. Il va au delà du prix du carburant pour l'automobile ou du titre de transport pour les transports en commun et considère par exemple le coût d'amortissement des véhicules, l'assurance, le temps passé en déplacement, l'inconfort ou les retards (Antoni et al., 2016).

Hägerstrand cherchait à appréhender les individus en tant qu'acteurs et à comprendre comment ils parviennent à réaliser leurs projets respectifs, suivant leurs intentions, leurs devoirs, et différents niveaux de contraintes (Sanders, 2007).

Les systèmes multi-agents (SMA) sont issus de l'intelligence artificielle distribuée (Gilbert and Troitzsch, 2005) et ont connu un essor aux Etats-Unis à partir des années 1990. Appliqués à de nombreuses disciplines telles que le transport, l'économie, la géographie ou la démographie, ils permettent de formaliser des situations complexes comportant des échelles multiples (Livet et al., 2014).

Un SMA est un ensemble d'agents qui partagent un environnement [3] commun (Michel, 2004), qui communiquent et collaborent pour achever des objectifs spécifiques personnels ou collectifs. Les agents sont selon Woodridge and Jennings (1995), des entités physiques ou virtuelles autonomes, capables de percevoir leur environnement, d'y réagir, d'interagir avec d'autres agents et d'adopter un comportement pour atteindre leurs objectifs. Dans l'analyse des mobilités, les modèles multi-agents considèrent chaque voyageur comme un agent. Ces modèles impliquent un grand nombre d'individus en interaction, avec des caractéristiques, des règles de comportement, des sources d'information différentes (Young, 2006 ; Kiesling et al., 2012).

Comparaison des deux approches

Des études ont comparé les performances des modèles multi-agents face aux modèles à quatre étapes (Walker, 2005 ; Vuk and Petersen, 2006 ; Griesenbeck and Garry, 2007 ; Lemp et al., 2007 ; Zhong et al., 2015 ; Kagho et al., 2020). Plus simples à implémenter, les modèles à quatre étapes tendent vers une approche pragmatique qui réduit les comportements de déplacements dans des composantes analytiques maniables et qui peuvent être traités avec des techniques relativement simples (Masson, 1998). Ils permettent une représentation globale des flux de trafic et demeurent un outil important de planification (Antoni, 2016). En revanche, les modèles à quatre étapes ne rendent pas suffisamment

3. Un environnement est un espace disposant généralement d'une métrique. La notion d'environnement dépend de l'objet de modélisation. Si les agents sont des individus, l'une des principales fonctions de l'environnement sera de fournir un contexte spatial (Gilbert and Troitzsch, 2005).

compte du comportement des individus et de leur mécanisme de choix en matière de mobilité (Rasouli and Timmermans, 2014 ; Kagho et al., 2020). Les modèles multi-agents présentent l'avantage de pallier cette limite des modèles à quatre étapes et de simuler des interactions dynamiques entre les individus dans leurs différents déplacements. Ces modèles sont toutefois assez complexes à mettre en œuvre car ils requièrent notamment des données détaillées sur les attributs (caractéristiques) des individus (données qui sont souvent indisponibles) et nécessitent un temps de calcul conséquent. Chaque approche possède des atouts mais également des inconvénients. Le choix d'une approche dépend davantage de l'objectif de la modélisation.

Quelle approche privilégier pour l'évaluation des externalités environnementales liées aux transports ?

Dans l'analyse des externalités associées aux transports, il est nécessaire de prendre en compte deux aspects essentiels de la mobilité.

1. Les comportements des individus. Sur une journée, les individus mènent différentes activités. Ils se rendent par exemple à l'école, au travail, vont chez le médecin, pratiquent des activités sportives...
2. Les mécanismes de choix des individus. Les individus planifient leurs activités et font des choix en tenant souvent compte des autres individus et de leur environnement. Ces choix peuvent être notamment relatifs au(x) mode(s) de transport à utiliser, à l'itinéraire à emprunter (pour éviter par exemple les embouteillages), au temps à consacrer pour une activité...

La vie des individus correspond à une succession d'activités spatialement localisées sur une période de temps, le plus souvent une journée. Par conséquent, le besoin de se déplacer trouve son origine dans la nécessité de mener une activité. La mobilité ne doit alors plus être appréhendée uniquement en termes de déplacements journaliers mais dans une optique de réalisation d'activités. Les individus peuvent être différemment exposés au bruit ou à la pollution selon l'activité menée, le lieu de cette activité et le moment de la journée dans lequel l'activité se déroule.

Les systèmes multi-agents constituent donc la meilleure approche pour mesurer cette exposition dynamique des individus aux externalités environnementales (bruit ou pollution) durant leur mobilité. La modélisation des mobilités est dans ce cas appréhendée par une classe spécifique de modèles multi-agents désignée sous le terme de modèles multi-agents basés sur les activités.

L'objectif fondamental des modèles multi-agents basés sur les activités est de prédire les séquences d'activités et de déplacements associées à tous les individus d'un ménage sous réserve notamment d'un ensemble de contraintes spatiales, temporelles et budgétaires (Rasouli and Timmermans, 2014). En simulant les séquences d'activités des individus d'une population, ces modèles fournissent des informations sur pourquoi, quand et où les gens se déplacent (Arentze and Timmermans, 2004).

Ces modèles restent toutefois des représentations, souvent simplifiées et moins détaillées de la réalité notamment à cause des contraintes rencontrées dans la modélisation (indisponibilité des données, formulation des relations, temps de calcul...). Malgré tout, ces modèles qui reposent sur les schémas d'activités des individus fournissent des informations pertinentes pour l'analyse de l'exposition des individus aux nuisances liées au transport (Shiftan, 2000 ; Shiftan et al., 2015).

Les principes des modèles multi-agents basés sur les activités

Les modèles multi-agents basés sur les activités ont fait l'objet d'une littérature scientifique abondante et ont été largement vulgarisés grâce à des plateformes logiciels telles que TRANSIMS (Dixon et al., 2007), MATSim (Horni et al., 2016), MobiSim (Antoni et al., 2016), SimMobility (Adnan et al., 2016) ou POLARIS (Auld et al., 2016). Il est ainsi possible de modéliser les comportements de mobilité dans différents contextes et à l'échelle d'un pays entier comme Singapour (Erath et al., 2012) et la Suisse (Bösch et al., 2016) ou de grandes agglomérations urbaines telles que Rouen (Vosooghi et al., 2019), Nantes (Le Bescond et al., 2021), Santiago de Chili (Kickhofer et al., 2016), Los Angeles et San Francisco (Balac, 2020).

Ces modèles impliquent une modélisation détaillée des comportements de mobilité

dans lesquels les individus interagissent entre eux et avec l'environnement dans le temps et l'espace. Les comportements de mobilité dépendent à la fois des attributs propres de l'individu et de sa situation familiale (Loo and Lam, 2013 ; Kalter and Geurs, 2016 ; Fournier et al., 2020). L'estimation de ces comportements est alors un processus qui se déroule en deux étapes principales :

1. Une détermination des attributs sociodémographiques et économiques de chaque individu du territoire étudié (sexe, âge, profession...) et de son ménage (taille du ménage, structure familiale, revenu...) ainsi qu'une localisation résidentielle précise ;
2. Une affectation à chaque individu d'un programme d'activités (types d'activités, horaires, localisation, modes de transport, itinéraires...). Ce programme décrit toute l'activité journalière de l'agent et dépend de ses attributs sociodémographiques et économiques, de la localisation de son domicile et de ses lieux d'activités.

L'utilisation d'un modèle multi-agents basé sur les activités nécessite donc une description détaillée des individus (pour appréhender les comportements de mobilité) à une échelle fine (plus la localisation des individus est précise, meilleure est l'estimation des comportements de mobilité). Toutefois, plus les données sont fines, plus la description des individus relève de la vie privée (Pivano, 2016).

Les problématiques de production de données précises

La production d'attributs détaillés pour les individus et les ménages constitue une pierre d'achoppement au niveau de la littérature. En effet, pour des raisons de coût et surtout de confidentialité, il n'existe pas de sources de données qui contiennent les attributs des individus et des ménages à une échelle géographique fine. L'unique alternative consiste à générer une population artificielle dite « population synthétique », représentative de la population réelle du territoire étudié. Au cours du processus de génération, les attributs de l'ensemble des individus et de l'ensemble des ménages d'un territoire sont déduits à partir des données disponibles, généralement composées d'un échantillon

d'individus et de ménages et de données agrégées du territoire.

Il existe des logiciels commerciaux ou open source qui permettent de générer automatiquement une population synthétique d'individus rattachés à leur ménage : PopSynIII (Vovsha et al., 2015), PopulationSim (Paul et al., 2018), TransCAD (Balakrishnaa et al., 2020)... De nombreuses études (Erath et al., 2012 ; Bösch et al., 2016 ; Kickhofer et al., 2016 ; Kamel et al., 2019 ; Ziemke et al., 2019 ; Delhoum et al., 2020 ; He et al., 2020 ; Moeckel et al., 2020)... ont également proposé différentes méthodes de génération de population synthétique. Le constat qui ressort pour la majorité de ces études est :

— une absence de données et de méthodologie claire qui ne permettent pas de tester les méthodes proposées et de les adapter selon les besoins ;

— une affectation aléatoire de certaines attributs sociodémographiques ;

— une localisation (allocation spatiale des résidences) de la population synthétique qui s'effectue à une échelle géographique qui n'offre pas une précision spatiale suffisante ;

— une utilisation de données qui ne sont pas librement accessibles.

Hörl and Balac (2021) ont créé une plateforme dénommée « Eqasim » qui offre une solution simple et rapide de génération d'une population de ménages et d'individus synthétiques assortis de leurs plans d'activités. A notre connaissance, il s'agit du premier exemple d'une plateforme de génération de population synthétique qui à la fois utilise des données librement accessibles et permet de tester des méthodes. Certains aspects méthodologiques [4] développés dans la plateforme nécessitent toutefois d'être approfondis. Cela concerne notamment :

— les méthodes de génération utilisées ;

— l'ajout d'attributs supplémentaires.

Ma thèse s'articule autour de ce constat. Je propose des solutions méthodologiques et appliquées qui permettent à terme de repousser certaines limites évoquées dans la géné-

4. Ces différentes aspects seront précisés dans le premier chapitre de la thèse.

ration de la population synthétique. Disposer d'une population synthétique d' individus et de ménages à une échelle spatiale fine permet par exemple d'évaluer :

— l'impact en termes de bruit ou d'émissions de polluants atmosphériques d'une nouvelle infrastructure de transport comme une nouvelle ligne de tramway ;

— une offre de mobilité innovante (véhicules autonomes) ou différents scénarios de comportement de mobilité.

Il est également possible de mener des analyses économiques (évaluation de politiques de transport), sociales (études des inégalités) et sanitaires (Kaddoura et al., 2017 ; Gurram et al., 2019 ; Kuehnel et al., 2021 ; Kocak et al., 2021).

Contributions de la thèse

La première contribution concerne la présentation et l'évaluation des principales méthodes de génération de population synthétique. Le choix d'une méthode de génération peut être délicat car conditionné aux données disponibles et aux attributs de la population à générer. De nombreuses professionnels utilisent le plus souvent les logiciels commerciaux à disposition ou les méthodes développées sans être nécessairement conscients de leur adéquation à leur cas d'étude (données disponibles, taille de la population à générer, taille de l'échantillon. . ..). Il ressort qu'il n'existe pas de méthodologie claire qui permette de privilégier une méthode au détriment des autres. Une comparaison détaillée des méthodes de génération est effectuée ; une procédure de décision traduite par un arbre de décision est également proposée. L'arbre permet de choisir la méthode adéquate en fonction des données disponibles. A travers cette étude détaillée des méthodes de génération, les chercheurs et les professionnels ont maintenant accès à un cadre standardisé et complet pour sélectionner la méthode appropriée en fonction de nombreux critères et de leurs objectifs de modélisation. Le travail de documentation des méthodes de génération de population synthétique à fait l'objet d'une valorisation scientifique par un article (Yaméogo et al., 2021) publié dans la revue scientifique Transportation Research Record (TRR) intitulé « Comparing methods for generating a two-layered synthetic population ».

La deuxième contribution est relative à la génération effective d'une population synthétique d'individus et de ménages sur un cas d'application, l'aire urbaine de Nantes. Quatre algorithmes de génération de population synthétique d'individus et de ménages sont analysés :

1. L'Iterative Proportional Update (Ye et al., 2009) ;
2. Le Hierarchical Iterative Proportional Fitting (Müller and Axhausen, 2011 ; Müller, 2017) ;
3. La minimisation de l'entropie (Bar-Gera et al., 2009 ; Lee and Fu, 2011) ;
4. Le calage sur marges (Deville et al., 1993).

Une comparaison de ces quatre algorithmes est réalisée dans un cadre conceptuel commun avec une harmonisation des notations et une description détaillée de chaque algorithme. Ce cadre commun facilite la compréhension de ces algorithmes et permet de mieux les positionner par rapport aux autres algorithmes ou méthodes qui existent. A partir des données du recensement français, une population synthétique d'individus et de ménages, représentative de l'aire urbaine de Nantes est générée avec chaque algorithme. Les résultats obtenus sont comparés à l'aide d'indicateurs de validation usuels. La diffusion des fichiers du recensement suit un format standard à l'échelle de la France métropolitaine. Il est donc possible de généraliser et d'adapter la démarche méthodologique de génération à d'autres zones géographiques françaises. L'ensemble du travail effectué a été évalué par les pairs sous la forme d'un article scientifique (Yameogo et al., 2021), publié dans la revue scientifique Journal of Artificial Societies and Social Simulation (JASSS) intitulé « Generating a two-layered synthetic population for French municipalities : Results and evaluation of four synthetic reconstruction methods ».

Une fois la population synthétique générée, une procédure habituelle consiste à rajouter des attributs supplémentaires aux individus ou aux ménages synthétiques à partir de sources de données externes. La problématique de l'ajout d'un attribut à une population synthétique à partir de données agrégées a toutefois reçu peu d'attention dans la littérature.

Une méthodologie générale et efficace qui permet de répondre à ce besoin pratique est proposée. La méthodologie intègre trois étapes distinctes dont la première modélise théoriquement le problème comme une distribution multinomiale. L'ajout d'un nouvel attribut est ensuite formulé comme une maximisation de l'entropie en utilisant les attributs disponibles dans la population synthétique et les données agrégées. La résolution de ce problème (dans ce cas d'application) n'est pas possible en raison du grand nombre de contraintes impliquées. La deuxième étape présente alors une heuristique apportant une solution pratique au problème. Cette heuristique combine le théorème de Bayes avec l'algorithme de minimisation de l'entropie croisée. En raison du nombre élevé de paramètres, certains des résultats obtenus s'avèrent non cohérents. Pour remédier à ce problème, une méthode de post-traitement est appliquée au cours d'une troisième étape pour garantir la cohérence des résultats. La méthodologie proposée est décrite en détail et appliquée à un cas réel : un niveau de vie (revenu) est affecté à chacun des 157 000 ménages synthétiques de la commune de Nantes sur la base des données agrégées du dispositif « Fichier Localisé Social et Fiscal »(FiLoSoFi). Le niveau de vie des ménages constitue un attribut important dans la prise en compte de nombreux aspects sociaux et économiques (pouvoir d'achat des ménages, politique de redistribution, politique fiscale, etc.). Le travail effectué a été soumis sous la forme d'un article intitulé « Methodology for Adding a Variable to a Synthetic Population from Aggregate Data : Example of the Income Variable » à la revue Journal of Artificial Societies and Social Simulation (JASSS) et est en cours de révision.

Une simulation multi-agents réaliste implique de pouvoir prendre en compte la localisation spatiale des agents d'une population synthétique. Suivant les besoins de la simulation, l'échelle géographique de localisation varie selon différents niveaux de précision : ville, commune, quartiers, bâtiments... Dans de nombreux cas, les lieux de résidence ou de travail des ménages synthétiques sont tirés au sort parmi tous les emplacements disponibles.

Une méthodologie d'affectation de ménages synthétiques d'une échelle spatiale supérieure à une échelle spatiale plus fine est développée. Cette méthodologie est décrite

en détail avec l'utilisation de données françaises issues du recensement de la population et des données fiscales du dispositif « Fichier Localisé Social et Fiscal »(FiLoSoFi). L'affectation des ménages synthétiques est alors modélisé comme un problème de programmation quadratique mixte en nombres entiers (MIQP problème en anglais). Il est démontré que l'applicabilité de cet algorithme est limitée aux populations synthétiques de petites tailles. Une heuristique est alors proposée pour tenir compte de populations synthétiques de taille plus importante. Les tests effectués montrent que l'heuristique donne des solutions quasi optimales en un temps de calcul rapide. La méthodologie considérée et l'heuristique proposée sont générales et permettent de répondre aux besoins des utilisateurs. Ce travail a été soumis à la revue scientifique Environment and Planning B : Urban Analytics and City Science, sous la forme d'un article intitulé « Allocating synthetic population to a finer spatial scale : a mixed-integer quadratic programming formulation » et est également en cours de révision.

Bien qu'appliquées à des données françaises, les méthodologies mises en œuvre restent génériques et transposables à d'autres contextes. Les sources de données exploitées sont librement accessibles. Elles sont clairement présentées et les liens permettant de les télécharger sont indiqués avec précision. Une attention particulière a été accordée au traitement des données afin de faciliter l'appropriation et l'adaptation des méthodologies. Les algorithmes développés dans la thèse ont été implémentés avec le langage R (logiciel libre) et les différents scripts associés seront mis à la disposition des utilisateurs.

Structure du document

Chapitre 1 : Modèles multi-agents basés sur les activités : présentation et intégration dans l'évaluation des externalités environnementales associées aux transports

Dans la première partie de ce chapitre les modèles de simulation multi-agents basés sur les activités sont présentés. Trois plateformes logiciel de cette classe de modèles (TRANSIMS, MATSim et SimMobility) sont également décrites. Dans la deuxième partie, l'intégration des modèles multi-agents basés sur les activités dans la prise en compte de l'exposition dynamique des individus aux externalités environnementales est illustrée.

Dans la troisième et dernière partie, la nécessité de disposer de données d'entrée précises est soulignée.

Chapitre 2 : Comparaison des méthodes de génération d'une population synthétique à deux niveaux

Le chapitre 2 dresse un état des lieux et passe en revue les principales méthodes de génération de population synthétique disponibles. Une description des différentes méthodes et des approches associées est proposée dans un premier temps. Par la suite, une comparaison de ces méthodes est établie. Cette comparaison illustre les avantages et les inconvénients d'utilisation de chaque méthode. Afin de faciliter le choix d'une méthode, un outil de décision (sous la forme d'un arbre de décision) est proposé à la fin du chapitre.

Chapitre 3 : Comparaison et application de méthodes de reconstruction synthétique pour générer une population synthétique à deux niveaux

Ce chapitre décrit la génération d'une population synthétique de ménages et d'individus sur un cas d'application français. Dans les deux premières parties, sont respectivement présentées le recensement français de la population et le cas d'étude. La troisième partie détaille le processus de génération et de validation d'une population synthétique. La quatrième partie résume et discute les résultats obtenus.

Chapitre 4 : Méthodologie d'ajout d'un attribut à une population synthétique à partir de données agrégées : exemple de l'attribut de revenu

Dans ce chapitre, la problématique de l'ajout d'un attribut à une population synthétique à partir de données agrégées est analysée. Dans un premier temps, les données sont présentées. Dans un second temps, le problème est mathématiquement formulé et les approches de résolution sont présentées. Un bilan des résultats obtenus est dressé à la fin du chapitre.

Chapitre 5 : Affectation d'une population synthétique à une échelle spatiale plus fine : résolution par une programmation mixte en nombres entiers

Le dernier chapitre de la thèse est consacré à la modélisation de l'affectation de ménages synthétiques à une zone spatiale plus fine. Les deux premières parties décrivent

le problème et la configuration des données. La troisième partie procède à la formulation mathématique du problème et introduit les variables de décision et les paramètres. La quatrième partie est consacrée à la présentation des approches de résolution du problème et des résultats obtenus. La dernière partie discute les résultats des analyses.

A la suite des chapitres, la conclusion générale sera l'occasion de faire le bilan des différents chapitres, de formuler des pistes d'amélioration et des perspectives que ce travail de thèse permet d'envisager.

Références bibliographiques

Adnan, M., Pereira, F. C., Azevedo, C. L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., and Ben-Akiva, M. (2016). SimMobility : A Multi-scale Integrated Agent-Based Simulation Platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*.

Antoni, J.-P. (2016). *Concepts, méthodes et modèles pour l'aménagement et les mobilités : l'aide à la décision face à la transition éco-énergétique*. Economica.

Antoni, J.-P., Lunardi, N., and Vuidel, G. (2016). Simuler les mobilités individuelles-les enjeux de l'information géographique. *Revue internationale de géomatique*, 26(2) :237–262.

Arentze, T. A. and Timmermans, H. J. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B : Methodological*, 38(7) :613–633.

Audard, F. (2006). *Modélisation de la mobilité : la génération de trafic à l'échelle régionale*. PhD thesis, Université de Franche-Comté.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., and Zhang, K. (2016). Polaris : Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C : Emerging Technologies*, 64 :101–116.

Balac, M. (2020). Agent-based scenarios of los angeles and san francisco. Technical report, Working paper. IVT, ETH Zurich, Zurich.

Balakrishnaa, R., Sundarama, S., and Lam, J. (2020). An enhanced and efficient population synthesis approach to support advanced travel demand models. In *99th Annual Meeting of the Transportation Research Board, Washington, DC*.

Bar-Gera, H., Konduri, K., Sana, B., Ye, X., and Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Bonnel, P. (2002). *Prévision de la demande de transport*. Habilitation à diriger des recherches, Université Lumière-Lyon II.

Bösch, P. M., Müller, K., and Ciari, F. (2016). The ivt 2015 baseline scenario. In *16th Swiss Transport Research Conference (STRC 2016)*. 16th Swiss Transport Research Conference (STRC 2016).

BruitParif (2019). Impacts sanitaires du bruit des transports dans la zone dense de la région ile-de-france.

Cerema (2015). *(Centre d'études et d'expertise sur les risques, l'environnement,la mobilité et l'aménagement). Etudes de simulation dynamique de trafic. Guide de réalisation*. Editions du Cerema,Collection Références.

CERTU (2003). *(Centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques). Modélisation des déplacements urbains de voyageurs : guide des pratiques*. CERTU, Collection Références.

Dale, L. M., Goudreau, S., Perron, S., Ragettli, M. S., Hatzopoulou, M., and Smargiassi, A. (2015). Socioeconomic status and environmental noise exposure in montreal, canada. *BMC public health*, 15(1) :205.

de Dios Ortuzar, J. and Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons.

Delhoum, Y., Belaroussi, R., Dupin, F., and Zargayouna, M. (2020). Activity-based demand modeling for a future urban district. *Sustainability*, 12(14) :5821.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423) :1013–1020.

Dixon, M., Chang, K., Keecheril, S., and Orton, B. (2007). Applying the transims modeling paradigm to the simulation and analysis of transportation and traffic control systems.

Emery, J., Marilleau, N., Thevenin, T., and Martiny, N. (2014). Du comptage ponctuel à laffectation par simulation multi-agents : Application à la circulation routière de la ville de Dijon. In *Conférence internationale de géomatique et d'analyse spatiale (SAGEO)*.

Erath, A., Fourie, P. J., van Eggermond, M. A., Ordonez Medina, S. A., Chakirov, A., and Axhausen, K. W. (2012). Large-scale agent-based transport demand model for singapore. *Arbeitsberichte Verkehrs-und Raumplanung*, 790.

Fournier, N., Christofa, E., Akkinepally, A. P., and Azevedo, C. L. (2020). Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, pages 1–27.

Gilbert, N. and Troitzsch, K. (2005). *Simulation for the social scientist*. McGraw-Hill Education (UK).

Griesenbeck, B. and Garry, G. (2007). Comparison of activity-based tour model to four-step model as a tool for metropolitan transportation planning. In *Proceedings of the National Transportation Planning Applications Conference, May*, pages 6–10.

Gurram, S., Stuart, A. L., and Pinjari, A. R. (2019). Agent-based modeling to estimate exposures to urban air pollution from transportation : Exposure disparities and impacts of high-resolution data. *Computers, Environment and Urban Systems*, 75 :22–34.

He, B. Y., Zhou, J., Ma, Z., Chow, J. Y., and Ozbay, K. (2020). Evaluation of city-scale built environment policies in new york city with an emerging-mobility-accessible synthetic population. *Transportation Research Part A : Policy and Practice*, 141 :444–467.

Hörl, S. and Balac, M. (2021). Synthetic population and travel demand for paris and île-de-france based on open and publicly available data. *Transportation Research Part C : Emerging Technologies*, 130 :103291.

Horni, A., Nagel, K., and Axhausen, K. W. (2016). Introducing matsim. In Horni, A., Nagel, K., and Axhausen, K. W., editors, *The Multi-Agent Transport Simulation MATSim*, pages 3–7. Ubiquity Press.

Kaddoura, I., Kröger, L., and Nagel, K. (2017). User-specific and dynamic internalization of road traffic noise exposures. *Networks and Spatial Economics*, 17(1) :153–172.

Kagho, G. O., Balac, M., and Axhausen, K. W. (2020). Agent-based models in transport planning : Current state, issues, and expectations. *Procedia Computer Science*, 170 :726–732.

Kalter, M.-J. O. and Geurs, K. T. (2016). Exploring the impact of household interactions on car use for home-based tours : a multilevel analysis of mode choice using data from the first two waves of the netherlands mobility panel. *European journal of transport and infrastructure research*, 16(4) :698–712.

Kamel, J., Vosooghi, R., Puchinger, J., Ksontini, F., and Sirin, G. (2019). Exploring the impact of user preferences on shared autonomous vehicle modal split : A multi-agent simulation approach. *Transportation Research Procedia*, 37 :115–122.

Kickhofer, B., Hosse, D., Turnera, K., and Tirachinic, A. (2016). Creating an open matsim scenario from open data : The case of santiago de chile. *http ://www. vsp. tuberline. de/publication : TU Berlin, Transport System Planning and Transport Telematics.*

Kickhöfer, B. and Nagel, K. (2016). Towards high-resolution first-best air pollution tolls. *Networks and Spatial Economics*, 16(1) :175–198.

Kiesling, E., Günther, M., Stummer, C., and Wakolbinger, L. M. (2012). Agent-based simulation of innovation diffusion : a review. *Central European Journal of Operations Research*, 20(2) :183–230.

Kocak, T. K., Gurram, S., Bertini, R. L., and Stuart, A. L. (2021). Impacts of a metropolitan-scale freeway expansion program on air pollution and equity. *Journal of Transport & Health*, 22 :101114.

Kuehnel, N., Huang, W.-C., and Moeckel, R. (2021). Environmental equity analysis in agent-based transport simulations : A study on causation and exposure. *Procedia Computer Science*, 184 :650–655.

Le Bescond, V., Can, A., Aumond, P., and Gastineau, P. (2021). Open-source modeling chain for the dynamic assessment of road traffic noise exposure. *Transportation Research Part D : Transport and Environment*, 94 :102793.

Lee, D.-H. and Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record*, 2255(1) :20–27.

Lemp, J. D., McWethy, L. B., and Kockelman, K. M. (2007). From aggregate methods to microsimulation : Assessing benefits of microscopic activity-based models of travel demand. *Transportation Research Record*, 1994(1) :80–88.

Livet, P., Phan, D., and Sanders, L. (2014). Diversité et complémentarité des modèles multi-agents en sciences sociales. *Revue francaise de sociologie*, 55(4) :689–729.

Loo, B. P. and Lam, W. (2013). A multilevel investigation of differential individual mobility of working couples with children : a case study of hong kong. *Transportmetrica A : Transport Science*, 9(7) :629–652.

Masson, S. (1998). Interactions entre système de transport et système de localisation. de l'héritage des modèles traditionnels à l'apport des modèles interactifs de transport et d'occupation des sols. *Les cahiers scientifiques du transport*, (33) :79–108.

Michel, F. (2004). *Formalisme, outils et éléments méthodologiques pour la modélisation et la simulation multi-agents*. PhD thesis, Montpellier II.

Moeckel, R., Kuehnel, N., Llorca, C., Moreno, A. T., and Rayaprolu, H. (2020). Agent-based simulation to improve policy sensitivity of trip-based models. *Journal of Advanced Transportation*, 2020.

Müller, K. (2017). *A generalized approach to population synthesis*. PhD thesis, ETH Zurich.

Müller, K. and Axhausen, K. W. (2011). Hierarchical ipf : Generating a synthetic population for switzerland. *paper presented at the 51st Congress of the European Regional Science Association*.

Park, Y. M. and Kwan, M.-P. (2017). Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health & place*, 43 :85–94.

Pascal, M., de Crouy Chanel, P., Corso, M., Medina, S., Wagner, V., Goria, S., Beaudeau, P., Bentayeb, M., Le Tertre, A., Ung, A., et al. (2016). Impacts de l'exposition chronique aux particules fines sur la mortalité en france continentale et analyse des gains en santé de plusieurs scénarios de réduction de la pollution atmosphérique. *Saint-Maurice : Santé publique France*.

Paul, B. M., Doyle, J., Stabler, B., Freedman, J., and Bettinardi, A. (2018). Multi-level population synthesis using entropy maximization-based simultaneous list balancing. In *97th Annual Meeting of the Transportation Research Board, Washington, DC*.

Pivano, C. (2016). *Désagrégation spatiale des données de mobilité du recensement de la population appliquée à l'Ile-de-France*. PhD thesis, Paris Est.

Rasouli, S. and Timmermans, H. (2014). Activity-based models of travel demand : promises, progress and prospects. *International Journal of Urban Sciences*, 18(1) :31–60.

Sanders, L. (2007). Objets géographiques et simulation agent, entre thématique et méthodologie. *Revue internationale de géomatique*, 17(2) :p135–160.

Shiftan, Y. (2000). The advantage of activity-based modelling for air-quality purposes : theory vs practice and future needs. *Innovation : The European Journal of Social Science Research*, 13(1) :95–110.

Shiftan, Y., Kheifits, L., and Sorani, M. (2015). Travel and emissions analysis of sustainable transportation policies with activity-based modeling. *Transportation Research Record*, 2531(1) :93–102.

Vosooghi, R., Kamel, J., Puchinger, J., Leblond, V., and Jankovic, M. (2019). Robotaxi service fleet sizing : assessing the impact of user trust and willingness-to-use. *Transportation*, 46(6) :1997–2015.

Vovsha, P., Hicks, J. E., Paul, B. M., Livshits, V., Maneva, P., and Jeon, K. (2015). New features of population synthesis. In *94th Annual Meeting on Transportation Research Board, Washington, DC*.

Vuk, G. and Petersen, E. (2006). Comparing a conventional travel demand model to an activity-based travel demand model : A case study of copenhagen. In *Proceedings of the European Transport Conference*.

Walker, J. L. (2005). Making household microsimulation of travel and activities accessible to planners. *Transportation research record*, 1931(1) :38–48.

Woodridge, M. and Jennings, N. R. (1995). Intelligent agents : Theory and practice. *The Knowledge Engineering Review*, 10(02) :115–121.

Yaméogo, B. F., Gastineau, P., Hankach, P., and Vandanjon, P.-O. (2021). Comparing methods for generating a two-layered synthetic population. *Transportation research record*, 2675(1) :136–147.

Yameogo, B. F., Vandanjon, P. O., Gastineau, P., and Hankach, P. (2021). Generating a two-layered synthetic population for french municipalities : Results and evaluation of four synthetic reconstruction methods. *JASSS-Journal of Artificial Societies and Social Simulation*, 24 :27p.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., and Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of

synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Young, H. P. (2006). Social dynamics : Theory and applications. *Handbook of computational economics*, 2 :1081–1108.

Zhong, M., Shan, R., Du, D., and Lu, C. (2015). A comparative analysis of traditional four-step and activity-based travel demand modeling : a case study of tampa, florida. *Transportation Planning and Technology*, 38(5) :517–533.

Ziemke, D., Kaddoura, I., and Nagel, K. (2019). The matsim open berlin scenario : A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data. *Procedia computer science*, 151 :870–877.

Chapitre 1

Modèles multi-agents basés sur les activités : présentation et intégration dans l'évaluation des externalités environnementales associées aux transports

Sommaire

Résumé

Les modèles multi-agents constituent un outil approprié pour modéliser les interactions complexes entre les individus et leur environnement dans leurs déplacements quotidiens. Ce chapitre est consacré aux modèles multi-agents basés sur les activités. Ces modèles désignent une classe de modèles multi-agents particulièrement adaptés pour la prise en compte les mobilités des individus. L'approche multi-agents basée sur les activités présente chaque individu comme un agent autonome durant les étapes du processus de modélisation des mobilités. Trois plateformes multi-agents basées sur les activités mentionnées dans la littérature sont décrites : TRANSIMS, MATSim et SimMobility. Les possibilités d'association des modèles multi-agents basés sur les activités avec des modèles de bruit ou d'émissions pour l'évaluation de l'exposition dynamique des individus sont également présentées. Dans leur application, ces modèles nécessitent comme données d'entrée, des informations détaillées relatives aux attributs démographiques, sociaux et économiques d'un nombre important d'individus et de leurs ménages. L'indisponibilité de ces informations requiert l'utilisation d'une population synthétique.

Mots clés

TRANSIMS, MATSim, SimMobility, externalités environnementales, population synthétique

Introduction

Le paradigme multi-agents permet de modéliser des comportements complexes pour une multitude d'agents dans le temps et dans un environnement précis. Les modèles multi-agents apportent une réponse adéquate à l'analyse et à la résolution de problèmes qui impliquent les dimensions spatiale, sociale, organisationnelle, politique, économique et financière (Antoni, 2010). Ces modèles ont gagné en popularité depuis les années 1990 et sont désormais appliqués dans de nombreux domaines tels que la santé (Tomintz et al., 2008 ; Edwards and Clarke, 2013), l'évaluation des politiques économiques (Avram et al., 2013 ; Sutherland and Figari, 2013), la géographie (O'Sullivan, 2008).

Dans le domaine des transports, les déplacements humains résultent du besoin de réaliser des activités (travail, école, santé, loisirs...). La modélisation des mobilités est alors appréhendée par une classe spécifique des modèles multi-agents désignée sous le terme de modèles multi-agents basés sur les activités. Ces modèles analysent le comportement des individus lorsqu'ils interagissent avec d'autres individus dans la réalisation de leurs activités quotidiennes ainsi que les implications de ces interactions dans le réseau de transport.

L'approche multi-agents basée sur les activités présente chaque individu comme un agent autonome durant toutes les étapes du processus de modélisation des mobilités (Figure 1.1). La première étape consiste à générer une population d'agents pour le territoire étudié. L'objectif de cette étape est de créer une base de données recensant les agents, auxquels sont attribués des attributs et éventuellement leur lieu de résidence et des lieux d'activités.

La deuxième étape consiste à générer pour chaque agent un plan d'activités (types d'activités, horaires, localisation, modes de transport, itinéraires...). Ce plan décrit toute l'activité journalière de l'individu et dépend donc des attributs de l'individu, de la localisation de son domicile et de ses lieux d'activités (emploi, études, autres). Ces programmes d'activités sont établis :

— à partir de l'exploitation des données d'enquêtes sur les mobilités qui permettent

de connaître les schémas de mobilité d'un échantillon d'individus sur un territoire donné. Dans les enquêtes sur les mobilités, se trouvent des attributs tels que les types d'activités, les horaires des activités, les modes de transport utilisés. Ces attributs sont le plus souvent affectés aux agents générés dans la première étape à travers une procédure d'appariement statistique ;

— à partir du réseau de transport qui traduit l'infrastructure sur laquelle les agents peuvent se déplacer. Le réseau de transport, généré le plus souvent avec des logiciels de système d'information géographique ou de bases de données cartographiques comme OpenStreetMap est constitué de l'ensemble des rues et des voies de circulation d'une ville, avec leur capacité, la vitesse maximale et les lignes de transport en commun. Ces éléments constituent les données du scénario de mobilité et leur ajout permet de déterminer les itinéraires possibles des agents dans la réalisation de leurs plans d'activités.

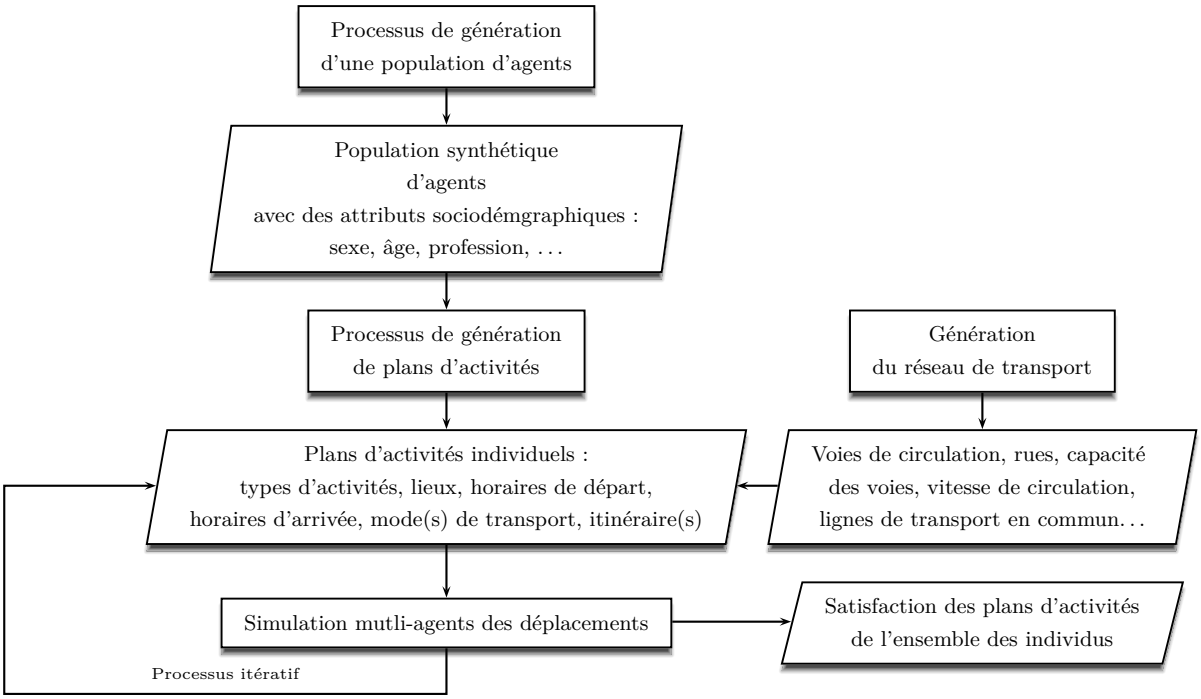
La troisième et dernière étape est l'optimisation itérative des déplacements dans une simulation de type multi-agents. L'optimisation concerne l'ensemble du programme d'activités et non pas chaque déplacement séparément. L'agent essaie alors d'exécuter son programme quotidien avec l'utilité la plus élevée. Cette dernière dépend de contraintes telles que les capacités des voies, les horaires d'ouvertures des établissements, les programmes des autres voyageurs. Certains agents peuvent alors voir certaines composantes de leur plan d'activités initial être modifié (un décalage des horaires ou un changement modal par exemple). Le processus d'optimisation se poursuit jusqu'à ce que la mobilité de chaque agent soit adéquate avec les contraintes de l'environnement et les mobilités des autres agents.

Comparativement aux modèles à quatre étapes, les modèles multi-agents basés sur les activités permettent notamment de mieux :

— représenter les liens entre activités et déplacements à l'échelle individuelle ;

— rendre compte du comportement des individus et leur mécanisme de choix en matière de mobilité (Rasouli and Timmermans, 2014 ; Kagho et al., 2020) ;

FIGURE 1.1 : Différentes étapes de la modélisation de la mobilité dans un modèle multi-agents basé sur les activités



— considérer les nouvelles offres de mobilités telles que le covoiturage ou l'autopartage et leurs défis opérationnels (Hörl and Balac, 2020) ;

— mesurer les externalités environnementales (bruit, pollution) associées aux transports (Shiftan et al., 2015 ; Kaddoura et al., 2017a).

Il existe de nombreuses plateformes logiciels de simulation multi-agents basés sur les activités. Le choix d'une de ces plateformes dépend en premier lieu de la problématique de recherche et d'autres critères tels que le nombre d'agents modélisables, les facilités de développement offertes (taille de la communauté, la disponibilité de guides et de supports d'aide), la capacité de la plateforme d'être associée à d'autres types de modèles, le critère open source (Othman, 2016).

Ce chapitre introduit quelques unes de ces plateformes et traite de leur intégration dans la prise en compte des externalités environnementales associées aux transports. Les différentes parties du chapitre sont organisées comme suit. Dans la prochaine section (section 1.1), trois plateformes logiciels sont présentées : TRANSIMS, MATSim et

SimMobility. La section suivante (section 1.2) illustre les possibilités d'association des modèles multi-agents basés sur les activités avec des modèles de bruit ou d'émissions pour l'évaluation de l'exposition dynamique des individus aux externalités environnementales. La troisième et dernière section (section 1.3) souligne la nécessité de disposer de données d'entrée précises pour une simulation multi-agents basée sur les activités réaliste.

1.1 Plateformes logiciels multi-agents basés sur les activités

1.1.1 TRANSIMS

La plateforme de simulation multi-agents basée sur les activités TRANSIMS (TRansportation ANalysis and SIMulation System) a été développée pour constituer une alternative solide aux modèles quatre étapes (Jeihani and Ardeshiri, 2014). La plateforme est open source [1] et codée en C++.

TRANSIMS combine la modélisation de la demande de déplacements et le comportement des flux dans un environnement d'analyse commun et simule les détails dynamiques qui contribuent à la complexité inhérente aux problèmes de transport (Dixon et al., 2007). La plateforme retrace et simule les mouvements de chaque voyageur à partir des activités menées sur une période d'une journée. Les plans d'activités de tous les agents (individus) sont exécutés simultanément dans le modèle selon le schéma suivant :

Conceptuellement, le modèle TRANSIMS se compose de quatre modules (Lawe et al., 2009) qui interagissent entre eux (les données de sortie de certains modules servent de données d'entrée à d'autres modules). Ces modules sont :

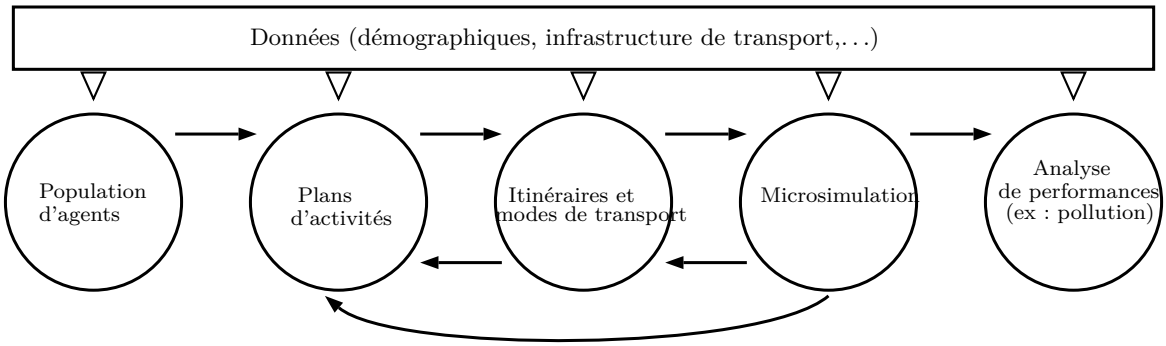
— le générateur de population ;

— le générateur d'activités ;

— le routeur ;

1. Le code source est disponible à partir du lien suivant : <`https://sourceforge.net/projects/transims/files`>, page consultée le 15 septembre 2021.

FIGURE 1.2 : Différentes étapes de l'estimation de la demande dans TRANSIMS



Source : Adapté de Rickert and Nagel (2001)

— le simulateur.

A travers le générateur de population, TRANSIMS commence par créer une population d'agents avec les attributs sociodémographiques renseignés, notamment à partir de données de recensement ou d'enquêtes. Le générateur d'activités attribue une liste d'activités que chaque agent peut réaliser en fonction de ses attributs sociodémographiques. Les heures, les lieux d'activités et les modes de transport sont également renseignés. Le générateur d'activités et le routeur calculent ensuite des plans combinés de trajet pour réaliser les activités souhaitées selon le mode de transport choisi.

Le dernier module (le simulateur) simule la dynamique du trafic qui en résulte selon le concept de rétroaction itérative : les temps de parcours dérivés d'une itération donnée sont utilisés pour ajuster de façon répétée les trajets jusqu'à ce que les trajets et les temps de parcours se stabilisent (Lawe et al., 2009). La technique utilisée par le simulateur est celle des automates cellulaires. Le mouvement des véhicules et les interactions entre voyageurs dans la zone d'étude sont alors simulés pour une période de temps d'une journée généralement.

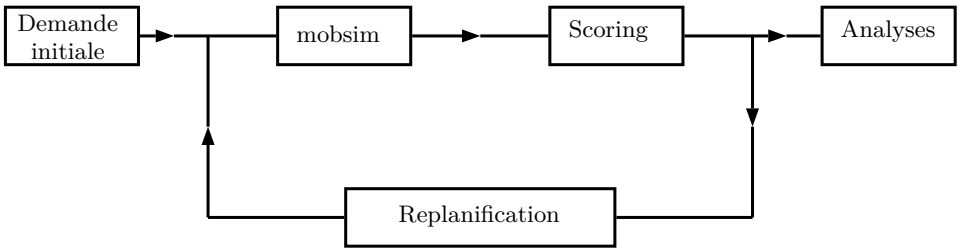
1.1.2 MATSim

MATSim (Multi-Agent transport Simulation) est une plateforme de simulation multi-agents basée sur les activités et implémenté en Java. Elle a été développée à partir de la

plateforme de simulation TRANSIMS (Illenberger et al., 2007). La plateforme est open source [2] et conçue pour une application à grande échelle (ville ou pays) (Horni et al., 2016).

La procédure de modélisation des déplacements dans MATSim est un processus d'optimisation basé sur le concept d'algorithmes évolutifs (Gao et al., 2010) dans lequel le plan de chaque agent est simulé. L'outil fonctionne de manière itérative. Une itération se décompose en trois étapes comme le montre la Figure 1.3.

FIGURE 1.3 : Différentes étapes de l'estimation de la demande dans MATSim



Source : Adapté de Horni et al. (2016)

La plateforme MATSim reçoit comme données d'entrée, des individus (agents) avec leurs attributs et les plans d'activités correspondant. Dans la première étape (mobsim), la demande de mobilité des individus est simulée. Cette étape s'appuie sur un modèle de simulation de file d'attente qui mesure les flux de trafic et estime les temps de parcours des agents dans l'accomplissement de leurs activités. Une fois que les limites de capacité d'une voie sont atteintes, le trafic ralentit et la congestion s'accumule sur les voies en amont. De cette façon, les choix effectués dans les plans des agents affectent directement les temps de déplacements (Hörl et al., 2018).

Dans MATSim, la qualité du plan d'activités d'un agent est décrit par un score d'utilité $S \in \mathbb{R}$ (Hörl, 2021). La seconde phase (scoring) consiste à déterminer ce score d'utilité à l'aide de paramètres variant en fonction du temps et des coûts équivalents.

2. Le code source est téléchargeable à partir du lien suivant : <`https://www.matsim.org`>, page consultée le 15 septembre 2021.

Des fonctions de score peuvent également être définies pour chaque agent en fonction des attributs sociodémographiques de l'agent ou des préférences de l'utilisateur (Vosooghi et al., 2019).

La dernière étape du processus itératif est la replanification (replaning). Les agents sont en concurrence les uns avec les autres, dans le temps et dans l'espace, et essayent de réaliser leurs plans d'activités (Horni et al., 2016). Les plans qui ont de mauvais scores sont éliminés et remplacés par d'autres plans (modification de l'heure de départ ou de l'itinéraire par exemple). L'affectation finale du trafic résulte de l'interaction du plan exécuté de chaque agent lorsque le système atteint un état stationnaire.

1.1.3 SimMobility

La plateforme SimMoblity a été conçue selon le paradigme de la modélisation basée sur les activités et intègre différents modèles qui prennent en compte les interactions entre l'utilisation des sols, le réseau de transport et de communication et les décisions de mobilité des agents (Marczuk et al., 2015). La plateforme est disponible en open source [3] et se compose de trois niveaux principaux de simulation : 1) le court terme, 2) le moyen terme et 3) le long terme.

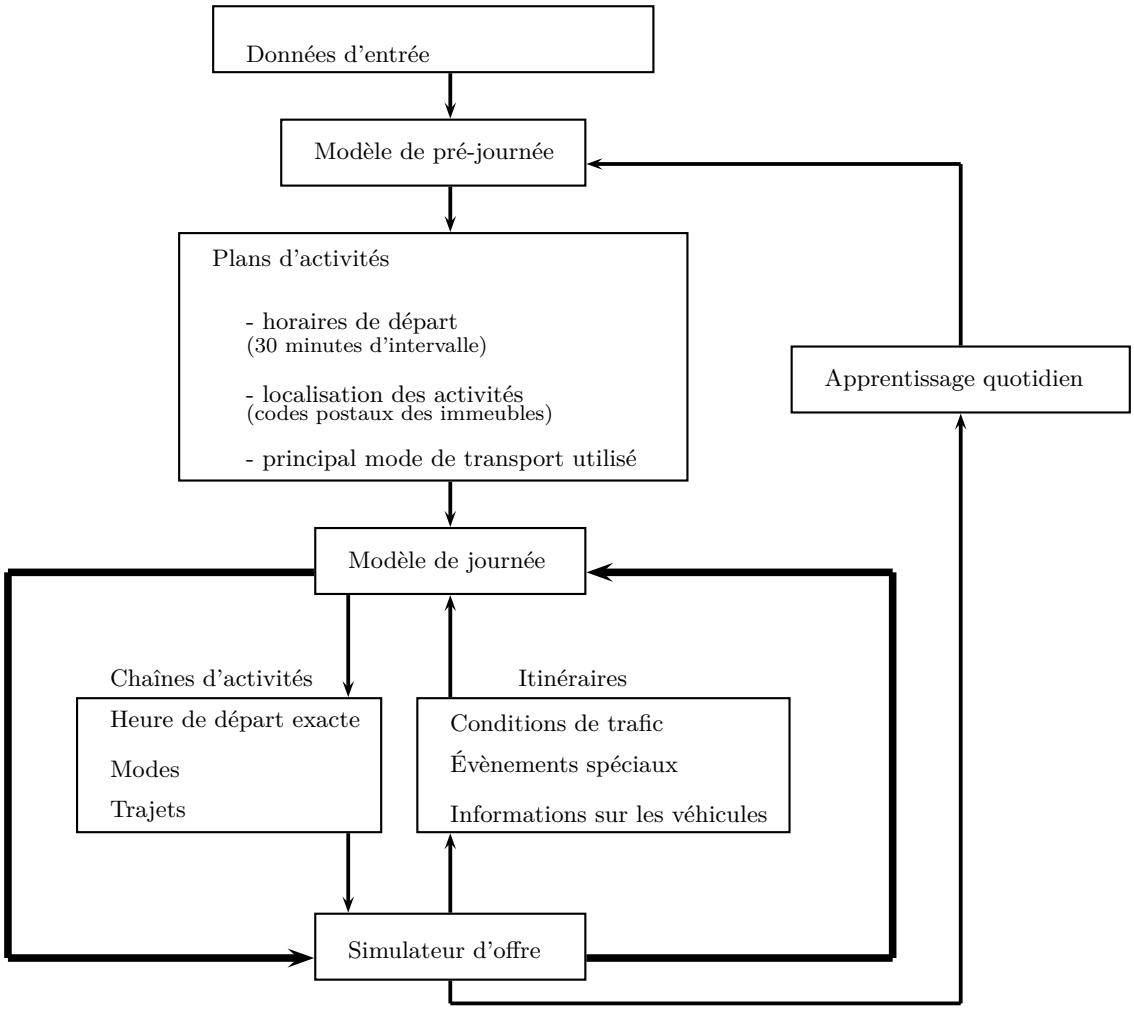
Le simulateur de long terme de SimMobility est un simulateur d'occupation des sols et de transport qui modélise les comportements des agents notamment sur le marché de l'immobilier et du travail afin de simuler les impacts de long terme des scénarios de mobilité futurs (Adnan et al., 2016). L'étape temporelle est de l'ordre de plusieurs mois voire quelques années.

Les données d'entrées du simulateur de moyen terme contiennent les attributs détaillés de chaque individu. Le simulateur se compose de trois éléments principaux : la pré-journée, la journée et le simulateur d'offre (Figure 1.4). Le module de pré-journée fournit le programme d'activités des individus avec les horaires correspondants ainsi que les modes de transport utilisés pour réaliser les activités. Le module de la journée est

3. Le code source est accessible à partir de ce lien :<`https://github.com/smart-fm/simmobility-prod`>, page consultée le 15 septembre 2021.

conçu pour simuler le choix de l'heure de départ et les décisions de choix d'itinéraire. Le simulateur d'offre fournit les attributs des réseaux des transports (Basu et al., 2018).

FIGURE 1.4 : Architecture du simulateur de moyen terme



Source : Adapté de Lu et al. (2015)

Le simulateur de court terme est un simulateur multimodal dans lequel les mouvements des individus sont capturés. Les mouvements étudiés comprennent le changement de voie, le freinage, la vitesse mais aussi le choix de la route. Ce simulateur permet également une intégration modulaire de comportements spécifiques associés à de nouveaux

services de mobilité et de modes de transport (Gao and Peh, 2014).

Les trois niveaux (long terme, moyen terme, court terme) fonctionnent de façon indépendante. Il est par exemple possible de procéder à une simulation de moyen terme sans auparavant recourir à une simulation de court terme. Malgré cette indépendance, il faut souligner que les trois niveaux utilisent les mêmes bases de données, la différence étant que l'information est exploitée en fonction des besoins de chaque niveau. Les trois niveaux restent donc liés et la plateforme de modélisation garde une trace des choix des agents à tous les niveaux (Adnan et al., 2016). A titre illustratif, la possession d'une voiture (attribut de long terme d'un agent) est potentiellement liée au choix modal (attribut de moyen terme) et éventuellement aux caractéristiques du comportement de conduite individuel, comme le temps de réaction ou la vitesse souhaitée (attributs de court terme).

Les trois plateformes présentées (TRANSIMS, MATSim et SimMobility) permettent l'analyse des mobilités quotidiennes des individus. TRANSIMS a surtout été appliquée aux Etats-Unis (Yang et al., 2020 ; Jeihani and Ardeshiri, 2014 ; Lawe et al., 2009 ; Dixon et al., 2007), La plateforme SimMobility est plus récente, comparativement aux deux autres. Les exemples d'application recensés concernent davantage les Etats-Unis et Singapour (Oh et al., 2021 ; Zhou et al., 2021 ; Alho et al., 2018 ; Basu et al., 2018 ; Azevedo et al., 2017 ; Adnan et al., 2016). La communauté de MATSim est la plus diversifiée avec un nombre important d'utilisateurs et des contextes d'application différents (Bean and Joubert, 2021 ; Diallo et al., 2021 ; Le Bescond et al., 2021 ; Balac, 2020 ; Vosooghi et al., 2019 ; Kickhofer et al., 2016). La section suivante présente les couplages de modèles multi-agents basés sur les activités avec des modèles de bruit et d'émission pour l'évaluation des externalités environnementales.

1.2 Evaluation de l'exposition des individus aux nuisances du trafic routier

La quantification et la réduction des expositions aux nuisances environnementales générées par le trafic routier, en particulier le bruit et les polluants atmosphériques, sont des enjeux centraux. [4] [5] Des couplages de modèles permettent d'estimer la répartition spatiale de ces nuisances et de représenter les niveaux de bruit ou les concentrations en polluants atmosphériques en une carte de récepteurs. Actuellement, l'approche consiste le plus souvent à estimer les expositions selon une modélisation statique reposant sur un croisement entre cette répartition spatiale des nuisances et les lieux de résidence (le voisinage est parfois utilisé comme aire d'exposition individuelle). Cette approche a toutefois plusieurs limites et n'assure pas une estimation robuste de l'exposition réelle des individus (Park and Kwan, 2017). D'une part, parce qu'elle ne tient pas compte des dynamiques temporelles de la mobilité des individus et des niveaux de pollution des lieux traversés par ces derniers. Cela peut entraîner des erreurs dans le calcul des expositions moyennes. D'autre part, parce que les expositions calculées sont statiques et s'appuient sur des indicateurs agrégés qui ne permettent pas de rendre compte de l'impact potentiel d'expositions brèves à des niveaux élevés.

Alors que l'approche classique ne permet de rendre compte que de l'exposition des individus dans leur logement, ou dans le voisinage de celui-ci, l'approche par les systèmes multi-agents permet de mesurer une exposition dynamique des individus, c'est-à-dire une exposition rendant compte des déplacements et changements de localisation vécus par les individus au cours de leur journée. Par ailleurs, puisque la modélisation de type multi-agents repose sur une description assez fine des individus (population synthétique) et de leurs activités (plans d'activités), cette démarche permet non seulement de quantifier plus précisément le nombre d'individus exposés en un endroit donné à un moment précis ; mais

4. En France, 48 000 décès par an pourraient être attribués à l'exposition chronique aux particules fines ($PM_{2.5}$) dont au moins la moitié dans les agglomérations de plus de 100 000 habitants (Pascal et al., 2016).

5. Au sein de la zone dense francilienne, chaque année, le bruit des transports serait responsable de la perte de 107 766 années de vie en bonne santé, ce qui représente une perte de 10,7 mois par habitant en moyenne au cours d'une vie (BruitParif, 2019).

aussi de mieux caractériser la population exposée (caractéristiques socioéconomiques) et les circonstances de l'exposition (au cours de quel déplacement ? de quelle activité ?). Ce faisant, elle permet d'approfondir les travaux existants qui reposent sur un croisement de données d'exposition moyenne et de localisation de lieux de résidence (Jerrett et al., 2001 ; Dale et al., 2015) ou sur des suivis de cohortes dans le cas d'études épidémiologiques (Havard et al., 2011).

En simulant les schémas d'activités et de déplacement des individus d'une population, les modèles de transport multi-agents basés sur les activités fournissent des informations sur pourquoi, quand et où les gens se déplacent (Arentze and Timmermans, 2004). Ces modèles permettent ainsi de prendre en compte les dynamiques spatio-temporelles et, ce faisant, d'améliorer l'estimation des expositions au bruit et aux polluants atmosphériques. La modélisation par le système multi-agents permet de mieux caractériser les sources d'émissions puisqu'elle permet d'identifier les véhicules sources d'émissions et donc de connaître leur(s) occupant(s) et le motif du déplacement. De même, elle permet de mieux caractériser l'exposition des individus, de certaines catégories de population ou de certaines zones et offre un cadre d'analyses prospectives (Houot et al., 2015). Ce faisant, elle permet potentiellement de mener des analyses plus fines de politiques de transport (restriction de la circulation, politiques de tarification...), d'infrastructures ou de nouvelles solutions de mobilité.[6] Il est nécessaire de prendre en compte les mobilités individuelles pour atteindre l'ensemble des personnes ayant besoin d'aide (au lieu de simplement cibler les personnes résidant dans les zones où la concentration (de polluants) est élevée) ou pour changer les caractéristiques négatives de la zone dont la population est effectivement exposée. Si les interventions ciblant les conditions de logement ou les initiatives implémentées durant la nuit peuvent être effectivement définies avec une approche basée sur les lieux de résidence (comme cela est fait la plupart du temps) ; cette logique ne peut être suivie pour les initiatives qui cherchent à changer l'environnement auquel les gens sont exposés quotidiennement ou pour des actions mises

6. Les effets négatifs des nuisances (i.e. les effets externes tels que la congestion, le bruit, la pollution locale) et les coûts (i.e. les coûts externes) qu'ils engendrent ne sont pas (ou pas complètement) pris en compte par l'usager du transport lorsqu'il prend ses décisions de déplacement. En d'autres termes, les coûts privés de déplacement ne représentent pas les coûts sociaux ressentis par la collectivité.

en œuvre pendant la journée (Vallée, 2017).

1.2.1 Chaîne de modélisation pour l'évaluation de l'exposition au bruit et à la pollution de l'air

Les estimations des impacts du bruit et de la pollution atmosphérique ne sont pas régies par les mêmes contraintes et ne suivent pas les mêmes objectifs (Can, 2019). Le Tableau 1.1 résume les différences en termes d'indicateurs à estimer, de granularités spatiales et temporelles. Par exemple, l'acousticien est particulièrement intéressé à savoir comment estimer les variations temporelles des niveaux de bruit, ainsi qu'à la distribution spatiale de ces niveaux. En outre, contrairement à l'exposition aux polluants atmosphériques, les séries temporelles des niveaux de bruit sont souhaitables en sortie des modèles. D'un autre côté, les erreurs dans l'estimation du bruit n'affectent pas les tranches de temps suivantes, contrairement aux polluants atmosphériques.

L'évaluation la plus complète de l'exposition individuelle à une nuisance (pollution de l'air ou bruit) reposant sur un modèle basé agent suit essentiellement 3 étapes (Figure 1.5) :

1. La modélisation des déplacements repose sur l'utilisation d'un modèle basé sur les activités. L'information sur les activités et les déplacements produite par la simulation du modèle comprend, pour chaque personne de la zone étudiée, la liste de toutes les activités hors domicile entreprises au cours d'une journée, ainsi que leur motif (travail, courses, loisirs, etc.), l'emplacement géocodé des secteurs (c'est-à-dire, la destination du trajet) et un certain nombre d'informations sur les déplacements à destination et en provenance de ces activités, y compris les heures de début et de fin du voyage estimées et le mode de transport (voiture à occupant unique, voiture à occupants multiples, autobus, marche, vélo). La génération de la population synthétique et des plans d'activités s'effectue le plus souvent en combinant différentes sources de données (recensement, enquêtes). La diversité des comportements retranscrite dans les plans d'activités est ensuite utilisée dans des modèles de

Tableau 1.1 : Comparaison des externalités environnementales.

	Bruit	**Pollution de l'air**
Impacts sur la santé	++ Les impacts concernent principalement la détérioration de la qualité de vie.	++ La morbidité liée à l'exposition aux polluants atmosphériques contitue un problème de santé publique.
Granularité spatiale	+++ Une granularité spatiale très fine est attendue. Idéalement, il faut estimer la distribution spatiale des niveaux de bruit avec une résolution spatiale d'environ 10 ou 20 m.	++ Une granularité spatiale relativement fine est attendue. Cependant, de nombreuses études se limitent aux quantités de polluants émises sur une zone assez large du territoire.
Granularité temporelle	+++ L'estimation de l'évolution temporelle des niveaux de bruit, l'estimation des événements sonores, sont des éléments cibles de la modélisation, car ils ont une grande influence sur la perception des environnements sonores.	+ L'évolution temporelle des niveaux de polluants estimés est d'une importance limitée, bien que certaines études tendent à montrer le danger associé à l'exposition à des pics de pollution très courts (quelques secondes).
Temps de calcul des émissions	+++ Le calcul des niveaux de puissance acoustique est très rapide. Il est négligeable par rapport aux calculs de propagation du son.	++ Les modèles agrégés suivent des lois simples donnant les émissions en fonction de la vitesse moyenne du flux ou éventuellement du trafic. Ils sont très rapides.
Rémanence	- Le phénomène n'est pas persistent : les erreurs commises sur une période de temps n'ont pas de répercussion.	++ Les niveaux de pollution observés sur un intervalle de temps donné peuvent être le résultat d'émissions survenues plusieurs heures ou des jours plus tôt.
Indicateurs cibles	Indicateurs calculés pour une carte réceptrice et une période donnée, basés sur l'évolution des niveaux sonores : indicateurs moyens, indicateurs statistiques, niveaux d'événements sonores.	Niveaux moyens émis, concentrations moyennes des polluants dans une carte de récepteurs pour une période donnée.

Source : Can (2019)

déplacement pouvant utiliser différents types de procédure d'affectation du trafic.

2. L'estimation des concentrations spatio-temporelles de bruit et d'un ou plusieurs polluants. Celle-ci se base sur l'estimation des flux de trafic, de leur composition et des vitesses moyennes par tronçon routier (simulés à l'étape 1).

   — 2.a. Dans le cas des polluants, il s'agit dans un premier temps d'estimer les émissions de polluants générées par le trafic puis de calculer leur dispersion, transport et concentration dans l'air. L'intégration d'un modèle basé sur les activités dans un cadre de modélisation de la qualité de l'air présente de nombreux avantages pour la modélisation du trafic, des émissions, de la dispersion et de l'exposition (Shiftan, 2000 ; Shiftan and Suhrbier, 2002). Ainsi, l'extraction de l'information sur les déplacements simulés peut fournir des

valeurs temporelles plus précises sur les déplacements et les émissions (Beckx et al., 2009b) qui dépendent non seulement de la distance et de la vitesse de conduite, du type de véhicule, mais aussi du nombre de trajets, du temps écoulé entre eux et du fait que le moteur était chaud ou froid au démarrage (Recker and Parimi, 1999). En fonction des données disponibles, plusieurs types de modèles peuvent être utilisés pour évaluer les quantités de polluants émises par le trafic routier : basé sur les vitesses moyennes, sur les situations de trafic, régression linéaire multiple, instantané (Andre et al., 2012 ; Hülsmann, 2014). Il existe également une grande diversité de modèles de la qualité de l'air qui peuvent être appliqués à des études de cas locales et urbaines. Ces modèles de pollution atmosphérique, qui diffèrent selon leur échelle, peuvent être physique, empirique, statistique ou déterministe (physico-chimique). [7]

— 2.b. Dans le cas du bruit, la démarche adoptée consiste, dans un premier temps, à calculer les émissions sonores en fonction du flux de trafic estimé grâce au modèle multi-agents, de la part des poids lourds et du niveau de vitesse. Les niveaux de puissance acoustique (i.e. les émissions) sont calculés pour chaque section de rue, en prenant en compte les facteurs de correction pour différents revêtements routiers et pentes longitudinales. Ces émissions sont discrétisées en points sources sur le réseau routier (environ un point tous les dix mètres). [8] Ensuite, un récepteur est généré pour chaque façade de bâtiment. Les niveaux sonores pour chaque récepteur, c'est-à-dire chaque façade de bâtiment, sont obtenus en tenant compte de la propagation entre chaque couple "point source / récepteur", et en sommant la contribution des

7. Notons que certains travaux reposent sur une carte de concentrations de polluants produites par d'autres modèles de transport (ou sur la base de comptage de trafic) ou générées en extrapolant des mesures en utilisant notamment des méthodes de régression (Dons et al., 2011a,b ; Park, 2020 ; Park and Kwan, 2020 ; Lu, 2021). D'autres ne mettent pas œuvre la chaîne complète de modélisation et s'arrêtent au stade de l'estimation des émissions (Shiftan et al., 2015 ; Elessa Etuman and Coll, 2018 ; Agarwal et al., 2020). Les premiers travaux ne permettent pas de faire l'analyse de scénarios (puisque les cartes de concentration sont indépendantes de la modélisation basée sur les activités) tandis que les seconds ne permettent pas d'évaluer l'exposition des individus puisqu'ils ne modélisent pas la concentration des polluants.

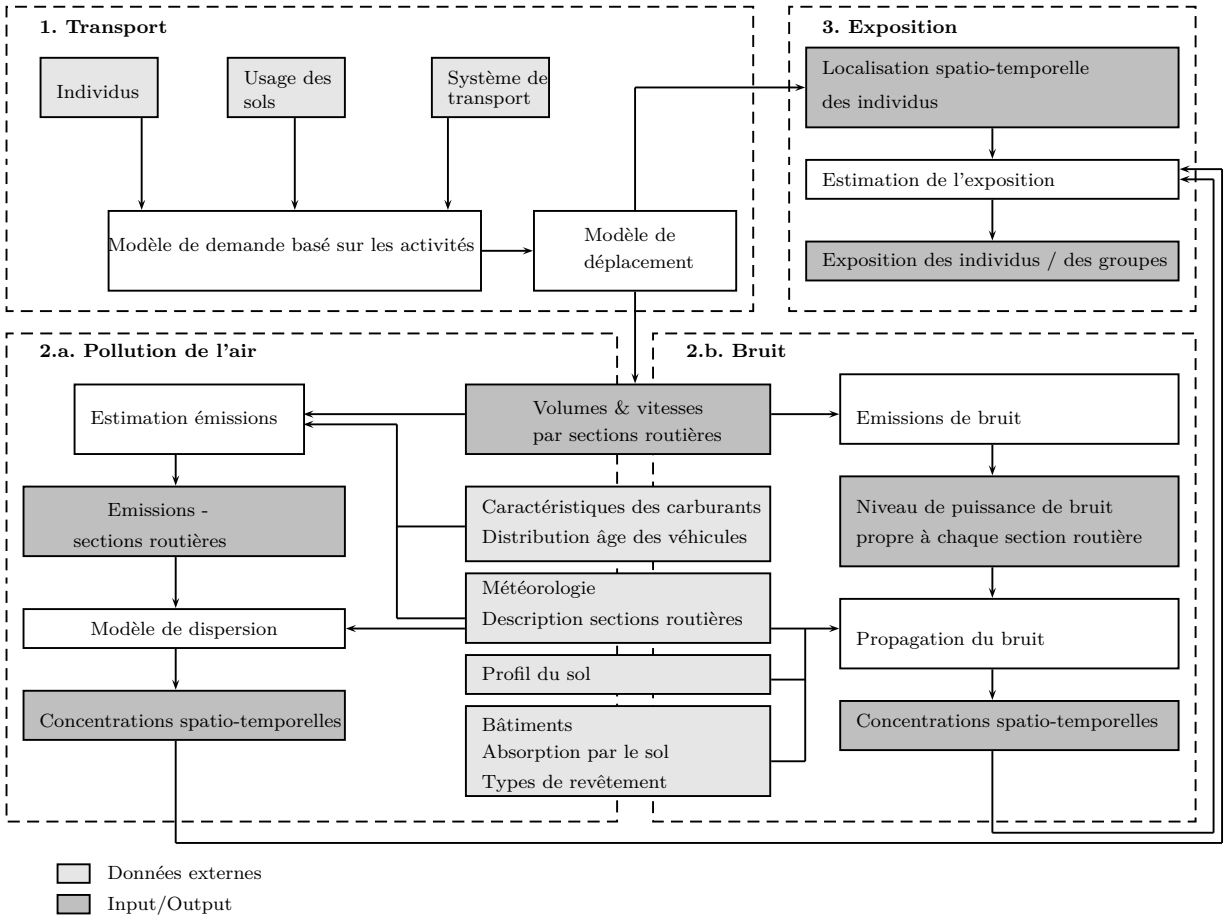
8. Pour une description des indicateurs acoustiques, nous renvoyons à la synthèse de Picaut (2011).

points sources.

3. La modélisation de l'exposition consiste à combiner les localisations spatio-temporelles des individus (simulées à l'étape 1) avec les distributions spatio-temporelles des concentrations de polluants (simulées à l'étape 2) pour estimer les mesures de l'exposition à l'échelle de l'individu. La modélisation de l'exposition individuelle repose sur la conservation de la trace des lieux d'activités de chaque individu ainsi que les heures de début et de fin de chacune des activités. En considérant que les activités se déroulent au récepteur le plus proche, le nombre d'individus qui peuvent être exposés au bruit ou à la pollution est enregistré pour chaque récepteur et chaque intervalle de temps (l'individu peut être partiellement pris en compte si son activité commence ou se termine pendant un intervalle de temps).

Bien que l'intérêt potentiel d'utiliser un modèle basé sur les activités pour estimer l'exposition humaine à la pollution de l'air ait été identifié il y a déjà une vingtaine d'années (Shiftan, 2000), les applications demeurent assez récentes. Deux études, l'une aux Pays-Bas (Beckx et al., 2009a,b,c) et l'autre au Canada (Hatzopoulou and Miller, 2010), constituent les premiers travaux de référence en matière d'utilisation de modèles de demande de transport basés sur l'activité pour estimer l'exposition à la pollution atmosphérique liée au trafic en modélisant la chaîne complète, i.e. modélisation basée agents couplée à un modèle de trafic afin d'estimer les émissions spécifiques aux liaisons routières, les concentrations ambiantes et les expositions individuelles. L'amélioration apportée par l'étude canadienne par rapport à l'étude néerlandaise consiste notamment en l'utilisation de MATSim afin de mieux prendre en compte les phénomènes de congestion (Hao et al., 2010 ; Hatzopoulou and Miller, 2010 ; Hatzopoulou et al., 2011). Par la suite Vallamsundar et al. (2016), Gurram et al. (2019) (repris ensuite par Kocak et al. (2021)) et Tayarani and Rowangould (2020) ont apporté un certain nombre d'améliorations aux travaux antérieurs en utilisant un modèle d'affectation dynamique du trafic. Vallamsundar et al. (2016) ont notamment montré l'intérêt de considérer d'une part différents types de micro-environnements et d'autre part un taux d'inhalation distinct selon l'âge et le genre. Gurram et al. (2019), eux, ont permis de comprendre les impacts

FIGURE 1.5 : Modélisation de l'exposition individuelle à la pollution de l'air et au bruit.



Source : Adapté de Gurram et al. (2019)

potentiels de l'intégration de données d'activité et de concentration à plus haute résolution, y compris pendant les déplacements. Enfin, Tayarani and Rowangould (2020) sont venus confirmer l'intérêt de prendre en compte l'exposition des individus au cours de leurs déplacements et de passer d'une modélisation statique de l'exposition à une modélisation dynamique. Ceux-ci démontrent notamment que, pour la région métropolitaine d'Atlanta, une analyse intégrant la mobilité des personnes aboutit à une mesure de l'exposition aux $PM_{2.5}$ 51% supérieure à celle d'une approche statique, où l'on suppose que toute l'exposition a lieu à la maison. En moyenne, l'exposition à la maison représente

43% de l'exposition quotidienne ; le travail et les déplacements représentent, eux, respectivement 27% et 18%. Cette dernière estimation est en ligne avec les observations faites par les études antérieures ayant montré que les expositions pendant les déplacements contribuent entre 6% et 24% des expositions quotidiennes (De Nazelle et al., 2013 ; Dons et al., 2011b, 2012 ; Gurram et al., 2015). C'est pourquoi les erreurs d'estimations induites par une approche statique sont beaucoup plus fortes dans les zones péri-urbaines et rurales où l'exposition est sous-estimée, tandis que l'exposition peut être surestimée près des routes à fort trafic et dans le centre urbain. Notons que les erreurs estimées par Tayarani and Rowangould (2020) sont beaucoup plus importantes que celles des études précédentes (Beckx et al., 2009a,b ; Dhondt et al., 2012 ; Hatzopoulou and Miller, 2010). Ces différences constatées peuvent s'expliquer en partie par le fait qu'elles se concentrent sur des régions urbaines plus ou moins compactes. Les différences estimées entre l'exposition statique et l'exposition dynamique peuvent également être fonction des activités et de leur localisation, de la résolution des données (de la concentration des polluants ou de la population synthétique) ou encore de la non prise en compte de l'exposition au cours du déplacement.

Les travaux ayant couplés un modèle de transport basé sur les activités avec un modèle de bruit sont moins nombreux que cela n'est le cas pour la pollution de l'air. Le premier travail a été celui de Gerike et al. (2012) qui ont couplé MATSim et le logiciel de cartographie du bruit IMMI (sans exemple de simulation numérique). [9] Par la suite, d'autres auteurs ont couplé un logiciel multi-agent de simulation des déplacements avec des logiciels d'acoustique environnementale (Houot et al., 2015 ; Le Bescond et al., 2021). Le cadre de modélisation ayant été le plus souvent utilisé est celui développé par Kaddoura et al. (2017a) qui repose sur MATSim et une de ses extensions dédiée spécifiquement à la modélisation du bruit. Ce cadre a essentiellement fait l'objet d'exploitation sur les cas de Berlin (Kaddoura et al., 2017a,b ; Kaddoura and Nagel, 2018 ; Kuehnel et al., 2021), Munich (Kuehnel et al., 2019 ; Kuehnel and Moeckel, 2020 ; Kaddoura

9. IMMI est un logiciel de prévision et de cartographie de bruit pour tout type de sources sonores (route, ferroviaire, industriel, activités de loisirs, aéroport) et pour tout type de cartes de bruit jusqu'aux cartes de bruit stratégiques pour les agglomérations.

et al., 2020) et Zurich (Zwick et al., 2021) pour l'évaluation de politiques de transport ou du déploiement de nouvelles solutions de mobilité. L'intérêt de prendre en compte la dynamique intrajournalière des différentes densités de population dans les différents quartiers d'une ville, comparativement à l'approche statique, a été notamment démontré par Kaddoura et al. (2017a) lors de leur estimation de l'exposition au bruit routier à Berlin. Selon leurs résultats, dans les zones résidentielles, moins de personnes sont exposées au bruit le jour que le soir ou la nuit tandis que l'on observe l'inverse dans le centre ville (où se concentrent notamment les lieux de travail et d'études). Ainsi, l'utilisation du nombre de résidents statiques entraînerait une surestimation (resp. sous-estimation) des dommages causés par le bruit pendant la journée dans les zones résidentielles (resp. dans le centre ville). Le Bescond et al. (2021) illustrent, eux, la variabilité des profils individuels d'exposition au bruit, et le biais induit par une approche statique (i.e. le niveau de dose de bruit mesuré avec une approche statique ne réflète pas le réel niveau de dose auquel est exposé l'individu), dans le cas de la ville de Nantes.

Les limites des cadres de modélisation aujourd'hui mis en oeuvre sont encore nombreuses. Certaines limites sont directement liées à la modélisation du transport. Ainsi certains types de trafic ou modes de transport (transit, fret, deux-roues motorisés) sont parfois imparfaitement pris en compte (voire même ignorés comme dans Hatzopoulou and Miller (2010)) bien qu'ils puissent avoir un impact important aussi bien du point de vue de l'exposition à la pollution ou au bruit (exemple des motards ou les travailleurs de la logistique urbaine). D'autres limites sont propres à la modélisation du bruit et de la pollution de l'air. Tout d'abord, le cadre de modélisation adopte une représentation des sources de pollution sous la forme d'un flux de trafic décrit par segment de route au moyen d'un débit et d'une vitesse moyenne. Cette approche peut conduire à des divergences près des intersections, où la vitesse des véhicules n'est pas uniforme et où les phases d'accélération affectent les émissions (aussi bien de polluants que de bruit). La cinématique des véhicules est généralement simplifiée. Les changements de voie, les phases d'accélération, les décisions des conducteurs, les dépassements ou les contrôles adaptatifs des feux de signalisation ne sont que quelques-uns des nombreux phénomènes

de circulation qui ne sont pas modélisés ici. De plus, dans le cas de la pollution, l'analyse se limite généralement à un polluant et aux seules émissions du trafic routier. Or les schémas de concentration des polluants peuvent être différents entre eux, i.e. l'étude d'un seul polluant ne peut donc pas rendre compte de l'exposition des individus à l'ensemble des polluants issus du trafic routier. Un certain nombre d'études négligent les expositions subies lors des déplacements qui peuvent notamment être importantes pour les usagers de modes de transport non motorisés (cyclistes, piétons ou passagers des transports publics attendant aux arrêts). Lorsqu'elles sont calculées, les expositions à la pollution ou au bruit le sont en utilisant les concentrations ambiantes simulées dans les zones traversées. De même, on attribue aux bâtiments (exposition intérieure) les concentrations ambiantes des zones dans lesquelles ils se situent (Tayarani and Rowangould, 2020) négligeant ainsi les sources sonores intérieures et la propagation du bruit à l'intérieur des bâtiments. De plus, la façon dont les récepteurs sont placés dans la simulation d'acoustique environnementale a un impact considérable sur les résultats. Si un récepteur d'activité est placé du côté rue d'un bâtiment, il aura une valeur d'exposition au bruit beaucoup plus élevée que si le même récepteur d'activité est placé du côté ombragé. Le cadre de modélisation considère les expositions en se concentrant exclusivement sur le bruit du transport routier, les impacts étant exprimés par des indicateurs de moyenne énergétique. Bien que recommandée par les normes de calcul, cette approche ne reflète pas pleinement les impacts acoustiques potentiellement observés dans les zones urbaines, la mise en œuvre actuelle ignore l'exposition des agents pendant les déplacements puisqu'elle se concentre sur les activités (Le Bescond et al., 2021). Notons que, par ailleurs, on ne peut pas distinguer en l'état des données disponibles, le temps passé à l'extérieur des bâtiments ou des véhicules. Enfin, la représentativité de la population d'agents, de leurs véhicules, des déplacements et des activités qui leur sont alloués est directement dépendante des données disponibles et des méthodes mises en œuvre pour générer la population synthétique et les plans d'activités. La plupart des enquêtes portant sur les jours ouvrés de la semaine ; aucun travail à notre connaissance ne rend compte de l'exposition des individus au bruit et à la pollution lors d'une journée représentative du

week-end ou des vacances. De plus, la description du parc de véhicules ou de bâtiments est souvent imprécise (par exemple, utilisation d'un parc de véhicules régional dans le cas de Vallamsundar et al. (2016)) ce qui a une conséquence directe sur la mesure des émissions de polluants (Tayarani and Rowangould, 2020). Enfin, la qualité de la mesure de l'exposition individuelle est directement liée à la précision de la localisation des individus ; les processus d'allocation spatiale des individus aux logements et la localisation des activités sont donc des étapes importantes et encore imparfaites. Par exemple, Dhondt et al. (2012) et Tayarani and Rowangould (2020) agrègent les données de localisation des agents et des ménages à l'échelle des zones d'analyse du trafic et estiment l'exposition sur la base de la concentration moyenne dans chacune des ces zones. Cette faible précision dans la localisation des individus introduit une certaine erreur dans les estimations d'exposition, notamment une surestimation de l'exposition au $PM_{2.5}$ dans les zones les plus grandes et les plus rurales où une grande partie de la population réside loin des routes à fort trafic (Tayarani and Rowangould, 2020).

1.2.2 Avantages de l'évaluation de l'exposition dynamique des individus aux nuisances des transports par la modélisation multi-agents

Les différents cadres de modélisation cités précédemment ont été utilisés pour étudier les variations d'exposition des individus aux nuisances du trafic routier consécutives à la mise en oeuvre d'une politique de transport tarifaire (péage urbain) ou réglementaire (zone 30), à l'implémentation d'une nouvelle technologie (Cucurachi et al., 2018) ou d'une nouvelle solution de mobilité (Zwick et al., 2021), à la construction de nouvelles infrastructures (Kocak et al., 2021), à la modification de l'offre de transport en commun ou à la désynchronisation des plannings d'activités des individus (Houot et al., 2015). A l'exception de Kaddoura et al. (2020), les études ne s'intéressent qu'à l'une des deux catégories de nuisances (pollution ou bruit) ; parfois cette nuisance est étudiée avec une autre externalité (congestion ou CO_2). La démarche adoptée la plus souvent consiste

à comparer les expositions d'un scénario de base à un scénario modélisant la mise en oeuvre d'une politique ou la modification de l'offre de transport.

Bien que les inégalités d'exposition aient été explorées auparavant, la modélisation multi-agent a jusqu'à présent été peu utilisée pour explorer des enjeux tels que la justice environnementale (Campbell et al., 2015). Au sein des travaux cités précédemment, l'évaluation de l'exposition de la population néerlandaise au NO_2 à différentes périodes, en différents lieux, pour différentes sous-populations (sexe, statut socioéconomique) et au cours de différentes activités (résidentielles, professionnelles, de transport, d'achat) par Beckx et al. (2009b) met en évidence l'intérêt de leur cadre de modélisation. Une analyse plus détaillée de l'exposition révèle les variations intra-journalières des estimations d'exposition et la présence de grandes différences d'exposition selon les activités (travail > achats > domicile) et entre les sous-populations (hommes > femmes, classe socioéconomique basse > classe socioéconomique élevée). Gurram et al. (2019), ont renseigné les disparités d'exposition au NO_x issu du trafic routier entre les sous-groupes de population de leur zone d'étude. Étudiant l'équité au regard de critères socioéconomiques des individus, de leurs activités et de leurs localisations, ils ont notamment montré que, dans le cas de Tampa (Floride, Etats-Unis), l'exposition moyenne de certaines catégories de la population (personnes vivant en dessous du seuil de pauvreté, personnes dont le temps de déplacement quotidien dépasse une heure...) était jusqu'à 16% supérieure à la moyenne de la population. Reposant sur le même cadre de modélisation, Kocak et al. (2021) contribuent à la compréhension de l'exposition potentielle à la pollution atmosphérique et des impacts sur les inégalités d'un programme d'infrastructure de transport à grande échelle impliquant l'expansion des routes et la mise en oeuvre d'un péage. Leurs résultats suggèrent que si de tels programmes peuvent permettre de légères réductions de l'exposition en moyenne (du moins à court terme), ils peuvent également accentuer les disparités existantes pour certains groupes vulnérables et creuser l'écart d'exposition entre les différents groupes sociaux avec des expositions plus élevées pour les personnes vivant sous le seuil de pauvreté et des expositions plus faibles pour les personnes à revenu élevé. Bien que Le Bescond et al. (2021) aient mon-

tré qu'il est possible d'obtenir des trajectoires individuelles d'exposition, la majorité des études réagrègent les observations individuelles au regard de critères sociaux, ethnique ou de localisation (résidentielle ou autre). Enfin, une prise en compte encore plus détaillée de la population permettrait d'approfondir l'analyse de l'exposition, par exemple, les activités professionnelles pourraient être différenciées en fonction du type de travail (exposition des travailleurs mobiles) ; les activités domestiques pourraient être différenciées sur la base d'attributs spécifiques à la personne pertinents pour la sensibilité au bruit (par exemple, les enfants par rapport aux adultes) ou la typologie des activités pourrait être plus détaillée (activités supplémentaires telles que les activités de loisirs peuvent être incorporées). Comme le soulignent Kuehnel et al. (2021), l'allocation simplifiée des résidents aux bâtiments demeure également une limite.

Notons que les cadres de modélisation présentés ici peuvent également se révéler utiles et pertinents (pour le décideur public) pour analyser également les impacts (Lefebvre et al., 2013 ; Kuehnel et al., 2021). Ce type de modélisation peut en effet permettre d'identifier les trajets, motifs de déplacements ou individus qui sont non seulement sources de pollution mais surtout à l'origine d'impact (i.e. exposition d'autres individus), ce que ne permet pas une simple simulation des émissions (de polluants ou de bruit). La valeur ajoutée de l'utilisation du modèle basé sur les activités dans la présente chaîne de modèles est qu'il ne prévoit pas seulement où et quand les personnes se déplacent, mais aussi pour quel motif. Par conséquent, Lefebvre et al. (2013) ont été en mesure d'analyser la contribution de chaque type de déplacement au cycle quotidien du NO_2. Récemment, Kuehnel et al. (2021) ont notamment montré, dans le cas du bruit du trafic routier à Berlin, que l'exposition et la causalité (i.e. l'impact) sont toutes les deux distribuées inégalement (selon l'indice de Gini) au sein de la population. De plus, puisque seule une partie de la population possède une voiture et conduit, tandis que la quasi totalité des habitations sont situées à proximité de routes, l'inégalité est bien plus importante pour la causalité qu'elle ne l'est pour l'exposition.

En identifiant des différences importantes d'exposition entre des groupes de population ayant des habitudes de déplacement ou des activités différentes, la prise en compte

de la mobilité des individus dans les études sur l'exposition aux émissions des véhicules peut permettre d'identifier des stratégies d'atténuation plus efficaces (Tayarani and Rowangould, 2020). Par conséquent, les décideurs politiques et les praticiens concernés par les impacts des programmes de transport à grande échelle sur la justice environnementale et l'équité devraient donc effectuer des analyses plus détaillées des impacts sur les inégalités lors de la planification des programmes, plutôt que des analyses se concentrant uniquement sur les émissions et expositions moyennes (Kocak et al., 2021). Ces approches dynamiques améliorées (i.e. prenant en compte la mobilité individuelle dans l'estimation de l'exposition) peuvent aider les planificateurs et les décideurs à identifier les populations défavorisées pour lesquelles les expositions sont généralement mal représentées et pourraient conduire à des politiques de planification mieux ciblées (Lu, 2021). Cette modélisation des mobilités individuelles nécessite comme données d'entrée, des informations détaillées relatives aux attributs démographiques, sociaux et économiques d'un nombre important d'individus et également de leurs ménages.

1.3 Production de données d'entrée précises

Il est généralement admis que les besoins de mobilité d'une personne dépendent à la fois de ses caractéristiques individuelles et de sa situation familiale (Loo and Lam, 2013 ; Kalter and Geurs, 2016 ; Fournier et al., 2020). Par exemple, les schémas de mobilité d'une personne à la retraite sont différents de ceux d'une personne professionnellement active. Egalement, les schémas de mobilité d'une personne vivant seule sont différents de ceux des membres d'une famille monoparentale. Il est alors nécessaire de disposer d'informations détaillées à deux niveaux :

1. Le premier niveau concerne les attributs relatifs aux individus (également appelés caractéristiques individuelles) tels que le sexe, l'âge, la profession, le temps de travail, la possession du permis de conduire...
2. Le second niveau est relatif aux attributs des ménages (également appelés caractéristiques des ménages ou de groupe). De tels attributs sont notamment la taille

> du ménage, le type de ménage, le niveau de vie, le nombre de voitures à disposition, la localisation résidentielle... Les individus d'un même ménage partagent les attributs du ménage.

Pour des raisons de confidentialité et de respect de vie privée, les instituts nationaux de statistique mettent rarement à disposition du public ou même des chercheurs une source de données qui contient les caractéristiques sociodémographiques des individus et de leurs ménages à une échelle géographique fine. Seules les caractéristiques d'un échantillon et des statistiques agrégées qui sont généralement des distributions marginales des caractéristiques de la population réelle sont disponibles.

Une étape intermédiaire nécessaire est la génération d'une « population synthétique », représentative de la population réelle à partir des différentes données disponibles. La population synthétique générée constitue alors une représentation microscopique simplifiée de la population réelle puisque seuls des attributs d'intérêt sont reproduits (Chapuis and Taillandier, 2019).

La génération d'une population synthétique d'individus et de ménages (population à deux niveaux) constitue donc la première étape de l'implémentation d'un modèle basé sur les activités. Le processus de génération d'une telle population demeure toutefois assez délicat et dans l'idéal doit respecter certaines conditions précises :

— maintenir la structure hiérarchique des données en associant les attributs des individus et des ménages de la manière la plus optimale possible ;

— refléter l'hétérogénéité de la distribution des ménages et des individus entre les zones géographiques (Münnich and Schürle, 2003) ;

— reproduire les interdépendances entre les agents dans le même ménage (Sun et al., 2018) ;

— correspondre aux données disponibles de la population réelle.

La population synthétique a fait l'objet d'une littérature scientifique importante. Le constat qui ressort pour la majorité de ces études est l'absence de données ou d'une méthodologie qui permettent de reproduire les méthodes proposées et de les adapter selon

les besoins. Les notations et les formulations mathématiques sont souvent différentes d'une étude à l'autre, ce qui ne facilite pas la comparaison des méthodes.

Des récentes études (Erath et al., 2012 ; Bösch et al., 2016 ; Kickhofer et al., 2016 ; Kamel et al., 2019 ; Vosooghi et al., 2019 ; Ziemke et al., 2019 ; Delhoum et al., 2020 ; He et al., 2020 ; Moeckel et al., 2020)... ont essayé d'intégrer les étapes de génération de la population synthétique à partir de données qui ne sont pas librement accessibles ou sans fournir suffisamment de détails pour reproduire les analyses.

A partir de données librement accessibles, Hörl and Balac (2021) ont créé une plateforme dénommée « Eqasim » qui intègre les différentes étapes de modélisation des mobilités quotidiennes (génération d'une population synthétique d'individus et de ménages, plans d'activités, réseau de transport). Eqasim représente une avancée importante pour les utilisatrices et utilisateurs car elle simplifie grandement l'utilisation des systèmes multi-agents basés sur les activités et notamment de MATSim. Eqasim a été initialement appliquée à l'Ile de France en utilisant notamment les données françaises du recensement de la population [10] (Hörl and Balac, 2020). Sallard et al. (2020) au Brésil, Balac and Hörl (2021) aux Etats-Unis, Diallo et al. (2021) et Le Bescond et al. (2021) en France, ont ensuite adapté cette plateforme à leur objet d'étude.

La plateforme offre une solution simple et rapide de génération d'une population de ménages et d'individus synthétiques assortis de leurs plans d'activités. Les concepteurs d'Eqasim reconnaissent toutefois que plusieurs voies d'amélioration sont possibles [11]. Au niveau de la population synthétique en particulier, certains aspects méthodologiques nécessitent d'être approfondis :

— la population synthétique est obtenue en utilisant la variable poids des ménages (variable IPONDI du recensement). La procédure de génération a consisté à directement multiplier chaque ménage et individu de l'échantillon du recensement par le poids de son ménage sans tenir compte des données agrégées du recensement

10. Les différents codes de programmation de Eqasim sont disponibles en ligne (page consultée le 23 septembre 2021) à l'adresse suivante : < `https://github.com/eqasim-org/ile-de-france/tree/v1.2.0`>

11. C'est une des raisons pour lesquelles la plateforme est en libre accès et modifiable par tout utilisateur ou utilisatrice.

pourtant disponibles ;

— l'affectation d'un lieu de résidence à la population synthétique s'effectue de manière simple. Pour chaque ménage synthétique, une adresse aléatoire est choisie parmi l'ensemble des adresses de la commune de résidence du ménage. Nous n'avons aucune garantie que les ménages seront affectés à leur véritable adresse. Le lieu de résidence représente pourtant un attribut important dans la différenciation sociale des ménages et des individus et influence leurs schémas de mobilité ;

— l'allocation d'un attribut supplémentaire tel que le revenu est réalisé de manière aléatoire en utilisant des statistiques agrégées. Une distribution des revenus par déciles est fournie pour chaque commune. Chaque ménage synthétique se voit affecter un décile de revenu de sa commune de résidence avec une probabilité d'allocation de 10%. Ce processus d'allocation qui repose sur une distribution d'échantillonnage aléatoire ne garantit pas la cohérence des données. De nombreux ménages peuvent en effet recevoir un revenu qui ne leur correspond pas au regard de leurs attributs sociodémographiques.

La suite du travail est consacrée à la proposition de solutions permettant à terme de repousser certaines limites évoquées à la fois dans Eqasim et dans la littérature sur le sujet.

Conclusion

La modélisation multi-agents basée sur les activités prône une approche désagrégée dans laquelle les mobilités quotidiennes des individus sont appréhendées dans une optique de réalisation d'activités. Ces modèles nécessitent de disposer des attributs des individus (sexe, âge, profession...) et de leurs ménages (taille, structure familiale, niveau de vie...). Au-delà des limites propres à la modélisation des phénomènes physiques (ou chimiques), une limite à la mesure de l'exposition individuelle aux polluants et au bruit issus du trafic routier en ayant recours à la modélisation multi-agent est actuellement liée au manque

de précision/fiabilité sur les caractéristiques des agents du modèle (caractéristiques socioéconomiques, localisation des résidences et des activités...). Une amélioration de cette partie de la modélisation permettrait d'approfondir l'étude de l'exposition des individus aux nuisances du trafic puisqu'au delà de l'étude de l'exposition d'un sous-groupe de la population (prédéfini selon des critères socioéconomiques, d'activités ou de localisation) sur une période donnée, elle permettrait l'étude des profils individuels d'exposition quotidienne aux nuisances du trafic (tel qu'illustré par Le Bescond et al. (2021) par exemple). La production de données précises à travers une population synthétique générée avec des approches appropriées, des attributs pertinents et une répartition spatiale fine est alors indispensable. La population synthétique a fait l'objet d'une production scientifique importante avec le développement de méthodes, de logiciels commerciaux ou open source. L'analyse de ces méthodes indiquent plusieurs voies d'amélioration notamment dans le processus de choix d'une méthode adéquate, l'ajout d'attributs supplémentaires et dans la répartition spatiale des lieux de résidence ou d'activités.

Les prochains chapitres de la thèse contribuent à ces voies d'amélioration. Dans un premier temps, les principales méthodes de génération de population synthétique à deux niveaux sont présentées de façon détaillée (chapitre 2). La recension de ces méthodes s'achève sur la proposition d'un arbre de décision. Cet arbre permet de faire un choix raisonné entre les différentes méthodes de génération qui existent. Une population synthétique d'individus répartis dans des ménages est créée dans un second temps en comparant différents algorithmes de génération (chapitre 3). La comparaison des algorithmes s'effectue dans un cadre conceptuel commun avec une harmonisation des notations une description détaillée de chaque algorithme. Une fois la population synthétique générée, une méthodologie d'affectation d'attributs supplémentaires à partir de données agrégées est développée (chapitre 4). Cette méthodologie offre la possibilité d'affecter un revenu à un nombre conséquent de ménages synthétiques comme dans la plateforme Eqasim mais de manière plus précise car les attributs des ménages synthétiques sont pris en compte dans le processus d'affectation. Enfin, une approche innovante de spatialisation d'une population synthétique à une échelle géographique plus fine est proposée (chapitre

5). Toutes les données utilisées sont librement accessibles et présentées de façon claire. Un soin particulier a été accordé à leur traitement, dans un souci d'appropriation et d'adaptation des solutions proposées.

Références bibliographiques

Adnan, M., Pereira, F. C., Azevedo, C. L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., and Ben-Akiva, M. (2016). SimMobility : A Multi-scale Integrated Agent-Based Simulation Platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*.

Agarwal, A., Ziemke, D., and Nagel, K. (2020). Bicycle superhighway : An environmentally sustainable policy for urban transport. *Transportation Research Part A : Policy and Practice*, 137 :519–540.

Alho, A. R., You, L., Lu, F., Cheah, L., Zhao, F., and Ben-Akiva, M. (2018). Next-generation freight vehicle surveys : Supplementing truck gps tracking with a driver activity survey. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2974–2979. IEEE.

Andre, M., Shorshani, M. F., Berger, C., Montenon, A., and Brutti-Mairesse, E. (2012). Évaluation des pdu-problématique du calcul des emissions de polluants du trafic.

Antoni, J.-P. (2010). *Modéliser la ville Forme urbaine et politiques de transport*. Economica, collection Méthodes et Approches.

Arentze, T. A. and Timmermans, H. J. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B : Methodological*, 38(7) :613–633.

Avram, S., Figari, F., Leventi, C., Levy, H., Navicke, J., Matsaganis, M., Militaru, E., Paulus, A., Rastringina, O., and Sutherland, H. (2013). The distributional effects of fiscal consolidation in nine eu countries. Technical report, Euromod working paper.

Azevedo, C. L., Deshmukh, N. M., Marimuthu, B., Oh, S., Marczuk, K., Soh, H., Basak, K., Toledo, T., Peh, L.-S., and Ben-Akiva, M. E. (2017). Simmobility short-term : An integrated microscopic mobility simulator. *Transportation Research Record*, 2622(1) :13–23.

Balac, M. (2020). Agent-based scenarios of los angeles and san francisco. Technical report, Working paper. IVT, ETH Zurich, Zurich.

Balac, M. and Hörl, S. (2021). Synthetic population for the state of california based on open-data : examples of san francisco bay area and san diego county. In *100th Annual Meeting of the Transportation Research Board (TRB)*.

Basu, R., Araldo, A., Akkinepally, A. P., Nahmias Biran, B. H., Basak, K., Seshadri, R., Deshmukh, N., Kumar, N., Azevedo, C. L., and Ben-Akiva, M. (2018). Automated Mobility-on-Demand vs. Mass Transit : A Multi-Modal Activity-Driven Agent-Based Simulation Approach. *Transportation Research Record*, 2672(8) :608–618.

Bean, W. L. and Joubert, J. W. (2021). An agent-based implementation of freight receiver and carrier collaboration with cost sharing. *Transportation Research Interdisciplinary Perspectives*, 11 :100416.

Beckx, C., Panis, L. I., Arentze, T., Janssens, D., Torfs, R., Broekx, S., and Wets, G. (2009a). A dynamic activity-based population modelling approach to evaluate exposure to air pollution : methods and application to a dutch urban area. *Environmental Impact Assessment Review*, 29(3) :179–185.

Beckx, C., Panis, L. I., Uljee, I., Arentze, T., Janssens, D., and Wets, G. (2009b). Disaggregation of nation-wide dynamic population exposure estimates in the netherlands : Applications of activity-based transport models. *Atmospheric Environment*, 43(34) :5454–5462.

Beckx, C., Panis, L. I., Vankerkom, J., Janssens, D., Wets, G., and Arentze, T. (2009c). An integrated activity-based modelling framework to assess vehicle emissions : approach and application. *Environment and Planning B : Planning and Design*, 36(6) :1086–1102.

Bösch, P. M., Müller, K., and Ciari, F. (2016). The ivt 2015 baseline scenario. In *16th Swiss Transport Research Conference (STRC 2016)*. 16th Swiss Transport Research Conference (STRC 2016).

BruitParif (2019). Impacts sanitaires du bruit des transports dans la zone dense de la région ile-de-france.

Campbell, H. E., Kim, Y., and Eckerd, A. M. (2015). *Rethinking environmental justice in sustainable cities : Insights from agent-based modeling*. Routledge.

Can, A. (2019). *Dynamic approaches for the characterization and mitigation of urban sound environments*. PhD thesis, Université du Maine.

Chapuis, K. and Taillandier, P. (2019). A brief review of synthetic population generation practices in agent-based social simulation. In *SSC2019, Social Simulation Conference*.

Cucurachi, S., Schiess, S., Froemelt, A., and Hellweg, S. (2018). Noise footprint from personal land-based mobility. *Journal of Industrial Ecology*.

Dale, L. M., Goudreau, S., Perron, S., Ragettli, M. S., Hatzopoulou, M., and Smargiassi, A. (2015). Socioeconomic status and environmental noise exposure in montreal, canada. *BMC public health*, 15(1) :205.

De Nazelle, A., Seto, E., Donaire-Gonzalez, D., Mendez, M., Matamala, J., Nieuwenhuijsen, M. J., and Jerrett, M. (2013). Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental pollution*, 176 :92–99.

Delhoum, Y., Belaroussi, R., Dupin, F., and Zargayouna, M. (2020). Activity-based demand modeling for a future urban district. *Sustainability*, 12(14) :5821.

Dhondt, S., Beckx, C., Degraeuwe, B., Lefebvre, W., Kochan, B., Bellemans, T., Panis, L. I., Macharis, C., and Putman, K. (2012). Health impact assessment of air pollution using a dynamic exposure profile : implications for exposure and health impact estimates. *Environmental impact assessment review*, 36 :42–51.

Diallo, A. O., Doniec, A., Lozenguez, G., and Mandiau, R. (2021). Agent-based simulation from anonymized data : An application to lille metropolis. *Procedia Computer Science*, 184 :164–171.

Dixon, M., Chang, K., Keecheril, S., and Orton, B. (2007). Applying the transims modeling paradigm to the simulation and analysis of transportation and traffic control systems.

Dons, E., Beckx, C., Arentze, T., Wets, G., and Panis, L. I. (2011a). Using an activity-based framework to determine effects of a policy measure on population exposure to nitrogen dioxide. *Transportation research record*, 2233(1) :72–79.

Dons, E., Panis, L. I., Van Poppel, M., Theunis, J., and Wets, G. (2012). Personal exposure to black carbon in transport microenvironments. *Atmospheric Environment*, 55 :392–398.

Dons, E., Panis, L. I., Van Poppel, M., Theunis, J., Willems, H., Torfs, R., and Wets, G. (2011b). Impact of time–activity patterns on personal exposure to black carbon. *Atmospheric Environment*, 45(21) :3594–3602.

Edwards, K. L. and Clarke, G. (2013). Simobesity : combinatorial optimisation (deterministic) model. In *Spatial Microsimulation : A reference guide for users*, pages 69–85. Springer.

Elessa Etuman, A. and Coll, I. (2018). Olympus v1. 0 : development of an integrated air pollutant and ghg urban emissions model–methodology and calibration over greater paris. *Geoscientific Model Development*, 11(12) :5085–5111.

Erath, A., Fourie, P. J., van Eggermond, M. A., Ordonez Medina, S. A., Chakirov, A., and Axhausen, K. W. (2012). Large-scale agent-based transport demand model for singapore. *Arbeitsberichte Verkehrs-und Raumplanung*, 790.

Fournier, N., Christofa, E., Akkinepally, A. P., and Azevedo, C. L. (2020). Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, pages 1–27.

Gao, J. H. and Peh, L.-S. (2014). Roadrunner : Infrastructure-less vehicular congestion control. *Proceedings of the 21st Intelligent Transport Systems World Congress*.

Gao, W., Balmer, M., and Miller, E. J. (2010). Comparison of matsim and emme/2 on greater toronto and hamilton area network, canada. *Transportation Research Record*, 2197(1) :118–128.

Gerike, R., Becker, T., Friedemann, J., Hülsmann, F., and Heidegger, F. (2012). Mapping external noise costs to the transport usersconceptual issues and empirical results. In *Proceedings of the Euronoise 9th European conference on noise control, Prague*, pages 10–13.

Gurram, S., Stuart, A. L., and Pinjari, A. R. (2015). Impacts of travel activity and urbanicity on exposures to ambient oxides of nitrogen and on exposure disparities. *Air Quality, Atmosphere & Health*, 8(1) :97–114.

Gurram, S., Stuart, A. L., and Pinjari, A. R. (2019). Agent-based modeling to estimate exposures to urban air pollution from transportation : Exposure disparities and impacts of high-resolution data. *Computers, Environment and Urban Systems*, 75 :22–34.

Hao, J. Y., Hatzopoulou, M., and Miller, E. J. (2010). Integrating an activity-based travel demand model with dynamic traffic assignment and emission models : Implementation in the greater toronto, canada, area. *Transportation Research Record*, 2176(1) :1–13.

Hatzopoulou, M., Hao, J. Y., and Miller, E. J. (2011). Simulating the impacts of household travel on greenhouse gas emissions, urban air quality, and population exposure. *Transportation*, 38(6) :871.

Hatzopoulou, M. and Miller, E. J. (2010). Linking an activity-based travel demand model with traffic emission and dispersion models : Transports contribution to air pollution in toronto. *Transportation Research Part D : Transport and Environment*, 15(6) :315–325.

Havard, S., Reich, B. J., Bean, K., and Chaix, B. (2011). Social inequalities in residential exposure to road traffic noise : an environmental justice analysis based on the record cohort study. *Occupational and environmental medicine*, 68(5) :366–374.

He, B. Y., Zhou, J., Ma, Z., Chow, J. Y., and Ozbay, K. (2020). Evaluation of city-scale built environment policies in new york city with an emerging-mobility-accessible synthetic population. *Transportation Research Part A : Policy and Practice*, 141 :444–467.

Hörl, S. (2021). Integrating discrete choice models with matsim scoring. *Procedia Computer Science*, 184 :704–711.

Hörl, S. and Balac, M. (2020). Reproducible scenarios for agent-based transport simulation. *Arbeitsberichte Verkehrs Raumplan*, 1499.

Hörl, S. and Balac, M. (2021). Synthetic population and travel demand for paris and île-de-france based on open and publicly available data. *Transportation Research Part C : Emerging Technologies*, 130 :103291.

Hörl, S., Balac, M., and Axhausen, K. W. (2018). A first look at bridging discrete choice modeling and agent-based microsimulation in matsim. *Procedia computer science*, 130 :900–907.

Horni, A., Nagel, K., and Axhausen, K. W. (2016). Introducing matsim. In Horni, A., Nagel, K., and Axhausen, K. W., editors, *The Multi-Agent Transport Simulation MATSim*, pages 3–7. Ubiquity Press.

Houot, H., Antoni, J.-P., Pujol, S., Mauny, F., and Lamiral, M. (2015). Les mobilités urbaines et leur impact sur lexposition au bruit : simulation de scénarios prospectifs appliqués à la ville de besançon. *Transports urbains*, (126) :16–20.

Hülsmann, F. (2014). *Integrated agent-based transport simulation and air pollution modelling in urban areas-the example of Munich*. PhD thesis, Technische Universität München.

Illenberger, J., Flotterod, G., and Nagel, K. (2007). Enhancing matsim with capabilities of within-day re-planning. In *2007 IEEE Intelligent Transportation Systems Conference*, pages 94–99. IEEE.

Jeihani, M. and Ardeshiri, A. (2014). Transims implementation for a small network and comparison with enhanced four-step model. *Journal of the Transportation Research Forum*, 53(1).

Jerrett, M., Burnett, R. T., Kanaroglou, P., Eyles, J., Finkelstein, N., Giovis, C., and Brook, J. R. (2001). A gis–environmental justice analysis of particulate air pollution in hamilton, canada. *Environment and Planning A*, 33(6) :955–973.

Kaddoura, I., Bischoff, J., and Nagel, K. (2020). Towards welfare optimal operation of innovative mobility concepts : External cost pricing in a world of shared autonomous vehicles. *Transportation Research Part A : Policy and Practice*, 136 :48–63.

Kaddoura, I., Kröger, L., and Nagel, K. (2017a). An activity-based and dynamic approach to calculate road traffic noise damages. *Transportation Research Part D : Transport and Environment*, 54 :335–347.

Kaddoura, I., Kröger, L., and Nagel, K. (2017b). User-specific and dynamic internalization of road traffic noise exposures. *Networks and Spatial Economics*, 17(1) :153–172.

Kaddoura, I. and Nagel, K. (2018). Simultaneous internalization of traffic congestion and noise exposure costs. *Transportation*, pages 1–22.

Kagho, G. O., Balac, M., and Axhausen, K. W. (2020). Agent-based models in transport planning : Current state, issues, and expectations. *Procedia Computer Science*, 170 :726–732.

Kalter, M.-J. O. and Geurs, K. T. (2016). Exploring the impact of household interactions on car use for home-based tours : a multilevel analysis of mode choice using data from the first two waves of the netherlands mobility panel. *European journal of transport and infrastructure research*, 16(4) :698–712.

Kamel, J., Vosooghi, R., Puchinger, J., Ksontini, F., and Sirin, G. (2019). Exploring the impact of user preferences on shared autonomous vehicle modal split : A multi-agent simulation approach. *Transportation Research Procedia*, 37 :115–122.

Kickhofer, B., Hosse, D., Turnera, K., and Tirachinic, A. (2016). Creating an open matsim scenario from open data : The case of santiago de chile. *http ://www. vsp. tuberline. de/publication : TU Berlin, Transport System Planning and Transport Telematics.*

Kocak, T. K., Gurram, S., Bertini, R. L., and Stuart, A. L. (2021). Impacts of a metropolitan-scale freeway expansion program on air pollution and equity. *Journal of Transport & Health*, 22 :101114.

Kuehnel, N., Huang, W.-C., and Moeckel, R. (2021). Environmental equity analysis in agent-based transport simulations : A study on causation and exposure. *Procedia Computer Science*, 184 :650–655.

Kuehnel, N., Kaddoura, I., and Moeckel, R. (2019). Noise shielding in an agent-based transport model using volunteered geographic data. *Procedia Computer Science*, 151 :808–813.

Kuehnel, N. and Moeckel, R. (2020). Impact of simulation-based traffic noise on rent prices. *Transportation Research Part D : Transport and Environment*, 78 :102191.

Lawe, S., Lobb, J., Sadek, A. W., Huang, S., and Xie, C. (2009). Transims implementation in chittenden county, vermont : development, calibration, and preliminary sensitivity analysis. *Transportation Research Record*, 2132(1) :113–121.

Le Bescond, V., Can, A., Aumond, P., and Gastineau, P. (2021). Open-source modeling chain for the dynamic assessment of road traffic noise exposure. *Transportation Research Part D : Transport and Environment*, 94 :102793.

Lefebvre, W., Degrawe, B., Beckx, C., Vanhulsel, M., Kochan, B., Bellemans, T., Janssens, D., Wets, G., Janssen, S., De Vlieger, I., et al. (2013). Presentation and evaluation of an integrated model chain to respond to traffic-and health-related policy questions. *Environmental modelling & software*, 40 :160–170.

Loo, B. P. and Lam, W. (2013). A multilevel investigation of differential individual

mobility of working couples with children : a case study of hong kong. *Transportmetrica A : Transport Science*, 9(7) :629–652.

Lu, Y. (2021). Beyond air pollution at home : Assessment of personal exposure to pm_2.5 using activity-based travel demand model and low-cost air sensor network data. *Environmental Research*, page 111549.

Lu, Y., Adnan, M., Basak, K., Pereira, F. C., Carrion, C., Saber, V. H., Loganathan, H., and Ben-Akiva, M. E. (2015). Simmobility mid-term simulator : A state of the art integrated agent based demand and supply model. In *94th Annual Meeting of the Transportation Research Board, Washington, DC*.

Marczuk, K. A., Hong, H. S. S., Azevedo, C. M. L., Adnan, M., Pendleton, S. D., Frazzoli, E., et al. (2015). Autonomous mobility on demand in simmobility : Case study of the central business district in singapore. In *2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pages 167–172. IEEE.

Moeckel, R., Kuehnel, N., Llorca, C., Moreno, A. T., and Rayaprolu, H. (2020). Agent-based simulation to improve policy sensitivity of trip-based models. *Journal of Advanced Transportation*, 2020.

Münnich, R. and Schürle, J. (2003). On the simulation of complex universes in the case of applying the german microcensus. *DACSEIS Research Paper Series No. 4*.

Oh, S., Lentzakis, A. F., Seshadri, R., and Ben-Akiva, M. (2021). Impacts of automated mobility-on-demand on traffic dynamics, energy and emissions : A case study of singapore. *Simulation Modelling Practice and Theory*, 110 :102327.

O'Sullivan, D. (2008). Geographical information science : agent-based models. *Progress in Human Geography*, 32(4) :541–550.

Othman, A. (2016). *Simulation multi-agent de l'information des voyageurs dans les transports en commun*. PhD thesis, Université Paris-Est.

Park, Y. M. (2020). Assessing personal exposure to traffic-related air pollution using individual travel-activity diary data and an on-road source air dispersion model. *Health & Place*, 63 :102351.

Park, Y. M. and Kwan, M.-P. (2017). Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. *Health & place*, 43 :85–94.

Park, Y. M. and Kwan, M.-P. (2020). Understanding racial disparities in exposure to traffic-related air pollution : Considering the spatiotemporal dynamics of population distribution. *International journal of environmental research and public health*, 17(3) :908.

Pascal, M., de Crouy Chanel, P., Corso, M., Medina, S., Wagner, V., Goria, S., Beaudeau, P., Bentayeb, M., Le Tertre, A., Ung, A., et al. (2016). Impacts de l'exposition chronique aux particules fines sur la mortalité en france continentale et analyse des gains en santé de plusieurs scénarios de réduction de la pollution atmosphérique. *Saint-Maurice : Santé publique France*.

Picaut, J. (2011). Projet anr eval-pdu : Indicateurs du bruit dans lenvironnement–synthèse bibliographique.

Rasouli, S. and Timmermans, H. (2014). Activity-based models of travel demand : promises, progress and prospects. *International Journal of Urban Sciences*, 18(1) :31–60.

Recker, W. W. and Parimi, A. (1999). Development of a microscopic activity-based framework for analyzing the potential impacts of transportation control measures on vehicle emissions. *Transportation Research Part D : Transport and Environment*, 4(6) :357–378.

Rickert, M. and Nagel, K. (2001). Dynamic traffic assignment on parallel computers in transims. *Future generation computer systems*, 17(5) :637–648.

Sallard, A., Balać, M., and Hörl, S. (2020). A synthetic population for the greater são paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, 1545.

Shiftan, Y. (2000). The advantage of activity-based modelling for air-quality purposes : theory vs practice and future needs. *Innovation : The European Journal of Social Science Research*, 13(1) :95–110.

Shiftan, Y., Kheifits, L., and Sorani, M. (2015). Travel and emissions analysis of sustainable transportation policies with activity-based modeling. *Transportation Research Record*, 2531(1) :93–102.

Shiftan, Y. and Suhrbier, J. (2002). The analysis of travel and emission impacts of travel demand management strategies using activity-based models. *Transportation*, 29(2) :145–168.

Sun, L., Erath, A., and Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B : Methodological*, 114 :199–212.

Sutherland, H. and Figari, F. (2013). Euromod : the european union tax-benefit microsimulation model. *International Journal of Microsimulation*, 6(1) :4–26.

Tayarani, M. and Rowangould, G. (2020). Estimating exposure to fine particulate matter emissions from vehicle traffic : Exposure misclassification and daily activity patterns in a large, sprawling region. *Environmental research*, 182 :108999.

Tomintz, M. N., Clarke, G. P., and Rigby, J. E. (2008). The geography of smoking in leeds : estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, 40(3) :341–353.

Vallamsundar, S., Lin, J., Konduri, K., Zhou, X., and Pendyala, R. M. (2016). A comprehensive modeling framework for transportation-induced population exposure assessment. *Transportation Research Part D : Transport and Environment*, 46 :94–113.

Vallée, J. (2017). Challenges in targeting areas for public action. target areas at the right place and at the right time.

Vosooghi, R., Kamel, J., Puchinger, J., Leblond, V., and Jankovic, M. (2019). Robotaxi service fleet sizing : assessing the impact of user trust and willingness-to-use. *Transportation*, 46(6) :1997–2015.

Yang, Y., Metcalf, S., and Mao, L. (2020). Modeling transit-assisted hurricane evacuation through socio-spatial networks. *International Journal of Geographical Information Science*, pages 1–18.

Zhou, M., Le, D.-T., Nguyen-Phuoc, D. Q., Zegras, P. C., and Ferreira Jr, J. (2021). Simulating impacts of automated mobility-on-demand on accessibility and residential relocation. *Cities*, 118 :103345.

Ziemke, D., Kaddoura, I., and Nagel, K. (2019). The matsim open berlin scenario : A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data. *Procedia computer science*, 151 :870–877.

Zwick, F., Kuehnel, N., Moeckel, R., and Axhausen, K. W. (2021). Agent-based simulation of city-wide autonomous ride-pooling and the impact on traffic noise. *Transportation Research Part D : Transport and Environment*, 90 :102673.

Chapitre 2

Comparaison des méthodes de génération d'une population synthétique à deux niveaux

Sommaire

Résumé

La littérature scientifique recense trois principales catégories de méthodes de génération de population synthétique à deux niveaux. Ces catégories sont la reconstruction synthétique (SR), l'optimisation combinatoire (CO) et l'apprentissage statistique (SL). Les méthodes SR et CO produisent des populations synthétiques à partir d'une réplication d'individus et de ménages, tandis que les méthodes SL génèrent une population synthétique à travers une estimation de probabilités jointes associant les attributs individuels et ménages. Le choix d'une de ces méthodes peut être délicat car conditionné aux données disponibles et aux caractéristiques de la population à générer. Les objectifs de ce chapitre sont alors 1) de fournir une description détaillée de ces méthodes, 2) de procéder à leur comparaison et 3) de proposer une procédure de décision permettant de choisir une méthode. La description et la comparaison des méthodes s'appuient sur de nombreux critères. Les avantages et les inconvénients de chaque méthode sont également illustrés. La procédure de décision est réalisée à travers un arbre de décision.

Mots clés

Génération de population synthétique, multiniveau, microsimulation, méthodes.

Introduction

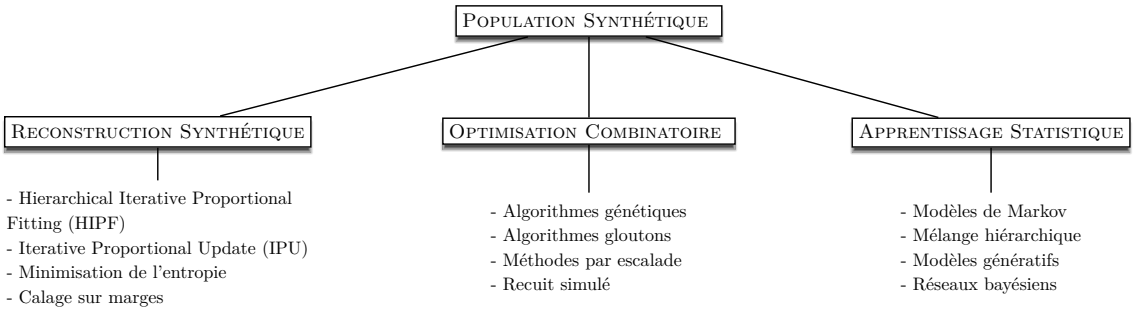
La littérature scientifique distingue deux approches méthodologiques différentes de génération d'une population synthétique à deux niveaux. La première approche, dite approche des méthodes sans échantillon (Gargiulo et al., 2010 ; Barthelemy and Toint, 2013 ; Lenormand and Deffuant, 2013 ; Huynh et al., 2016) ne nécessite pas d'échantillon et ne prend en compte que des données agrégées. Dans la deuxième approche désignée sous le terme de méthodes avec échantillon, un échantillon d'individus et de ménages est utilisé pour contrôler la distribution jointe des attributs tout en générant la population. La grande majorité des méthodes de génération de la littérature scientifique exploitent un échantillon auquel il est possible ou non d'associer des données agrégées.

Diverses classifications de méthodes issues de cette approche ont alors été proposées (Templ et al., 2017 ; Sun et al., 2018 ; Chapuis and Taillandier, 2019). La catégorisation adoptée dans ce chapitre est dérivée de celle de Sun et al. (2018) qui considèrent trois grands groupes ou catégories de génération : la reconstruction synthétique (Synthetic Reconstruction en anglais, abrégée SR), l'optimisation combinatoire (Combinatorial Optimisation ou CO en anglais) et l'apprentissage statistique (Statistical Learning ou SL en anglais).

Les méthodes SR et CO produisent des populations synthétiques à partir d'une réplication d'individus et de ménages, tandis que les méthodes SL génèrent une population synthétique à travers une estimation de probabilités jointes associant les attributs individuels et ménages. Les méthodes CO et SL sont stochastiques contrairement aux méthodes SR qui sont déterministes. La Figure 2.1 présente les principales catégories ainsi que les méthodes associées.

L'objectif de ce chapitre est de présenter et d'évaluer les différentes méthodes de génération à deux niveaux. Dans la littérature scientifique, quelques études ont procédé à un examen de ces méthodes. Hermes and Poulsen (2012) ont analysé les méthodes de génération mais seulement à un seul niveau : soit le ménage, soit l'individu. Ils n'ont pas également considéré les méthodes SL. Müller and Axhausen (2010, 2012) ont passé

FIGURE 2.1 : Méthodes de génération d'une population synthétique à deux niveaux (ménages et individus)



en revue les méthodes à deux niveaux mais n'ont couvert que les méthodes SR. Enfin, Chapuis and Taillandier (2019) ont également analysé les méthodes de génération à deux niveaux en les classant en deux catégories principales (ils regroupent les méthodes SR et SL en une seule catégorie et les méthodes CO dans une seconde catégorie distincte). Cependant, leurs recherches se sont limitées aux travaux uniquement publiés dans la revue JASSS (Journal of Artificial Societies and Social Simulation) sur une période de 5 ans. Aucune de ces différentes études ne proposait de comparaison détaillée ni de procédure de décision entre les méthodes SR, CO et SL. Notre contribution peut se résumer en trois points essentiels. Il s'agit de :

1. fournir une description détaillée des principales méthodes et des algorithmes relatifs ;
2. procéder à leur comparaison ;
3. proposer une procédure de décision permettant de choisir une méthode.

La description et la comparaison des méthodes s'appuient sur différents critères : disponibilité des données agrégées ou données marginales, taille de l'échantillon disponible, nombre d'attributs (caractéristiques) potentiels pouvant être générés, taille de la population à générer... Les avantages et les inconvénients de chaque méthode sont illustrés et leurs performances respectives sont également évaluées. La procédure de décision est réalisée à travers un arbre de décision. Les chercheurs et les praticiens ont maintenant accès à un cadre standardisé et complet pour sélectionner la méthode appropriée en

fonction de nombreux critères et de leurs objectifs de modélisation.

Le reste de ce chapitre est organisé comme suit. Les méthodes seront décrites selon leur classification en trois catégories (une section consacrée à chaque catégorie). Dans les deux dernières sections (section 2.4 et section 2.5), une comparaison des différentes méthodes est établie et un arbre de décision est proposé.

2.1 Méthodes de Reconstruction synthétique (SR)

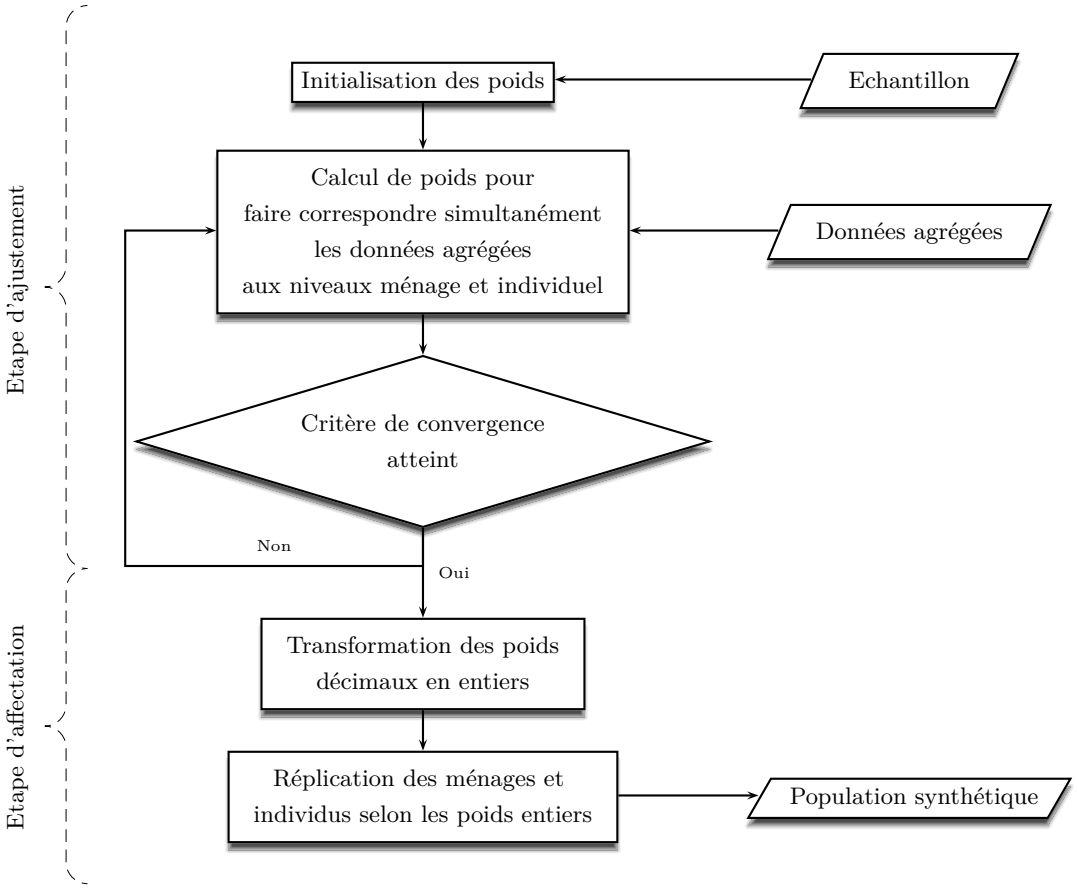
2.1.1 Présentation générale

Les méthodes SR sont les plus utilisées pour générer des populations synthétiques. Elles reposent sur une procédure de génération en deux étapes : *l'ajustement* et *l'affectation*. L'étape d'ajustement consiste à attribuer des poids positifs aux individus et aux ménages d'un échantillon de la population à générer, de telle sorte que la somme de ces poids corresponde aux données marginales (ou agrégées) de la population. Les poids résultants sont généralement non entiers (c'est-à-dire des fractions d'individus et de ménages). Au cours de l'étape d'affectation, ces poids non entiers sont convertis en poids entiers et les individus sont répliqués proportionnellement à leurs poids. Par exemple, un individu qui se voit attribuer un poids de 8 sera répliqué huit fois dans la population synthétique. Une fois cette procédure terminée, chaque agent synthétique a des attributs clairement définis. Ces deux étapes constituent le fondement de l'approche SR (Templ et al., 2017). La Figure 2.2 présente un diagramme simplifié de la mise en oeuvre de ces étapes.

Les méthodes SR sont déterministes, ce qui signifie que selon l'échantillon utilisé, les poids obtenus lors de la phase d'ajustement ne varient pas. La plupart de ces méthodes nécessitent donc de disposer à la fois d'un échantillon et de données agrégées. L'hypothèse sous-jacente est que l'échantillon traduit la représentation réelle de la population (Farooq et al., 2013), et que les associations entre les attributs individuels et des ménages sont dans une large mesure préservées pour les agents synthétiques (Müller and Axhausen, 2010).

Pour cela, l'échantillon doit être représentatif (permettre une description valide de la population réelle) et complet (contenir au moins une observation pour chaque type de ménages et d'individus de la population réelle). Cette dernière contrainte permet d'éviter le problème des observations présentes dans la population et absentes dans l'échantillon [1].

FIGURE 2.2 : Diagramme simplifié des méthodes de reconstruction synthétique



1. Dans la littérature scientifique, ce problème est appelé « zero-cell problem ».

2.1.2 L'ajustement itératif proportionnel (IPF) et ses adaptations dans la génération d'une population à deux niveaux

L'ajustement itératif proportionnel (en anglais Iterative Proportional Fitting en abrégé IPF) demeure l'algorithme SR par excellence. Il consiste à ajuster et faire correspondre un tableau de contingence d'attributs (construit à partir de l'échantillon) aux distributions marginales. Il s'agit d'un processus itératif qui a pour objectif de minimiser les écarts entre le tableau et les valeurs marginales. Les lecteurs intéressés sont invités à consulter Beckman et al. (1996) et Pritchard and Miller (2012) pour une description complète de cet algorithme. L'IPF présente plusieurs avantages, le premier étant la garantie que les attributs agrégés de la population synthétique corresponde aux données agrégées. La structure de la population est donc préservée (Rich and Mulalic, 2012). Le second avantage est relatif au fait que les attributs qui n'entrent pas dans le processus de génération sont raisonnablement estimés (Beckman et al., 1996). L'algorithme est relativement simple, rapide et précis (Lovelace and Ballas, 2013 ; Choupani and Mamdoohi, 2016). Toutefois, dans sa formulation originale, l'IPF ne peut pas estimer simultanément les attributs au niveau du ménage et de l'individu.

Certaines approches basées sur l'IPF ont cherché à contourner ce problème. Arentze et al. (2007) ont proposé une procédure en deux étapes : dans la première étape, ils utilisent l'IPF pour générer une population synthétique d'individus (à partir d'un échantillon d'individus). Les distributions marginales des individus obtenues sont converties en distribution marginales des ménages (pour quelques attributs) en utilisant une matrice de relation individus-ménages. Dans la seconde étape, ces auteurs utilisent un échantillon de ménages et toujours avec l'IPF, génèrent une population synhtétique de ménages qui respectent les distributions marginales des ménages obtenues dans l'étape 1. Ils obtiennent à la fin une population synthétique de ménages avec quelques caractéristiques individuelles. Guo and Bhat (2007) et Auld and Mohammadian (2010) ont estimé les distributions jointes des attributs des ménages et des individus séparément avec l'IPF. De façon itérative, ils associent ménages synthétiques et individus synthétiques tout en s'assurant que les données marginales individuelles et ménages sont respectées. Dans le

processus de génération proposé par Pritchard and Miller (2012), une population synthétique d'individus et de ménages est également construite séparément avec l'IPF. Un ensemble d'attributs communs aux niveaux individuel et ménage sont sélectionnés. Les auteurs utilisent l'IPF et génèrent une population de ménages qui respecte les marginaux des attributs partagés. A l'étape de l'affectation, les ménages et les individus sont appariés par le biais de ces attributs partagés, et une méthode de Monte Carlo conditionnelle est utilisée pour affecter les personnes aux ménages (Zhu and Ferreira Jr, 2014).

Il ressort de toutes ces études que la distribution jointe des attributs au niveau des ménages et des individus n'est jamais ajustée simultanément. L'ajustement entre les deux niveaux s'effectue soit séparément, soit de façon itérative, ce qui ne garantit pas la cohérence entre ces deux niveaux. Egalement le nombre d'attributs individuels et ménages obtenu à la fin du processus est souvent limité.

2.1.3 Algorithmes de génération simultanée à deux niveaux

L'Iterative Proportional Update (IPU), le Hierarchical IPF (HIPF), la minimisation de l'entropie (ent) et le calage sur marges (GR pour Generalized Raking en anglais) sont des algorithmes SR qui prennent en compte de façon simultanée le niveau individuel et ménage dans le processus de génération de la population synthétique.

L'IPU (Ye et al., 2009) et l'HIPF (Müller and Axhausen, 2012 ; Müller, 2017) génèrent des populations synthétiques d'individus rattachés à des ménages qui respectent à la fois les données marginales des deux niveaux. Ces deux algorithmes estiment des poids d'individus et de ménages de manière itérative avec une catégorisation croisée des attributs individuels et des ménages (Chapuis and Taillandier, 2019). L'algorithme HIPF passe constamment du domaine des ménages à celui des personnes, à travers une étape d'ajustement qui consiste à optimiser l'entropie (Müller and Axhausen, 2011). Dans l'algorithme IPU, les poids sont d'abord ajustés pour satisfaire les contraintes au niveau du ménage, puis actualisés pour satisfaire les contraintes au niveau des individus.

La génération d'une population synthétique peut également être formulée comme un problème d'optimisation sous contraintes. L'objectif de cette formulation est d'attri-

buer un poids à chaque ménage de telle sorte que les distributions des caractéristiques des individus et des ménages dans l'échantillon pondéré correspondent aux distributions des individus et des ménages de la population totale (Bar-Gera et al., 2009). Ce problème d'optimisation peut être résolu en utilisant deux méthodes : une minimisation de l'entropie croisée [2] (ent), appelée Minxent dans la littérature ou un calage sur marges.

La procédure de minimisation de l'entropie croisée recherche la configuration la plus adéquate d'éléments dans une situation sous contraintes (Lee and Fu, 2011). Cette méthode optimise directement une métrique de similarité des poids à partir des données disponibles (Müller, 2017). Le calage sur marges (GR) permet de pondérer des observations (individus ou ménages) d'un échantillon en utilisant une information auxiliaire disponible sur certains attributs appelés attributs de calage (Deville et al., 1993). Dans le processus de génération d'une population synthétique, les attributs de calage représentent les données agrégées de la population réelle. Une fonction de distance qui mesure l'écart entre les poids initiaux et les poids finaux sous les contraintes des attributs de calage est choisie. Pour une attribut de calage qualitatif (exemple de l'attribut sexe de l'individu), les effectifs des catégories (hommes, femmes) dans l'échantillon seront après pondération égaux aux effectifs des hommes et femmes de la population. Les fonctions de distance peuvent être de type linéaire, raking ratio, logit ou linéaire tronquée. Dans le cadre d'une génération de population à deux niveaux, l'agorithme GR ajuste des poids pour directement satisfaire les contraintes au niveau des individus et des ménages.

Une grande partie des générateurs de population synthétique à deux niveaux utilisent des algorithmes SR. A titre illustratif, PopulationSim (Paul et al., 2018) et PopSynIII (Vovsha et al., 2015) sont tous deux basés sur la procédure d'optimisation de l'entropie ; TransCAD (Balakrishnaa et al., 2020) utilise une version améliorée de l'IPU. Pour une connaissance plus approfondie des caractéristiques de certains générateurs, les lecteurs intéressés peuvent consulter Müller (2017) (chapitre 2).

2. Nous reprenons le terme d'entropie croisée qui est utilisée dans la littérature mais la minimisation porte sur la divergence de Küllback-Leibler.

2.2 Méthodes d'optimisation combinatoire(CO)

2.2.1 Présentation générale

La deuxième catégorie de méthodes fait référence à l'optimisation combinatoire (CO). Les méthodes CO peuvent générer directement une liste de ménages et de personnes synthétiques (Ma and Srinivasan, 2015). Tout comme les méthodes SR, ces méthodes nécessitent de disposer à la fois d'un échantillon et de données agrégées de la population réelle. La population synthétique est également obtenue par une réplication des individus et ménages. En revanche, il existe des différences entre ces deux méthodes : les méthodes CO sont stochastiques ; elles génèrent directement des individus et des ménages entiers. Il n'est donc pas nécessaire de procéder à une étape d'affectation comme dans les méthodes SR. Egalement, les exigences en matière de données restent moins restrictives que celles des méthodes SR (Templ et al.,2017). Un inconvénient majeur des méthodes CO est leur complexité de calcul pour les grandes tailles de population. En effet, lorsque la taille de la population augmente, les méthodes CO ne peuvent pas garantir une solution optimale et nécessitent un temps de calcul conséquent (Lee and Fu, 2011). Pour cette raison, ces méthodes ne sont pas largement utilisées.

2.2.2 Description des techniques de génération

Une description générale du processus de génération des méthodes CO a été donnée par Voas and Williamson (2000) et Templ et al. (2017). L'échantillon d'invidus et de ménages est dans un premier temps divisé en groupes mutuellement exclusifs pour lesquels des données agrégées (ou données marginales) existent. La constitution de ces groupes peut éventuellement se réaliser sur le critère géographique (on regroupe tous les ménages et individus d'une même zone géographique). Le processus devient alors itératif :

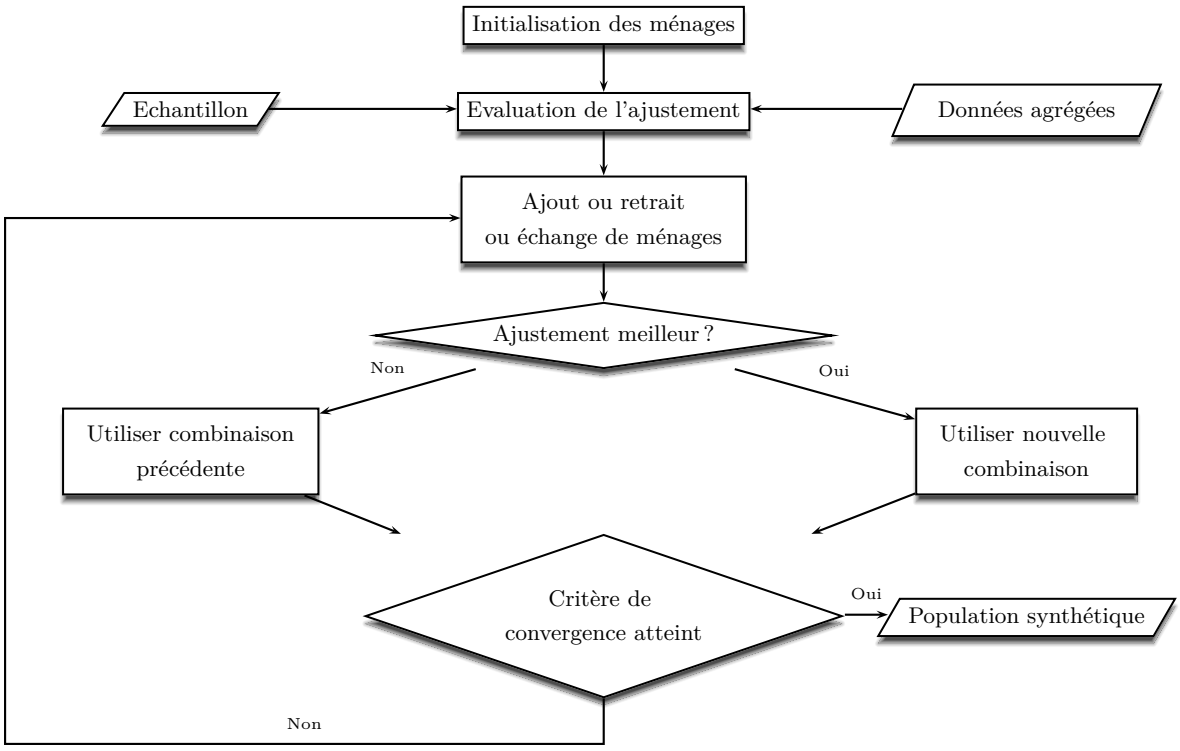
— un ensemble de ménages est aléatoirement choisi dans chaque groupe de manière à correspondre à la taille de la population du groupe. On évalue l'ajustement des ménages sélectionnés aux données agrégées avec une mesure statistique. Une mesure statistique de l'ajustement qui pourrait être utilisé est la somme relative

globale des scores Z au carré (RSSZ) proposée par Voas and Williamson (2001) et utilisée par Ryan et al. (2009).

— un ménage est soit ajouté, soit retiré, soit remplacé par un autre ménage du groupe. La qualité de l'ajustement est de nouveau évaluée. Si cette procédure améliore l'ajustement, ce ménage est conservé ; dans le cas contraire, ce ménage est retiré du groupe. Cette procédure est répétée jusqu'à ce que le critère de convergence fixé est atteint. Il est également possible d'arrêter la procédure à partir d'un certain nombre d'itérations.

Divers algorithmes de génération par optimisation combinatoire ont été proposés dans la littérature : algorithmes génétiques (Birkin et al., 2006), algorithmes gloutons, (Srinivasan et al., 2008), méthode par escalade (Abraham et al., 2012), et le recuit simulé (Harland et al., 2012). La Figure 2.3 présente un diagramme simplifié du fonctionnement des méthodes CO.

FIGURE 2.3 : Diagramme simplifié des méthodes d'optimisation combinatoire



La plupart des études utilisant les méthodes CO n'ont pas cherché à générer une population synthétique à deux niveaux. Les travaux se sont concentrés sur un niveau donné avec un nombre restreint d'attributs générés. Par exemple, Ryan et al. (2009) ont montré que les méthodes CO produisent des populations plus précises que l'IPF, bien que leur étude ait été menée sur une population relativement petite (30 000 individus) et seulement trois attributs de niveau individuel. Abraham et al. (2012) avec la méthode par escalade, Ma and Srinivasan (2015) avec un algorithme personnel (Fitness-Based Syntesis) et Murata et al. (2017) avec l'algorithme de recuit simulé ont cherché à créer des populations synthétiques en contrôlant à la fois le niveau ménage et individuel sans garantie que la cohérence entre individus et ménages est respectée.

2.3 Apprentissage statistique

2.3.1 Présentation générale

La troisième catégorie de méthodes permettant de générer une population synthétique à deux niveaux est l'apprentissage statistique (SL). Egalement désignée sous le nom d'approche basée sur la simulation, les méthodes SL s'intéressent à la distribution jointe de tous les attributs dans l'échantillon en estimant directement une probabilité pour chaque combinaison de ménages et d'individus, y compris celles qui ne sont pas observées dans l'échantillon (Sun et al., 2018). Farooq et al. (2013) et Sun et al. (2018) énoncent les principes des méthodes SL. Dans la population réelle, les individus sont représentés par des attributs discrets ou continus avec une distribution jointe unique. Pour des raisons de confidentialité et de disponibilité des données, seules des vues partielles (échantillon, données marginales simples ou croisées) de cette distribution jointe existent. Les méthodes SL cherchent à exploiter les informations partielles disponibles pour construire une population synthétique dont la distribution empirique doit être aussi proche que la distribution jointe des individus de la population réelle.

Ces méthodes offrent une plus grande flexibilité que les méthodes SR et CO par rapport aux données d'entrée requises. Elles fournissent de bons résultats lorsqu'il s'agit

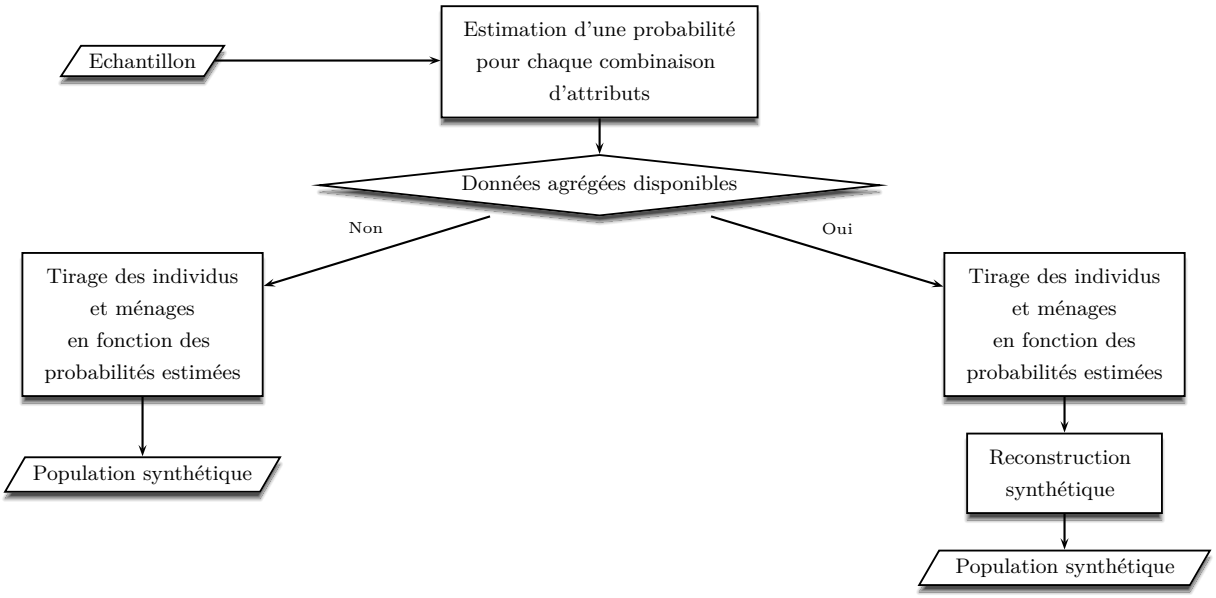
de traiter à la fois le problème de l'absence d'hétérogénéité (Sun et al., 2018) et les échantillons de petite taille. Ces méthodes sont également capables de reproduire des agents synthétiques qui ne sont pas présents dans l'échantillon. Les méthodes SL peuvent également générer une population synthétique à partir de données issues uniquement de l'échantillon lorsque l'information agrégée n'est pas disponible.

Cependant, un inconvénient majeur des méthodes SL est de ne pas parvenir à faire correspondre les distributions conditionnelles et les distributions marginales (données agrégées) de tous les attributs simultanément. Lorsque les données marginales de la population réelle sont disponibles, il est pourtant indispensable que les données agrégées de la population synthétique soient le plus proche possible des données agrégées de la population réelle. Dans certaines configurations, une combinaison des méthodes SL et SR s'avère être le moyen le plus pertinent pour générer une population synthétique. Les méthodes SL peuvent être appliquées en premier lieu pour construire un population synthétique de grande taille (Sun and Erath, 2015 ; Sun et al., 2018). Les méthodes SR sont par la suite introduites pour obtenir une population synthétique correspondant aux marginaux. La Figure 2.4 présente un diagramme simplifié du fonctionnement des méthodes SL. Différentes méthodes SL ont été proposées pour la génération de populations synthétiques.

2.3.2 Méthodes basées sur les modèles de Markov

Farooq et al. (2013) ont suggéré d'utiliser l'algorithme de l'échantillonneur de Gibbs et une méthode de Monte-Carlo par chaînes de Markov (méthodes MCMC pour Markov Chain Monte Carlo en anglais) pour générer la population synthétique. Cette méthode utilise les distributions marginales conditionnelles disponibles pour simuler le tirage de la distribution jointe des attributs de la population réelle. Les auteurs ont employé des modèles de choix discrets pour construire ces distributions marginales conditionnelles. Une fois les distributions obtenues, l'échantillonneur de Gibbs est exécuté et une population synthétique est obtenue en tirant le nombre d'individus ou de ménages correspondant à la taille de la population requise.

FIGURE 2.4 : Diagramme simplifié des méthodes d'apprentissage statistique



Saadi et al. (2016) ont proposé un modèle de Markov caché (HMM pour Hidden Markov model en anglais). Les auteurs considèrent que chaque catégorie d'un attribut représente un état caché latent. Par exemple, l'attribut sexe représenté par deux catégories (homme, femme) possède deux états. Les attributs continus (âge, revenu...) sont discrétisés. Le nombre d'états est donc supposé connu. Ces états sont formés dans un processus markovien : chaque état est entièrement déterminé par un état précédent. Les taux de transition (passage) entre différents états sont mesurés par des tables de probabilités conditionnelles. Les attributs considérés dans le processus de génération sont alors positionnés en séquences suivant un ordre précis. Considérons par exemple la séquence suivante : âge, sexe, profession. Dans une première étape, l'attribut âge est considéré. Toutes les états de cet attribut sont alors déterminés à partir d'une distribution de probabilité de l'âge. Dans une deuxième étape, le sexe de l'individu est considéré. La transition de l'âge au sexe est déterminée par des tables de probabilités conditionnelles âge×sexe. Il est par exemple possible d'estimer la probabilité d'être un homme sachant que l'âge est compris entre 5 et 10 ans. Toutes les probabilités associant l'âge et le sexe sont alors déterminées. Ce processus se poursuit pour toutes les catégories présentes. Le

principal inconvénient de cette méthode réside dans l'ordre des attributs pour définir les séquences d'états. Les résultats obtenus peuvent en effet changer selon les séquences considérées. Les auteurs recommandent de positionner les attributs par ordre décroissant du nombre de catégories.

Les deux techniques basées sur le processus de Markov (MCMC et HMM) ne tiennent pas compte du lien entre ménage et individu (Saadi et al., 2016 ; Sun et al., 2018). Pour générer une population synthétique à deux niveaux, Casati et al. (2015) ont proposé une extension du modèle MCMC de Farooq et al. (2013) pour tenir compte de la structure hiérarchique des ménages et des individus ; ils ont nommé ce processus hiérarchical MCMC (ou hMCMC). Le hMCMC consiste à ordonner les agents vivant dans le même ménage en fonction de leurs rôles dans le ménage. Trois rôles types sont ainsi définis : les propriétaires (personnes ayant le revenu le plus élevé dans leur ménage ou choisies en vertu d'un autre critère de sélection), les types intermédiaires (conjoints et enfants), et les autres. Pour générer un ménage, tous les propriétaires sont d'abord générés selon la méthodologie utilisée par Farooq et al. (2013), prenant ainsi en compte les variables caractérisant le ménage (puisque tous les membres d'un ménage partagent les mêmes attributs du ménage). Si la taille du ménage est supérieure à 1, les autres membres du ménage sont générés comme suit :

1. Les probabilités conditionnelles du conjoint sont générées à partir de certains attributs des propriétaires ;
2. Les probabilités conditionnelles des enfants sont générées à partir de certains attributs des propriétaires et des conjoints ;
3. Les probabilités conditionnelles des autres membres sont générées à partir des attributs de tous les autres types de membres (propriétaires, conjoints et enfants).

En procédant ainsi, les attributs des individus et des ménages sont combinés simultanément ; le lien entre individus et ménages est également préservé. Toutefois, cette méthode présente quelques limites. Pour certains types de ménages comme les ménages monoparentaux ou les personnes en colocation, il est difficile d'associer les attributs des ménages

et des individus selon la méthodologie proposée. Il est également difficile de savoir quels attributs doivent être considérés pour générer les probabilités conditionnelles des autres membres du ménage. Une autre incertitude réside dans les implications du choix d'un ensemble particulier d'attributs par rapport à un autre.

2.3.3 Réseaux bayésiens

Une autre méthode SL appliquée à la génération de populations synthétiques est le réseau bayésien (BN pour Bayesian Network en anglais) (Sun and Erath, 2015 ; Zhang et al., 2019). Les réseaux bayésiens forment une classe de modèles graphiques qui fournissent, à l'aide d'un graphe acyclique orienté (DAG pour Directed Acyclic Graph en anglais) une représentation intuitive de la structure probabiliste d'un ensemble d'attributs (Denis and Scutari, 2014). Dans le cadre de la génération de populations synthétiques, les réseaux bayésiens sont capables de reproduire les dépendances et interactions complexes qui existent entre les individus et ménages (Sun and Erath, 2015). Dans un BN, les noeuds d'un DAG représentent les attributs et les flèches associées traduisent les dépendances probabilistes entre les différents attributs. Un BN s'estime en deux étapes :

— l'apprentissage de la structure du DAG qui permet d'identifier les noeuds qui sont reliés entre eux. Cette étape permet de définir la structure du réseau qui décrit l'indépendance conditionnelle des attributs ;

— l'estimation des distributions conditionnelles des attributs compte tenu de la structure du DAG.

La population synthétique est ensuite générée par échantillonnage des probabilités conditionnelles obtenues. L'avantage des BN se traduit par le respect de la structure hiérarchique des données (ménages et individus). Néanmoins, l'apprentissage d'un BN peut constituer une tâche ardue, en particulier lorsque de nombreux attributs doivent être considérés. L'espace des DAG possibles est en effet immense et croit de façon exponentielle avec le nombre de noeuds. Comme le soulignent Sun and Erath (2015), un réseau à six noeuds contient environ 3 millions de DAGs possibles. Zhang et al. (2019)

ont suggéré de garder le BN aussi simple que possible et de créer une « liste blanche » qui permet d'établir un certain nombre de relations garanties entre les attributs. Cela réduira le nombre de DAG candidats. Un tel procédé impliquerait la construction du réseau sur la base de connaissances d'experts. Même en procédant ainsi, des erreurs peuvent toujours être commises et certaines interactions entre attributs peuvent ne pas être prises en compte.

2.3.4 Mélange hiérarchique

A la suite de leurs travaux sur les réseaux bayésiens, Sun et al. (2018) ont proposé une méthodologie de génération SL qui s'appuie sur un cadre de modélisation de mélange hiérarchique (HM pour Hierarchical Mixture). Dans cette approche, les auteurs ont considéré l'existence de classes latentes multiples au niveau du ménage. Pour chaque classe latente au niveau du ménage, des classes latentes existent également au niveau de l'individu. L'interaction entre ces deux niveaux peut être prise en compte en modélisant la probabilité d'appartenir à une classe latente et en modélisant la distribution des caractéristiques conditionnelles d'appartenance à cette classe.

Le cadre proposé intègre trois composantes pour reproduire la distribution statistique sous-jacente entre les deux niveaux (techniques de factorisation de tableaux multidimensionnels appelés tenseurs, modèle de classes latentes multi-niveaux, échantillonnage de rejet). Cependant, lorsque le nombre d'attributs augmente, la méthode HM devient très difficile à appliquer. Un autre défi de cette approche concerne sa robustesse par rapport à la sélection des variables latentes (Borysov et al., 2019).

2.3.5 Modèles génératifs

Plus récemment, Borysov et al. (2019) ; Garrido et al. (2020) ont proposé une approche de modélisation générative profonde (DGM pour Deep Generative Modeling) pour construire une population synthétique à deux niveaux. Les modèles génératifs sont des extensions des modèles graphiques probabilistes qui utilisent des réseaux de neurones. Les différents auteurs utilisent un auto-encodeur variationnel (VAE), qui est un

modèle de variable latente constitué de deux réseaux neurones : un encodeur et un décodeur. L'hypothèse de départ considère que chaque individu est représentée par un vecteur d'attributs aléatoires. L'objectif du VAE est d'apprendre la distribution jointe des attributs des individus de l'échantillon dans l'encodeur. Dans le décodeur, de nouveaux individus (non présents dans l'échantillon de départ) et d'anciens individus (similaires à ceux présents dans l'échantillon) sont générés à partir de la distribution de probabilité jointe apprise. Une population synthétique cohérente de ménages et d'individus avec de nombreux attributs peut ainsi être générée. L'utilisation de telles méthodes est très récente et leur mise en œuvre peut nécessiter des compétences informatiques poussées.

2.4 Comparaison des méthodes de génération de population synthétique à deux niveaux

Les principales caractéristiques, les avantages et les inconvénients des trois catégories de méthodes (SR, CO et SL) sont résumés dans les lignes suivantes :

— les méthodes SR combinent les informations provenant d'un échantillon d'individus rattachés à des ménages et des statistiques agrégées (données marginales) relatives aux individus et aux ménages pour calculer des poids. Ces poids traduisent le nombre d'unités que chaque individu et ménage de l'échantillon représentent dans la population totale. Les méthodes CO utilisent également l'échantillon d'individus et de ménages ainsi que les données marginales pour sélectionner une combinaison appropriée de ménages qui correspondent le mieux aux données marginales. Les méthodes SL ne considèrent que l'échantillon d'individus et de ménages. Elles se focalisent sur la distribution jointe des attributs de l'échantillon en estimant une probabilité pour chaque combinaison d'attributs ;

— les méthodes SR et CO génèrent une population synthétique en reproduisant un échantillon de référence. Par conséquent, les combinaisons d'attributs qui ne font pas partie de l'échantillon ne peuvent pas être générées. Ce problème est connu sous le nom de « zero-cell problem » et se produit souvent sur des zones géogra-

phiques de taille réduite. Il est uniquement présent dans les méthodes SR et CO. Une procédure de prétraitement pour éliminer le problème de la cellule zéro est souvent nécessaire. Il peut éventuellement s'agir d'un regroupement des catégories d'un attribut afin d'avoir des effectifs plus importants. Cela peut éventuellement entrainer une perte de précision. Les méthodes SL, en raison de l'estimation de la distribution jointe de tous les attributs de l'échantillon permettent de générer des individus et des ménages qui ne sont pas présent dans l'échantillon. Par conséquent, elles traitent efficacement le « zero-cell problem » ;

— les résultats des méthodes CO et SR correspondent directement aux marginaux ; avec les méthodes SL un post-traitement utilisant éventuellement des méthodes SR est necéssaire pour assurer une correspondance avec les données marginales ;

— les méthodes SL peuvent produire des résultats cohérents même si la taille de l'échantillon est faible. En revanche, la taille de l'échantillon est critique pour les méthodes CO et SR. Par exemple, dans leur étude de cas, Sun and Erath (2015) ont montré que pour que les résultats de l'IPF soient similaires à ceux du BN, la taille de l'échantillon doit être supérieure à 20% de la taille de la population totale ;

— de nombreux articles font état de l'application des méthodes SR dans la génération de populations synthétiques à deux niveaux de taille importante et qui intègrent un grand nombre d'attributs et de catégories (Ye et al., 2009 ; Müller, 2017 ; Bar-Gera et al., 2009 ; Lee and Fu, 2011). Les méthodes CO à deux niveaux peuvent être limitées par leur complexité de calcul. La majorité des applications de ces méthodes ont été restreintes à des populations de taille réduite (Abraham et al., 2012 ; Ma and Srinivasan, 2015). Certaines méthodes SL ne peuvent être utilisées lorsque le nombre d'attributs à générer est important. Pour les réseaux bayésiens (BN) et le mélange hiérarchique (HM), il est ainsi préférable que le nombre d'attributs soit faible, comme le soulignent Borysov et al. (2019) et Zhang et al. (2019). Un réseau bayésien avec six noeuds contient environ 3 millions de DAG possibles, et ce nombre atteint 1,1 milliard avec un réseau à sept noeuds (Sun and Erath, 2015).

En revanche, les modèles génératifs (DGM) et le modèle hiérarchique par chaînes de Markov (hMCMC) ne sont pas limités par la taille de la population et le nombre d'attributs.

Le Tableau 2.1 ci après résume les différents aspects évoqués précedemment.

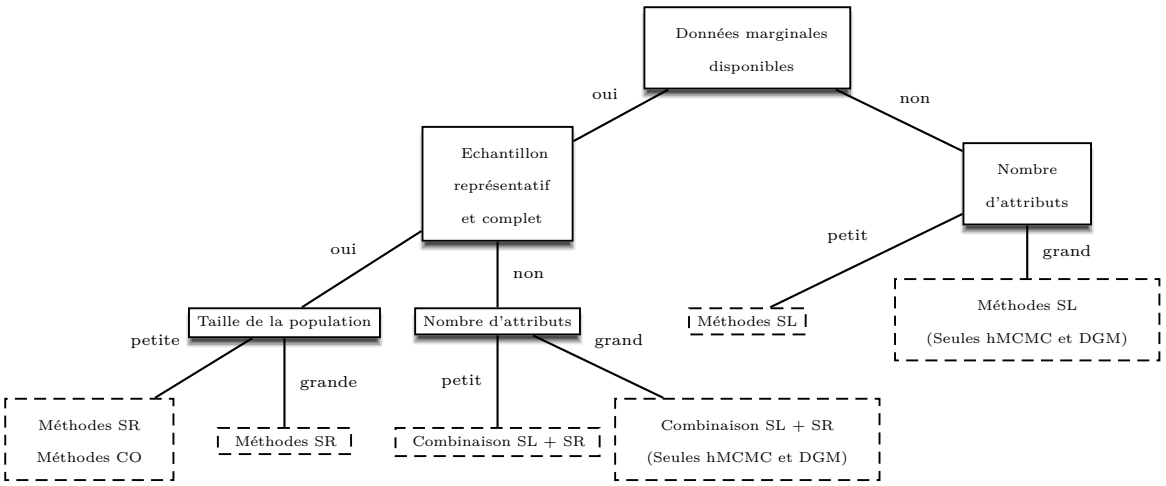
Tableau 2.1 : Un bref résumé des différentes méthodes de génération de populations synthétiques à deux niveaux

Caractéristiques	**Reconstruction synthétique** minimisation de l'entropie, IPU, HIPF, calage sur marges (GR)	**Optimisation combinatoire**	**Apprentissage statistique**			
			Modèle hiérarchique par chaînes de Markov (hMCMC)	Réseaux bayésiens (BN)	Mélange Hiérarchique (HM)	Modèles génératifs (DGM)
Algorithmes de traitement	Un algorithme déterministe avec une procédure en deux étapes : ajustement et affectation.	Un algorithme stochastique basé sur une combinaison d'individus et de ménages pour satisfaire un critère de convergence.	Définition de différents types d'agents, calcul de la distribution conditionnelle de chaque type d'agent à partir d'un échantillonneur de Gibbs afin d'estimer la distribution de probabilité jointe.	Constitution d'un graphe acyclic orienté (DAG) et estimation de la distribution de probabilité jointe en fonction du DAG constitué.	Un cadre intégrant trois composantes pour estimer la distribution de probabilité jointe.	Un algorithme utilisant un auto-encodeur variationnel (VAE) pour estimer la distribution de probabilité jointe.
			+ Génération d'individus et de ménages selon la distribution jointe.			
Données nécessaires	Échantillon et données agrégées au niveau des individus et des ménages.	Échantillon et données agrégées au niveau des individus et des ménages.	Seulement un échantillon d'individus et de ménages.			
Correspondance avec les données agrégées	Oui	Oui	Nécessite une étape de post-traitement avec des méthodes SR.			
Taille de l'échantillon	Produit de bons résultats si la taille de l'échantillon est suffisamment grande.	Produit de bons résultats si la taille de l'échantillon est suffisamment grande.	Possibilité d'utiliser un échantillon de petite taille.			
Nombre d'attributs potentiels pouvant être générés	Important	Important	Important	Limité		Important
Zero-cell problem	Possible	Possible	Pas de zero-cell problem			
Diffusion	Largement utilisé	Peu utilisé	Rarement utilisé	Peu utilisé	Rarement utilisés : ces méthodes ont été mises en oeuvre récemment et n'ont été testées que par leurs concepteurs.	

2.5 Procédure de prise de décision

Le choix d'une méthode de génération dépend étroitement de la quantité, du type et de la qualité (représentativité et exhaustivité) des données disponibles (Rich, 2018). En plus de la comparaison des différentes méthodes, un arbre de décision est proposé (Figure 2.5). L'arbre vise à faciliter le processus de décision en identifiant les méthodes de génération adéquates en fonction des paramètres suivants : données d'entrée disponibles (échantillon et marginaux), nombre d'attributs à générer et taille de la population synthétique.

FIGURE 2.5 : Arbre de décision



Note : SR : reconstruction synthétique ; CO : optimisation combinatoire ; SL : apprentissage statistique ; hMCMC : modèle hiérarchique par chaînes de Markov ; DGM : modèles génératifs.

L' hypothèse de départ concerne l'existence d'un échantillon présentant à la fois les caractéristiques du ménage et de la personne (données du recensement ou des enquêtes sur les déplacements des ménages), comme dans la plupart des études. Lorsque des données marginales sont disponibles et que l'échantillon est représentatif (au moins 5% de la population), les méthodes de reconstruction synthétique (SR) sont préférées pour générer une population de taille importante (des millions de personnes et des ménages avec un nombre élevé d'attributs comme relevé par (Müller and Axhausen, 2011 ; Vovsha et al., 2015 ; Fournier et al., 2020).

Les méthodes d'optimisation combinatoire (CO) sont limitées à la génération de populations de petite taille. Cela est dû à la complexité de calcul de ces méthodes. Par exemple, Hafezi and Habib (2014) et Ma and Srinivasan (2015) génèrent des populations avec moins de 60 000 personnes et 30 000 ménages.

Lorsque l'échantillon est de petite taille (moins de 5%) mais que des données marginales sont disponibles, il est plus pertinent de combiner les méthodes d'apprentissage statistiques (SL) et de reconstruction synthétique (SR). Les méthodes SL peuvent être appliquées en premier lieu pour construire un échantillon de taille appropriée ; les méthodes SR peuvent être appliquées ensuite pour générer une population correspondant aux données marginales. Sur ce principe, Sun et al. (2018) à Singapour et Borysov et al. (2019) au Danemark ont généré une population à partir d'échantillons d'enquêtes sur les ménages qui correspondent respectivement à 1% et 2,5% de la population totale.

Lorsque les données marginales ne sont pas disponibles, seules les méthodes SL peuvent être appliquées pour générer la population.

Conclusion

Dans la modélisation multi-agents, la capacité de générer une population synthétique à deux niveaux (individus et ménages) est essentielle parce que de nombreuses décisions (choix de mobilités, choix de résidence, achats de biens et de services,...) d'un individu dépendent à la fois de ses attributs personnels et de ceux de son ménage. La génération d'une telle population constitue un problème épineux (Fournier et al., 2020), notamment en raison du nombre élevé de méthodes de génération qui existent. Dans ce chapitre, les principales méthodes de génération qui utilisent un échantillon sont passées en revue. En adoptant la catégorisation de Sun et al. (2018), ces méthodes ont été classées en trois groupes : Reconstruction synthétique (SR), Optimisation combinatoire (CO) et Apprentissage statistique (SL). La catégorie SR combine les informations d'un échantillon d'une population (individus et ménages) et de données agrégées afin de calculer des poids qui reflètent la représentativité de chaque ménage de l'échantillon. La catégorie CO utilise

également un échantillon d'individus et de ménages associé à des données agrégées afin de sélectionner une combinaison appropriée de ménages qui correspond le mieux aux marginaux. La dernière catégorie (SL) considère uniquement un échantillon d'individus et de ménages et se concentre sur la distribution jointe de tous les attributs en estimant une probabilité pour chaque combinaison.

Pour chaque catégorie, une description détaillée des différentes méthodes et approches associées a été proposée dans un premier temps. Par la suite, une comparaison de ces méthodes a été établie à partir d'un certain nombre de critères : caractéristiques de l'algorithme, données d'entrée nécessaires, sorties obtenues, taille de l'échantillon, nombre d'attributs pouvant être générés, zero-cell problem et diffusion de la méthode auprès des utilisateurs. Cette comparaison illustre les avantages et les inconvénients d'utilisation de chaque méthode.

Afin de faciliter le choix d'une catégorie voire d'une méthode, un outil de décision (sous la forme d'un arbre de décision) a été proposé à la fin du chapitre. Cet arbre représente une contribution à la littérature scientifique sur les méthodes de génération de population synthétique à deux niveaux. Confrontés à un problème de génération d'une population synthétique à deux niveaux, chercheurs et les professionnels ont maintenant accès à un cadre standardisé et complet pour sélectionner la méthode appropriée en fonction de nombreux critères et de leurs objectifs de modélisation. Le second chapitre procède à la génération effective d'une population synthétique à deux niveaux.

Références bibliographiques

Abraham, J. E., Stefan, K. J., and Hunt, J. D. (2012). Population synthesis using combinatorial optimization at multiple levels. In *91th Annual Meeting of the Transportation Research Board, Washington, DC*.

Arentze, T., Timmermans, H., and Hofman, F. (2007). Creating synthetic household populations : problems and approach. *Transportation Research Record*, 2014(1) :85–91.

Auld, J. and Mohammadian, A. (2010). Efficient methodology for generating synthetic populations with multiple control levels. *Transportation Research Record*, 2175(1) :138–147.

Balakrishnaa, R., Sundarama, S., and Lam, J. (2020). An enhanced and efficient population synthesis approach to support advanced travel demand models. In *99th Annual Meeting of the Transportation Research Board, Washington, DC*.

Bar-Gera, H., Konduri, K., Sana, B., Ye, X., and Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Barthelemy, J. and Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, 47(2) :266–279.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A : Policy and Practice*, 30(6) :415–429.

Birkin, M., Turner, A., and Wu, B. (2006). A synthetic demographic model of the uk population : Methods, progress and problems. In *Regional Science Association International British and Irish Section, 36th Annual Conference*.

Borysov, S. S., Rich, J., and Pereira, F. C. (2019). How to generate micro-agents ? a deep generative modeling approach to population synthesis. *Transportation Research Part C : Emerging Technologies*, 106 :73–97.

Casati, D., Müller, K., Fourie, P. J., Erath, A., and Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1) :107–116.

Chapuis, K. and Taillandier, P. (2019). A brief review of synthetic population generation practices in agent-based social simulation. In *SSC2019, Social Simulation Conference*.

Choupani, A.-A. and Mamdoohi, A. R. (2016). Population synthesis using iterative proportional fitting (ipf) : A review and future research. *Transportation Research Procedia*, 17 :223–233.

Denis, J. and Scutari, M. (2014). *Réseaux Bayésiens avec R : Élaboration, Manipulation et Utilisation en Modélisation Appliquée.* EDP Sciences.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423) :1013–1020.

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B : Methodological*, 58 :243–263.

Fournier, N., Christofa, E., Akkinepally, A. P., and Azevedo, C. L. (2020). Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, pages 1–27.

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PloS one*, 5(1) :e8828.

Garrido, S., Borysov, S. S., Pereira, F. C., and Rich, J. (2020). Prediction of rare feature combinations in population synthesis : Application of deep generative modelling. *Transportation Research Part C : Emerging Technologies*, 120 :102787.

Guo, J. Y. and Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014(1) :92–101.

Hafezi, M. H. and Habib, M. A. (2014). Synthesizing population for microsimulation-based integrated transport models using atlantic canada micro-data. *Procedia Computer Science*, 37 :410–415.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales : A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, 15(1).

Hermes, K. and Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, 36(4) :281–290.

Huynh, N. N., Barthelemy, J., and Perez, P. (2016). A heuristic combinatorial optimisation approach to synthesising a population for agent-based modelling purposes. *Journal of Artificial Societies and Social Simulation*, 19(4) :11.

Lee, D.-H. and Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record*, 2255(1) :20–27.

Lenormand, M. and Deffuant, G. (2013). Generating a synthetic population of individuals in households : Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation*, 16(4) :12.

Lovelace, R. and Ballas, D. (2013). truncate, replicate, sample : A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41 :1–11.

Ma, L. and Srinivasan, S. (2015). Synthetic population generation with multilevel controls : A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, 30(2) :135–150.

Müller, K. (2017). *A generalized approach to population synthesis*. PhD thesis, ETH Zurich.

Müller, K. and Axhausen, K. W. (2010). Population synthesis for microsimulation : State of the art. *Arbeitsberichte Verkehrs-und Raumplanung*, 638.

Müller, K. and Axhausen, K. W. (2011). Hierarchical ipf : Generating a synthetic population for switzerland. *paper presented at the 51st Congress of the European Regional Science Association.*

Müller, K. and Axhausen, K. W. (2012). Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung*, 821.

Murata, T., Sugiura, S., and Harada, T. (2017). Income allocation to each worker in synthetic populations using basic survey on wage structure. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE.

Paul, B. M., Doyle, J., Stabler, B., Freedman, J., and Bettinardi, A. (2018). Multi-level population synthesis using entropy maximization-based simultaneous list balancing. In *97th Annual Meeting of the Transportation Research Board, Washington, DC.*

Pritchard, D. R. and Miller, E. J. (2012). Advances in population synthesis : fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39(3) :685–704.

Rich, J. (2018). Large-scale spatial population synthesis for denmark. *European Transport Research Review*, 10(2) :63.

Rich, J. and Mulalic, I. (2012). Generating synthetic baseline populations from register data. *Transportation Research Part A : Policy and Practice*, 46(3) :467–479.

Ryan, J., Maoh, H., and Kanaroglou, P. (2009). Population synthesis : Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, 41(2) :181–203.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., and Cools, M. (2016). Hidden markov model-based population synthesis. *Transportation Research Part B : Methodological*, 90 :1–21.

Srinivasan, S., Ma, L., and Yathindra, K. (2008). Procedure for forecasting household characteristics for input to travel-demand models. project report of university of florida, gainesville ; florida department of transportation. Technical report, TRC-FDOT-64011-2008.

Sun, L. and Erath, A. (2015). A bayesian network approach for population synthesis. *Transportation Research Part C : Emerging Technologies*, 61 :49–62.

Sun, L., Erath, A., and Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B : Methodological*, 114 :199–212.

Templ, M., Meindl, B., Kowarik, A., and Dupriez, O. (2017). Simulation of synthetic complex data : The r package simpop. *Journal of Statistical Software*, 79(10) :1–38.

Voas, D. and Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5) :349–366.

Voas, D. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2) :177–200.

Vovsha, P., Hicks, J. E., Paul, B. M., Livshits, V., Maneva, P., and Jeon, K. (2015). New features of population synthesis. In *94th Annual Meeting on Transportation Research Board, Washington, DC.*

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., and Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC.*

Zhang, D., Cao, J., Feygin, S., Tang, D., Shen, Z.-J. M., and Pozdnoukhov, A. (2019). Connected population synthesis for transportation simulation. *Transportation Research Part C : Emerging Technologies*, 103 :1–16.

Zhu, Y. and Ferreira Jr, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record*, 2429(1) :168–177.

Chapitre 3

Comparaison et application de méthodes de reconstruction synthétique pour générer une population synthétique à deux niveaux

Sommaire

Résumé

Ce chapitre décrit la génération d'une population synthétique de ménages et d'individus sur un cas d'application français. Les sources françaises disponibles sont caractérisées par des données agrégées associées à un échantillon de taille importante, qui correspond à environ 30% de la population. L'échantillon et les données agrégées proviennent d'une même source, le recensement français, ce qui garantit une cohérence des données. Cette configuration des données favorise l'utilisation des méthodes de reconstruction synthétique (méthodes SR). A partir des données du recensement de l'aire urbaine de Nantes, quatre algorithmes de reconstruction synthétique (HIPF, IPU, minimisation de l'entropie et calage sur marges) associés à deux méthodes probabilistes de conversion de poids (probabilités proportionnelles et TRS) sont testés. Une description approfondie de chaque algorithme est proposée dans un premier temps à travers un cadre harmonisé de notations. Une comparaison de ces algorithmes est par la suite effectuée sur la base de différents indicateurs de validation. Les résultats obtenus montrent que les algorithmes testés produisent des populations synthétiques réalistes. En termes de performance, la minimisation de l'entropie et l'HIPF combinés à la méthode TRS génèrent des populations synthétiques de ménages et d'individus dont les valeurs sont les plus proches des valeurs de la population réelle.

Mots clés

Recensement français, IPU, HIPF, entropie, calage sur marges, probabilités proportionnelles, TRS.

Introduction

Les différentes méthodes de génération de population synthétique à deux niveaux diffèrent selon l'approche employée et les données requises. La plupart des instituts nationaux de statistique diffusent deux types de données. Le premier est une source de données désagrégées caractérisant un échantillon de la population. Un tel échantillon est généralement constitué à partir de données de recensement et fournit des informations sur les attributs (caractéristiques) de certains individus ou ménages de la population à une échelle géographique spécifique. Ces attributs sont par exemple le sexe, la profession, la taille du ménage, le revenu du ménage... Le second type de données disponibles représente des valeurs agrégées qui traduisent les distributions marginales des attributs des individus et des ménages. Ces valeurs agrégées sont également appelées données marginales ou attributs de contrôle (Templ et al., 2017).

Les données françaises diffusées par l'Institut national de la statistique et des études économiques (Insee) se distinguent de celles diffusées dans les autres pays sur deux aspects : la taille de l'échantillon disponible est importante. Elle correspond à 30% de la population totale contre souvent moins de 5% ailleurs ; également l'échantillon et les données agrégées proviennent d'une même source, le recensement français de la population, ce qui assure une cohérence entre ces deux types de données.

A partir de la revue des méthodes de génération effectuée dans le chapitre 2 et en prenant en compte la configuration des données du recensement français, il ressort que les méthodes de reconstruction synthétique (SR) sont les plus adéquates pour obtenir une population synthétique à deux niveaux. Dans ce chapitre, différents algorithmes de reconstruction synthétique sont décrits et comparés sur la base d'un cadre conceptuel commun avec une harmonisation des notations. Il s'agit de : l'Iterative Proportional Update (IPU) (Ye et al., 2009), le Hierarchical Iterative Proportional Fitting (HIPF) (Müller and Axhausen, 2011 ; Müller, 2017), la minimisation de l'entropie (ent) (Bar-Gera et al., 2009 ; Lee and Fu, 2011) et le calage sur marges (GR) (Deville et al., 1993).

Dans le processus de génération d'une population synthétique, ces algorithmes pro-

duisent des fractions de ménages et d'individus. Il est alors nécessaire de transformer les fractions en nombres entiers à travers l'étape d'affectation (Figure 2.2). Deux méthodes probabilistes de conversion des poids sont également introduites : la méthode des probabilités proportionnelles (PP) (Lovelace et al., 2015 ; Joubert, 2018) et la méthode Truncate Replicate Sample (TRS) (Lovelace and Ballas, 2013). Les différentes associations (algorithmes et méthodes de conversion des poids) sont testées et évaluées à travers la génération d'une population synthétique à deux niveaux représentative de l'aire urbaine de Nantes.

Le reste du chapitre se compose de quatre sections. Dans les deux prochaines sections (section 3.1 et section 3.2), le recensement français de la population et le cas d'étude sont présentés. La section 3.3 décrit chaque algorithme et chaque méthode de conversion des poids. La quatrième section (section 3.4) résume et discute les résultats obtenus.

3.1 Source de données : le recensement français de la population

3.1.1 Présentation générale

Le recensement français de la population est réalisé par l'Institut national de la statistique et des études économiques (Insee). Il constitue la source de données par excellence pour appréhender la dynamique démographique et la structure sociale des territoires. Le dernier dénombrement exhaustif de la population vivant en France a eu lieu en 1999. Depuis 2004, des enquêtes annuelles de recensement (EAR) se sont substituées à l'opération décennale de comptage exhaustif. L'utilisation d'enquêtes annuelles permet de disposer de données statistiques récentes et régulières concernant la population vivant en France (Insee, 2020 [1]). Les EAR sont organisées chaque début d'année selon des modalités qui distinguent les communes en fonction d'un seuil de population fixé à 10 000 habitants :

1. Insee : RECENSEMENT DE LA POPULATION. CONSEILS D'UTILISATION - SYNTHÈSE (Note synthétique, juin 2020) [en ligne] (page consultée le 19 juillet 2021) < https://www.insee.fr/fr/statistiques/fichier/2383177/fiche-conseils_synthese_2020-06-29.pdf>

— toutes les communes de moins de 10 000 habitants sont regroupées en cinq groupes. Chaque année, une EAR exhaustive est menée sur toutes les communes d'un groupe. Au bout de cinq années successives, l'intégralité des communes de moins de 10 000 sont enquêtées ;

— pour chaque commune de 10 000 habitants ou plus, une EAR est menée chaque année sur un échantillon d'adresses représentant environ 8 % des logements de la commune. Il s'agit d'un sondage qui s'appuie sur un échantillonnage du répertoire des immeubles localisés (RIL) qui est une base de données exhaustive des adresses et mise en jour en continu. Au bout de cinq années successives, 40% des logements de chaque commune de 10 000 habitants sont enquêtés. En revanche, les communautés, les habitations mobiles et les personnes sans abris sont enquêtées exhaustivement.

En cumulant les informations collectées au bout de cinq EAR successives, l'Insee élabore et diffuse chaque année des résultats complets de recensement. Toutefois, pour assurer une égalité de traitement entre communes, les informations collectées sur cinq années sont ramenées à une même date de référence (Insee, 2019a [2]). Cette date est fixée au 1er janvier de l'année médiane des cinq années d'enquêtes. Le terme de recensement millésimé est alors utilisé. Par exemple, le recensement millésimé de 2015 correspond aux EAR de 2013 à 2017. Un millésime du recensement comprend deux phases d'exploitation :

— une exploitation *principale* qui est exhaustive pour toutes les communes de moins de 10 000 habitants et concerne 40% des logements des communes de 10 000 habitants ou plus. Les données diffusées portent sur la majorité des attributs recueillis directement lors de l'administration du questionnaire : âge, sexe, caractéristiques du logement, migrations résidentielles, diplômes et mobilités scolaires, mobilités professionnelles...

— une exploitation *complémentaire*, destinée à produire des attributs dont l'élaboration est délicate et longue. L'exploitation complémentaire permet notamment de

2. Insee : Présentation du recensement de la population [en ligne] (page consultée le 19 juillet 2021)< `https://www.insee.fr/fr/information/2383265`>

décrire la structure familiale des ménages [3] et de définir les liens qui existent entre les personnes d'un même ménage. Les catégories socioprofessionnelles des individus ainsi que les secteurs d'activité sont également déterminés avec une meilleure précision que dans l'exploitation principale (Brilhault and Caron, 2016). En revanche, l'exploitation complémentaire est plus coûteuse et est réalisée sur un sous-échantillon dont la construction est précisée dans le Tableau 3.1.

Tableau 3.1 : Procédure d'échantillonnage de l'exploitation complémentaire du recensement français de la population

	Communes de moins de 10 000 habitants	Communes de 10 000 habitants ou plus
Ménages	20%	100% des ménages enquêtés (40% du total des ménages)
Individus des communautés	20%	
Habitations mobiles terrestres et personnes sans abri	20%	100%
Habitations mobiles fluviales (bateliers)	100%	

Source : Insee (2019b) : RECENSEMENT DE LA POPULATION. LES EXPLOITATIONS PRINCIPALE ET COMPLÉMENTAIRE (fiche thématique, juin 2019) [en ligne] (page consultée le 19 juillet 2021)< `https://www.insee.fr/fr/statistiques/fichier/2383265/fiche-exploitation.pdf`>

3.1.2 Données du recensement diffusées

Les données d'un recensement millésimé (année n) sont généralement disponibles sur le site de l'Insee à partir du mois de juin de l'année n+3. Pour toutes les données diffusées, issues de l'exploitation principale ou complémentaire, l'échantillonnage pratiqué entraîne une marge d'incertitude sur les résultats. Cette imprécision varie d'une com-

3. Un ménage selon l'Insee, désigne l'ensemble des occupants d'une résidence principale (logement occupé de façon habituelle et à titre principal) qu'ils aient ou non des liens de parenté.

mune à l'autre en fonction du taux de sondage qui dépend du type de commune, du type d'exploitation, et de l'effectif de l'attribut d'intérêt (plus l'effectif obtenu est réduit, plus l'imprécision risque d'être grande [4]).

Les résultats du recensement de la population sont disponibles à différents échelles géographiques : région, aire urbaine, département, arrondissement (pour les grandes villes comme Paris, Lyon, Marseille), canton, communes ... Au niveau infracommunal, la plupart des communes d'au moins 5 000 habitants sont découpées en Ilots Regroupés pour l'Information Statistique (IRIS [5]).

Les IRIS constituent la plus petite échelle géographique sur laquelle des informations du recensement sont disponibles. Construits à partir de critères géographiques et statistiques, les IRIS doivent être homogènes du point de vue de l'habitat. Ils sont assimilables à des quartiers dont la population est de l'ordre de 2 000 habitants (Godinot, 2005). Les IRIS offrent l'outil le plus élaboré à ce jour pour décrire la structure interne des communes. Afin d'assurer une certaine homogénéité du découpage, une commune non irisée (ayant une taille de population insuffisante pour être découpée en IRIS) est assimilable à un IRIS.

L'Insee permet un usage personnalisé des données du recensement en fournissant des bases de données volumineuses intitulées fichiers détail. Ces fichiers représentent des échantillons d'individus, de ménages et de logements avec leurs attributs détaillés. Les fichiers détail se répartissent en six grandes thématiques comme le montre le Tableau 3.2.

4. Pour davantage de détails, consulter Insee (2017) : RECENSEMENT DE LA POPULATION. LA PRÉCISION DES RÉSULTATS DU RECENSEMENT, (fiche thématique, mars 2017) [en ligne] (page consultée le 19 juillet 2021) < `https://www.insee.fr/fr/statistiques/fichier/2383177/fiche-precision.pdf`>

5. Pour davantage de détails, consulter Insee (2016a) : IRIS, [en ligne] (page consultée le 19 juillet 2021)< `https://www.insee.fr/fr/metadonnees/definition/c1523`>

Tableau 3.2 : Fichiers détail du recensement

Thématique	Nom	Description	Type d'exploitation et champ	Echelon géographique
Logement	Fichier détail « logement »	Chaque observation est un logement décrit selon ses attributs et ceux des ménages qui y résident.	Exploitation principale (résidences principales).	IRIS et commune non irisée.
Individu (deux fichiers détail disponibles)	1) Fichier détail « Individus - localisés à la région » ; 2) Fichier détail « Individus - localisation au canton-ou-ville »	Chaque observation est un individu décrit selon ses attributs sociodémographiques, ceux de son ménage et ceux de sa résidence principale.	Exploitation complémentaire (tous les individus).	1) Région (ou département et *EPCI, pour les départements et EPCI de 700 000 habitants ou plus) ; 2) IRIS et cantons de résidences pour les communes non irisées.
Migrations résidentielles (trois fichiers détail disponibles)	1) Fichier détail « commune de résidence/commune de résidence antérieure » ; 2) Fichier détail « commune de résidence/pays de résidence antérieure » ; 3) Fichier détail « département de résidence/pays de résidence antérieure »	Chaque observation est un individu décrit selon sa résidence actuelle et sa résidence antérieure (1 an auparavant à compter du millésime 2013), ses principaux attributs sociodémographiques et ceux de son ménage.	Exploitation complémentaire (individus âgés de 1 an ou plus).	1) commune de résidence / commune de résidence antérieure ; 2) Commune de résidence (5 000 habitants ou plus) / pays de résidence antérieure ; 3) Département de résidence / pays de résidence antérieure (étranger).
Mobilités professionnelles	Fichier détail « mobilités professionnelles »	Chaque observation est un individu décrit selon ses déplacements domicile-travail, ses principaux attributs sociodémographiques et ceux de son ménage.	Exploitation complémentaire (individus actifs ayant un emploi âgés de 15 ans ou plus)	Commune et arrondissement municipal uniquement pour Paris, Lyon, Marseille.
Mobilités scolaires	Fichier détail « mobilités scolaires »	Chaque observation est un individu décrit selon ses déplacements domicile-études, ses principaux attributs sociodémographiques et ceux de son ménage.	Exploitation complémentaire (individus inscrits dans un établissement d'enseignement et âgés de 2 ans ou plus)	Commune et arrondissement municipal uniquement pour Paris, Lyon, Marseille.
Activités professionnelles	Fichier détail « activité professionnelle des individus »	Chaque observation est un individu localisé au lieu de travail décrit selon les caractéristiques de son emploi ainsi que ses principaux attributs sociodémographiques.	Exploitation complémentaire (individus actifs ayant un emploi, âgés de 15 ans ou plus et travaillant en France hors Mayotte)	Zone d'emploi et région du lieu de travail.

Note : *EPCI= Etablissements publics de coopération intercommunale

Source : Insee (2018) : Documentation sur les fichiers détail [en ligne] (page consultée le 19 juillet 2021)< `https://www.insee.fr/fr/information/2383306`>

Seul le fichier détail « logement » provient de l'exploitation principale. Dans ce fichier, chaque observation représente un logement. En revanche, dans les autres fichiers, chaque observation représente un individu. Il est également important de signaler que dans tous les fichiers détail issus de l'exploitation complémentaire, les individus sont identiques.

Le fichier détail « Individus - localisation au canton-ou-ville » désigné par INDCANT est particulièrement intéressant. Ce fichier comporte des informations sur les individus, sans restriction d'âge et de statut professionnel. Il renseigne également sur la composition des ménages, les secteurs d'activité et sur d'autres attributs sociodémographiques qui se retrouvent en partie dans les autres fichiers détail. En termes de localisation, pour les communes d'au moins 5 000 habitants, les lieux de résidence sont plus précis. En effet, l'échelle géographique de localisation des individus, de leur ménage et de leur logement est l'IRIS. En revanche pour les communes non irisées, les individus, ménages et logements sont localisés au niveau du canton qui englobe plusieurs communes. Une procédure de jointure (présentée dans la prochaine section) permet de passer du canton à la commune de résidence.

En plus des fichiers détail, l'Insee met à disposition du public, des bases de données infracommunales agrégées. Ces bases de données comportent les effectifs totaux des individus, de leur ménage, et de leur logement en fonction de différents attributs. Le niveau géographique renseigné est l'IRIS (pour les communes découpées en IRIS), et la commune (pour les communes non irisées, i.e., les communes non découpées en IRIS). Il existe cinq bases de données infracommunales :

1. La base infracommunale « Population » fournit des données agrégées sur les attributs de la population selon le sexe, l'âge, la catégorie socioprofessionnelle et la catégorie de nationalité ;
2. La base infracommunale « Activité des résidents » fournit des données agrégées sur les attributs des actifs (sexe, âge, catégorie socioprofessionnelle), des salariés et non-salariés (sexe et âge) ;
3. La base infracommunale « Couples - Familles - Ménages » fournit des données agrégées relatifs aux attributs des ménages et des personnes qui les composent (âge,

statut conjugal, catégorie socioprofessionnelle) et les caractéristiques des familles (nombre d'enfants) ;

4. La base infracommunale « Diplôme - Formation » fournit des données agrégées sur l'âge des personnes scolarisées et les niveaux de diplôme des personnes non scolarisées ;
5. La base infracommunale « Logement » fournit des données agrégées sur les caractéristiques des résidences principales, la date d'emménagement, la possession d'une voiture ainsi que le parc de logements, maisons et appartements.

Les bases infracommunales diffusées sont accompagnées d'informations sur la qualité des données. Chaque IRIS et commune non irisée est alors caractérisé par un label de qualité. En fonction de ce label de qualité, il peut arriver que la précision et la stabilité de l'estimation de la population ne soient pas d'une qualité suffisamment bonne pour que les données fournies en l'état soient utilisables. La préconisation est de regrouper les IRIS ou communes ayant un faible label de qualité avec d'autres IRIS pour gagner en précision et en stabilité des données.

En résumé, l'Insee diffuse à travers les fichiers détail, des échantillons d'individus avec leurs attributs sociodémographiques, les attributs de leur ménage et ceux de leur logement au niveau IRIS et commune. A ces fichiers détail, sont associées des données agrégées pour certains attributs individuels, ménages et de logement figurant dans les fichiers détail. Ces différentes données seront utilisées dans la suite du travail pour générer une population synthétique d'individus et de ménages, représentative de la population de l'aire urbaine de Nantes.

3.2 Cas d'étude : aire urbaine de Nantes

3.2.1 Présentation générale

Selon la définition de l'Insee, une aire urbaine (AU) correspond à un ensemble de communes d'un seul tenant et sans enclave, constitué par un pôle urbain et par des communes rurales ou unités urbaines, dont au moins 40 % de la population résidente

ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci [6]. Le zonage en aires urbaines 2010 permet d'avoir une vision des aires d'influence des unités urbaines sur le territoire français. Trois catégories d'aires sont ainsi définies : « grande », « moyenne » et « petite » pour une couronne et un pôle urbain respectivement d'au moins 10 000 emplois, 5 000 à 10 000 emplois et 1 500 à 5 000 emplois.

L'AU de Nantes est une grande aire urbaine constituée de 108 communes (107 communes en Loire-Atlantique et 1 commune en Maine-et-Loire) réparties en 29 communes irisées et 79 communes non irisées. Dans le cas d'application de cette thèse, la population synthétique est générée à l'échelle de l'IRIS ou de la commune (pour les communes non irisées). Les différentes sources de données utilisées proviennent du recensement millésimé de 2015, diffusées en octobre 2018 et disponibles en libre accès sur le site de l'Insee (liens consultés le 22 juillet 2021) :

1. le fichier détail « INDCANT »(Individus - localisation au canton-ou-ville), disponible à `https://www.insee.fr/fr/statistiques/3625223?sommaire=3558417` ;
2. la base « Population », disponible à `https://www.insee.fr/fr/statistiques/3627376#consulter` ;
3. la base « Activité des résidents », disponible à `https://www.insee.fr/fr/statistiques/3627009#consulter` ;
4. la base « Couples - Familles - Ménages », disponible à `https://www.insee.fr/fr/statistiques/3627367#consulter` ;
5. la base « Diplômes - Formation », disponible à `https://www.insee.fr/fr/statistiques/3627372#consulter` ;
6. la base « Logement », disponible à `https://www.insee.fr/fr/statistiques/3627374#consulter`.

Dans le fichier INDCANT, l'échelle géographique de localisation résidentielle des individus est l'IRIS pour les individus habitant dans les communes irisées. En revanche,

6. Insee (2016b) : Aire urbaine [en ligne] (page consultée le 21 juillet 2021)<`https://www.insee.fr/fr/metadonnees/definition/c2070`>

pour les individus des communes non irisées, l'échelle géographique de résidence mentionné est le canton. Une procédure pour passer du canton à la commune est proposée dans les lignes qui suivent.

3.2.2 Procédure pour rattacher les individus à leur commune de résidence

Au fichier détail INDCANT est associé le fichier détail « commune de résidence/-commune de résidence antérieure » (confère Tableau 3.2) que nous désignerons par INDCOM par soucis de simplification. Dans INDCOM [7], le lieu de résidence de l'individu est sa commune, qu'il appartienne à une commune irisée ou non irisée. Par rapport à INDCANT, le niveau de localisation est plus précis dans INDCOM pour les individus habitant dans des communes non irisées. En revanche, les attributs sociodémographiques des individus, ménages et logements sont moins nombreux que dans INDCANT.

INDCANT et INDCOM portant sur les mêmes individus [8], il existe dans ces deux fichiers détail des attributs communs. L'idée est de se servir de ces attributs communs pour retrouver la commune de résidence des individus de INDCANT. Il s'agit d'un cas de jointure classique entre deux bases de données.

Dans INDCANT, les individus âgés de moins de 1 an sont retirés. Il reste alors 47 382 individus dont il faut retrouver la commune. Les attributs communs aux deux fichiers ont été recensés par la suite. Il s'agit notamment des attributs AGEREVQ (âge révolu de l'individu avec 25 catégories), NPERR (nombre de personnes du ménage en 7 catégories). Pour chaque individu des deux fichiers, une clé d'appariement (concaténation des catégories liées à 22 attributs communs) est constituée. Avec cette clé d'appariement, il est possible de rattacher de façon certaine 46 128 individus de INDCANT à leur commune, soit environ 97% des individus. Pour les 1 254 individus restants répartis dans 08 communes, la jointure s'est avérée plus délicate car il existait des doublons. Les

7. Fichier détail disponible à <https://www.insee.fr/fr/statistiques/3566042?sommaire=3558417>, (page consultée le 22 juillet 2021)

8. Une précision est nécessaire. INDCOM prend en compte les individus âgés d'au moins un an alors que INDCANT n'opère pas de distinction d'âge (Tableau 3.2).

doublons ont été supprimés en considérant toutes les configurations possibles. A la fin de la procédure, une commune est affectée à chaque individu de INDCANT. Pour vérifier la qualité de l'appariement, nous avons croisé les données des deux fichiers détail avec trois attributs communs à INDCOM et INDCANT mais non utilisés dans la concaténation. Les chiffres obtenus coïncidaient parfaitement, même pour les 1 254 individus qui avaient des doublons.

La dernière étape consiste à réintégrer dans la base INDCANT, les individus âgés de moins d'une année et à leur affecter la commune de résidence de leur ménage en utilisant les attributs NUMMI (numéro du ménage dans le canton) et CANTVILLE (numéro du canton). L'échantillon de données (totalité des communes de l'AU de Nantes) est constitué d'environ 294 000 individus et 136 000 ménages. A partir de cet échantillon, nous devons estimer les attributs d'une population totale de 949 000 individus et 418 000 ménages.

En résumé, les données du recensement français disponibles sont caractérisées par un échantillon d'individus associés à leur ménage avec des attributs personnels, des attributs du ménage et du logement. Cet échantillon équivaut à plus de 30% de la population totale. Des données agrégées (bases infracommunales) sont également fournies. Elles traduisent les effectifs globaux de certains attributs individuel, ménage et logement à l'échelle de l'IRIS ou de la commune. Une cohérence existe entre les données agrégées et l'échantillon qui proviennent tous les deux de la même source. Par rapport aux critères de choix de l'arbre de décision (Figure 2.5), les méthodes de reconstruction synthétique sont plus adaptées pour générer la population synthétique d'individus et de ménages souhaitée pour l'aire urbaine.

La prochaine section présente le processus de génération d'une population synthétique avec des méthodes de reconstruction synthétique.

3.3 Méthodologie de génération de populations synthétiques à deux niveaux avec des méthodes de reconstruction synthétique

La génération d'une population synthétique avec les méthodes de reconstruction synthétique (SR) implique une procédure en deux étapes : ajustement et affectation. Dans la première sous-section, quatre algorithmes d'ajustement SR à deux niveaux sont décrits. Il s'agit de l'Iterative Proportional Update (IPU), du Hierarchical Iterative Proportional Fitting (HIPF), de la minimisation de l'entropie (ent) et du calage sur marges (GR). La deuxième sous-section décrit ensuite deux méthodes d'affectation (approche des probabilités proportionnelles et méthode Truncate Replicate Sample) qui convertissent les poids non entiers (obtenus dans l'étape d'ajustement) en poids entiers avant réplication des individus et des ménages.

3.3.1 Etape d'ajustement des poids

L'objectif de cette étape est de trouver un vecteur de poids des ménages $\boldsymbol{W} = (w_h)$, avec $h = 1 \dots n_H^s$. n_H^s représente le nombre de ménages dans l'échantillon et w_h est un réel positif qui représente le poids d'un ménage. Ce poids sera utilisé dans l'étape d'affectation pour répéter ou tirer le ménage. Certaines données agrégées (également appelées attributs de contrôle) relatives aux attributs individuels et ménages constituent les contraintes du vecteur des poids des ménages. Afin de permettre une comparaison entre les différents algorithmes, le problème d'ajustement est présenté sous la forme d'un problème inverse mal posé à résoudre [9]. L'objectif recherché est de trouver $\boldsymbol{W}$ qui satisfasse les contraintes marginales [10] :

9. Il existe une abondante littérature scientifique sur ce sujet depuis les travaux de Tikhonov (Tikhonov and Arsenin, 1977)

10. Les contraintes marginales désignent les catégories des attributs. Par exemple, si l'attribut taille du ménage comporte les catégories suivantes : 1 personne, 2 personnes, 3 personnes ou plus, cela équivaut à trois contraintes marginales.

$$\begin{aligned} \boldsymbol{O^H} \cdot \boldsymbol{W} &= \boldsymbol{M^H} \\ \boldsymbol{O^I} \cdot \boldsymbol{W} &= \boldsymbol{M^I} \\ \boldsymbol{W} &\geq 0 \end{aligned} \tag{3.1}$$

$\boldsymbol{M^H}$[11] (resp. $\boldsymbol{M^I}$) est le vecteur $n_{mH} \times 1$ (resp. $n_{mI} \times 1$) des valeurs des contraintes marginales des ménages (resp. des individus) ; n_{mH} et n_{mI} désignant respectivement le nombre de contraintes marginales ménages et individuelles ; $\boldsymbol{O^H}$ (resp. $\boldsymbol{O^I}$) représente la matrice d'occurrence des ménages (resp. des individus) selon les contraintes marginales des ménages (resp. des individus) et de taille $n_{mH} \times n_H^s$ (resp. $n_{mI} \times n_H^s$).

Le problème est mal posé dans la mesure où il existe plus de ménages que de contraintes ($n_m = n_{mH} + n_{mI} < n_H^s$). De plus, les contraintes ne sont pas forcément liées entre elles. Le problème doit donc être régularisé. Une régularisation intuitive consiste à chercher un vecteur solution $\widehat{\boldsymbol{W}}$ proche d'un vecteur de poids initiaux $\boldsymbol{W^{prior}}$, associés aux observations de l'échantillon. Ces poids initiaux représentent les poids de sondage. L'équation suivante traduit la formulation mathématique de l'idée exprimée.

$$\widehat{\boldsymbol{W}} = \underset{\substack{\boldsymbol{W} \geq 0 \\ \boldsymbol{O^H} \cdot \boldsymbol{W} \approx \boldsymbol{M^H} \\ \boldsymbol{O^I} \cdot \boldsymbol{W} \approx \boldsymbol{M^I}}}{\arg\min} \ \text{dist}(\boldsymbol{W}, \boldsymbol{W^{prior}}) \tag{3.2}$$

avec dist, une mesure des écarts entre le vecteur $\boldsymbol{W}$ et le vecteur $\boldsymbol{W^{prior}}$. Sans information cohérente sur la méthode d'échantillonnage utilisée, nous considérons que tous les poids de sondage sont identiques et égaux à l'unité ($\boldsymbol{W^{prior}} = \boldsymbol{1}$). Les techniques de résolution proposées tolèrent quelques petites différences par rapport aux contraintes marginales. De ce fait, nous n'avons plus d'égalités comme dans l'équation 3.1. Dans la suite du document, nous considérons la matrice $\boldsymbol{O}$ comme étant la concaténation des matrices d'occurrence $\boldsymbol{O^H}$ et $\boldsymbol{O^I}$, $\boldsymbol{O} = \left(\boldsymbol{O^{Ht}}\ \boldsymbol{O^{It}}\right)^t$; également $\boldsymbol{M} = \left(\boldsymbol{M^{Ht}}\ \boldsymbol{M^{It}}\right)^t$.

Les algorithmes de reconstruction synthétique décrits dans ce chapître (IPU, HIPF, ent, et GR) proposent différentes approches de résolution du problème régularisé (équa-

11. **H** est l'abréviation de household (traduction de ménage en anglais)

tion 3.2). Pour ent et GR, la minimisation est explicite bien que les mesures des écarts sont différentes. Pour l'IPU et l'HIPF, la minimisation est implicite :

— l'*IPU* propose un point de vue géométrique de l'équation 3.2. L'algorithme part de l'échantillon et considère que les poids de sondage (poids initiaux) sont uniformes. Le vecteur des poids de sondage est projeté sur l'hyperplan correspondant aux contraintes des ménages avant d'être projeté sur un second hyperplan correspondant aux contraintes des individus. La recherche d'un vecteur solution proche des poids initiaux et cohérent avec les contraintes des niveaux ménage et individuels passe par la satisfaction des contraintes ménages dans un premier temps puis des contraintes individuelles par la suite. La solution trouvée peut être considérée comme une heuristique du problème 3.2. L'IPU présente certaines limites lors de la génération simultanée d'une population synthétique d'individus et de ménages. En particulier, Ye et al. (2020) ont montré que théoriquement, l'IPU est incapable de converger vers une distribution optimale de la population qui satisfasse simultanément les contraintes des deux niveaux. Cependant, dans le cas d'étude, l'IPU génère des solutions appropriées parce que l'échantillon de départ a une grande taille.

— l'*HIPF* part de l'échantillon et considère que les poids intiaux sont également uniformes. Dans la recherche d'une solution à l'équation 3.2, les écarts sont minimisés à la fois sur le niveau ménage et le niveau individuel, l'objectif étant d'avoir des poids finaux proches des poids initiaux ;

— *ent* résout l'équation 3.2 par une approche probabiliste. L'objectif est de déterminer une probabilité p_h, associée à chaque ménage qui peut être interprétée comme le poids w_h, divisé par le nombre de ménages de la population totale n_H. La solution recherchée doit satisfaire les contraintes marginales (à la fois au niveau ménage et individuel) et minimiser l'entropie relative par rapport à une distribution de probabilité initiale uniforme [12]. En divisant les vecteurs de poids dans l'équation 3.2

12. Pour être exact, c'est la divergence de Kullback-Leilbler qui est minimisée. Le terme d'entropie relative est gardé car c'est celui qui est utilisé dans la littérature.

par la taille de la population cible et en remplaçant l'opérateur dist($\mathrm{p}_h, \mathrm{p}_h^{prior}$) par $\mathrm{p}_h \log\left(\frac{\mathrm{p}_h}{\mathrm{p}_h^{prior}}\right)$, il devient alors possible de minimiser l'entropie relative.

— *GR* propose une approche de résolution de l'équation 3.2 à travers un processus d'optimisation qui minimise les écarts entre les poids initiaux et les poids finaux sous les contraintes des marginaux des niveaux ménage et individuel.

Les quatre algorithmes introduits sont présentés de manière détaillée dans la partie suivante.

Iterative Proportional Update (IPU)

L'IPU developpé par Ye et al. (2009), est un algorithme heuristique itératif qui contrôle simultanément les contraintes marginales au niveau des individus et des ménages pendant la procédure d'ajustement. Le problème d'optimisation mathématique correspondant peut être formulé avec la fonction objectif suivante (Ye et al., 2009) :

$$\min_{w_h} \quad \sum_{j=1}^{n_m} \left[\left(\sum_h o_{jh} w_h - m_j \right) / m_j \right]^2 \tag{3.3}$$

sous la contrainte que $w_h \geqslant 0$.

w_h représente le poids d'un ménage h ($h = 1 \dots n_H^s$) ; j désigne une contrainte et n_m le nombre total de contraintes ; o_{jh} représente l'occurrence de la contrainte j dans le ménage h ; m_j est la valeur de la contrainte j dans la population.

La fonction objectif mesure les différences entre l'échantillon pondéré et les contraintes marginales. À la première itération, tous les ménages ont un poids initial de 1. L'IPU commence par ajuster les poids pour satisfaire les contraintes des ménages, puis les contraintes individuelles. À la fin de chaque itération, un indicateur de qualité de l'ajustement δ est calculé :

$$\delta = \frac{\sum_j \left[\left| \left(\sum_h o_{jh} w_h - m_j \right) \right| / m_j \right]}{n_m} \tag{3.4}$$

Le gain d'ajustement entre deux itérations consécutives est ensuite calculé ($\Delta =$

$|\delta_a - \delta_b|$). L'ensemble du processus se poursuit jusqu'à ce que le gain d'ajustement soit négligeable ou inférieur à un seuil de tolérance prédéfini. Ce seuil de tolérance est le critère de convergence qui permet à l'algorithme d'arrêter le calcul (Ye et al., 2009).

Hierarchical Iterative Proportional Fitting (HIPF)

L'algorithme HIPF (Müller and Axhausen, 2011 ; Müller, 2017) agit simultanément sur les niveaux ménage et individuel en procédant à la conversion des poids ménages en poids individuels et vice versa. Les notations suivantes sont introduites :

h désigne un ménage de l'échantillon et i un individu de l'échantillon ;

$\mathbf{H}^s$, l'ensemble des ménages de l'échantillon ($\mathbf{H}^s = \{1 \dots n_H^s\}$) ;

$\mathbf{I}^s$, l'ensemble des individus de l'échantillon ($\mathbf{I}^s = \{1 \dots n_I^s\}$) avec n_I^s le nombre d'individus de l'échantillon ;

h_i spécifie le ménage auquel chaque individu i de l'échantillon appartient ;

n_{hI} représente le nombre d'individus dans un ménage h ;

$\mathbf{I_h}$ représente les individus associés à leur ménage ($\mathbf{I_h} = \{1_h \dots n_{I_h}^s\}$).

Le processus d'ajustement est également itératif avec une boucle comportant cinq étapes.

— $k \leftarrow 0$

$w_h^0 \leftarrow 1 \quad \forall \quad h \in \mathbf{H}^s$

répeter l'opération

— étape 1 : $w_h^{(k+1)} \leftarrow \text{FIT}(w_h^{(k)}, m_{j1}^H) \quad \forall \quad h \in \mathbf{H}^s$

— étape 2 : $w_i^{(k+2)} \leftarrow w_{h_i}^{(k+1)} \quad \forall \quad i \in \mathbf{I}^s$

— étape 3 : $w_i^{(k+3)} \leftarrow \text{FIT}(w_i^{(k+2)}, m_{j2}^I) \quad \forall \quad i \in \mathbf{I}^s$

— étape 4 : $w_h^{(k+4)} \leftarrow \frac{1}{n_{hI}} \sum_{i \in \mathbf{I_h}} w_i^{(k+3)} \quad \forall \quad h \in \mathbf{H}^s$

— étape 5 : estimer $w_h^{(k+5)}$ à partir de $w_h^{(k+4)}$ avec une minimisation de l'entropie relative.

— $k \leftarrow k + 5$

jusqu'à convergence

$w_h^{(k)}$ désigne le poids d'un ménage à l'itération k. Contrairement à l'IPU où m_j ($j = 1 \dots n_m$) représente la valeur d'une contrainte marginale (ménage comme individuel), au niveau de l'HIPF, une distinction est établie entre les contraintes marginales de niveau ménage et celles de niveau individuel. m_j se décompose en m_{j1}^H (avec $j1 = 1 \dots n_{mH}$) et m_{j2}^I (avec $j2 = 1 \dots n_{mI}$) pour désigner respectivement la valeur d'une contrainte marginale de niveau ménage et la valeur d'une contrainte marginale de niveau individuel.

Lors de la première itération ($k = 0$), tous les ménages ont un poids initial de 1. Les poids ménages sont par la suite modifiés pour correspondre [13] aux contraintes marginales des ménages m_{j1}^H (étape 1). Dans l'étape 2, ces poids modifiés sont transformés en poids individuels selon le principe suivant : chaque individu se voit affecter le poids de son ménage. Les poids individuels obtenus sont modifiés pour correspondre cette fois aux contraintes marginales des individus m_{j2}^I (étape 3). Dans l'étape 4, les poids individuels sont transformés une seconde fois en poids ménages $w_h^{(k+4)}$ selon le principe suivant : le poids d'un ménage équivaut à la moyenne des poids des individus qui composent ce ménage. Dans la dernière étape (étape 5), de nouveaux poids $w_h^{(k+5)}$ sont estimés en minimisant l'entropie relative à partir des poids $w_h^{(k+4)}$ selon la formule suivante :

$$\mathrm{D}\left(w_h^{(k+5)} || w_h^{(k+4)}\right) = \sum_{h \in \mathbf{H}^s} w_h^{(k+5)} \ln \frac{w_h^{(k+5)}}{w_h^{(k+4)}} \tag{3.5}$$

en fonction des deux contraintes suivantes :

$$\sum_{h \in \mathbf{H}^s} w_h^{(k+5)} = n_H \tag{3.6}$$

La somme des poids des ménages de l'échantillon doit correspondre au total des ménages de la population réelle (n_H).

$$\sum_{h \in \mathbf{H}^s} n_{hI} w_{h_i}^{(k+5)} = n_I \tag{3.7}$$

La somme des poids individuels de tous les ménages doit correspondre au total des

13. Nous avons gardé l'expression initiale FIT en anglais car elle est la mieux appropriée pour transcrire l'idée de correspondance aux contraintes marginales.

individus de la population réelle (n_I). L'ensemble du processus (étape 1 à 5) est ainsi répété jusqu'à convergence de l'algorithme.

Minimisation de l'entropie (ent)

Bar-Gera et al. (2009) et Lee and Fu (2011) ont proposé une formulation mathématique de minimisation de l'entropie relative pour l'étape d'ajustement des poids de l'échantillon. La méthode d'optimisation de l'entropie (ent) décrite ci-après s'inspire de la formulation de Lee and Fu (2011), plus détaillée que celle de Bar-Gera et al. (2009). Toutefois, les notations sont nombreuses et complexes. Pour faciliter la compréhension, les paramètres introduits sont illustrés par des exemples.

Considérons une population de n_H ménages et n_I individus. Cette population est représentée par différents attributs. Au niveau ménage :

1. La taille du ménage avec les catégories taille 1, taille 2, taille 3 ;
2. Le statut de propriété (locataire, propriétaire) ;
3. Le type de logement (maison, appartement).

Au niveau individuel :

1. Le sexe (hommes, femmes) ;
2. La profession (étudiant, retraité, autre) ;
3. L'âge (0-20 ans, plus de 20 ans).

Soient :

— nc_{mH}, le nombre d'attributs de niveau ménage. Dans l'exemple, $nc_{mH} = 3$ (taille, statut, type) ;

— n_{mI}, le nombre de catégories individuelles. Dans l'exemple, $n_{mI} = 7$ (hommes, femmes, étudiants,..., plus de 20 ans) ;

— $\boldsymbol{\alpha}$, certains attributs ménages. $\boldsymbol{\alpha} \in \mathscr{P}(\{1 \dots nc_{mH}\})$. $\mathscr{P}$ représente l'ensemble des parties. Par exemple, $\boldsymbol{\alpha} = \{\text{taille et statut}\}$;

— $\boldsymbol{\beta}$, certaines catégories individuelles, $\boldsymbol{\beta} \in \mathscr{P}(\{1 \dots n_{mI}\})$. Dans notre exemple, $\boldsymbol{\beta} = \{$hommes, femmes, étudiant, retraité, autre$\}$;

— M_a^H, un attribut de niveau ménage avec $a = \{1 \dots nc_{mH}\}$. $M_3^H =$ type ;

— M_b^I, une catégorie de niveau individuel avec $b = \{1 \dots n_{mI}\}$. $M_7^I =$ plus de 20 ans ;

— M_α^H, un des attributs ménages associé à $\boldsymbol{\alpha}$ avec $M_\alpha^H = (M_a^H)_{a \in \boldsymbol{\alpha}}$. $M_1^H =$ taille ;

— M_β^I, une des catégories individuelles associée à $\boldsymbol{\beta}$ avec $M_\beta^I = (M_b^I)_{b \in \boldsymbol{\beta}}$. $M_1^I =$hommes ;

— $\mathbf{u}$, vecteur des différentes catégories de M_a^H. Pour M_3^H, $\mathbf{u} = \{$maison, appartement$\}$;

— $\mathbf{r}$, vecteur du nombre possible de personnes dans un ménage avec une catégorie individuelle M_b^I avec $\mathbf{r} = \{x_1 \dots x_{n_{mI}}\}$;

— u_a, une catégorie de M_a^H avec $u_a \in \Omega_a$ (l'ensemble des catégories de M_a^H) et $a = \{1 \dots nc_{mH}\}$. Ainsi pour M_3^H, $u_a =$maison ou appartement ;

— u_α, une catégorie de M_α^H avec $u_\alpha \in \prod_{a \in \alpha} \Omega_a$;

— r_b nombre de personnes dans un ménage associées à M_b^I avec $r_b \in \mathbb{N}$ et $b = \{1 \dots n_{mI}\}$. Ainsi, pour M_6^I, $r_6 =$ nombre de personnes de 0-20 ans ;

— r_β nombre de personnes dans un ménage associés à M_β^I avec $r_\beta \in \prod_{b \in \beta} \mathbb{N}$;

— $\widetilde{\mathrm{p}}_\alpha(u_\alpha)$, distribution jointe des catégories ménages associées à M_α^H. $\widetilde{\mathrm{p}}_\alpha(u_\alpha) = m_\alpha^H / n_H$ avec m_α^H, la valeur des différentes catégories de ménages associées à $\boldsymbol{\alpha}$ et n_H^s, la population totale de ménages. Par exemple, $\widetilde{p}_1(u_1) = \frac{\text{Nombre de ménages de taille 1}}{\text{Population totale de ménages}}$;

— $\widetilde{E}_\beta(r_\beta)$, la proportion d'individus attendus dans un ménage pour les catégories individuelles associées à M_β^I. $\widetilde{E}_\beta(r_\beta) = m_\beta^I / n_I$ avec m_β^I, la valeur des différentes catégories individuelles associées à $\boldsymbol{\beta}$, n_I la population des individus, $r_\beta \in \prod_{b \in \beta} \mathbb{N}$ et $\sum_\beta \widetilde{E}_\beta(r_\beta) = n_I$. Par exemple, $\widetilde{E}_1(r_1) = \frac{\text{Nombres d'hommes}}{\text{Population totale de ménages}}$;

— $\mathrm{p}_{[\mathbf{u},\mathbf{r}]}^{prior}$, distribution jointe initiale des catégories des ménages et des catégories individuelles au niveau de l'échantillon avec $\mathbf{u}$, $\mathbf{r}$ et a définis précédemment.

— $\mathrm{p}_{[\mathbf{u},\mathbf{r}]}$, distribution jointe initiale des catégories des ménages et des catégories individuelles au niveau de la population totale ;

L'objectif est de minimiser l'entropie relative (également appelée divergence de Kullback-Leibler) entre $\mathrm{p}_{[\mathbf{u},\mathbf{r}]}$ et $\mathrm{p}^{prior}_{[\mathbf{u},\mathbf{r}]}$. Il est en effet admis que la distribution jointe des attributs de la population totale peut être approchée par la distribution jointe des attributs de l'échantillon. La fonction objectif est formulée comme suit :

$$\min_{\mathrm{p}_{[\mathbf{u},\mathbf{r}]}} \quad D(\mathrm{p}_{[\mathbf{u},\mathbf{r}]} \,||\, \mathrm{p}^{prior}_{[\mathbf{u},\mathbf{r}]}) = \sum_{\mathbf{u},\mathbf{r}} \mathrm{p}_{[\mathbf{u},\mathbf{r}]} \ln\left(\frac{\mathrm{p}_{[\mathbf{u},\mathbf{r}]}}{\mathrm{p}^{prior}_{[\mathbf{u},\mathbf{r}]}}\right) \tag{3.8}$$

sous les contraintes :

$$\sum_{\{\mathbf{u},\mathbf{r}|a\notin\alpha\}} \mathrm{p}_{[\mathbf{u},\mathbf{r}]} = \widetilde{\mathrm{p}}_{\alpha}(u_{\alpha}) \quad \forall u_a \in \Omega_a, \quad a = 1\ldots nc_{mH}, u_{\alpha} \in \prod_{a\in\alpha} \Omega_a \tag{3.9}$$

$$\sum_{r_\beta} r_\beta \left(\sum_{\{\mathbf{u},\mathbf{r}|b\notin\beta\}} \mathrm{p}_{[\mathbf{u},\mathbf{r}]} \right) = \widetilde{E}_{\beta}(r_{\beta}) \quad \forall r_b \in \mathbb{N}, \qquad b = 1\ldots n_{mI}, \quad r_{\beta} \in \prod_{b\in\beta} \mathbb{N} \tag{3.10}$$

$$\mathrm{p}_{[\mathbf{u},\mathbf{r}]} \geq 0 \quad \sum_{\mathbf{u},\mathbf{r}} \mathrm{p}_{[\mathbf{u},\mathbf{r}]} = 1 \tag{3.11}$$

Les équations 3.9 et 3.10 représentent respectivement les contraintes sur les ménages et les individus. l'équation 3.8 constitue une implémentation de l'équation 3.2, avec $\mathrm{p}_{[\mathbf{u},\mathbf{r}]}$ qui remplace $\mathrm{p}_h = \frac{w_h}{n_H}$ et l'opérateur $\mathrm{dist}(\mathrm{p}_h, \mathrm{p}^{prior}_h)$ remplacé par $\mathrm{p}_{[\mathbf{u},\mathbf{r}]} \ln\left(\frac{\mathrm{p}_{[\mathbf{u},\mathbf{r}]}}{\mathrm{p}^{prior}_{[\mathbf{u},\mathbf{r}]}}\right)$

Calage sur marges (GR)

Les techniques de calage sur marges (GR) développées par Deville et al. (1993) permettent de générer une population synthétique d'individus et de ménages. Ces techniques permettent d'ajuster les poids initiaux d'un échantillon afin de correspondre aux données marginales individuelles et ménages disponibles. La formalisation mathématique de GR s'inspire de Deville et al. (1993) et Müller (2017). Soient :

— un échantillon s constitué de n^s_H ménages et de n^s_I individus ;

— m_j, la valeur d'une contrainte marginale (ménage comme individuel) avec $j = 1\ldots n_m$ (n_m est le nombre total de contraintes) ;

— w_h^{prior}, les poids initiaux associés aux observations de l'échantillon. Il s'agit des poids de sondage ;

— o_{jh} , l'occurrence de la contrainte j dans le ménage h.

L'objectif est de chercher de nouvelles pondérations appelées poids de calage et notées w_h qui soient aussi proches que possible, au sens d'une certaine « fonction de distance G », des pondérations initiales w_h^{prior} et qui correspondent aux valeurs des différentes contraintes marginales j. Les w_h doivent vérifier l'équation suivante :

$$\forall j = 1 \dots n_m, \quad \sum_{h \in s} w_h o_{jh} = m_j \tag{3.12}$$

La fonction de distance G, d'argument $r = w_h / w_h^{prior}$ utilisée pour mesurer les distances entre les w_h et les w_h^{prior} est positive, convexe et vérifie G(1)=0. Les poids finaux obtenus w_h doivent minimiser la quantité $D = \sum_{h \in s} w_h^{prior} G(w_h / w_h^{prior})$ sous les contraintes de calage exprimées dans l'équation 3.12. Il existe quatre fonctions de distance différentes (linéaire, raking-ratio, logit et linéaire tronquée). [14]

Les algorithmes SR décrits (IPU, HIPF, ent et GR) caractérisent des poids ménages et individuels décimaux. Il n'est pas donc pas possible de répliquer les différentes observations de l'échantillon en fonction de leur poids. Pour générer la population synthétique souhaitée, il est nécessaire de rendre ces poids entiers.

3.3.2 Etape d'affectation des ménages et des individus

Deux méthodes de conversion probabilistes sont testées : probabilités proportionnelles et l'approche Truncate Replicate Sample (troncature, réplication, échantillonnage). Selon Lovelace et al. (2015) ces deux méthodes sont plus précises que les méthodes déterministes de simple arrondi ou d'approche par les seuils.

14. Pour plus de précisions sur ces fonctions de distance, se référer à Deville et al. (1993).

L'approche des probabilités proportionnelles

L'approche des probabilités proportionnelles (PP) considère les poids décimaux obtenus comme des probabilités (Lovelace et al., 2015 ; Joubert, 2018). Par exemple, la probabilité p_h qu'un ménage donné se trouve dans la population synthétique finale se traduit par la probabilité $p_h = w_h / \sum w_h$. Ainsi donc, plus le poids décimal est élevé, plus la probabilité qu'un individu/ménage se trouve dans la population finale est grande. Par conséquent, un individu avec un poids très élevé peut être répliqué plusieurs fois, tandis qu'un individu avec un poids très faible peut ne pas être inclus dans la population synthétique finale.

L'approche Truncate Replicate Sample (TRS)

L'approche Truncate Replicate Sample (TRS) (Lovelace and Ballas, 2013) combine un échantillonnage déterministe et probabiliste afin de générer des poids entiers selon un processus en trois étapes : troncature, réplication et échantillonnage.

1. L'étape de troncature permet d'obtenir des valeurs entières en supprimant toutes les informations à droite du point décimal. Les restes décimaux (entre 0 et 1) sont alors conservés. À titre d'exemple, un ménage dont le poids est de 4,65 aura une valeur tronquée de 4. Son reste décimal est de 0,65.
2. Lors de la deuxième étape (réplication), les individus/ménages sont répliqués en fonction de leurs poids entiers obtenus lors de l'étape de troncature. Seuls les poids tronqués supérieurs à 0 sont répliqués. Par exemple, le ménage avec un poids de 4,65 sera répliqué 4 fois. Un autre ménage avec un poids de 0,99 ne sera pas répliqué dans cette étape (sa valeur tronquée étant 0). Lors de la troncature et de la réplication, il n'y a aucun risque de sur-échantillonnage (la somme de tous les poids entiers sera toujours inférieure à la taille de la population).
3. Au cours de la dernière étape (échantillonage), on estime la population de ménages/individus restants $n_H - \sum_{h \in s} \lfloor w_h \rfloor$. On procède ensuite à un tirage aléatoire sans remise à partir des restes décimaux pour compléter la population d'individus

et de ménages. Les probabilités de sélection sont égales aux restes de poids décimaux. Dans notre exemple, le ménage avec le poids de départ de 4,65 aura une probabilité de 0,65 d'être choisi à nouveau, tandis que l'autre ménage aura une probabilité de 0,99.

3.3.3 Mesures de validation utilisées

Deux éléments peuvent être considérés dans l'évaluation de la précision d'une population synthétique : la validation interne et la validation externe. La validation interne consiste à comparer les marginaux de la population synthétique avec les marginaux de la population réelle afin de tester la fiabilité des données générées. Par exemple, la distribution estimée de la structure familiale dans la population synthétique correspond-elle à la distribution réelle (distribution des données de recensement) ? Une validation interne teste donc la capacité de la population à s'ajuster aux données agrégées. Une validation est externe si les attributs estimés de la population synthétique sont comparés à une autre source de données non utilisée dans le processus d'estimation. Dans ce cas d'étude, il n'existe pas de source externe de données au recensement français au niveau de l'IRIS ou de la commune. Par conséquent, seule une validation interne peut être réalisée. Selon la littérature, la validation interne peut être effectuée sur les attributs agrégés (les données agrégées de la population sont comparées aux données agrégées de la population synthétique), sur les observations ou sur la population synthétique entière. Il existe de nombreuses mesures de validation interne (Lovelace et al., 2015 ; Timmins et al., 2016). Les mesures suivantes ont été considérées :

— le coefficient de détermination R^2 qui correspond au carré du coefficient de corrélation de Pearson. Il s'agit d'un indicateur qui varie entre 0 et 1 et révèle dans quelle mesure les valeurs simulées correspondent aux données réelles. Une valeur de R^2 de 1 dénote d'un ajustement parfait entre données simulées et données réelles tandis qu'une valeur de R^2 proche de zéro suggère une absence de correspondance entre les données réelles et données simulées (Lovelace et al., 2015) ;

— l'erreur absolue totale (TAE) et l'erreur absolue standardisée (SAE). Le TAE est la

somme de la différence entre les valeurs agrégées simulées et les données marginales de la population et le SAE correspond au TAE divisé par la population totale ;

— l'erreur quadratique moyenne standardisée (SRMSE). Cet indicateur porte sur la dispersion des erreurs et est utilisé pour évaluer la qualité de l'ajustement entre la population synthétique estimée et les marginaux. Le SRMSE est l'un des indicateurs les plus couramment utilisés pour évaluer une population synthétique (Lee and Fu, 2011 ; Lovelace et al., 2015 ; Sun and Erath, 2015 ; Saadi et al., 2016). Une valeur nulle indique une correspondance parfaite entre les données de recensement et la population synthétique, tandis qu'une valeur élevée traduit une mauvaise correspondance ;

— la méthode de Bland-Altman, méthode graphique pouvant être utilisée en complément des autres inddicateurs définis précedemment (Timmins et al., 2016). Les graphiques de Bland-Altman permettent une mesure de la concordance d'une même grandeur selon des techniques différentes. Dans le cas des générations de populations synthétiques, ils permettent d'apprécier l'écart entre valeurs simulées et valeurs réelles et de déduire le biais, la précision et les limites de l'intervalle de confiance à 95 %.

3.4 Résultats et discussion

Une population synthétique d'individus et de ménages est générée pour chaque algorithme SR (IPU, HIPF, entr, GR) associé à une méthode de conversion des poids (probabilités proportionnelles et TRS). Les résultats obtenus sont par la suite évalués afin de déterminer la population synthétique la plus proche de la population réelle. Dans le processus de génération, les individus hors ménage (personnes résidant dans une communauté, les personnes vivant dans des habitations mobiles et les personnes sans-abri) sont exclus car ils ne possèdent pas de caractéristiques ménages. Certains IRIS et communes ont également été regroupés afin de respecter les labels de qualité édictés par l'Insee et de disposer de données stables et robustes. L'échantillon final d'individus et

de ménages est alors constitué exactement de 287 017 individus répartis dans 136 501 ménages et 307 IRIS et communes. La taille moyenne d'un IRIS/commune de l'AU de Nantes est de 1 363,3 ménages ($\pm 631,3$) [15] et 3 092,2 individus ($\pm 1\,334,6$) ; l'échantillon contient 32,30% de ménages ($\pm 11,36$) et 30,87% d'individus (± 9.96) de la population réelle. Le Tableau 3.3 décrit les attributs et les contraintes marginales (catégories) utilisées dans le processus de génération. Les Figures 3.1 et 3.2 traduisent les distributions (en pourcentage) des différentes contraintes au niveau individuel et ménage à l'intérieur des 307 IRIS. Pour la plupart de ces distributions, on observe une assez grande variabilité.

15. ± 631,3 correspond à plus ou moins un écart type de 631,3 ménages

Tableau 3.3 : Contraintes marginales utilisées dans le processus de génération de la population synthétique

Niveau	Attribut	Définition [nombre de contraintes]	Contraintes (catégories)
Ménage	SFM_agg	Structure familiale [5]	Personnes vivant seules ; Couple sans enfants ; Couple avec enfants ; Famille monoparentale ; Autre composition
	CSP_agg	Profession de la personne de référence* [7]	Agriculteurs, artisans ; Cadres et professions intellectuelles supérieures ; Professions intermédiaires ; Employés ; Ouvriers ; Retraités ; Autres personnes sans activités professionnelles
	NPERR_bin	Taille du ménage [2]	Une personne ; Deux personnes ou plus
	VOIT_rec	Nombre de voitures [3]	Aucune voiture ; Une ; Deux ou plus
Individu	AGEREVband	Age [12]	0-2 ; 3-5 ; 6-10 ; 11-14 ; 15-17 ; 18-24 ; 25-29 ; 30-39 ; 40-54 ; 55-64 ; 65-79 ; 80/+
	SEXE_rec	Sexe [2]	Femmes ; Hommes
	LPRM1	Lien avec la personne de référence du ménage [2]	Personne de référence ; Autre membre du ménage
	CSP	Profession [7]	Agriculteurs, artisans ; Cadres et professions intellectuelles supérieures ; Professions intermédiaires ; Employés ; Ouvriers ; Retraités ; Autres personnes sans activités professionnelle
	EMPL_rec	Condition d'emploi [7]	En emploi à durée déterminée ; Emploi à durée indéterminée ; Indépendants ; Élèves, étudiants, stagiaires non rémunérés de 14 ans ou plus ; En situation de chômage ; Moins de 15 ans ; Autres inactifs
	TP_rec	Temps de travail [3]	Temps complet ; Temps partiel ; Non concerné

Note : *La personne de référence (PR) du ménage est déterminée à partir de la structure familiale du ménage et des attributs des individus qui le composent. Les critères permettant de désigner une PR ont évolué au fil des recensements. La définition actuelle ne prend plus en compte le critère du sexe. Les critères actuels sont l'activité, le fait d'avoir un conjoint, le fait d'avoir un enfant et l'âge.

FIGURE 3.1 : Distributions (en pourcentage) des différentes contraintes au niveau ménage à l'intérieur des 307 IRIS et communes

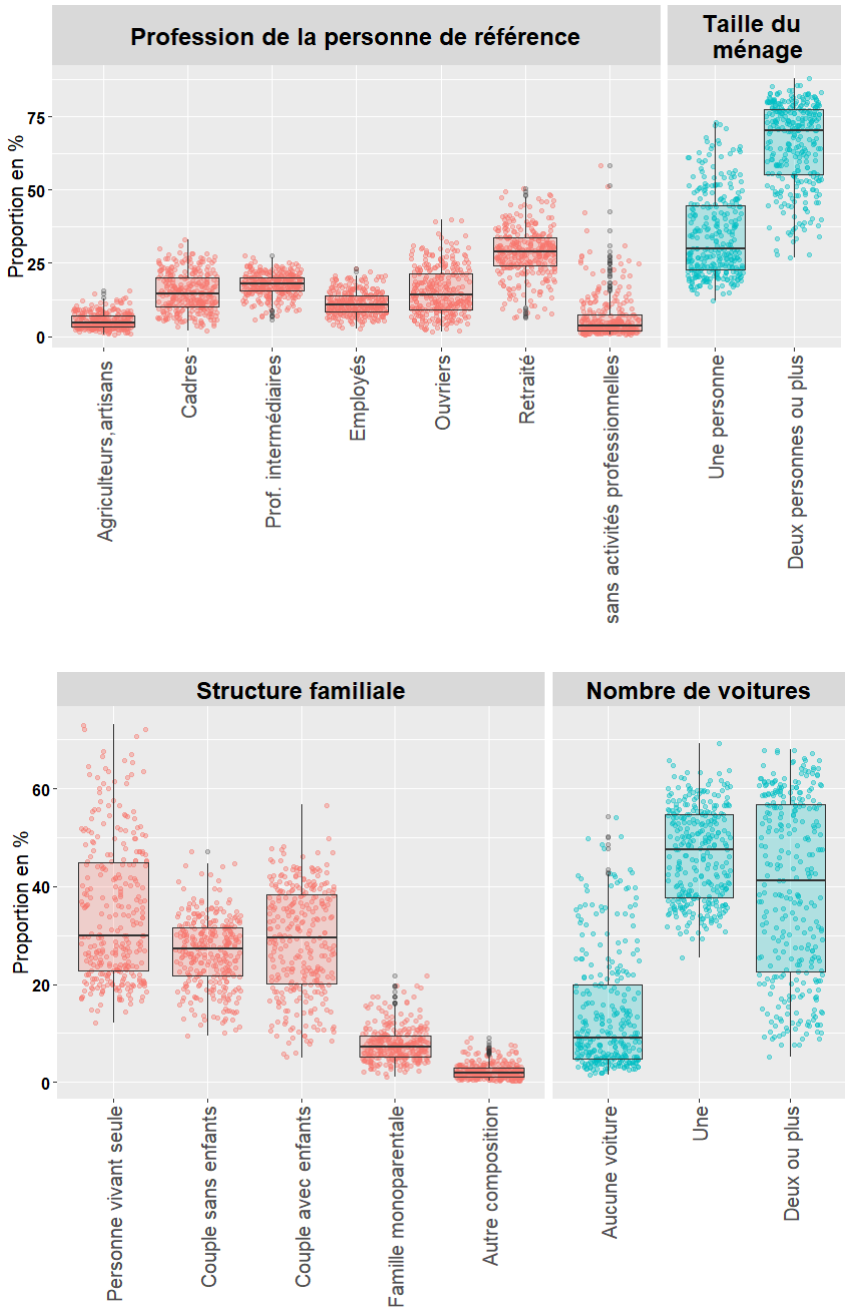
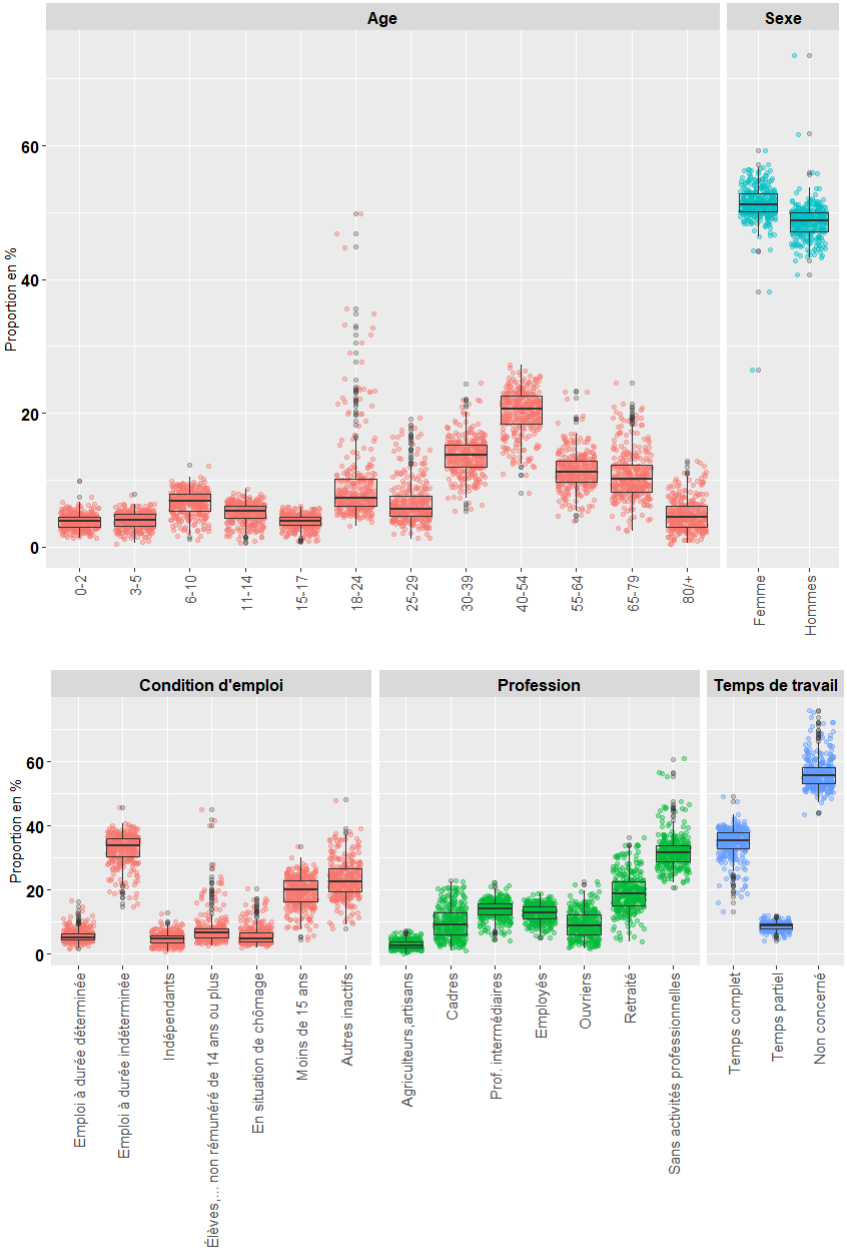


FIGURE 3.2 : Distributions (en pourcentage) des différentes contraintes au niveau individuel à l'intérieur des 307 IRIS et communes



3.4.1 Validation interne sur les attributs

Toutes les analyses ont été effectuées avec le logiciel et le langage de programmation R (version 4.0.3), sur une machine Windows 10 professionnel, Intel(R) Core (TM) i7-8665U CPU @ 1.90GHz, 2.11 GHz et 16 GB de RAM. L'implémentation des algorithmes de reconstruction synthétique a été réalisée avec la librairie R « mlfit »(Müller 2021) [16] disponible sur le « Comprehensive R Archive Network »(CRAN). En ce qui concerne la technique GR, la convergence n'a été atteinte qu'avec la fonction de distance logit. Les temps de convergence (en minutes) sont de 2,50 ; 6,19 ; 44,24 et 5,44 respectivement pour l'IPU, l'HIPF, ent et GR (logit).

Les résultats de la validation montrent que toutes les associations proposées (algorithmes et méthodes de conversion) produisent des populations synthétiques représentatives de la population réelle mais certaines s'avèrent plus efficaces. En considérant le premier indicateur R^2 :

— HIPF ou IPU associé avec la méthode TRS (HIPF+TRS ou IPU+TRS) donnent des coefficients supérieurs à 0,99 ;

— HIPF ou IPU associé avec la méthode des probabilités proportionnelles (HIPF+PP ou IPU+PP) et la minimisation de l'entropie associée avec le TRS or les probabilités proportionnelles (ent+TRS or ent+PP) donnent des coefficients supérieurs ou égaux à 0,98 ;

— GR associé avec le TRS ou les probabilités proportionnelles (GR+TRS ou GR+PP) donnent des coefficients supérieurs ou égaux à 0,97.

La méthode TRS donne de meilleurs résultats que la méthode des probabilités proportionnelles ; l'algorithme GR est moins précis que les trois autres algorithmes de génération. La mesure de validation du R^2 ne fournit qu'une indication de l'ajustement et peut être influencée par d'éventuelles valeurs aberrantes. Une analyse plus poussée, basée sur les trois autres indicateurs (TAE, SAE et SRMSE) est présentée dans le Tableau 3.4. Ces résultats confirment que toutes les méthodes sont globalement efficaces,

16. Disponible à < `https://cran.r-project.org/web/packages/mlfit/index.html`>, page consultée le 25 juillet 2021.

mais que la minimisation de l'entropie et l'HIPF sont plus performantes que les autres.

Tableau 3.4 : Synthèse des résultats obtenus avec les indicateurs TAE, SAE et SRMSE

	Niveau individuel			Niveau ménage		
Méthode	TAE	SAE (%)	SRMSE	TAE	SAE (%)	SRMSE
IPU+TRS	87 046	1,53	0,0024	17 082	1,02	0,0012
IPU+PP	188 191	3,30	0,0032	56 529	3,37	0,0027
HIPF+TRS	53 436	0,94	0,0013	14 134	0,84	0,0007
HIPF+PP	176 612	3,10	0,0027	54 564	3,26	0,0025
ent+TRS	49 658	0,87	0,0007	16 031	0,96	0,0009
ent+PP	168 732	2,96	0,0024	55 692	3,33	0,0026
GR+TRS	250 362	4,39	0,0093	69 867	4,17	0,0098
GR+PP	326 076	5,72	0,0093	106 861	6,38	0,0127

Note : IPU : Iterative Proportional Update ; HIPF : Hierarchical Iterative Proportional Fitting ; GR : Generalized Raking ; ent : minimisation de l'entropie ; TRS : Truncation, Replication, Sampling ; PP : probabilités proportionnelles.

A partir du Tableau 3.4, les méthodes peuvent être classées dans l'ordre suivant, de la plus précise à la moins précise :

— minimisation de l'entropie, HIPF, IPU et GR pour le niveau individuel ;

— HIPF, minimisation de l'entropie, IPU et GR pour le niveau ménage ;

— TRS et probabilités proportionnelles.

Selon tous les indicateurs de validation considérés (R^2, TAE, SAE et SRMSE), de légères différences ressortent entre la minimisation de l'entropie et l'HIPF. Ces deux algorithmes sont toutefois plus performants que l'IPU et le calage sur marges. En ce qui concerne les méthodes de conversion des poids, la méthode TRS est meilleure que celle des probabilités proportionnelles. Les algorithmes HIPF et la minimisation de l'entropie

combinés au TRS fournissent donc la meilleure approximation possible de la population réelle.

3.4.2 Analyse au niveau IRIS

En parallèle à l'analyse relative aux contraintes, une analyse IRIS par IRIS a été réalisée afin d'identifier les zones présentant les erreurs les plus élevées (IRIS dont les valeurs SAE sont les plus élevées). Pour chaque association testée, que ce soit au niveau de l'individu ou du ménage, trois IRIS ressortent. Le Tableau 3.5 présente les valeurs de ces trois IRIS pour les deux meilleurs algorithmes (HIPF et ent) associées au TRS.

Tableau 3.5 : IRIS avec les valeurs les plus élevées pour le SAE

		Niveau individuel	Niveau ménage
Méthode	Iris Id	SAE (%)	SAE (%)
HIPF+TRS	136	35,25	6,63
	237	13,71	2,82
	162	10,08	3,09
ent+TRS	136	14,28	36,70
	237	3,71	10,37
	162	2,40	7,51

Note : HIPF : Hierarchical Iterative Proportional Fitting ; ent : minimisation de l'entropie ; TRS : Truncation, Replication, Sampling,

Une analyse qualitative des contraintes marginales de ces trois IRIS souligne leurs caractéristiques particulières. Les IRIS 136 et 237 sont des zones d'activités qui comptent un petit nombre de ménages et d'individus. La population de l'IRIS 136 (resp. 237) est de 359 (resp. 327) ménages et 964 (resp. 810) individus. Dans ces deux IRIS, 67% (resp. 51%) des ménages sont des ménages individuels ; la plupart des habitants de ces IRIS sont des hommes (73%) (resp. 62%) et appartiennent à la tranche d'âge 18-54 ans. L'IRIS 162 est une zone résidentielle comptant 1 163 ménages et 2 133 individus. Cependant, une partie importante du territoire est occupée par un hôpital psychiatrique. Dans cet IRIS, 65% des ménages sont composés d'une seule personne et 65 % des individus ont entre 15

et 64 ans. En conclusion, les simulations s'avèrent proches des données réelles pour tous les IRIS, sauf pour quelques uns d'entre eux, en raison de la répartition particulière de la population de ces IRIS.

3.4.3 Approche de Bland-Altman

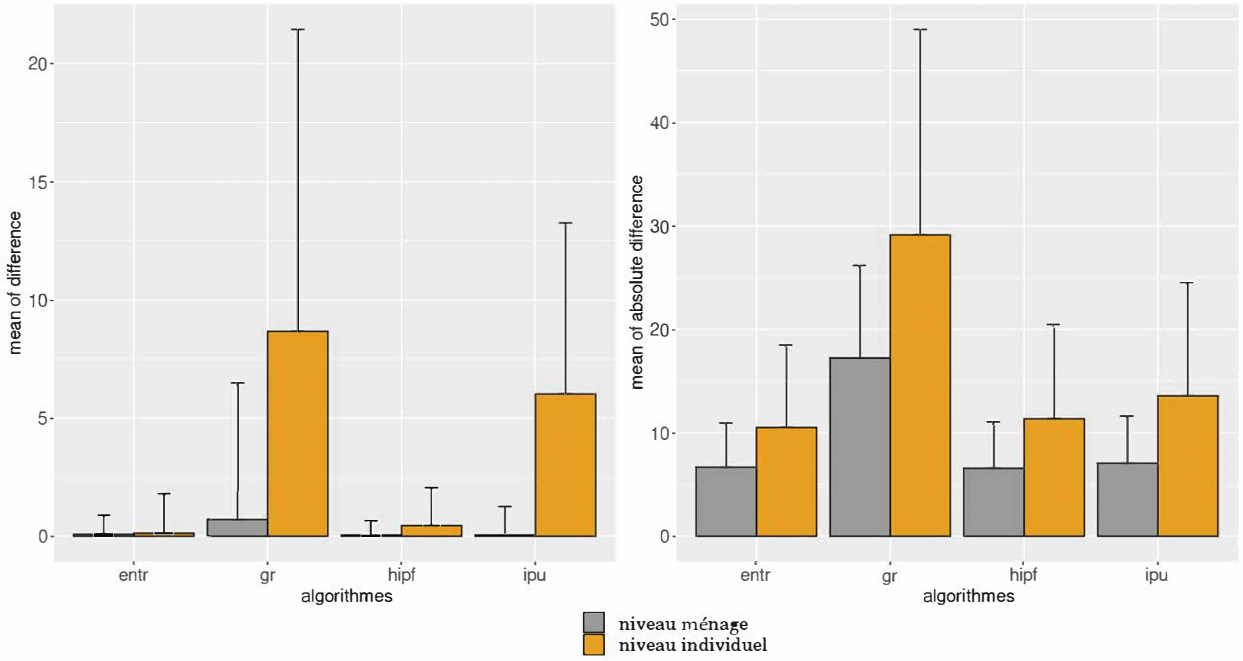
La méthode de Bland-Altman permet de comparer les valeurs simulées de la population synthétique à celles du recensement. Cette méthode consiste à réaliser un graphe étudiant en ordonnées la différence entre deux valeurs et en abscisses la moyenne de deux valeurs obtenues (Bland and Altman, 1999). Il est alors possible de tracer deux seuils de limites de concordance définis par $\pm$ 1.96 $\times$ l'écart type de la moyenne. Cette analyse graphique aide à identifier certaines anomalies telles qu'une surestimation ou une sous-estimation systématique des valeurs obtenues par des méthodes SR (Kalra et al., 2017).

La méthode a été appliquée à chaque contrainte (Tableau 3.3) et en considérant chaque algorithme SR associé à une méthode de conversion (50 $\times$ 4 $\times$ 2), ce qui donne 400 cas possibles. La moyenne des différences (en termes réels et absolus) entre les valeurs simulées et les valeurs de recensement a été calculée. La Figure 3.3 montre la moyenne et l'écart-type de ces moyennes pour chaque algorithme avec le TRS.

Il ressort que la moyenne des différences entre les valeurs simulées et les valeurs du recensement, exprimée en termes réels, est proche de zéro pour les méthodes HIPF et ent, avec un écart-type plutôt faible. Les différentes valeurs de la Figure 3.3 confirment que les méthodes HIPF et ent sont plus précises que GR et IPU.

La Figure 3.4 présente les valeurs de Bland-Altman pour les cinq catégories de la variable structure familiale (Tableau 3.3). Nous avons un total de 307 points, chaque point représentant un IRIS. En ordonnées, les différences entre les valeurs simulées et les valeurs réelles sont représentées. L'axe des abscisses représente les moyennes des deux valeurs. Les traits en pointillés correspondent aux limites de concordance (pouvant être assimilées à des intervalles de confiance des différences). Selon l'IRIS, les valeurs simulées sont dans certains cas supérieures et dans d'autres cas inférieures aux valeurs

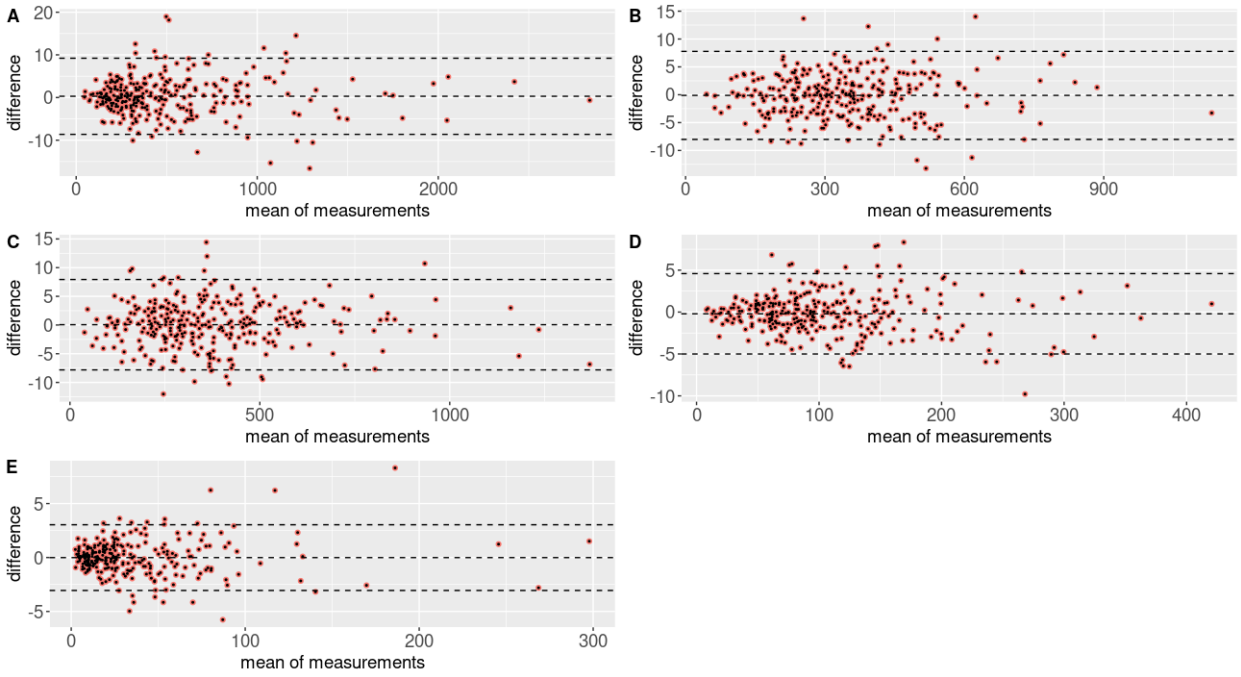
FIGURE 3.3 : Approche de Bland-Altman : moyenne de la différence (en termes réels et absolus) entre les valeurs simulées et les valeurs de recensement pour chaque méthode d'ajustement associée au TRS



Note : mean of difference : moyenne de la différence (en termes réels) ; mean of absolute difference : moyenne de la différence en termes absolus.

du recensement. Toutefois, la quasi-totalité des points se retrouvent compris entre les deux limites de concordance et la grande majorité autour de zéro, ce qui dénote dune bonne estimation. Pour la catégorie couples sans enfants par exemple (notée B dans la Figure 3.4), sur des effectifs de 600 ménages, l'erreur d'estimation est d'environ de 15 ménages.

FIGURE 3.4 : Approche de Bland-Altman : différences entre les valeurs de recensement et les valeurs simulées générées avec la technique HIPF associée au TRS pour les cinq catégories de la variable structure familiale



Note : A : Personne vivant seule ; B : Couple sans enfants ; C : Couple avec enfants ; D : Famille monoparentale ; E : Autre composition.

Conclusion

Ce chapitre a été consacré à la génération effective d'une population synthétique d'individus et de ménages. Pour un cas d'application français, il ressort que le recensement de la population présente d'importants avantages comme source de données à utiliser. Les résultats du recensement sont disponibles à différentes échelles géographiques et sont diffusés sous deux types :

1. Des échantillons d'individus avec leurs attributs sociodémographiques, les attributs de leur ménage et ceux de leur logement appelés fichiers détail. Ces échantillons correspondent à plus de 30% de la population totale ;
2. des données agrégées qui traduisent les effectifs globaux de certains attributs indi-

viduel, ménage et logement.

Ces données sont dans leur grande majorité librement accessibles et sont identiques à l'échelle de la France, ce qui facilite les possibilités d'utilisation. La configuration des données du recensement montre que les méthodes de reconstruction synthétique sont plus adaptées pour générer la population synthétique souhaitée. La suite du chapitre a consisté :

— en une description détaillée de quatre algorithmes de reconstruction synthétique : l'Iterative Proportional Update (IPU), le Hierarchical Iterative Proportional Fitting (HIPF), la minimisation de l'entropie (ent) et le calage sur marges (GR) ;

— en une présentation de deux méthodes de conversion des poids décimaux en poids entiers à savoir l'approche des probabilités proportionnelles (PP) et l'approche TRS (Truncate Replicate Sample) pour troncature, réplication, échantillonnage ;

— en une évaluation des méthodes et algorithmes présentés.

Le cas d'application a porté sur la génération d'une population synthétique pour l'aire urbaine (AU) de Nantes composée de 418 000 ménages et 949 000 individus à partir d'un échantillon de 136 501 ménages et 287 017 individus. L'échelle géographique considérée pour la génération est la commune ou l'IRIS de résidence de l'individu. La population synthétique a été générée avec quatre attributs au niveau des ménages et six attributs au niveau individuel pour un total de cinquante catégories. L'Aire Urbaine compte 307 IRIS et communes. Le nombre total de contraintes est alors de 15 350. Les résultats ont été validés à l'aide de différents indicateurs : R^2, TAE, SAE, SRMSE et l'approche de Bland-Altman. Les résultats obtenus ont montré que toutes les associations (algorithmes et méthodes de conversion) considérées génèrent une population synthétique à deux niveaux dont les caractéristiques sont proches des données du recensement. Les résultats les plus précis sont toutefois obtenus en combinant la minimisation de l'entropie et le Hierarchical Iterative Proportional Fitting avec la méthode TRS.

La comparaison des différentes algorithmes effectuée dans ce chapitre avec un cadre théorique commun et des notations communes facilitera leur meilleure diffusion. Ce cadre

commun permet également de mieux positionner ces algorithmes par rapport aux autres méthodes de génération qui existent et peut éventuellement stimuler le développement de nouvelles méthodes.

L'étape suivante consiste à enrichir la population synthétique de nouveaux attributs.

Références bibliographiques

Bar-Gera, H., Konduri, K., Sana, B., Ye, X., and Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2) :135–160.

Brilhault, G. and Caron, N. (2016). Le passage à une collecte par sondage : quel impact sur la précision du recensement ? *Économie et statistique*, 483(1) :23–40.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423) :1013–1020.

Godinot, A. (2005). Pour comprendre le recensement de la population. *INSEE méthodes*.

Joubert, J. W. (2018). Synthetic populations of south african urban areas. *Data in brief*, 19 :1012–1020.

Kalra, A. et al. (2017). Decoding the bland–altman plot : basic review. *Journal of the Practice of Cardiovascular Sciences*, 3(1) :36.

Lee, D.-H. and Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record*, 2255(1) :20–27.

Lovelace, R. and Ballas, D. (2013). truncate, replicate, sample : A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41 :1–11.

Lovelace, R., Birkin, M., Ballas, D., and van Leeuwen, E. (2015). Evaluating the performance of iterative proportional fitting for spatial microsimulation : new tests for an established technique. *Journal of Artificial Societies and Social Simulation*, 18(2).

Müller, K. (2017). *A generalized approach to population synthesis*. PhD thesis, ETH Zurich.

Müller, K. and Axhausen, K. W. (2011). Hierarchical ipf : Generating a synthetic population for switzerland. *paper presented at the 51st Congress of the European Regional Science Association*.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., and Cools, M. (2016). Hidden markov model-based population synthesis. *Transportation Research Part B : Methodological*, 90 :1–21.

Sun, L. and Erath, A. (2015). A bayesian network approach for population synthesis. *Transportation Research Part C : Emerging Technologies*, 61 :49–62.

Templ, M., Meindl, B., Kowarik, A., and Dupriez, O. (2017). Simulation of synthetic complex data : The r package simpop. *Journal of Statistical Software*, 79(10) :1–38.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of ill posed problems*. V. H. Winston and Sons (Wiley), New York.

Timmins, K. A., Edwards, K. L., et al. (2016). Validation of spatial microsimulation models : a proposal to adopt the bland-altman method. *International Journal of Microsimulation*, 9(2) :106–122.

Ye, P., Tian, B., Lv, Y., Li, Q., and Wang, F.-Y. (2020). On iterative proportional updating : Limitations and improvements for general population synthesis. *IEEE Transactions on Cybernetics*.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., and Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Chapitre 4

Méthodologie d'ajout d'un attribut à une population synthétique à partir de données agrégées : exemple de l'attribut de revenu

Sommaire

Résumé

Ce chapitre analyse la problématique de l'ajout d'un attribut à une population synthétique à partir de données agrégées. Il s'agit d'un aspect qui a reçu peu d'attention dans la littérature. Le travail présenté ici enrichit donc la littérature existante en proposant une méthodologie nouvelle et efficace pour répondre à ce besoin pratique. La méthodologie intègre trois étapes distinctes dont la première modélise théoriquement le problème comme une distribution multinomiale. L'ajout d'un nouvel attribut est formulé comme une maximisation de l'entropie en associant les attributs disponibles dans la population synthétique et les données agrégées. La résolution de ce problème (dans ce cas d'étude spécifique) n'est pas possible en raison du grand nombre de contraintes impliquées. La deuxième étape présente alors une heuristique apportant une solution pratique au problème. Cette heuristique combine le théorème de Bayes avec l'algorithme de minimisation de l'entropie croisée. En raison du nombre élevé de paramètres, certains des résultats obtenus s'avèrent non cohérents. Pour remédier à ce problème, une méthode de post-traitement est appliquée au cours d'une troisième étape pour garantir la cohérence des résultats. La méthodologie est décrite en détail et des exemples sont fournis pour une meilleure compréhension de ces trois étapes. De plus, cette méthodologie est appliquée à un cas réel. Un niveau de vie a été attribué à chacun des 157 000 ménages synthétiques de la commune de Nantes sur la base des données agrégées du dispositif FiLoSoFi. La méthodologie proposée reste générale et est applicable à d'autres attributs pour lesquels des données agrégées sont disponibles.

Mots clés

Données agrégées, entropie, revenu, population synthétique, FiLoSoFi.

Introduction

Pour générer une population synthétique (ménages et/ou individus), les méthodes les plus utilisées sont basées sur des échantillons. Lorsque ces méthodes sont appliquées, il n'est possible de générer une population synthétique qu'à partir des attributs présents dans l'échantillon. Par exemple, si un attribut de revenu n'est pas inclus dans l'échantillon, une population synthétique avec un revenu ne peut être directement générée. En fonction des besoins de génération, il est probable que tous les attributs d'intérêt ne soient pas présents dans l'échantillon. Dans de telles situations, il peut s'avérer nécessaire d'enrichir la population synthétique avec des informations supplémentaires. L'approche traditionnelle fait appel à l'appariement statistique (statistical matching) qui s'appuie sur des informations disponibles dans différentes sources de données. Selon D'Orazio et al. (2006), l'appariement statistique vise à intégrer deux (ou plusieurs) sources de données caractérisées par le fait que : 1) les différentes sources de données contiennent des informations sur un ensemble d'attributs communs et des attributs qui ne sont pas observés conjointement, et 2) les unités observées dans les sources de données sont différentes.

Dans le cadre d'un ajout d'attributs à une population synthétique par appariement statistique, la source de données supplémentaire est le plus souvent un échantillon de données désagrégées moins important que l'échantillon du recensement mais qui comporte des attributs absents de l'échantillon du recensement. Il est par exemple possible de combiner les données du recensement et les données des Enquêtes Ménages Déplacements (EMD). Le processus d'appariement nécessite au préalable une opération de sélection. Dans cette opération, une liste d'attributs communs aux données de la population synthétique et à l'échantillon supplémentaire est définie. Ces attributs communs sont appelés attributs d'appariement et peuvent être : le sexe, l'âge, la profession ou la taille du ménage. Après cette étape, une opération d'appariement est menée entre la population synthétique et l'échantillon supplémentaire sur ces attributs communs : pour chaque observation synthétique, seules les observations de l'échantillon supplémentaire

identiques sur ces attributs sont éligibles pour l'appariement (Bösch et al., 2016). À la fin du processus, les observations synthétiques héritent des attributs de l'échantillon supplémentaire. Plus le nombre d'attributs communs est élevé, plus l'opération d'appariement est précise.

Ce schéma de sélection et d'appariement a été utilisé pour affecter des attributs supplémentaires (précisément des plans d'activités et de déplacements quotidiens) à une population synthétique générée pour la Suisse (Hackl and Dubernet, 2019), Jakarta (Indonésie) (Ilahi and Axhausen, 2019), Rouen (France) (Vosooghi et al., 2019), Carinthia (Autriche) (Felbermair et al., 2020), New York (Etats-Unis) (He et al., 2020) et Sao Paulo (Brésil) (Sallard et al., 2020). Les plans d'activités et de déplacements provenaient d'enquêtes sur les déplacements des ménages.

Zhang et al. (2019) ont utilisé une autre méthode pour attribuer une communauté (définie par un ensemble d'individus possédant des liens au sein d'un groupe) à chaque individu synthétique. Ils utilisent des données de recensement et un échantillon de journaux d'appels collectés au niveau des opérateurs de téléphonie mobile. Cependant, en raison de la protection de la vie privée, aucune information personnelle n'était disponible dans l'échantillon d'appels. La méthode d'appariement statistique ne pouvant être utilisée, les auteurs ont formulé l'affectation des communautés comme un problème de programmation en nombres entiers.

L'appariement statistique ou la programmation en nombres entiers ne s'appliquent qu'à des situations où les nouvelles sources de données introduites contiennent des données désagrégées. Toutefois, de telles sources de données ne sont pas toujours disponibles. En revanche, les chercheurs et les praticiens ont souvent accès à des bases de données agrégées pour des attributs d'intérêt comme par exemple la distribution du revenu ou du niveau d'éducation. Lorsqu'on ajoute des attributs à une population synthétique à partir de données agrégées, les deux méthodes décrites précédemment ne peuvent plus être utilisées.

Quelques approches ont été développées pour ajouter de nouveaux attributs à une population synthétique à partir de sources de données agrégées. Bösch et al. (2016), Hackl

and Dubernet (2019) ont attribué un lieu d'activité (lieux de travail et d'enseignement) à partir de statistiques agrégées sur les navetteurs. Pour chaque employé synthétique, le lieu de travail est dérivé comme suit : pour toutes les personnes d'une commune ayant un mode de déplacement identique, les communes de lieu de travail sont échantillonnées de manière aléatoire à partir de la matrice observée des navetteurs, pondérée par leurs fréquences relatives. Les lieux d'enseignement les plus proches sont affectés aux personnes concernées selon leur niveau d'éducation.

Murata et al. (2017) ont attribué un revenu à chaque travailleur d'une population synthétique issue d'une source de données agrégées en utilisant une procédure en deux étapes. Ils ont utilisé une méthode de recuit simulé pour attribuer un statut professionnel (employé ou chômeur) à chaque membre d'un ménage synthétique, puis un type de secteur d'activité à chaque employé. En fonction de la situation professionnelle et du type de secteur d'activité, la deuxième étape a consisté à attribuer un salaire moyen à chaque travailleur en utilisant une autre source de données agrégées et en contrôlant par le sexe, l'âge et le type d'industrie.

Plus récemment, Hörl and Balac (2020) ont développé une approche permettant d'associer un revenu à chaque ménage de la population synthétique au moyen de statistiques agrégées. Une distribution des revenus par déciles est fournie au niveau de la commune, et pour chaque ménage synthétique, la commune de résidence est connue. Le processus d'affectation (en deux étapes) est assez simple : chaque ménage est affecté à un décile avec une probabilité de 10% ; une valeur aléatoire de revenu est par la suite échantillonnée avec une probabilité uniforme dans la plage comprise entre les limites inférieure et supérieure du décile de chaque ménage.

Le travail mené dans ce chapitre se fonde sur l'analyse de cet état de l'art. L'allocation d'attributs à une population synthétique à partir de données agrégées est relativement peu explorée dans la littérature. Dans la majorité des cas, ce processus d'allocation repose sur une distribution d'échantillonnage aléatoire, ce qui ne garantit pas la cohérence des données. Certains agents (qu'il s'agisse d'individus ou de ménages) peuvent en effet recevoir des attributs qui ne leur correspondent pas.

Une méthodologie efficace et cohérente d'ajout d'attributs est proposé avec un exemple à l'appui. Dans cet exemple, un revenu (plus précisément un niveau de vie) est affectée à une population synthétique de la commune de Nantes. Ce choix se justifie pour trois raisons principales. Le revenu constitue un attribut de microsimulation essentiel pour prendre en compte de nombreux aspects sociaux et économiques (pouvoir d'achat des ménages, politique de redistribution, politique fiscale, etc.). Le revenu sera également utilisé pour affiner la localisation spatiale des individus et ménages synthétiques dans le prochain chapitre. Enfin, la méthodologie proposée étant au stade de validation, les analyses ont été limitées à une taille de population raisonnable. Plutôt que de prendre en compte toute l'aire urbaine de Nantes, seule la commune de Nantes (plus grande commune de l'aire urbain)e avec une population totale de 295 000 individus et 157 000 ménages est considérée. Dans un premier temps des probabilités conditionnelles sont estimées à partir d'une combinaison d'attributs synthétiques des ménages. Dans un second temps, nous optimisons les probabilités conditionnelles obtenues afin d'attribuer les valeurs appropriées aux ménages synthétiques. La méthodologie proposée est générale et est applicable à d'autres attributs pour lesquels des données agrégées sont disponibles.

Le reste du chapitre est organisé comme suit. Les deux prochaines sections (section 4.1 et section 4.2) décrivent la configuration des données et le problème de façon générale. La section 4.3 procède à la formulation mathématique du problème et introduit les variables de décision et les paramètres. La section 4.4 est consacrée à la présentation des approches de résolution du problème et des résultats obtenus sur un jeu de données simulées. La section 4.5 présente les résultats de la meilleure approche sur la commune de Nantes. Dans la dernière section, le bilan des analyses est dressé et des pistes d'amélioration sont proposées.

4.1 Configuration des données

Pour la commune de Nantes, nous avons librement accès à des informations agrégées sur les conditions de vie des ménages (indicateurs de revenu, d'inégalité et de pauvreté).

Ces données proviennent d'une source de données appelée dispositif « Fichier Localisé Social et Fiscal »(FiLoSoFi). Nous présentons cette source de données dans la partie suivante.

4.1.1 Le dispositif FiLoSoFi

Le dispositif FiLoSoFi , mis en œuvre par l'Institut national de la statistique et des études économiques (Insee) à partir du millésime de revenus 2012 [1] succède aux dispositifs « Revenus Fiscaux Localisés » des ménages (RFL) et « Revenus Disponibles Localisés » des ménages (RDL). Ce dispositif effectue un rapprochement des fichiers fiscaux et sociaux, ce qui permet une meilleure estimation des ressources des ménages à un échelon géographique infra départemental. FiLoSoFi est établi à partir des fichiers suivants :

— le fichier des déclarations de revenus fiscaux qui contient les données relatives aux déclarations des revenus des contribuables ;

— le fichier de la taxe d'habitation dont on extrait les observations correspondant aux redevables de la taxe d'habitation reliés à des logements d'habitation ;

— le fichier des personnes physiques qui répertorie tous les déclarants aux services fiscaux ;

— les fichiers sociaux contenant les données exhaustives sur les prestations sociales versées par les principaux organismes gestionnaires de ces prestations.

Le champ couvert dans FiLoSoFi est celui des ménages fiscaux. Ce terme désigne un ménage constitué par le regroupement des foyers fiscaux répertoriés dans un même logement. A titre illustratif, un couple de concubins, où chacun remplit sa propre déclaration de revenus constitue un seul ménage fiscal parce que les membres du couple sont répertoriés dans le même logement, même s'ils constituent deux contribuables distincts au sens de l'imposition sur les revenus. Les personnes vivant dans des structures collectives (maisons de retraite, centres d'hébergements, foyers de travailleurs, communautés

1. Le millésime FiLoSoFi de l'année N est élaboré à partir des revenus perçus l'année N mais déclarés l'année N+1 et de la taxe d'habitation au 1er janvier de l'année N+1.

religieuses, cités universitaires, prisons, etc.), sans domicile fixe et sans abri ne sont pas retenues dans le champ de Filosofi.

Sur le champ des ménages fiscaux, les indicateurs produits permettent de décrire la distribution des revenus (quartiles, déciles), la structure de ces revenus (part des salaires, pensions, retraites et des prestations sociales dans le revenu fiscal) et la pauvreté monétaire (taux de pauvreté, intensité de la pauvreté). Différents types de revenus sont ainsi renseignés dans FiLoSoFi.

Tableau 4.1 : Types de revenus renseignés dans le dispositif FiLoSoFi

Type	Définition	Composantes
Revenu déclaré	Ressources déclarées par les contribuables	Somme des différents types de revenus déclarés, nets de cotisations sociales
Revenu disponible	Revenu à disposition du ménage pour consommer et épargner	Revenu déclaré + prestations sociales + revenus financiers non déclarés - impôts
Revenu déclaré par unités de consommation (UC)	Revenu déclaré qui tient compte de la composition du ménage	Revenu déclaré divisé par le nombre d'UC du ménage
Niveau de vie	Revenu disponible qui tient compte de la composition du ménage	Revenu disponible divisé par le nombre d'UC du ménage

Source : Insee (2018) : Sources et méthodes (juin 2018), Fichier Localisé Social et Fiscal (Filosofi) [en ligne] (page consultée le 06 août 2021)< `https://www.insee.fr/fr/metadonnees/source/fichier/Source_methodes_Filosofi_20180518_DSRL_PRFS_20180613.pdf`>

L'Insee définit le revenu par unité de consommation d'un ménage comme le revenu divisé par un coefficient, dénommé unités de consommation (UC), qui dépend de la taille et de l'âge des membres du ménage. Les UC proviennent d'un système de pondération qui attribue à chaque membre d'un ménage, un poids en rapport avec sa part supposée dans la consommation du ménage. Les pondérations suivantes (connues sous le nom d'échelle de l'OCDE) sont appliquées :

— 01 UC pour le premier adulte du ménage ;

— 0,5 UC pour chaque autre personne du ménage âgée de 14 ans ou plus ;

— 0,3 UC pour chaque enfant de moins de 14 ans.

La détermination d'un revenu par unité de consommation présente l'avantage de prendre en compte les diverses compositions des ménages et les économies d'échelle liées

à la vie en groupe. Dans ce chapitre, le revenu considéré est le niveau de vie des ménages (Tableau 4.1). Le niveau de vie apparaît comme le revenu le plus pertinent pour prendre en compte la réalité économique du ménage.

Il traduit en effet ce dont dispose un individu pour vivre, compte tenu de la composition du ménage auquel il appartient ; également, tous les individus d'un même ménage possèdent le même niveau de vie. Il devient dès lors possible de comparer le revenu de ménages ayant des tailles et des compositions différentes. Par exemple, une famille monoparentale d'un adulte et deux enfants de moins de 14 ans avec un revenu de 1 500 euros a un niveau de vie moins élevé qu'une personne célibataire avec un revenu de 1 500 euros. Pour le cas d'application, les données exploitées proviennent du dispositif FiLoSoFi de 2015.

4.1.2 Population synthétique de ménages

Les données du dispositif FiLoSoFi portent uniquement sur les attributs des ménages. Une population synthétique a donc été uniquement générée au niveau ménage pour chaque IRIS de la commune de Nantes avec les données du recensement millésimé de 2015 [2].

Cette population synthétique de ménages a été générée de manière simple, uniquement à partir de l'échantillon du fichier détail « Individus - localisation au canton-ouville »(INDCANT) restreint à la commune de Nantes. Cela correspond à un échantillon de 62 000 ménages pour une population totale de 157 000 ménages à atteindre. Le processus de génération a été réalisé en deux étapes :

— chaque ménage de INDCANT a été pondéré en utilisant la variable poids du ménage dénommée « IPONDI » ;

— les poids obtenus n'étant pas entiers, nous avons appliqué la méthode Truncate Replicate Sample (TRS) pour convertir ces poids décimaux en poids entiers et les ménages ont été répliqués sur la base de leurs poids entiers obtenus.

2. Il est important de mentionner que les données du recensement et de du dispositif FiLoSofi utilisées proviennent de la même année (2015).

A la fin de ce processus, une population synthétique d'environ 157 000 ménages est générée.

4.1.3 Présentation des attributs communs aux données du recensement et au dispositif FiLoSoFi

Dans le dispositif FiLoSoFi, les distributions du niveau de vie sont données en déciles (du premier au neuvième). Il est à signaler qu'aucun revenu minimal et maximal n'est mentionné. Pour la commune de Nantes, ces déciles [3] sont fournis pour l'ensemble de la population des ménages mais également pour certains attributs sociodémographiques des ménages. ces attributs sont :

— la taille du ménage ;

— la structure familiale du ménage ;

— l'âge du référent fiscal dans le ménage ;

— le statut de propriété du logement.

Le Tableau 4.2 présente les catégories associées à ces attributs.

Tableau 4.2 : Catégories de ménages pour lesquelles les distributions des niveaux de vie sont mentionnées.

Variable	Définition [nombre de catégories]	Catégories
T	Taille du ménage [5]	1 personne ; 2 personnes ; 3 personnes ; 4 personnes ; 5 personnes ou plus
C	Structure familiale du ménage [6]	Femmes vivant seules ; Hommes vivant seuls ; Couples sans enfant ; Couples avec enfants ; Familles monoparentales ; Ménages complexes
A	Age du référent fiscal dans le ménage [6]	Moins de 30 ans ; 30-39 ans ; 40-49 ans ; 50-59 ans ; 60-74 ans ; 75 ans et plus
B	Statut de propriété du logement [2]	Propriétaires ; Locataires

3. La base de données utilisée est accessible à partir de ce lien (page consultée le 05 août 2021) : < `https://www.insee.fr/fr/statistiques/3560118`>, Base niveau communes en 2015 - y compris arrondissements municipaux

Le référent fiscal du ménage peut également être assimilé à la personne de référence du ménage dans les données du recensement. Toutes les catégories renseignées dans ce tableau se trouvent également dans la population synthétique. Il est donc possible d'établir un rapprochement entre ces deux sources de données. Toutefois, un certain nombre de différences relatives à la notion de ménage et de membres existent entre les deux sources.

Dans FiLoSoFi, le terme de personne est préféré à celui d'habitant pour signifier le fait qu'une personne fiscalement rattachée à un ménage a la possibilité de ne pas résider au sein de ce ménage. C'est le cas par exemple d'un nombre important d'étudiants, physiquement absents du ménage parental mais fiscalement rattachés à celui-ci. Il est également possible d'avoir des personnes majeures, physiquement présentes dans un ménage mais non répertoriées au sens fiscal dans ce ménage (personnes au pair ou hébergées par exemple). En revanche, dans le recensement, seul le lieu de résidence effectif de la personne est prise en compte. En raison de telles particularités, le nombre de personnes de FiLoSoFi ne coïncide pas avec la population des ménages issue du recensement de la population. Des écarts d'effectifs entre les deux sources peuvent donc être constatés.

4.2 Formulation générale du problème

Il s'agit d'estimer le niveau de vie des ménages en associant les différents attributs sociodémographiques présents dans FiLoSoFi (Tableau 4.2). Par souci de simplification, le niveau de vie sera désigné par le revenu et noté R dans la suite du document. Le Tableau 4.3 présente la distribution en déciles du niveau de vie des ménages de la commune de Nantes. L'objectif est d'attribuer un niveau de vie (revenu R) à toutes les différentes combinaisons des catégories des quatre attributs $T_g C_f A_y B_l$, avec :

$g = 1, 2, ..., 5$: le nombre de catégories de la taille du ménage ;

$f = 1, 2, ..., 6$: le nombre de catégories de la structure familiale du ménage ;

$y = 1, 2, ..., 6$: le nombre de catégories de l'âge du référent fiscal dans le ménage ;

$l = 1, 2$: le nombre de catégories du statut de propriété.

Ainsi, $T_4C_5A_2B_2$ représente dans la population synthétique la catégorie croisée suivante : ménage de taille 4, famille monoparentale, référent fiscal âgé de 30-39 ans avec un statut de locataire.

Tableau 4.3 : Distribution des déciles du niveau de vie des ménages de la commune de Nantes pour l'ensemble de la population et certaines catégories.

Catégories	Déciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Ensemble	10 303	13 336	16 024	18 631	21 263	24 188	27 774	32 620	41 308
1 personne	9 794	12 961	14 914	16 865	18 687	20 763	23 357	27 069	33 514
2 personnes	12 176	15 553	18 356	20 919	23 435	26 331	30 140	35 136	44 134
3 personnes	10 584	13 656	16 489	19 145	21 893	24 891	28 440	33 432	42 079
4 personnes	10 740	14 130	17 207	20 138	22 955	26 148	29 644	34 238	42 998
5 personnes/+	8 758	10 990	12 879	15 467	18 991	23 164	27 638	33 238	43 292
Femmes vivant seules	10 714	13 334	15 332	17 186	19 031	21 111	23 715	27 360	33 480
Hommes vivant seuls	9 016	12 224	14 288	16 388	18 268	20 305	22 908	26 696	33 551
Couples sans enfant	14 417	18 066	20 791	23 225	25 785	28 911	32 718	37 961	47 273
Couples avec enfants	10 822	14 238	17 646	20 665	23 596	26 837	30 528	35 573	44 977
Familles monoparentales	8 702	10 367	11 915	13 557	15 179	17 135	19 370	22 761	28 733
Ménages complexes	8 692	11 052	13 063	15 207	17 648	20 452	23 853	27 843	35 179
Moins de 30 ans	8 371	11 117	13 501	15 678	17 572	19 557	21 803	24 513	28 920
30-39 ans	9 985	12 872	15 539	18 122	20 688	23 463	26 704	30 771	37 300
40-49 ans	9 827	12 733	15 226	17 993	20 839	24 055	27 842	32 837	42 018
50-59 ans	10 371	13 512	16 617	19 509	22 561	26 030	30 095	35 710	46 658
60-74 ans	12 474	15 582	18 641	21 412	24 360	27 778	32 049	37 751	48 548
75 ans et plus	14 005	16 389	18 583	20 869	23 275	26 028	29 648	34 849	43 945
Propriétaire	16 543	19 966	22 545	25 022	27 626	30 612	34 336	39 727	50 060
Locataire	8 764	10 912	12 748	14 433	16 265	18 266	20 615	23 870	29 860

Note : le premier décile (D1) est le revenu au dessous duquel se situent 10 % des revenus ; le neuvième décile (D9) est le revenu au dessous duquel se situent 90 % des revenus.

Source : Insee, Structure et distribution des revenus, inégalité des niveaux de vie en 2015 [en ligne] (page consultée le 05 août 2021)< `https://www.insee.fr/fr/statistiques/3560118`>, Base niveau communes en 2015 - y compris arrondissements municipaux

La population synthétique comporte un total de 19 catégories $(5 + 6 + 6 + 2)$, ce qui donne 360 catégories croisées potentielles $(5 \times 6 \times 6 \times 2)$. Parmi ces catégories croisées, certaines sont évidemment irréalistes comme par exemple la catégorie croisée $T_1C_3A_yB_l$, ménage d'une personne, couple sans enfant. Par conséquent, la probabilité d'avoir une telle combinaison est nulle quelles que soient les catégories de l'âge du référent et du statut de propriété ; autrement dit $\mathrm{P}(T_1C_3A_yB_l) = 0$. Sur les 360 catégories croisées po-

tentielles, 187 ont des probabilités non nulles. L'affectation du revenu consiste à estimer la distribution de probabilité $\mathrm{P}(R|T_gC_fA_yB_l)$.

Remarques importantes

Aucune information entre deux déciles de revenus n'est mentionnée. L'estimation de la probabilité continue $\mathrm{P}(R)$ est remplacée par une estimation de E probabilités discrètes $\mathrm{P}(R_{e-1} < R < R_e)$; R_e étant la suite croissante de tous les déciles de revenus allant de $e = 1$ à $E = 171$ (19 catégories $\times$ 9 déciles). Les déciles de l'ensemble de la population (première ligne du Tableau 4.3) ne sont pas pris en compte car cette information est redondante avec celle des autres catégories.

La valeur des revenus minimal et maximal pour chaque catégorie de déciles est également indisponible. Nous supposons alors une distribution linéaire des revenus et nous fixons pour chaque catégorie, un revenu minimal $R_0 = 0$ et un revenu maximal R_{max} égale à D9 $\times 1,5$.

Soit $\boldsymbol{R}$ un vecteur colonne de revenus dont la composante e se situe dans l'intervalle $]R_{e-1}, R_e[$ avec e = 1 à E=190. Ce vecteur inclut les 171 déciles et les 19 revenus maximaux. Il existe donc en tout 190 valeurs de revenu. Les probabilités suivantes sont alors estimées :

$$\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l) \tag{4.1}$$

Cela représente 35 530 (187 $\times$ 190) catégories croisées avec revenu à estimer.

4.3 Formulation mathématique du problème

L'Equation 4.1 peut être déterminée de deux manières distinctes :

— en utilisant la définition de la probabilité conditionnelle

$$\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l) = \mathrm{P}((R_{e-1} < R < R_e) \cap (T_gC_fA_yB_l)) \times \frac{1}{\mathrm{P}(T_gC_fA_yB_l)} \tag{4.2}$$

— ou en appliquant le théorème de Bayes

$$\mathrm{P}(R_{e-1} < R < R_e | T_g C_f A_y B_l) = \mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e) \times \frac{\mathrm{P}(R_{e-1} < R < R_e)}{\mathrm{P}(T_g C_f A_y B_l)} \tag{4.3}$$

avec :

$g = 1, 2, ..., 5$

$f = 1, 2, ..., 6$

$y = 1, 2, ..., 6$

$l = 1, 2$

Une résolution du problème à partir de l'équation 4.2 est proposée dans une premier temps. Chaque terme de cette équation est estimé séparément.

4.3.1 Variables de décision et paramètres

Etape 1 : Estimation des probabilités $\mathrm{P}(T_g C_f A_y B_l)$

$\mathrm{P}(T_g C_f A_y B_l)$ représente les probabilités jointes correspondantes aux différentes catégories croisées de ménages dans la population synthétique. Ces probabilités jointes sont calculées à partir de la population synthétique. Par exemple :

$\mathrm{P}(T_4 C_5 A_2 B_2) = \frac{\text{nombre de ménages } T_4 C_5 A_2 B_2}{\text{nombre total de ménages}}$

Etape 2 : Estimation des probabilités $\mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l))$

$$\mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)) = \frac{\text{nombre de ménages } (R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)}{\text{nombre total de ménages}} \tag{4.4}$$

Le nombre de ménages $(R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)$ n'est pas connu mais peut être estimé. Afin de simplifier l'écriture de l'Equation 4.4, de nouvelles notations sont

introduites.

$(R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)$ est une catégorie croisée avec revenu parmi les 35 530 modalités possibles et sera indexé par k ; v_k désigne le nombre de ménages associé à la catégorie croisée avec revenu k ; $\boldsymbol{P}$ est le vecteur colonne représentant les probabilités des catégories croisées avec revenu ; p_k est la probabilité associée à la catégorie croisée avec revenu k ; n_H désigne le nombre total de ménages. Avec les conventions, l'Equation 4.4 devient

$$\mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)) = p_k = \frac{v_k}{n_H} \tag{4.5}$$

4.3.2 Fonction objectif et contraintes

Introduisons maintenant le vecteur aléatoire $\boldsymbol{V} = (v_1 \dots v_{n_{mH}})^t$ avec n_{mH}, le nombre total de catégories croisées avec revenu. $\boldsymbol{V}$ décrit un état de la population où v_1 ménages possèdent la catégorie croisée avec le revenu le plus faible,... et $v_{n_{mH}}$ ménages possèdent la catégorie croisée avec le revenu le plus élevé. La distribution de probabilité de $\boldsymbol{V}$ est donc décrite par la distribution multinomiale :

$$\mathrm{P}(\boldsymbol{V}|n_H, n_{mH}, \boldsymbol{Q}) = n_H! \prod_{k=1}^{n_{mH}} \frac{q_k^{v_k}}{v_k!} \tag{4.6}$$

où q_k est la probabilité initiale associée à la catégorie croisée avec revenu k. Pour simplifier, nous remplaçons $\mathrm{P}(\boldsymbol{V}|n_H, n_{mH}, \boldsymbol{Q})$ par $\mathrm{P}(\boldsymbol{V})$.

Le vecteur des probabilités $\boldsymbol{V}$ doit respecter différentes contraintes relatives au nombre total de ménages n_H, et à la fréquence jointe calculée à partir de la population synthétique et les déciles :

1. $\sum_k v_k = n_H$ est la contrainte naturelle.
2. $\sum_{k \in M_{T_g C_f A_y B_l}} v_k =$ Nombre de ménages synthétiques dont la catégorie croisée dans la population synthétique est $T_g C_f A_y B_l$.

   $M_{T_g C_f A_y B_l}$ est l'ensemble des indices k du vecteur des probabilités jointes catégories croisées avec revenu qui ont la même catégorie croisée $T_g C_f A_y B_l$ quel que soit le

revenu. Ces contraintes assurent la cohérence entre $\boldsymbol{V}$ et la population synthétique.

3. $\sum_{k \in M_{(R_{d-1,T_g} < R < R_{d,T_g})}} v_k = 10\% \times v_{T_g}$

   R_{d,T_g} représente la valeur du revenu correspondant au décile d pour la catégorie T_g. Le Tableau 4.3 présente ces valeurs.

   $M_{(R_{d-1,T_g} < R < R_{d,T_g})}$ est l'ensemble des indices k du vecteur des probabilités jointes catégories croisées avec revenus qui appartiennent au même intervalle de revenu $T_g C_f A_y B_l$ quelle que soit la catégorie croisée. ; v_{T_g} est le nombre de ménages de la catégorie T_g.

   Cette contrainte est répétée pour toutes les catégories de tous les attributs et pour tous les déciles. Cet ensemble de contraintes garantit la compatibilité de $\boldsymbol{V}$ avec les informations sur le revenu.

Ces contraintes sont linéaires et résumées par :

$$\boldsymbol{D} \cdot \boldsymbol{V} = \mathbf{x} \tag{4.7}$$

Par convention, la première ligne de $\boldsymbol{D}$ correspond à la contrainte naturelle ; la somme de ses composantes est égale à 1. Cette propriété sera utilisée par la suite. La meilleure solution à notre problème est de rechercher la réalisation la plus probable du vecteur $\boldsymbol{V}$ tel que :

$$\boldsymbol{V}^* = \underset{\boldsymbol{D} \cdot \boldsymbol{V} = \mathbf{x}}{\arg\max} \, \mathrm{P}(\boldsymbol{V}) \tag{4.8}$$

Selon Niven (2005), ce problème d'optimisation peut être approximé par un problème de minimisation de l'entropie croisée, qui est beaucoup plus facile à résoudre. Le logarithme de $\mathrm{P}(\boldsymbol{V})$ est introduit afin de transformer les multiplications $\prod$ de l'Equation 4.6 en une somme Σ. Le problème d'optimisation référencé devient alors :

$$\boldsymbol{V}^* = \underset{\boldsymbol{D} \cdot \boldsymbol{V} = \mathbf{x}}{\arg\max} \log(\mathrm{P}(\boldsymbol{V})) \tag{4.9}$$

$\boldsymbol{V}$ est considéré comme un vecteur aléatoire réel et non comme un vecteur aléatoire entier, ce qui signifie que les composantes de ce vecteur peuvent être non entières. Pour déterminer $\boldsymbol{V}^*$, nous appliquons la règle de Fermat : si une fonction f est minimisée ou maximisée en un point a et si f est dérivable en a, alors la dérivée $f'(a)$ de f en a est nulle. Dans notre cas, la dérivée du lagrangien L à $\boldsymbol{V}^*$ est nulle. Nous commençons par différencier le lagrangien :

$$\frac{\partial L}{\partial v_k} = \frac{\partial}{\partial v_k}\left(\log(n_H!) - \log(v_k!)\right) + \log(q_k) + \boldsymbol{\lambda} \cdot \boldsymbol{D}_k \tag{4.10}$$

avec $\boldsymbol{\lambda}$, les multiplicateurs de Lagrange, et $\boldsymbol{D}_k$ la k$^{\text{ème}}$ colonne de la matrice $\boldsymbol{D}$.

A partir de l'approximation de Stirling,

$$\log(v_k!) \approx v_k \log(v_k) - v_k$$

Grâce à cette formule d'approximation et en rappelant que $n_H = \sum_k v_k$, l'Equation 4.10 devient :

$$\frac{\partial L}{\partial v_k} \approx \log(n_H) - \log(v_k) + \log(q_k) + \boldsymbol{\lambda} \cdot \boldsymbol{D}_k$$

En appliquant la règle de Fermat :

$$\log(q_k) - \log(\frac{v_k^*}{n_H}) + \boldsymbol{\lambda} \cdot \boldsymbol{D}_k = 0 \tag{4.11}$$

La première composante du vecteur $\boldsymbol{D}_k$, $\boldsymbol{D}_{k1}$ est égale à 1 (confère commentaires relatifs à l'Equation 4.7). Pour un $\boldsymbol{\lambda}$, nous avons pu définir $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$ sauf pour sa première composante.

Ainsi, $\boldsymbol{\lambda}'_1 = \boldsymbol{\lambda}_1 + 1$, tel que :

$$\log(q_k) - \log(\frac{v_k^*}{n_H}) + \boldsymbol{\lambda}' \cdot \boldsymbol{D}_k = 1$$

Cette équation définit la solution du problème de minimisation suivant :

$$\boldsymbol{P}^* = \underset{\boldsymbol{D}\cdot\boldsymbol{P}=\mathbf{x}/n_H}{\arg\min}\ \boldsymbol{P}^t \cdot \log(\frac{\boldsymbol{P}}{\boldsymbol{Q}}) \tag{4.12}$$

$\boldsymbol{P}$ est un vecteur colonne dont la composante k est p_k ; $\log(\frac{\boldsymbol{P}}{\boldsymbol{Q}})$ correspond au vecteur colonne dont la composante k est $\log(\frac{p_k}{q_k})$.

En remplaçant la variable d'intérêt v_k par $p_k = \frac{v_k}{n_H}$ (avec n_H, le nombre total de ménages), le problème d'optimisation de l'Equation 4.12 représente une formulation de divergence de Kullback-Leibler, ou minimisation de l'entropie croisée et désignée par MinxEnt dans la littérature [4], avec une distribution de probabilité initiale représentée par $\boldsymbol{Q}$. Résoudre ce problème d'optimisation équivaut à résoudre le problème de maximisation de l'Equation 4.9 en utilisant l'approximation de Stirling. Puisque la solution du problème 4.9 est la réalisation la plus probable, la solution de MinxEnt sera appelée la distribution la plus probable en utilisant l'approximation de Stirling.

Nous ne disposons d'aucune information préalable sur les probabilités initiales associées aux catégories croisées avec revenu. Selon le principe de la raison suffisante de Laplace, nous supposons que ces probabilités sont identique et égales à $q = \frac{1}{n_{mH}}$ avec n_{mH}, le nombre total de catégories croisées avec revenu. Le problème de minimisation de l'entropie croisée devient dans ce cas un problème de maximisation de l'entropie, désigné par MaxEnt dans la littérature.

$$\boldsymbol{P}^* = \underset{\boldsymbol{D}\cdot\boldsymbol{P}=\mathbf{x}/n_H}{\arg\max}\ -\boldsymbol{P}\log(\boldsymbol{P}) \tag{4.13}$$

4.4 Résolution du problème

4.4.1 Résolution avec la méthode de maximisation de l'entropie

Il s'agit d'estimer $\mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l))$ (Equation 4.2) en fonction de contraintes à la fois sur les catégories croisées de la population synthétique (Equa-

4. Rigoureusement, c'est un problème de minimisation de la divergence de Kullback-Leibler. Cependant, le nom de l'algorithme est MinxEnt qui fait référence à l'entropie croisée et c'est le terme utilisé dans la littérature. C'est ce terme qui a été retenu dans ce manuscrit.

tion 4.16) et sur les déciles de revenus (Equation 4.17). La méthode de maximisation de l'entropie est appliquée :

$$Max \; - \sum_{g,f,y,l} \mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)). \ln \mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)) \tag{4.14}$$

en fonction des contraintes suivantes :

La contrainte naturelle :

$$\sum_{g,f,y,l} \mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)) = 1 \tag{4.15}$$

Contraintes dérivées de la population synthétique :

$$\sum_{e} \mathrm{P}((R_{e-1} < R < R_e) \cap (T_g C_f A_y B_l)) = \mathrm{P}(T_g C_f A_y B_l) \tag{4.16}$$

Cette contrainte traduit le fait que pour une catégorie croisée donnée, la somme de toutes les probabilités de revenu est égale à la probabilité de la catégorie croisée.

Les contraintes sur les déciles :

$$\begin{aligned}
&\sum_{f,y,l} \mathrm{P}((T_g C_f A_y B_l) \cap (R < R_{d,T_g})) = d \\
&\sum_{g,y,l} \mathrm{P}((T_g C_f A_y B_l) \cap (R < R_{d,C_f})) = d \\
&\sum_{g,f,l} \mathrm{P}((T_g C_f A_y B_l) \cap (R < R_{d,A_y})) = d \\
&\sum_{g,f,y} \mathrm{P}((T_g C_f A_y B_l) \cap (R < R_{d,B_l})) = d
\end{aligned} \tag{4.17}$$

avec d= 0,1 ; 0,2 ; ... ; 0,9 et R_d représente la valeur du revenu correspondant au décile d pour une catégorie donnée de la population synthétique. Au total, 19 catégories sont présentes dans la population synthétique ; pour chaque catégorie il y a 9 déciles et un revenu maximal, ce qui donne 190 contraintes de ce type.

Appuyons-nous sur Kapur and Kesavan (1992) et Mattos and Veiga (2004) pour résoudre numériquement l'équation 4.13. L'algorithme numérique est basé sur la solution

de l'Equation 4.11 :

$$p_k = q_k \exp(\boldsymbol{\lambda} \cdot \boldsymbol{D}_k) \tag{4.18}$$

Cette équation constitue la clé de réécriture du problème d'optimisation (équation 4.13) sous sa forme duale :

$$\boldsymbol{\lambda}^* = \arg\max \sum_k (\log(q_k) + \boldsymbol{\lambda} \cdot \boldsymbol{D}_k) q_k \exp(\boldsymbol{\lambda} \cdot \boldsymbol{D}_k) + \boldsymbol{\lambda} \cdot (\boldsymbol{D} \cdot \boldsymbol{P} - \mathbf{x}/n_H) \tag{4.19}$$

Les variables d'intérêt dans cette optimisation duale sont les multiplicateurs de Lagrange, dont le nombre est considérablement inférieur au nombre de variables d'intérêt dans l'optimisation primale. Autrement dit, le nombre de contraintes est bien inférieur à la dimension de la distribution de probabilité. Ce problème d'optimisation dual est résolu par un algorithme de Newton, qui est bien adapté et efficace (Kapur and Kesavan, 1992).

Avant de lancer l'algorithme d'optimisation, il est recommandé de vérifier la cohérence des contraintes. Cette vérification implique la recherche d'une distribution de probabilité initiale $\boldsymbol{P}^0$ cohérente avec les contraintes. La distribution initiale a été trouvée en résolvant le problème de programmation linéaire :

$$(\boldsymbol{P}^0 \ \boldsymbol{0}) = \underset{\boldsymbol{D}\cdot\boldsymbol{P}+\boldsymbol{\Delta}=\mathbf{x}/n_H, \boldsymbol{\Delta}\geq 0}{\arg\min} \boldsymbol{1} \cdot \boldsymbol{\Delta} \tag{4.20}$$

avec : $\boldsymbol{\Delta}$, un vecteur colonne de dimension égale au nombre de contraintes, $\boldsymbol{1}$, un vecteur ligne constitué de 1 et de dimension égale à celle de $\boldsymbol{\Delta}$; $\boldsymbol{0}$, un vecteur colonne constitué de 0 et de dimension égale à celle de $\boldsymbol{\Delta}$. $(\boldsymbol{P}^0 \ \boldsymbol{0})$ est la concaténation de $\boldsymbol{P}^0$ et de $\boldsymbol{0}$. Les variables d'intérêt dans cette configuration de programmation linéaire sont $\boldsymbol{P}$ et $\boldsymbol{\Delta}$. Le programme est initalisé avec $\boldsymbol{P} = 0$ et $\boldsymbol{\Delta} = \mathbf{x}/n_H$. Si le système est cohérent, le minimum est atteint avec une distribution de probabilité $\boldsymbol{P}^0$ qui vérifie : $\boldsymbol{D} \cdot \boldsymbol{P}^0 = \mathbf{x}/n_H$ et $\boldsymbol{\Delta} = \boldsymbol{0}$.

La mise en œuvre de MaxEnt été effectuée avec le logiciel et le langage de program-

mation R (version 4.0.3), sur une machine Windows 10 professionnel, Intel(R) Core (TM) i7-8665U CPU @ 1.90GHz, 2.11 GHz et 16 GB de RAM. Certaines implémentations de l'algorithme MaxEnt exigent que la matrice $\boldsymbol{D}$ soit de plein rang. Si ce n'est pas le cas, un sous-ensemble de $\boldsymbol{D}$ est sélectionné en utilisant la décomposition QR (Golub et al., 1996). Ce point n'est pas fondamental pour la compréhension de l'approche mais est mentionné ici afin de faciliter la compréhension des scripts R diffusés.

L'utilisation de Maxent dans le cas d'application se révèle impossible à cause d'un problème numérique. En effet, le problème d'optimisation comprend 378 (187+190+1) contraintes (un peu moins en réalité, puisque toutes les contraintes ne sont pas indépendantes) et 35 530 variables de décision. L'Equation 4.18 devient numériquement instable à cause de la fonction exponentielle, qui dépend des nombreuses valeurs de $\boldsymbol{\lambda}$. L'échec du calcul de cette équation entraîne l'échec de l'algorithme maxEnt. Une heuristique permettant de trouver une solution pratique est proposée ci-dessous. Le problème d'optimisation globale sera divisé en plusieurs sous-problèmes d'optimisation afin de réduire le nombre de contraintes.

4.4.2 Minimisation de l'entropie croisée

La résolution du problème (équation 4.1) s'effectue maintenant à partir de l'équation 4.3 (théorème de Bayes)

$$\mathrm{P}(R_{e-1} < R < R_e | T_g C_f A_y B_l) = \mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e) \times \frac{\mathrm{P}(R_{e-1} < R < R_e)}{\mathrm{P}(T_g C_f A_y B_l)}$$

avec :

$g = 1, 2, ..., 5$

$f = 1, 2, ..., 6$

$y = 1, 2, ..., 6$

$l = 1, 2$

Les différents termes de cette équation son estimés séparément.

Etape 1 : estimation des probabilités $\mathrm{P}(T_g C_f A_y B_l)$, confère section 4.3.1.

Etape 2 : estimation des probabilités $\mathrm{P}(R_{e-1} < R < R_e)$

R_e est une suite croissante de tous les déciles (171 au total) et des 19 revenus maximaux (confère section 4.2). Une extrapolation linéaire de ces 190 revenus est effectuée, à partir des déciles de la population totale (première ligne du Tableau 4.3). Par exemple, si R_e est compris entre les déciles R_{d-1} et R_d de la population totale, nous estimons :

$$\mathrm{P}(R < R_e) = \mathrm{P}(R < R_{d-1}) + \frac{R_e - R_{d-1}}{R_d - R_{d-1}}(\mathrm{P}(R_d) - \mathrm{P}(R_{d-1})) \tag{4.21}$$

$\mathrm{P}(R < R_{e-1})$ est estimée de manière similaire. Par conséquent, les probabilités à estimer $\mathrm{P}(R_{e-1} < R < R_e) = \mathrm{P}(R < R_e) - \mathrm{P}(R < R_{e-1})$. La troisième étape consiste à rechercher la distribution $\mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e)$ composée de 190 probabilités.

Etape 3 : estimation des probabilités $\mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e)$

La distribution $\mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e)$ est inconnue. Nous supposons qu'elle peut être approchée à partir de la distribution de $\mathrm{P}(T_g C_f A_y B_l)$ en minimisant l'entropie croisée (MinxEnt) entre les deux distributions.

La formulation de MinxEnt s'exprime comme suit :

$$\min \sum_{g,f,y,l} \mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e) \ln \frac{\mathrm{P}(T_g C_f A_y B_l | R_{e-1} < R < R_e)}{\mathrm{P}(T_g C_f A_y B_l)} \tag{4.22}$$

selon les contraintes suivantes :

$$
\begin{aligned}
\mathrm{P}(T_g|R_{e-1} < R < R_e) &= \mathrm{P}(R_{e-1} < R < R_e|T_g) \times \frac{\mathrm{P}(T_g)}{\mathrm{P}(R_{e-1} < R < R_e)} \\
\mathrm{P}(C_f|R_{e-1} < R < R_e) &= \mathrm{P}(R_{e-1} < R < R_e|C_f) \times \frac{\mathrm{P}(C_f)}{\mathrm{P}(R_{e-1} < R < R_e)} \\
\mathrm{P}(A_y|R_{e-1} < R < R_e) &= \mathrm{P}(R_{e-1} < R < R_e|A_y) \times \frac{\mathrm{P}(A_y)}{\mathrm{P}(R_{e-1} < R < R_e)} \\
\mathrm{P}(B_l|R_{e-1} < R < R_e) &= \mathrm{P}(R_{e-1} < R < R_e|B_l) \times \frac{\mathrm{P}(B_l)}{\mathrm{P}(R_{e-1} < R < R_e)}
\end{aligned}
\tag{4.23}
$$

Les probabilités $\mathrm{P}(T_g)$, $\mathrm{P}(C_f)$, $\mathrm{P}(A_y)$ et $\mathrm{P}(B_l)$ sont calculées à partir des données de la population synthétique.

Les probabilités conditionnelles $\mathrm{P}(R_{e-1} < R < R_e|T_g)$, $\mathrm{P}(R_{e-1} < R < R_e|C_f)$, $\mathrm{P}(R_{e-1} < R < R_e|A_y)$ et $\mathrm{P}(R_{e-1} < R < R_e|B_l)$ ont également été estimées par extrapolation linéaire pour chaque catégorie, de la même manière qu'à l'étape 2. Au final, il existe 19 contraintes et 187 variables d'intérêt.

La cohérence de chaque optimisation est testée au moyen de la programmation linéaire en utilisant la même méthode que celle décrite dans l'approche précédente. L'insuffisance de rang de la matrice des contraintes est également traitée en utilisant la décomposition QR. L'implémentation de MinXent a été réalisée avec la librairie R « minxent »(Senay Asma, 2015) [5] disponible sur le « Comprehensive R Archive Network »(CRAN).

L'heuristique de minimisation de l'entropie croisée est utilisée pour affecter un revenu aux ménages synthétiques de la commune de Nantes.

4.5 Application de l'heuristique à la commune de Nantes

Dans l'heuristique proposée et appliquée à la commune de Nantes, les probabilités $\mathrm{P}(T_g C_f A_y B_l)$ (étape 1) ; $\mathrm{P}(R_{e-1} < R < R_e)$ (étape 2) ; $\mathrm{P}(T_g C_f A_y B_l|R_{e-1} < R < R_e)$ (étape 3) ont été estimées. Il est alors possible de déduire la distribution de probabilités

5. Disponible à < `https://cran.r-project.org/web/packages/minxent/index.html`>, page consultée le 09 août 2021.

$\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l)$ (distribution formulée dans l'Equation 4.3).

4.5.1 Résultats obtenus

A l'issue du processus d'estimation des probabilités $\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l)$, des probabilités incorrectes ont été identifiées. Elles sont de deux types :

$\mathrm{P}(R < R_e|T_gC_fA_yB_l) > 1$

$\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l) < 0$

Ces probabilités non valides sont dues à la fois au nombre relativement important de paramètres à estimer et aux très faibles différences qui peuvent exister entre deux déciles de revenu successifs. Par exemple, dans le Tableau 4.3, entre les déciles D8 pour les ménages complexes (27 843 euros) et D7 pour la catégorie d'âge 40-49 ans (27 842 euros), la différence s'élève à seulement un euro. Une procédure de post-traitement est appliquée afin de corriger ces probabilités invalides. Cela permet d'obtenir des tranches de revenus plus larges tout en assurant la cohérence des données. La procédure de post-traitement est décrite ci-dessous :

— Pour chaque $\mathrm{P}(R < R_e|T_gC_fA_yB_l) > 1$, nous fixons $\mathrm{P}(R_{e-1} < R|T_gC_fA_yB_l) = 0$; le fait de renormaliser la distribution de probabilité $\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l)$ a pour but d'éviter la sensibilité au revenu maximum R_{max}, fixé arbitrairement à $1,5$ fois le 9^e décile ($1.5 \times D9$).

— Pour chaque $\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l) < 0$, la tranche de revenu est augmentée jusqu'à ce que
$\mathrm{P}(R_{e-1} < R < R_{e+\gamma}|T_gC_fA_yB_l) \geq 0$ avec $\gamma \geq 1$. Cela signifie que les revenus e à e+γ-1 sont retirés du vecteur colonne $\boldsymbol{R}$.

Les probabilités finales $\mathrm{P}(R_{e-1} < R < R_e|T_gC_fA_yB_l)$ sont des probabilités de tranches de revenus en fonction des catégories croisées du ménage. Dans la population synthétique, le nombre de ménages pour les différentes catégories croisées est connu. Par exemple il existe 14 613 ménages de taille 1, femme vivant seule, moins de 30 ans et locataire.

Une tranche de revenu est alors affectée à chaque ménage à partir de cette distribution de probabilités et du nombre de ménages d'une catégorie croisée donnée. Un revenu spécifique est par la suite attribué de manière aléatoire à chaque ménage selon une distribution uniforme continue ; ce revenu se situe entre les limites inférieure et supérieure de chaque tranche de revenu. A l'issue de ce processus, chaque ménage de la population synthétique dispose d'un revenu spécifique.

4.5.2 Validation des résultats obtenus

Pour évaluer la précision de l'heuristique proposée, de nouveaux déciles (déciles simulés) sont recalculés à partir des revenus affectés. Le Tableau 4.4 présente les déciles simulés après l'affectation d'un revenu à chaque ménage synthétique de la commune de Nantes.

Tableau 4.4 : Distribution des déciles simulés pour la commune de Nantes

Catégories	Déciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Ensemble	9 993	13 333	16 053	18 632	21 249	24 152	27 684	32 500	41 549
1 personne	9 699	12 664	14 972	17 160	19 361	21 842	24 969	29 408	37 239
2 personnes	11 086	15 692	18 803	21 545	24 237	24 411	31 302	36 758	48 706
3 personnes	9 746	13 397	16 649	19 713	22 840	25 931	29 685	35 263	47 193
4 personnes	10 441	14 085	17 833	21 019	24 052	27 275	30 603	35 619	46 409
5 personnes/+	8 867	11 120	12 758	16 048	20 769	24 991	29 652	35 879	49 401
Femmes vivant seules	10 293	13 106	15 295	17 376	19 556	22 016	25 177	29 475	36 872
Hommes vivant seuls	9 093	12 340	14 528	16 902	19 147	21 612	24 721	29 356	37 835
Couples sans enfant	14 381	18 363	21 276	23 837	26 554	29 777	33 981	39 419	54 456
Couples avec enfants	10 837	14 475	18 359	21 648	24 663	27 889	31 737	37 760	51 965
Familles monoparentales	8 703	10 617	12 029	13 782	15 818	18 248	21 331	25 680	32 993
Ménages complexes	9 085	13 338	15 809	17 848	20 301	22 927	25 887	30 109	35 901
Moins de 30	8 463	11 465	13 820	16 049	18 104	20 246	22 700	26 016	31 803
30-39	10 355	13 206	16 002	18 701	21 425	24 464	27 706	31 898	39 302
40-49	10 049	13 002	15 594	18 577	21 637	24 943	28 894	33 657	44 341
50-59	10 719	13 849	17 042	20 082	23 333	26 831	30 972	37 151	50 267
60-74	12 564	15 818	19 060	22 002	25 108	28 716	32 658	39 107	54 787
75/+	14 055	16 638	18 967	21 299	23 833	26 691	30 533	35 770	46 406
Propriétaire	17 025	20 499	23 125	25 598	28 221	31 306	35 196	41 078	53 325
Locataire	8 529	11 692	13 724	15 528	17 466	19 676	22 377	26 311	32 716

Note : le premier décile (D1) est le revenu au dessous duquel se situent 10 % des revenus ; le neuvième décile (D9) est le revenu au dessous duquel se situent 90 % des revenus.

Ces déciles simulés sont comparés avec les déciles originaux (déciles observés) du jeu de données (Tableau 4.3). Deux méthodes de validation sont alors utilisées .

La première méthode est une représentation graphique des déciles simulés et observés avec des diagrammes quantile-quantile (Q-Q plots). Les Q-Q plots permettent d'évaluer graphiquement l'ajustement des déciles simulés aux déciles réels et de détecter facilement d'éventuelles observations aberrantes (outliers). Dans ces graphiques, chaque point représente un décile simulé et la ligne de référence tracée est la première bissectrice. Si les deux distributions sont identiques, les points du graphique doivent exactement s'aligner sur la ligne. La Figure 4.1 compare la distribution des déciles simulés à la distribution des déciles observés pour chaque catégorie, ainsi que pour la population entière.

Les Q-Q plots générés indiquent un ajustement parfait entre les déciles simulés et les déciles observés de FiLoSoFi pour l'ensemble de la population. Pour toutes les autres catégories, la majorité des points (plus de 80 %) s'alignent le long de la première bissectrice. Cela montre ainsi que les déciles simulés s'ajustent très bien aux déciles observés, ce qui démontre que la population synthétique a un revenu simulé global, cohérent avec les données agrégées de FiLoSoFi. En revanche, l'ajustement des catégories ménages complexes, 5 personnes/+ et familles monoparentales est moins précis ; également, les déciles D8 et surtout D9 s'écartent davantage de la ligne de référence.

La deuxième méthode de validation estime la précision des résultats obtenus avec deux métriques généralement utilisées dans le contexte de la microsimulation : l'erreur absolue $|R_e - R_n|$ et l'erreur relative $|\frac{R_e - R_n}{R_e}|$ avec R_e and R_n qui représentent respectivement les déciles observés et simulés pour chaque catégorie. Les Tableaux 4.5 et 4.6 présentent respectivement les erreurs absolues et relatives qui existent entre les deux distributions.

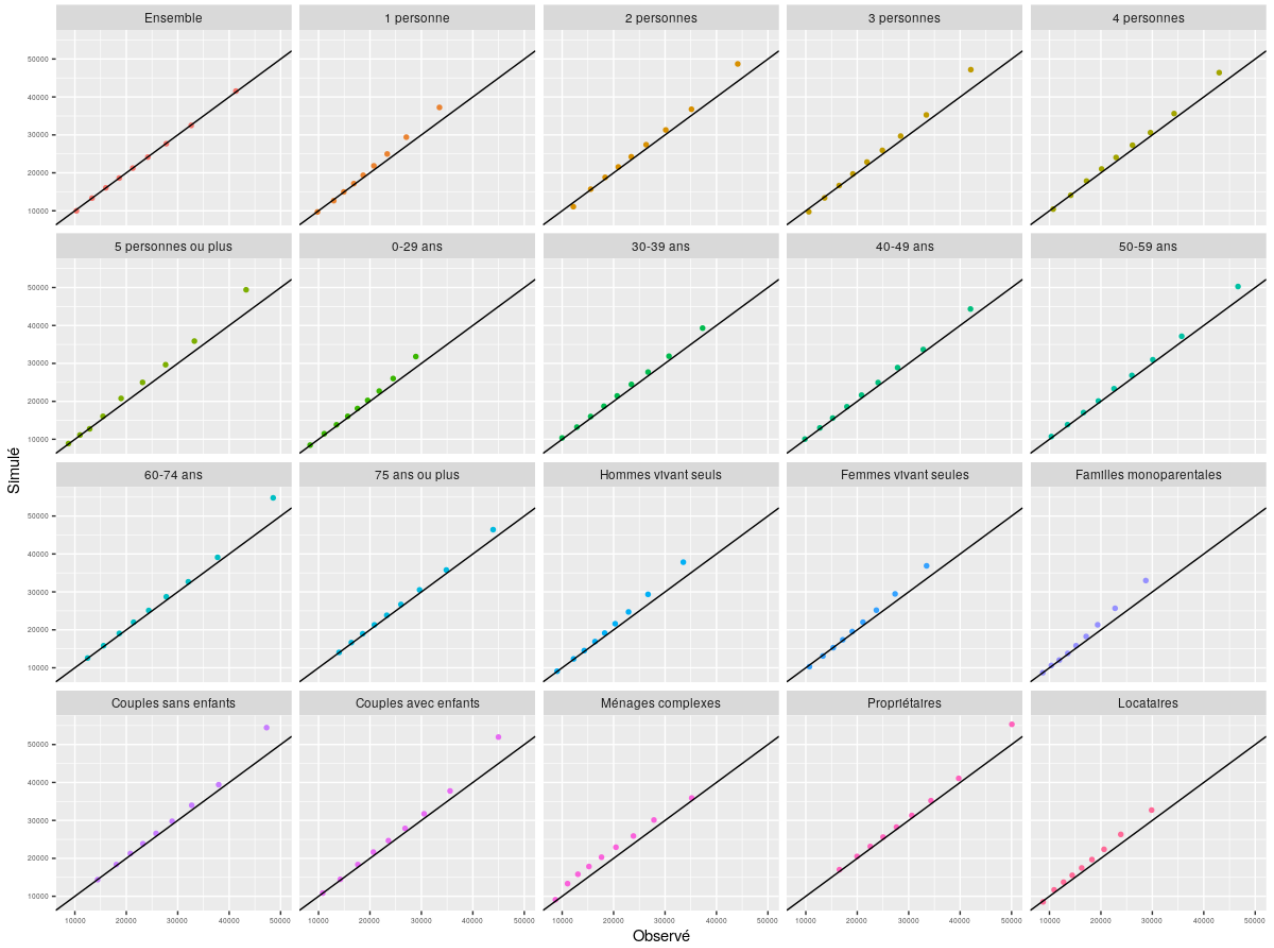


FIGURE 4.1 : Diagrammes Quantile-Quantile (Q-Q plots) entre déciles simulés et observés pour la commune de Nantes

Pour les déciles de l'ensemble de la population, les différences entre les déciles simulés et observés sont très faibles (environ 3% pour le premier décile D1 et moins de 0,6% pour les autres déciles), avec des différences absolues de 3 euros et 1 euro respectivement pour les déciles D2 et D4 (Tableau 4.5).

Pour les autres catégories, de petites différences sont également observées : 88% des déciles simulés présentent des différences de moins de 10% avec les déciles réels et environ 69% des déciles simulés présentent des différences de moins de 5%. Les erreurs relatives les plus faibles sont observées pour les attributs âge de la personne de référence et taille du ménage, avec des différences inférieures à 1% pour le décile D1 des modalités 75/+,

Tableau 4.5 : Erreurs absolues observées entre déciles observés et simulés pour la commune de Nantes

Catégories	Déciles (euros)								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Ensemble	310	3	29	1	14	36	90	120	241
1 personne	95	297	58	295	674	1 079	1 612	2 339	3 725
2 personnes	1 090	139	447	626	802	1 080	1 162	1 622	4 572
3 personnes	838	259	160	568	947	1 040	1 245	1 831	5 114
4 personnes	299	45	626	881	1 097	1 127	959	1 381	3 411
5 personnes/+	109	130	121	581	1 778	1 827	2 014	2 641	6 109
Femmes vivant seules	421	228	37	190	525	905	1 462	2 115	3 392
Hommes vivant seuls	77	116	240	514	879	1 307	1 813	2 660	4 284
Couples sans enfant	36	297	485	612	769	866	1 263	1 458	7 183
Couples avec enfants	15	237	713	983	1 067	1 052	1 209	2 187	6 988
Familles monoparentales	1	250	114	225	639	1 113	1 961	2 919	4 260
Ménages complexes	393	2 286	2 746	2 641	2 653	2 475	2 034	2 266	722
Moins de 30	92	348	319	371	532	689	897	1 503	2 883
30-39	370	334	463	579	737	1 001	1 002	1 127	2 002
40-49	222	269	368	584	798	888	1 052	820	2 323
50-59	348	337	425	573	772	801	877	1 441	3 609
60-74	90	236	419	590	748	938	609	1 356	6 239
75/+	50	249	384	430	558	663	885	921	2 461
Propriétaire	482	533	580	576	595	694	860	1 351	5 265
Locataire	235	780	976	1 095	1 201	1 410	1 762	2 441	2 856

Note : le premier décile (D1) est le revenu au dessous duquel se situent 10 % des revenus ; le neuvième décile (D9) est le revenu au dessous duquel se situent 90 % des revenus.

60-74 et 1 personne. De façon générale, les résultats montrent que les revenus simulés sont plus cohérents avec les données FiLoSoFi pour les déciles D1 à D7. Plus de 80 % de ces déciles ont une différence relative de moins de 5 % avec les données réelles.

En revanche, seuls 71% des déciles D8 et D9 simulés présentent des différences relatives inférieures à 10% avec les déciles réels. Comme indiqué précédemment dans la Figure 4.1, des différences plus importantes sont observées pour l'attribut structure familiale du ménage. Pour les catégories ménages complexes et familles monoparentales en particulier, les différences relatives respectives s'élèvent à plus de 15% (pour les déciles D2 à D4) et à plus de 10% (pour les déciles D7 à D9). Les écarts observés sont attribuables aux différences méthodologiques entre les deux sources de données utilisées et aux hypothèses formulées dans le processus d'estimation et d'affectation du revenu. Cet

Tableau 4.6 : Erreurs relatives observées entre déciles observés et simulés pour la commune de Nantes (%)

Catégories	Déciles								
	D1	D2	D3	D4	D5	D6	D7	D8	D9
Ensemble	3.01	0.02	0.18	0.01	0.07	0.15	0.32	0.37	0.58
1 personne	0.97	2.29	0.39	1.75	3.61	5.20	6.90	8.64	11.11
2 personnes	8.95	0.89	2.44	2.99	3.42	4.10	3.86	4.62	10.36
3 personnes	7.92	1.90	0.97	2.97	4.33	4.18	4.38	5.48	12.15
4 personnes	2.78	0.32	3.64	4.37	4.78	4.31	3.24	4.03	7.93
5 personnes/+	1.24	1.18	0.94	3.76	9.36	7.89	7.29	7.95	14.11
Femmes vivant seules	3.93	1.71	0.24	1.11	2.76	4.29	6.16	7.73	10.13
Hommes vivant seules	0.85	0.95	1.68	3.14	4.81	6.44	7.91	9.96	12.77
Couples sans enfant	0.25	1.64	2.33	2.64	2.98	3.00	3.86	3.84	15.19
Couples avec enfants	0.14	1.66	4.04	4.76	4.52	3.92	3.96	6.15	15.54
Familles monoparentales	0.01	2.41	0.96	1.66	4.21	6.50	10.12	12.82	14.83
Ménages complexes	4.52	20.68	21.02	17.37	15.03	12.10	8.53	8.14	2.05
Moins de 30	1.10	3.13	2.36	2.37	3.03	3.52	4.11	6.13	9.97
30-39	3.71	2.59	2.98	3.20	3.56	4.27	3.75	3.66	5.37
40-49	2.26	2.11	2.42	3.25	3.83	3.69	3.78	2.50	5.53
50-59	3.36	2.49	2.56	2.94	3.42	3.08	2.91	4.04	7.74
60-74	0.72	1.51	2.25	2.76	3.07	3.38	1.90	3.59	12.85
75/+	0.36	1.52	2.07	2.06	2.40	2.55	2.99	2.64	5.60
Propriétaire	2.91	2.67	2.57	2.30	2.15	2.27	2.50	3.40	10.52
Locataire	2.68	7.15	7.66	7.59	7.38	7.72	8.55	10.23	9.56

Note : le premier décile (D1) est le revenu au dessous duquel se situent 10 % des revenus ; le neuvième décile (D9) est le revenu au dessous duquel se situent 90 % des revenus.

aspect sera mieux explicité dans la conclusion.

Conclusion

Ce chapitre a proposé une méthodologie pour aborder le problème de l'ajout d'attributs à une population synthétique. Comme exemple, un niveau de vie a été affecté à chacun des 157 000 ménages de la commune de Nantes sur la base des données agrégées du dispositif FiLoSoFi. Ce dispositif constitue une source de données importante pour estimer les ressources économiques et le niveau de vie en fonction de certains attributs sociodémographiques des ménages. Il est alors possible d'établir un rapprochement entre les données agrégées de FiLoSoFi et le fichier détail du recensement français qui a permis de constituer la population des ménages. La méthodologie utilisée intègre trois étapes distinctes.

La première étape a consisté à modéliser théoriquement le problème au moyen d'une distribution multinomiale. Il s'agissait ici d'identifier la probabilité la plus probable du revenu en considérant les catégories croisées de la population synthétique. Autrement dit, étaient recherchées, les probabilités conditionnelles de tranches de revenus en fonction de la combinaison des différentes catégories des ménages. L'estimation de ces probabilités conditionnelles pouvait être formulée à travers un problème de maximisation de l'entropie (MaxEnt), basé sur les attributs disponibles à la fois dans la population synthétique et dans les données agrégées. Dans ce cas d'étude spécifique, des problèmes numériques se sont posés lors de l'application de la méthode d'optimisation MaxEnt en raison du nombre important de paramètres à estimer. Cette méthode pourrait être appliquée à des problèmes comportant moins de contraintes et de variables d'intérêt, comme le suggèrent certaines simulations préliminaires effectuées sur des problèmes à plus petite échelle.

Dans la deuxième étape, une heuristique offrant une solution pratique au problème est proposée. Le problème d'optimisation a été divisé en plusieurs sous problèmes d'optimisation afin de réduire les contraintes. Cette heuristique a combiné le théorème de Bayes et l'algorithme de minimisation de l'entropie croisée (MinxEnt). En considérant toutefois le grand nombre de paramètres à estimer, certains des résultats obtenus n'étaient pas valides. Pour remédier à cela, une méthode de post-traitement a été appliquée lors d'une

troisième étape afin d'assurer la cohérence de nos résultats.

Les résultats obtenus suggèrent que la méthodologie proposée produit des résultats cohérents avec la plupart des données agrégées de départ, malgré le fait qu'un certain nombre de revenus aient été retirés dans l'étape de post-traitement. Toutefois, des différences plus importantes ont été constatées pour certaines catégories de ménage ainsi que pour les déciles les plus élevés (D8 et D9). Les différences entre les catégories sont dues aux différences relatives entre les données du recensement et celles de FiLoSoFi et aux hypothèses formulées.

Comme souligné dans la section 4.1.3, l'unité d'enquête dans FiLoSoFi est le ménage fiscal (tous les individus figurant sur une même déclaration de revenus). Une personne peut donc être affiliée à un ménage du point de vue fiscal sans nécessairement résider dans ce ménage. Dans le recensement, seuls les résidents d'un ménage donné sont pris en compte. Il existe alors des différences méthodologiques entre les deux sources et un même ménage peut se retrouver dans deux catégories différentes. Par conséquent, le nombre de ménages et la taille des ménages peuvent diverger entre les deux sources. Cela est particulièrement vrai pour la variable structure familiale du ménage. Par exemple, du point de vue du recensement, un ménage monoparental est composé d'un parent isolé et d'un ou plusieurs enfants (qui ne sont pas parents). En revanche, dans la base de données FiLoSoFi, un ménage fiscal peut être considéré comme un ménage monoparental s'il est composé de plusieurs personnes dont le principal déclarant fiscal est célibataire, divorcé ou veuf. Ce ménage monoparental dans FiLoSoFi sera considéré comme un ménage complexe dans le recensement. Ces différences méthodologiques ne peuvent pas être corrigées par l'heuristique proposée.

Aucune information n'étant disponible entre deux déciles de revenu, nous avons supposé une distribution linéaire des revenus. Pour chaque catégorie, un revenu maximal $R_{max} = D9 \times 1,5$ est fixé afin de procéder à une approximation linéaire. Cette hypothèse peut cependant être incorrecte pour les déciles les plus élevés et donc conduire à certains biais. Pour remédier à ce problème, il est nécessaire d'acquérir une meilleure connaissance à priori de la distribution des revenus et, en particulier, des revenus extrêmes en

se référant notamment à la littérature économique. La précision de l'approximation de Stirling dépend du nombre de ménages d'une catégorie croisée avec revenu. Si ce nombre est supérieur à 20, l'approximation est sous-estimée de moins de 10%. Pour résoudre numériquement le problème d'optimisation, le nombre de ménages d'une catégorie croisée avec revenu doit être supérieur à 20. La prudence est donc de mise lors de la manipulation des résultats relatifs à un petit nombre de ménages. Cette dernière remarque fournit une justification a posteriori de l'étape de post-traitement, qui a permis d'obtenir des tranches de revenus plus larges, et donc des effectifs plus importants pour chaque tranche.

Une amélioration de la méthode pourrait consister à trouver la bonne discrétisation des revenus afin de produire des intervalles de revenus plus larges qui conduisent à un nombre plus approprié de ménages pour une catégorie croisée. Ce changement entraînerait moins de probabilités invalides.

Dans cet exemple d'application, les données agrégées prennent la forme de déciles. Notre heuristique a été conçue pour traiter ce type d'information. Cependant, la méthodologie reste générale et est capable d'intégrer différents types de données agrégées : revenu moyen, nombre de mètres carrés par type de ménage, etc. À titre d'exemple, il est possible d'utiliser des données agrégées typiques telles que les données de mobilité disponibles en Suisse [6] ou aux Etats-Unis [7], pour déterminer un mode de transport ou l'état de mobilité d'individus synthétiques.

Une attention particulière a été accordée au traitement des données pour permettre une appropriation de la méthode par d'autres utilisateurs. Les scripts de programmation (scripts R) utilisés seront tous librement disponibles dès que l'article issu de ce chapitre sera publié. L'allocation d'un revenu aux ménages synthétiques offre la possibilité d'affecter la population synthétique à une échelle géographique plus petite que la commune et l'IRIS.

6. <https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/tables.assetdetail.7226558.html>, page consultée le 10 août 2021.

7. <https://www.census.gov/content/census/en/data/tables/2020/demo/geographic-mobility/cps-2020.html>, page consultée le 10 août 2021.

Références bibliographiques

Bösch, P. M., Müller, K., and Ciari, F. (2016). The ivt 2015 baseline scenario. In *16th Swiss Transport Research Conference (STRC 2016)*. 16th Swiss Transport Research Conference (STRC 2016).

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical matching : Theory and practice*. John Wiley & Sons.

Felbermair, S., Lammer, F., Trausinger-Binder, E., and Hebenstreit, C. (2020). Generating synthetic population with activity chains as agent-based model input using statistical raster census data. *Procedia Computer Science*, 170 :273–280.

Golub, G. H., Loan, C. F. V., Loan, C. F. V., and Golub (1996). *Matrix Computations*. JHU Press.

Hackl, J. and Dubernet, T. (2019). Epidemic spreading in urban areas using agent-based transportation models. *Future Internet*, 11(4) :92.

He, B. Y., Zhou, J., Ma, Z., Chow, J. Y., and Ozbay, K. (2020). Evaluation of city-scale built environment policies in new york city with an emerging-mobility-accessible synthetic population. *Transportation Research Part A : Policy and Practice*, 141 :444–467.

Hörl, S. and Balac, M. (2020). Reproducible scenarios for agent-based transport simulation. *Arbeitsberichte Verkehrs Raumplan*, 1499.

Ilahi, A. and Axhausen, K. W. (2019). Integrating bayesian network and generalized raking for population synthesis in greater jakarta. *Regional Studies, Regional Science*, 6(1) :623–636.

Kapur, J. N. and Kesavan, H. K. (1992). Entropy Optimization Principles and Their Applications. In Singh, V. P. and Fiorentino, M., editors, *Entropy and Energy Dissipation in Water Resources*, Water Science and Technology Library, pages 3–20. Springer Netherlands, Dordrecht.

Mattos, R. and Veiga, A. (2004). Entropy optimization : Computer implementation of the maxent and minexent principles. Technical report, Working Paper. Universidade Federal de Juiz de Fora, Brazil.

Murata, T., Sugiura, S., and Harada, T. (2017). Income allocation to each worker in synthetic populations using basic survey on wage structure. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE.

Niven, R. K. (2005). Combinatorial Information Theory : I. Philosophical Basis of Cross-Entropy and Entropy.

Sallard, A., Balać, M., and Hörl, S. (2020). A synthetic population for the greater são paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, 1545.

Vosooghi, R., Puchinger, J., Jankovic, M., and Vouillon, A. (2019). Shared autonomous vehicle simulation and service design. *Transportation Research Part C : Emerging Technologies*, 107 :15–33.

Zhang, D., Cao, J., Feygin, S., Tang, D., Shen, Z.-J. M., and Pozdnoukhov, A. (2019). Connected population synthesis for transportation simulation. *Transportation Research Part C : Emerging Technologies*, 103 :1–16.

Chapitre 5

Affectation d'une population synthétique à une échelle spatiale plus fine : résolution par une programmation mixte en nombres entiers

Sommaire

Résumé

Une simulation multi-agents réaliste implique donc de pouvoir prendre en compte la localisation spatiale des agents d'une population synthétique. Suivant les besoins de la simulation, l'échelle géographique de localisation varie selon différents niveaux de précision : ville, commune, quartier, bâtiment... Ce chapitre analyse le problème de l'affectation d'agents synthétiques à une échelle spatiale plus fine. Deux niveaux géographiques différents sont introduits.

Le premier est un niveau géographique sur lequel les données de la population synthétique sont disponibles. Ce niveau est le niveau supérieur et est appelé zone statistique conteneur (ZSC). Les agents synthétiques doivent être affectés à plusieurs niveaux géographiques inférieurs (plus petits que le niveau supérieur) pour lesquels des données agrégées sont uniquement disponibles. Chaque niveau géographique inférieur est désigné sous le terme de zone statistique élémentaire (ZSE). Les contours des ZSE sont bien définis et ne se chevauchent pas ; leur regroupement permet de reconstituer la ZSC à laquelle elles sont rattachées.

Une méthodologie d'affectation spatiale des agents synthétiques d'une ZSC vers plusieurs ZSE est proposée. La méthodologie repose sur des attributs communs entre la ZSC et la ZSE, et des attributs supplémentaires permettant de différencier les observations synthétiques. Ces attributs sont appelés attributs d'intérêt et sont estimés à partir de sources de données externes. La méthodologie est décrite en détail avec l'utilisation de données françaises issues du recensement de la population et des données fiscales.

L'affectation des agents synthétiques est modélisé comme un problème de programmation quadratique mixte en nombres entiers (MIQP problème en Anglais). Il est démontré que l'applicabilité de cet algorithme est limitée aux populations synthétiques de petites tailles. Une heuristique est par la suite proposée pour tenir compte de populations synthétiques de taille plus importante. Les tests effectués montrent que l'heuristique donne des solutions quasi optimales en un temps de calcul rapide. La méthodologie considérée et l'heuristique proposée sont générales et permettent de répondre aux besoins des utilisateurs.

Mots clés

Spatialisation, zones statistiques, agents synthétiques, programmation mixte en nombres entiers.

Introduction

Dans de nombreuses simulations multi-agents, les comportements des agents synthétiques sont fortement déterminés par leurs attributs et leur localisation. Ainsi, l'hétérogénéité spatiale des caractéristiques des agents doit être prise en compte avec une granularité (niveau de détail) qui dépend des objectifs du modèle et des données spatiales disponibles (Zhu and Ferreira Jr, 2014 ; Chapuis et al., 2018). Le manque de spécificité géographique a été identifié dans la littérature scientifique comme un problème majeur (Su et al., 2010 ; Anderson et al., 2014 ; Long and Shen, 2015 ; Ji and Wan, 2021). Après la génération de la population synthétique, les agents constituant la population (individus, ménages) sont souvent répartis dans des zones de grande taille (villes, communes, quartiers...). Ces unités de surface n'offrent pas une précision spatiale suffisante et ne permettent pas dans certaines situations de réaliser des analyses satisfaisantes en matière de transport et d'urbanisme (Thomson et al., 2018). Une information spatiale plus précise peut donc améliorer considérablement la pertinence et la qualité des analyses.

Dans de nombreux cas, les lieux de résidence ou de travail des ménages synthétiques sont tirés au sort parmi tous les emplacements disponibles (Bösch et al., 2016 ; Hackl and Dubernet, 2019 ; Sallard et al., 2020).

D'autres approches appliquent des méthodes de génération de population synthétique comme par exemple l'ajustement proportionnel itératif (IPF) et considèrent la localisation de la population de la même manière que les autres attributs des agents (Zhu and Ferreira Jr, 2014).

Harada and Murata (2017) ont proposé sur un cas japonais, une méthode qui s'appuie sur des marginaux fournis à des unités d'échelle plus fines. Ils génèrent une population synthétique à l'échelle d'une ville, puis dans un premier temps affectent aléatoirement chaque ménage synthétique dans un district en fonction du nombre de ménages par types de famille et de la taille des ménages. Dans un deuxième temps, ils sélectionnent aléatoirement deux ménages identiques (même type de famille et même taille) mais affectés à deux districts distincts et les échangent. A partir de la distribution par sexe

et par âge des districts, ils évaluent le résultat de l'échange et répètent cette procédure afin d'avoir la meilleure affectation possible.

Chapuis et al. (2018) cherchent à affecter une population synthétique d'individus générée au niveau IRIS à des bâtiments. Pour y parvenir, ils utilisent des images satellitaires sur l'occupation des sols et des données sur la géométrie des bâtiments. A partir des images satellitaires, ils procèdent à un découpage de leur terrain d'étude (ville de Rouen) en différentes catégories : zones densément peuplées, zones moins peuplés, espaces verts, routes, points d'eau... Les individus synthétiques localisés au niveau IRIS sont ensuite affectés à des grilles de 30 m^2 par interpolation spatiale. Dans une dernière étape, les individus sont affectés aux bâtiments se trouvant dans les grilles de 30 m^2 à partir des caractéristiques des bâtiments notamment leur capacité. Si aucun bâtiment ne se trouve dans la grille, l'individu est affecté au bâtiment le plus proche.

Dans ce chapitre, une approche méthodologique générique d'affectation d'une population synthétique est proposée. Cette approche est motivée par la configuration des données dont disposent souvent les modélisateurs : d'une part, les informations détaillées nécessaires pour générer une population synthétique sont disponibles avec une faible précision spatiale. D'autre part, des données agrégées (également appelées données marginales) sont souvent renseignées à une échelle spatiale plus fine. Ces données agrégées peuvent être exploitées pour améliorer la précision de la localisation des agents synthétiques. Le processus d'affectation que considéré nécessite deux niveaux géographiques différents :

— le niveau géographique supérieur pour lequel les données de la population synthétique sont disponibles. Ce niveau est désigné par le terme « zone statistique conteneur (ZSC) » ;

— les niveaux géographiques inférieurs qui contiennent les données marginales. Chaque niveau est désigné par le terme « zone statistique élémentaire (ZSE) ». Les contours des ZSE sont bien définis et ne se chevauchent pas et leur regroupement permet de reconstituer la ZSC à laquelle elles sont rattachées.

Afin de répartir des agents synthétiques d'une zone statistique de niveau supérieur à

des zones statistiques plus petites (ZSC vers ZSE), deux types d'attributs sont utilisés : des attributs communs et des attributs supplémentaires d'intérêt calculés à partir de sources de données externes. Les attributs communs sont disponibles à la fois au niveau des ZSC et des ZSE. En revanche, les attributs d'intérêt sont renseignés dans l'un des deux niveaux, mais peuvent être estimés pour l'autre niveau.

La méthodologie proposée s'appuie sur un exemple comme dans le chapitre précédent. Dans cette exemple, des ménages synthétiques (générés au niveau IRIS) sont répartis dans des carreaux. Les données carroyées sont extraites du dispositif FiLoSoFi. L'affectation des ménages est modélisée comme un problème de programmation mixte en nombres entiers à termes quadratiques (Mixed Integer Quadratic Problem ou MIQP problem en anglais). Dans un premier temps, un algorithme exact est appliqué pour résoudre le problème. Cependant, l'applicabilité de cet algorithme est limitée aux petites populations synthétiques. Par conséquent, une heuristique est proposée pour tenir compte de populations synthétiques de taille plus importante. Les tests effectués montrent que l'heuristique donne des solutions quasi optimales en un temps de calcul rapide.

Le reste du chapitre est organisé comme suit. Les deux prochaines sections (section 5.1 et section 5.2) décrivent le problème et la configuration des données. La section 5.3 procède à la formulation mathématique du problème et introduit les variables de décision et les paramètres. La section 5.4 est consacrée à la présentation des approches de résolution du problème et des résultats obtenus. La section 5.5 discute les résultats de nos analyses et est suivie d'une conclusion offrant des perspectives de recherche.

5.1 Formulation générale du problème

Ce chapitre traite du problème de l'affectation des ménages d'une zone statistique conteneur (ZSC) à plusieurs zones statistiques élémentaires (ZSE) qui ne se chevauchent pas. Habituellement, le nombre d'attributs disponibles au niveau des ZSE est moins important qu'au niveau des ZSC. Il existe trois types d'attributs entre les deux niveaux :

1. Les attributs communs à la ZSC et à la ZSE ;

2. Les attributs uniquement présents dans la ZSC ;
3. Les attributs uniquement présents dans la ZSE.

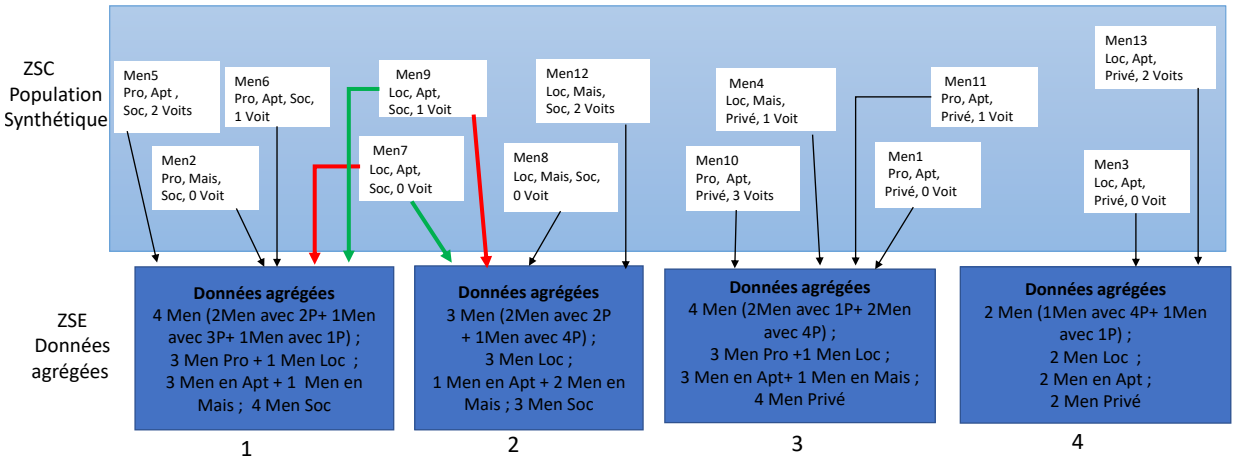
Afin d'obtenir une localisation plus précise de la population, chaque ménage d'une ZSC doit être attribué à exactement une ZSE. Les ménages doivent également être répartis de telle sorte que les valeurs agrégées de leurs attributs soient égales aux attributs agrégés de la ZSE. Le problème d'affectation ainsi formulé peut être résolu à l'aide du paradigme de la programmation par contraintes (Rossi et al., 2008), dans lequel les contraintes sont définies de manière à garantir que les données agrégées de chaque ZSE sont respectées pendant l'étape d'affectation des ménages. Une solution à ce problème de satisfaction de contraintes (CSP) consiste à une affectation qui satisfasse toutes les contraintes (Freuder and Mackworth, 2006). La qualité de la solution de ce problème dépend à la fois du nombre et de la nature discriminante (capacité à différencier les ménages) des attributs communs. Par exemple, lorsque le nombre d'attributs communs est faible et que ces attributs ne sont pas suffisamment discriminants, la satisfaction des contraintes peut donner lieu à des solutions multiples dont le nombre peut être important.

La Figure 5.1 présente un exemple de 13 ménages synthétiques d'une ZSC à répartir entre 4 ZSE (numérotées de 1 à 4). Dans cet exemple, 5 attributs des ménages sont disponibles : 3 attributs communs et 2 attributs uniquement présents dans l'un des deux niveaux. Les attributs communs aux deux niveaux sont les suivants :

1. Le statut de propriété du ménage avec comme catégories dans la ZSC, propriétaire ou locataire. Dans chaque ZSE, nous avons le nombre de ménages propriétaires et le nombre de ménages locataires ;
2. Le type de logement (ménage appartement ou ménage dans une maison dans la ZSC) et les données agrégées du type de logement dans chaque ZSE (le nombre de ménages en appartements et le nombre de ménages en maisons) ;
3. Le statut du logement (social ou privé dans la ZSC) et les données agrégées du statut du logement dans chaque ZSE (nombre de logements sociaux et nombre de

logement privés).

FIGURE 5.1 : Illustration du problème d'affectation des ménages (1/2).



Note : **Men** : ménage ; **P** : personne ; **Pro** : ménage propriétaire ; **Loc** : ménage locataire ; **Apt** : ménage dans un appartement ; **Mais** : ménage dans une maison ; **Soc :** ménage dans un logement social ; **Privé** : ménage dans un logement non social ; **Voit :** voiture.

Deux solutions distinctes sont possibles. Les flèches de couleur rouge et verte traduisent le fait que les ménages 7 et 9 peuvent être affectés soit dans la ZSE 1 ou soit la ZSE 2.

Les deux attributs restants, à savoir la possession d'une voiture et la composition du ménage ne sont présents respectivement que dans la ZSC et la ZSE. Dans cet exemple, la procédure d'affectation peut être effectuée en prenant en compte les attributs communs. Une solution consiste alors en une affectation qui satisfait les données agrégées (contraintes) de chaque ZSE.

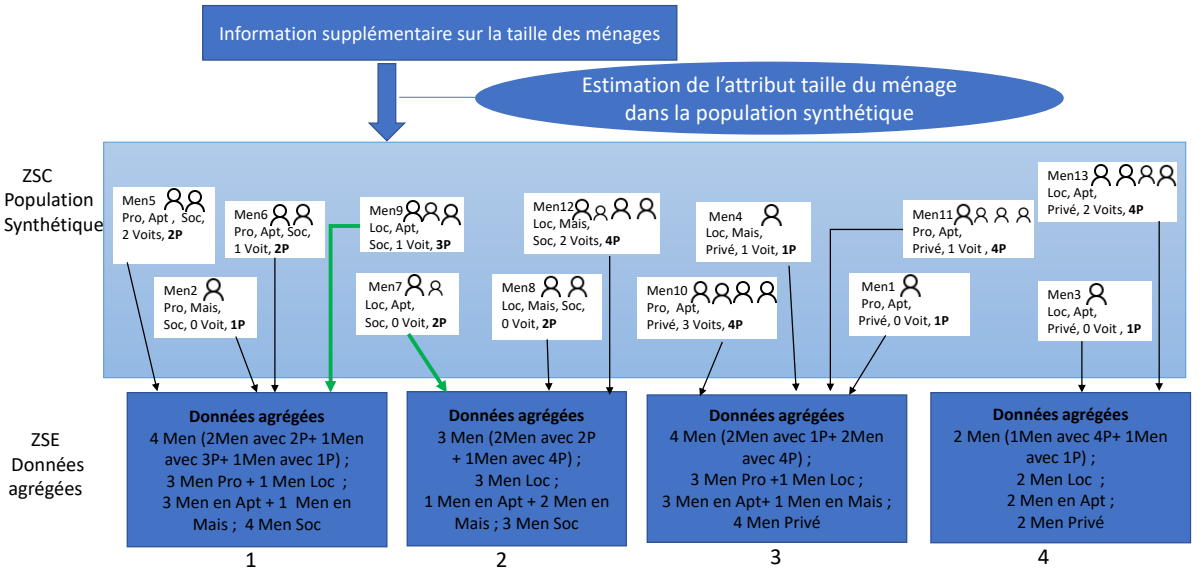
Comme le montre la Figure 5.1, en s'appuyant sur les trois attributs communs, deux solutions distinctes sont possibles. Les flèches vertes et rouges illustrent le fait que les ménages Men7 et Men9 peuvent êtres affectés soit à la ZSE 1 ou soit à la ZSE 2. Dans des configurations avec un nombre élevé de ménages, le nombre de solutions possibles peut être beaucoup plus important.

Dans de telles situations, il peut souvent s'avérer nécessaire d'enrichir les attributs

de la population synthétique (niveau ZSC) ou les attributs agrégés (niveau ZSE) avec des attributs supplémentaires, estimés à partir d'autres sources de données. Lorsqu'ils sont utilisés dans le processus d'allocation, ces attributs supplémentaires permettent de mieux différencier (discriminer) les ménages. Il est donc intéressant de les considérer. Ces attributs sont des « attributs d'intérêt » et seront désignés sous ce terme dans tout le chapitre.

Dans l'exemple traduit par la Figure 5.1, la composition du ménage est un attribut spécifique, uniquement présent au niveau de la ZSE. Supposons que le nombre d'individus puisse être estimé pour chaque ménage synthétique en utilisant des sources de données externes. L'ajout de ces informations à la population synthétique peut fournir une solution unique pour l'affectation des ménages, comme l'illustre la Figure 5.2.

FIGURE 5.2 : Illustration du problème daffectation des ménages (2/2).



Note : **Men** : ménage ; **P** : personne ; **Pro** : ménage propriétaire ; **Loc** : ménage locataire ; **Apt** : ménage dans un appartement ; **Mais** : ménage dans une maison ; **Soc :** ménage dans un logement social ; **Privé** : ménage dans un logement non social ; **Voit :** voiture.
L'ajout de l'attribut taille du ménage dans la population synthétique permet d'avoir une solution unique d'allocation des ménages.

Contrairement aux attributs communs, les attributs d'intérêt sont estimés pour au

moins un des deux niveaux. Il est donc probable que leurs valeurs ne soient pas égales dans les deux niveaux. Par conséquent, utiliser ces attributs dans le processus d'affectation comme contraintes à respecter pourrait ne pas être approprié. Il conviendrait plutôt de minimiser les différences entre les attributs d'intérêt dans les deux niveaux en formulant une ou plusieurs fonctions objectif. L'affectation des ménages est alors résolue comme un problème d'optimisation mono-objectif ou multi-objectif formulé sur les attributs estimés et des contraintes formulées sur les attributs communs. La section suivante procède à une description du cas d'étude.

5.2 Configuration des données

Pour illustrer la méthodologie proposée, des données françaises sont utilisées. Ces données sont diffusées à deux niveaux géographiques différents. Le premier niveau est l'Ilot Regroupé pour l'Information Statistique (IRIS) qui représente la zone statistique conteneur ; le deuxième niveau est le carreau qui représente la zone statistique élémentaire. Les données carroyées diffusées sont des données agrégées, relatives aux attributs des ménages et à leur logement.

5.2.1 Population synthétique de ménages : niveau IRIS (ZSC)

Les IRIS constituent la plus petite échelle géographique sur lequel des informations du recensement français de la population sont disponibles. Ils offrent l'outil le plus élaboré à ce jour pour décrire la structure interne des communes. Dans le chapitre précédent, une population synthétique de ménages a été générée pour chaque IRIS de la commune de Nantes, et un revenu a été alloué à chaque ménage synthétique. Il s'agit alors d'affecter cette population synthétique de ménages vers les carreaux en utilisant les attributs disponibles dans les deux niveaux.

5.2.2 Données marginales agrégées : niveau carreau (ZSE)

Le carroyage est une technique de quadrillage consistant à découper le territoire en carrés appelés carreaux dont les côtés peuvent aller de 200 mètres jusqu'à plusieurs kilomètres. Les carreaux présentent l'avantage d'être stables au fil du temps et permettent de s'affranchir des limites administratives habituelles. Ils peuvent être assemblés pour construire ou approcher n'importe quelle zone d'intérêt. Il est ainsi possible de regrouper les carreaux pour retrouver les IRIS auxquels ils correspondent. La découpage en carreaux s'effectue dans le cadre de la directive européenne INSPIRE [1] (Directive 2007/2/CE) qui vise à harmoniser les données spatiales dans les pays européens pour une meilleure diffusion, disponibilité et exploitation de l'information géographique (Bartha and Kocsis, 2011). Grâce à une norme standard et compatible, les données carroyées françaises peuvent être juxtaposées ou combinées aux données carroyées allemandes ou italiennes (Darriau, 2020).

Les données carroyées françaises sont diffusées par l'Institut national de la statistique et des études économiques (Insee), à partir du dispositif « Fichier Localisé Social et Fiscal »(FiLoSoFi) présenté dans le chapitre précédent (section 4.4.1). Les données carroyées s'appliquent donc aux ménages fiscaux et permettent de disposer de nombreuses informations sur les ménages et leurs conditions économiques (taille, statut familial, niveau de vie, situation de pauvreté...). Ces informations sont précieuses pour identifier et planifier les besoins en infrastructures de transport ou en équipements publics. Elles sont aussi utiles aux acteurs économiques pour conforter ou adapter leur stratégie de développement.

Sur certains carreaux, le nombre de ménages fiscaux peut être inférieur à 11, seuil de confidentialité pour le dispositif FiLoSoFi. Dans ce cas, les informations sont imputées (pondérées) avant d'êtres diffusées. La procédure consiste à regrouper les carreaux entre eux jusqu'à obtenir des groupes de carreaux qui rassemblent au moins 11 ménages. Une fois les groupes déterminés, on impute à chaque carreau, et pour tous les attributs

1. European Commission : About INSPIRE [en ligne] (page consultée le 18 août 2021)< `https://inspire.ec.europa.eu/about-inspire/563`>

agrégés, la valeur moyenne du groupe pondérée par le nombre d'individus [2]. L'Insee applique également des précautions supplémentaires sur les informations relatives à la pauvreté et au niveau de vie des ménages :

— pour les carreaux ayant plus de 11 ménages mais dont plus de 80 % des ménages sont pauvres, le nombre de ménages pauvres a été ramené à la valeur de 80% ;

— pour la distribution des niveaux de vie, les valeurs extrêmes ont fait l'objet d'un traitement particulier, appelé winsorisation [3]. La winsorisation désigne une technique statistique de lissage qui consiste à remplacer les niveaux de vie les plus faibles et les plus élevés par des valeurs seuils, calculées au niveau du département. Dans un département donné, le niveau de vie d'un ménage est rabaissé au 95^{e} centile de la distribution départementale des niveaux de vie si son niveau de vie est supérieur à ce seuil. Inversement, le niveau de vie d'un ménage est ramené au 5^{e} centile de la distribution départementale si son niveau de vie est inférieur à ce seuil. Si le niveau de vie se situe entre ces deux seuils, aucun traitement n'est effectué.

Les carreaux concernés par toutes ces modifications possèdent des variables indicatrices qui mentionnent le type d'imputation appliquée. Après tous ces différents traitements statistiques, les données carroyées sont diffusées sur trois niveaux de grille :

1. la grille de niveau naturel qui correspond à un partitionnement du territoire en carreaux de différentes tailles (de 200 m jusquà 32 km) afin de diffuser toutes les informations sans imputation des données, tout en respectant le secret statistique. On commence par couvrir le territoire avec des carreaux de 32 km. Les divisions successives permettent de passer de carreaux de 32 km à 16 km, puis 8 km, 4 km, 2 km, 1 km et enfin 200 m. Les divisions s'arrêtent lorsque les carreaux obtenus sont de taille 200 m ou lorsque la prochaine division entraînerait qu'un ou plusieurs carreaux ne respectent pas le seuil de confidentialité fixé à 11 ménages ;

2. Insee (2019c) : Les données carroyées de l'Insee [en ligne] (page consultée le 18 août 2021)<`https://www.insee.fr/fr/statistiques/fichier/4176290/documentation_DonneesCarroyees.pdf`>

3. Nous utilisons cet anglicisme pour nous conformer au vocabulaire de l'Insee. Le terme français correspondant est fenêtrage.

2. la grille de niveau 1 km qui correspond à un pavage du territoire français par des carreaux de 1 km de côté. Lorsque le nombre de ménages fiscaux est inférieur à 11, le carreau a sa variable indicatrice de l'imputation « I_est_1km » égale à 1 ;
3. la grille de niveau 200 mètres qui correspond à un pavage du territoire français par des carreaux de 200 mètres de côté. Lorsque le nombre de ménages fiscaux est inférieur à 11, le carreau a sa variable indicatrice de l'imputation « I_est_cr » égale à 1.

Dans la suite du travail, sont utilisées les données carroyées issues du dispositif Fi-LoSoFi de 2015, disponibles sur le site de l'Insee en juin 2019 [4].

5.2.3 Présentation des attributs communs aux données du recensement et aux données carroyées

Le processus de répartition des ménages synthétiques des IRIS vers les carreaux est effectué en utilisant deux types d'attributs des ménages :

1. Les attributs non estimés, communs aux IRIS et aux carreaux. Ces attributs sont principalement liés aux caractéristiques des ménages et de leur logement ;
2. Un attribut d'intérêt, le niveau de vie des ménages. Dans les données carroyées, l'information sur le niveau de vie diffusée représente « la somme des niveaux de vie de tous les ménages vivant dans le carreau ». Il s'agit du niveau de vie agrégé de tous les ménages du carreau. Dans le chapitre précédent (chapitre 4), un niveau de vie a été attribué à chaque ménage synthétique de la commune de Nantes.

Le Tableau 5.1 résume les attributs communs et l'attribut d'intérêt. Il présente également les catégories associées aux attributs. La section suivante fournit une description détaillée de la méthodologie d'affectation que proposée.

4. Insee (2019d) : Données carroyées - Carreau de 200m [en ligne] (page consultée le 18 août 2021)<https://www.insee.fr/fr/statistiques/4176290?sommaire=4176305>

Tableau 5.1 : Attributs utilisés pour l'affectation de la population synthétique des IRIS aux carreaux.

Attributs des ménages	**Catégories de la Population synthétique**	**Marginaux des Carreaux**
Taille	1 personne	Nombre de ménages d'une personne
	2-4 personnes	Nombre de ménages entre 2 et 4 personnes
	5 personnes ou plus	Nombre de ménages de 5 personnes plus
Statut de propriété	Propriétaire	Nombre de ménages propriétaires
	Locataire	Nombre de ménages locataires
Structure familiale	Ménage monoparental	Nombre de ménages monoparentaux
	Ménage non monoparental	Nombre de ménages non monoparentaux
Statut économique	Ménage pauvre	Nombre de ménages pauvres
	Ménage non pauvre	Nombre de ménages non pauvres
Niveau de vie	*Attribut simulé (euros)*	Niveau de vie total (euros)
Type de logement	Maison	Nombre de ménages en maison
	Appartement	Nombre de ménages en appartement
Date de construction du logement	Avant 1945	Nombre de logements construits avant 1945
	Entre 1945-1989	Nombre de logements construits entre 1945-1989
	A partir de 1990	Nombre de logements construits à partir de 1990
Statut du logement	Logement social	Nombre de logements sociaux
	Logement non social	Nombre de logements non sociaux
Ménage	Ménage à affecter	Nombre de ménages

5.3 Formulation mathématique du problème

5.3.1 Variables de décision et paramètres

A partir des informations du tableau précédent (Tableau 5.1), nous introduisons les notations suivantes :

Ensembles et indices

Soit $\mathbf{H}$ l'ensemble des ménages synthétiques : $\mathbf{H} = \{1, 2, \ldots, n_H\}$

Soit $\mathbf{Z}$ l'ensemble des carreaux : $\mathbf{Z} = \{1, 2, \ldots, n_Z\}$

Chaque ménage $h \in \mathbf{H}$ est affecté à un carreau unique $z \in \mathbf{Z}$.

Variables de décision

Nous définissons $n_H \times n_Z$ avec

$$x_{zh} = \begin{cases} 1 & \text{si le ménage } h \text{ est affecté au carreau z} \\ 0 & \text{sinon} \end{cases} \quad , z \in \mathbf{Z}, h \in \mathbf{H}$$

Paramètres

n_{Hz} : nombre de ménages présents dans le carreau z ;

TR_z : niveau de vie total dans le carreau z ;

R_h : niveau de vie du ménage h ;

Les autres paramètres du problème sont énumérés dans le Tableau 5.2.

Tableau 5.2 : Liste des autres paramètres du problème

Description	**Paramètres des ménages niveau IRIS**	**Paramètres des marginaux niveau carreau**
Taille/structure familiale		
Ménages entre 2-4 personnes	T_{2h}	NT_{2z}
Ménages $>$ 4 personnes	T_{5h}	NT_{5z}
Ménages monoparentaux	C_{5h}	NC_{5z}
Propriété/type		
Propriétaire	B_{1h}	NB_{1z}
Maison	G_{1h}	NG_{1z}
Date de construction du logement		
Avant 1945	D_{1h}	ND_{1z}
A partir de 1990	D_{3h}	ND_{3z}
Conditions économiques		
Ménage social	E_{1h}	NE_{1z}
Ménage pauvre	F_{1h}	NF_{1z}

Tous les paramètres du Tableau 5.2 sont binaires et prennent la valeur 1 si le ménage h possède la caractéristique ou 0 sinon. Par exemple :

$$T_{2h} = \begin{cases} 1 & \text{si la taille du ménage } h \text{ est comprise entre 2 ou 4 personnes} \\ 0 & \text{autrement} \end{cases}$$

5.3.2 Fonction objectif et contraintes

Comme souligné précédemment, le paramètre R_h (revenu du ménage) h a été estimé après génération de la population synthétique. Par conséquent, l'égalité suivante $\sum_{h=1}^{\mathbf{H}} R_h x_{zh} = TR_z$ n'est pas forcément respectée. Nous cherchons alors à affecter les ménages dans un carreau z telle que la somme des niveaux de vie des ménages affectés

soit aussi proche que possible du niveau de vie total du carreau. Cette minimisation de l'écart peut être modélisée comme une fonction objectif. En revanche, les données agrégées des autres attributs (attributs communs) peuvent être respectées.

Le problème d'affectation de la population synthétique dans les carreaux est alors formulé de la façon suivante :

$$\min \sum_{z \in \mathbf{Z}} \left(\sum_{h \in \mathbf{H}} R_h x_{zh} - TR_z \right)^2 \tag{5.1}$$

sous les contraintes suivantes :

$$\sum_{z \in \mathbf{Z}} x_{zh} = 1, \quad h \in \mathbf{H} \tag{5.2}$$

$$\sum_{h \in \mathbf{H}} x_{zh} = n_{Hz}, \quad z \in \mathbf{Z} \tag{5.3}$$

$$\sum_{h \in \mathbf{H}} x_{zh} T_{2h} = NT_{2z}, \quad z \in \mathbf{Z} \tag{5.4}$$

$$\sum_{h \in \mathbf{H}} x_{zh} T_{5h} = NT_{5z}, \quad z \in \mathbf{Z} \tag{5.5}$$

$$\sum_{h \in \mathbf{H}} x_{zh} C_{5h} = NC_{5z}, \quad z \in \mathbf{Z} \tag{5.6}$$

$$\sum_{h \in \mathbf{H}} x_{zh} B_{1h} = NB_{1z}, \quad z \in \mathbf{Z} \tag{5.7}$$

$$\sum_{h \in \mathbf{H}} x_{zh} G_{1h} = NG_{1z}, \quad z \in \mathbf{Z} \tag{5.8}$$

$$\sum_{h \in \mathbf{H}} x_{zh} D_{1h} = N D_{1z}, \quad z \in \mathbf{Z} \tag{5.9}$$

$$\sum_{h \in \mathbf{H}} x_{zh} D_{3h} = N D_{3z}, \quad z \in \mathbf{Z} \tag{5.10}$$

$$\sum_{h \in \mathbf{H}} x_{zh} E_{1h} = N E_{1z}, \quad z \in \mathbf{Z} \tag{5.11}$$

$$\sum_{h \in \mathbf{H}} x_{zh} F_{1h} = N F_{1z}, \quad z \in \mathbf{Z} \tag{5.12}$$

$$x_{zh} \in \{0, 1\}, \quad z \in \mathbf{Z}, h \in \mathbf{H} \tag{5.13}$$

L'équation 5.1 représente la fonction objectif ; les autres équations traduisent les contraintes qui peuvent être classées en deux types :

— les contraintes de validité (équations 5.2 et 5.13) qui s'appliquent à chaque ménage $h \in \mathbf{H}$. Ces contraintes expriment l'appartenance d'un ménage à un unique carreau ;

— les contraintes sur les carreaux (équations 5.3 à 5.12) qui garantissent que les attributs communs sont respectés.

Dans la formulation du modèle, les variables de décision x_{zh} sont binaires ; la fonction objectif est quadratique et toutes les contraintes sont linéaires. Le modèle se présente sous la forme d'un problème de programmation quadratique mixte en nombres entiers (MIQP) qui se caractérise par une fonction objectif de type quadratique, des contraintes linéaires et éventuellement des variables de décisions binaires (limitées aux valeurs 0 et 1) [5].

5. Pour une étude formelle sur les MIQP problèmes, le lecteur intéressé peut consulter Billionnet and Elloumi (2007) et Del Pia et al. (2017).

5.4 Résolution du problème

Cette section est consacrée à la résolution pratique du problème d'affectation des ménages. Les analyses sont effectuées sur un jeu de données simulées, représentatif de la configuration des données françaises. L'utilisation d'un jeu de données simulées permet une meilleure évaluation des résultats et des performances de calcul des algorithmes de résolution.

5.4.1 Construction du jeu de données

Dans la commune de Nantes, le nombre maximal de ménages dans un IRIS est d'environ 4 200. Nous sélectionnons aléatoirement 6 000 ménages synthétiques parmi l'ensemble des ménages synthétiques de la commune de Nantes. Chaque ménage synthétique sélectionné possède les attributs énumérés dans le Tableau 5.1. Ces 6 000 ménages appartiennent à un IRIS constitué de six carreaux numérotés de 1 à 6. Chaque ménage synthétique est affecté à un carreau de manière à avoir des carreaux hétérogènes en termes de taille et de composition. Par exemple, certains carreaux ont un niveau de vie plus élevé tandis que d'autres ont un grand nombre de logements sociaux. Les six carreaux sont ainsi représentatifs de l'hétérogénéité des carreaux de la commune de Nantes. Après affectation des ménages, les carreaux 1, 2, 3, 4, 5 et 6 contiennent respectivement 1 332, 540, 1 518, 1 044, 840 et 726 ménages. Afin de contrôler la taille du problème et les résultats attendus, l'approche suivante est adoptée :

— dans une première étape, nous sélectionnons aléatoirement un nombre de ménages dans chaque carreau. Par exemple, nous choisissons X_1 ménages parmi les 1 332 ménages que compte le carreau 1 ($X_1 \leq 1\,332$), X_2 ménages parmi les 540 ménages du carreau 2 ($X_2 \leq 540$) et ainsi de suite pour les autres carreaux. A la fin du processus, nous obtenons un iris simulé de X ménages synthétiques (avec $X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$) ;

— dans une deuxième étape, nous calculons les données agrégées des carreaux pour les X_i ménages sélectionnées dans la première étape. Les données agrégées d'un

carreau sont obtenues en additionnant les valeurs des attributs des ménages appartenant à ce carreau. Ainsi, les données agrégées du carreau 1 sont obtenues en additionnant les valeurs des attributs des X_1 ménages (y compris le niveau de vie) ;

— une fois les données agrégées des carreaux calculées, nous supprimons (dans une troisième étape) le carreau d'appartenance du ménage afin d'obtenir la même configuration que celle des données réelles. Nous cherchons alors à réaffecter les ménages dans leur carreaux d'origine.

Grâce au jeu de données et à la procédure de construction appliquée, tous les paramètres du modèle sont contrôlés. Nous avons également la certitude que le niveau de vie d'un carreau correspond parfaitement à la somme des niveaux de vie des ménages qui le composent. Le minimum de la fonction objectif est donc de zéro. Dans le cas réel, le minimum n'est pas égal à zéro car la somme des niveaux de vie des ménages du carreau n'est pas égale au niveau du vie du carreau (le revenu étant un attribut estimé au niveau de la population synthétique). A partir de données simulées qui correspondent exactement sur les deux niveaux, il est possible de tester la précision des résultats lorsque l'algorithme de résolution n'est pas exact. Ceci est particulièrement important car les algorithmes efficaces en termes de calcul ne sont pas des algorithmes exacts. A travers les X ménages ($X = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$) à affecter, il est également possible de faire varier la taille des ménages à répartir dans les carreaux. Cela donne l'opportunité de tester les performances de calcul et déceler les limites des algorithmes de résolution. Le Tableau 5.3 présente les marginaux des carreaux calculés en prenant en compte l'ensemble des 6 000 ménages ($X_1 = 1\ 332$, $X_2 = 540, \ldots, X_6 = 726$) ;

Tableau 5.3 : Marginaux des carreaux calculés à partir des attributs des 6 000 ménages

Paramètres	Carreaux (z)					
	Carreau 1	Carreau 2	Carreau 3	Carreau 4	Carreau 5	Carreau 6
n_H	1 332	540	1 518	1 044	840	726
TR	61 083 330	33 507 367	66 008 163	50 128 765	46 573 815	24 869 918
TR/n_H	45 858,35	62 050,68	43 483,64	48 016,06	55 445,02	34 256,085
NF_1	333	67	269	247	107	134
NT_2	573	277	668	553	487	249
NT_5	174	54	58	107	43	6
NB_1	53	341	474	40	461	143
NC_5	275	33	189	226	79	58
NG_1	23	282	174	47	448	75
ND_1	20	226	14	29	50	59
ND_3	258	103	628	155	129	154
NE_1	1233	28	769	986	241	332

Note : n_H : nombre de ménages ; TR : niveau de vie total ; TR/n_H : niveau de vie moyen ; NF_1 : nombre de ménages pauvres ; NT_2 : nombre de ménages entre 2 et 4 personnes ; NT_5 :nombre de ménages de 5 personnes ou plus ; NB_1 : nombre de ménages propriétaires ; NC_5 : nombre de ménages monoparentaux ; NG_1 : nombre de ménages en maison ; ND_1 : nombre de logements construits avant 1945 ; ND_3 : nombre de logements construits à partir de 1990 ; NE_1 : nombre de logements sociaux.

5.4.2 Solveur de programmation quadratique mixte en nombres entiers (MIQP solveur)

Forme matricielle du MIQP problème

L'utilisation d'un MIQP solveur implique d'écrire la fonction objectif et les équations (équations 5.1 à 5.13) sous une forme matricielle (Bonami et al., 2018) :

$$\min \quad \frac{1}{2}x^T Q x + c^T x \tag{5.14}$$

$$Ax = b \tag{5.15}$$

$$x_{zh} \in \{0,1\}, \quad z \in \mathbf{Z}, h \in \mathbf{H} \tag{5.16}$$

— l'équation 5.14 traduit la forme matricielle de la fonction objectif quadratique de l'équation 5.1, avec $c \in \mathbb{R}^n$ et Q (une matrice réelle symétrique $n \times n$) ;

— l'équation 5.15 traduit toutes les contraintes linéaires présentes dans le problème (équations 5.2 à 5.12), avec $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$ qui représente le terme de droite des contraintes ;

— l'équation 5.16 précise que les variables de décision sont binaires.

Après avoir modélisé le problème sous une forme matricielle en spécifiant les quatre composantes Q, c, A, b, un solveur dédié à l'implémentation du MIQP problème est utilisé. Il s'agit du solveur ILOG CPLEX Optimization Studio, logiciel commercial permettant de modéliser et de résoudre des problèmes d'optimisation (Laborie et al., 2018). CPLEX est gratuit pour un usage académique. La procédure d'affectation des ménages est exécutée avec le logiciel et le langage de programmation R (version 4.0.3) et ILOG CPLEX Optimization Studio (version 20.1.0), sur une machine Windows 10 professionnel, Intel(R) Core (TM) i7-8665U CPU @ 1.90GHz, 2.11 GHz et 16 GB de RAM. L'association de CPLEX avec R est réalisée grâce à la librairie « Rcplex »(Bravo et al.,2016) [6], qui est une interface de R avec CPLEX et disponible sur le « Comprehensive R Archive Network »(CRAN).

Résultats

Nous faisons progressivement varier le nombre de ménages à affecter. Quel que soit le nombre de ménages considéré, une quinzaine de simulations est effectuée pour apprécier le temps de calcul. Le solveur a pu calculer la solution optimale (minimum de la fonction objectif=0) pour les tailles des ménages suivants :

— pour une taille $X = 30$ ménages ($X_1 = 7, X_2 = 4, X_3 = 6, X_4 = 3, X_5 = 6, X_6 = 4$),

6. Disponible à <`https://cran.r-project.org/web/packages/Rcplex/index.html`>, page consultée le 17 août 2021.

le solveur a pu résoudre le MIQP problème en moins d'une seconde ;

— pour $X = 60$ ménages ($X_1 = 17, X_2 = 8, X_3 = 14, X_4 = 5, X_5 = 12, X_6 = 4$), le temps de résolution varie entre 1,4 minutes et 2 heures ;

— à partir de $X = 120$ ménages ($X_1 = 32, X_2 = 17, X_3 = 21, X_4 = 15, X_5 = 25, X_6 = 10$), le temps de résolution dépasse cinq heures.

L'affectation d'un nombre important de ménages ne peut être réalisée en un temps raisonnable avec un solveur MIQP. Une heuristique permettant d'obtenir une solution pratique et rapide est proposée.

5.4.3 Heuristique proposée pour une affectation d'un nombre important de ménages

Relaxation de la programmation mixte à termes quadratiques

Dans l'heuristique proposée, les équations 5.14 et 5.15 sont conservées :

$$\begin{aligned} \min \quad & \frac{1}{2} x^T Q x + c^T x \\ & Ax = b \end{aligned}$$

En revanche, l'équation 5.16 (qui spécifie que les variables de décision sont binaires)

$$x_{zh} \in \{0, 1\}, \quad z \in \mathbf{Z}, h \in \mathbf{H}$$

est remplacée par les équations 5.17 et 5.18 :

$$x_{zh} \in [0, 1] \quad z \in \mathbf{Z}, h \in \mathbf{H} \tag{5.17}$$

$$\sum_{h \in \mathbf{H}} x_{zh} = 1, \quad z \in \mathbf{Z} \tag{5.18}$$

La contrainte binaire appliquée sur les variables de décision x_{zh} permettait de s'assurer qu'un ménage h ne peut être affecté qu'à un carreau. Dans l'heuristique, nous

estimons des probabilités d'appartenance d'un ménage h à chaque carreau en fonction du niveau de vie et des autres attributs considérés. Le MIQP problème est ainsi transformé en un problème de programmation quadratique (QP) problème (Gill and Wong, 2015), plus facile à résoudre.

Résultats

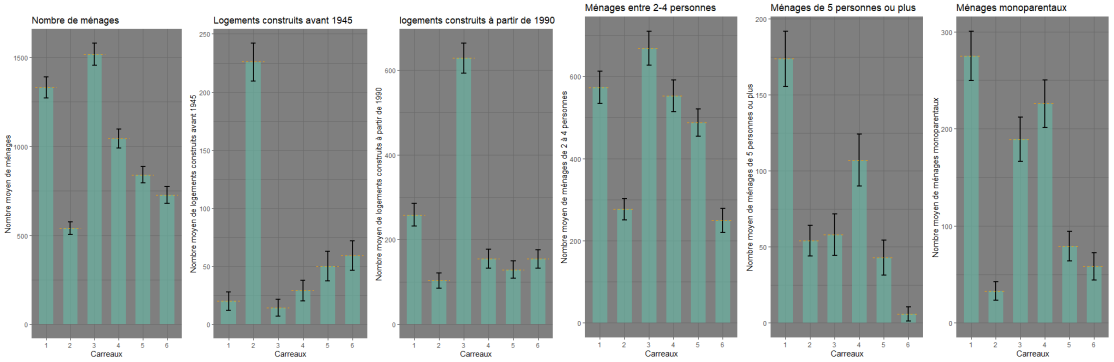
Le temps de résolution du QP problème pour les 6 000 ménages est de 43 secondes. A la fin du processus de résolution, nous obtenons 6 probabilités pour chaque ménage, chaque probabilité représentant la chance d'un ménage d'être affecté à un carreau. Chaque ménage est ensuite affecté à un carreau en fonction de ces probabilités. L'heuristique proposée ne donne toutefois pas de solutions exactes comme dans le MIQP problème (il est possible d'avoir plusieurs probabilités non nulles pour un ménage), les résultats doivent donc être examinés avec attention. Nous avons alors procédé à 10 000 tirages différents. Afin d'évaluer la performance de l'heuristique proposée, les valeurs moyennes des marginaux des carreaux issus des 10 000 tirages (Tableau 5.4) sont comparées avec les valeurs réelles des marginaux des carreaux (Tableau 5.3).

Tableau 5.4 : Valeurs moyennes des marginaux des carreaux obtenues après 10 000 tirages

Paramètres	Carreaux (z)					
	Carreau 1	Carreau 2	Carreau 3	Carreau 4	Carreau 5	Carreau 6
n_H	1332	540	1518	1044	840	726
TR	61 095 905	33 481 980	66 019 380	50 110 860	46 586 833	24 876 401
TR/n_H	45 859.75	62 042.82	43 484.25	48 021.23	55 449.09	34 256.56
NF_1	333	67	269	247	107	134
NT_2	573	277	668	553	487	249
NT_5	174	54	58	107	43	6
NB_1	53	341	474	40	461	143
NC_5	275	33	189	226	79	58
NG_1	23	282	174	47	448	75
ND_1	20	226	14	29	50	59
ND_3	258	103	628	155	129	154
NE_1	1233	28	769	986	241	332

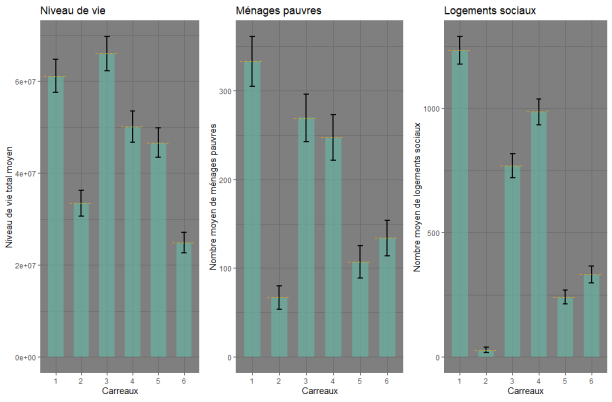
Les Tableaux 5.3 et 5.4 sont quasiment identiques, ce qui signifie qu'il n'y a pas de biais dans le processus d'estimation. La Figure 5.3 présente les valeurs moyennes obtenues après 10 000 tirages pour chaque attribut et chaque carreau.

FIGURE 5.3 : Histogrammes et barres d'erreurs des intervalles de confiance à 95% des valeurs moyennes des marginaux des carreaux obtenues après 10 000 tirages

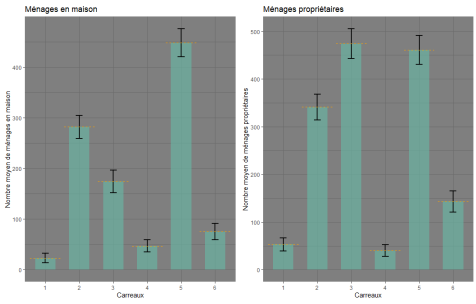


(a) Taille du ménage et type de logement

(b) Structure familiale



(c) Conditions économiques



(d) Propriété

Les traits en pointillées de cette Figure 5.3 représentent les valeurs réelles des marginaux des carreaux (indiquées dans le Tableau 5.3). Les valeurs moyennes et les valeurs réelles sont identiques ; pour chaque carreau et chaque attribut, la valeur réelle appartient toujours à l'intervalle de confiance de 95%. La longueur des intervalles de confiance est également faible. Cela implique qu'une grande proportion des populations synthé-

tiques affectées ont des valeurs proches des valeurs réelles.

5.5 Discussion

5.5.1 Sélection d'une population synthétique

Les résultats suggèrent que l'heuristique proposée produit des résultats globaux, cohérents et très proches des données réelles. Toutefois, l'utilisateur a besoin d'une population synthétique unique affectée à des carreaux. Un exemple de méthode de décision est présentée. Cette méthode permet de sélectionner une seule population synthétique affectée dans des carreaux, parmi les 10 000 populations qui ont été générées :

— Une tolérance, notée tol, est définie autour de l'attribut d'intérêt. Dans le cas d'étude, une tolérance de 5% est fixée autour de l'attribut niveau de vie noté TR. Seules les populations synthétiques dont le niveau de vie du carreau appartient à l'intervalle $[(1-tol)\times TR, (1+tol)\times TR]$ sont conservées. En procédant ainsi, un premier groupe de populations synthétiques est sélectionné ;

— un critère $\mu = \sum_{z\in \mathbf{Z}} |\hat{A}_z - A_z|$ est défini avec $\hat{A}_z$, le marginal simulé obtenu et A_z, le marginal réel. Le critère μ est calculé pour chaque population synthétique ayant passé le filtre de l'étape précédente. Ce critère correspond à la somme de la valeur absolue des différences entre les données agrégées obtenues après affectation des ménages dans les carreaux et les données agrégées réelles des carreaux ;

— la population synthétique qui possède le critère μ minimal est choisie.

5.5.2 Évaluation de l'heuristique

Une population synthétique répartie dans des carreaux a été sélectionnée selon les critères définis dans la méthode de décision. Le Tableau 5.5 présente les différences absolues entre les données agrégées des carreaux de cette population et les données agrégées réelles des carreaux. De très petites différences ont été trouvées. Les différences les plus importantes sont de 25 ménages (NB_1, carreau 5), 19 ménages (NB_1, carreau

2), 14 ménages (n_H, carreau 3). Cependant, ces différences doivent être relativisées car elles concernent des attributs qui sont de taille importante respectivement 461 ménages, 341 ménages, et 1 518 ménages.

Tableau 5.5 : Différences absolues entre les marginaux des carreaux de la population synthétique qui possède le critère μ minimal et les marginaux initiaux des carreaux

Paramètres	Carreaux (z)					
	Carreau 1	Carreau 2	Carreau 3	Carreau 4	Carreau 5	Carreau 6
n_H	3	10	14	3	0	2
NF_1	5	4	4	8	8	5
NT_2	2	0	0	3	3	4
NT_5	14	0	3	8	1	4
NB_1	3	19	2	5	25	6
NC_5	3	2	1	7	7	2
NG_1	3	4	9	14	11	9
ND_1	4	4	6	6	7	7
ND_3	2	4	4	3	7	14
NE_1	5	2	9	6	10	0

Conclusion

L'affectation des ménages ou individus synthétiques à une zone spatiale plus fine est très importante dans une simulation multi-agents notamment dans l'analyse et la modélisation des questions relatives à la gestion des systèmes environnementaux et urbains. Dans ce chapitre, l'affectation d'une population synthétique d'une zone statistique conteneur (ZSC) vers plusieurs zones statistiques élémentaires (ZSE) est étudiée. Dans les ZSC, sont diffusées des données adéquates pour la génération de la population. Dans les ZSE, seules des données agrégées sont disponibles. Les contours des ZSE sont bien définis et ne se chevauchent pas. Ce processus de désagrégation spatiale (passage d'une zone géographique supérieure à une zone géographique inférieure) doit être réalisé à partir de deux types d'attributs : 1) des attributs communs disponibles dans les deux ni-

veaux et 2) des attributs supplémentaires d'intérêt disponibles à un niveau mais pouvant être estimés dans l'autre niveau à partir de sources de données externes. Les instituts nationaux de statistique mettent souvent à disposition du public de telles données.

Dans l'exemple d'application, nous utilisons des données françaises pour affecter des ménages synthétiques du niveau IRIS (ZSC) vers les niveaux carreaux (ZSE). Le problème d'affectation a été modélisé dans une première approche, comme un problème de programmation quadratique mixte en nombres entiers (MIQP) mono-objectif. Il s'agissait d'affecter les ménages synthétiques dans des carreaux, en minimisant l'écart entre la valeur agrégée de l'attribut d'intérêt (estimée au niveau de l'IRIS) et la valeur agrégée réelle diffusée dans le carreau. Les attributs présents à la fois dans l'IRIS et dans le carreau (attributs communs) sont utilisées comme contraintes du MIQP problème. L'attribut d'intérêt retenu était le niveau de vie du ménage et les attributs communs utilisés comprenaient notamment la taille du ménage, la structure familiale, le type de logement.... Un MIQP solveur a été appliqué sur un jeu de données simulées représentatif de la configuration des données réelles : 6 000 ménages synthétiques d'un IRIS de la commune de Nantes à allouer à 6 carreaux. Cette méthode de résolution ne pouvait pas être utilisé pour affecter un nombre important de ménages en un temps de calcul raisonnable.

Une heuristique modélisant le problème comme un problème de programmation quadratique (QP) a été développée dans une seconde approche. Cette heuristique a permis d'obtenir pour la totalité des ménages du jeu de données, une résolution efficace avec un temps de calcul rapide et des résultats proches des données réelles. La méthodologie et l'heuristique proposée restent génériques et peuvent être utilisées pour allouer une population synthétique dans un large éventail de cas.

Références bibliographiques

Anderson, W., Guikema, S., Zaitchik, B., and Pan, W. (2014). Methods for estimating population density in data-limited areas : Evaluating regression and tree-based models in peru. *PloS one*, 9(7) :e100037.

Bartha, G. and Kocsis, S. (2011). Standardization of geographic data : The european inspire directive. *European Journal of Geography*, 2(2) :79–89.

Billionnet, A. and Elloumi, S. (2007). Using a mixed integer quadratic programming solver for the unconstrained quadratic 0-1 problem. *Mathematical Programming*, 109(1) :55–68.

Bonami, P., Lodi, A., and Zarpellon, G. (2018). Learning a classification of mixed-integer quadratic programming problems. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 595–604. Springer.

Bösch, P. M., Müller, K., and Ciari, F. (2016). The ivt 2015 baseline scenario. In *16th Swiss Transport Research Conference (STRC 2016)*. 16th Swiss Transport Research Conference (STRC 2016).

Chapuis, K., Taillandier, P., Renaud, M., and Drogoul, A. (2018). Gen* : a generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science*, 32 :1194–1210.

Darriau, V. (2020). Les données carroyées, des outils et méthodes innovants pour percevoir la réalité des territoires. *Courrier des statistiques*, (5) :53–73. Retrievied 13 April 2021 : `https://www.insee.fr/fr/information/5008679?sommaire=5008710`.

Del Pia, A., Dey, S. S., and Molinaro, M. (2017). Mixed-integer quadratic programming is in np. *Mathematical Programming*, 162(1) :225–240.

Freuder, E. C. and Mackworth, A. K. (2006). Constraint satisfaction : An emerging paradigm. In *Handbook of constraint programming*, volume 2, pages 13–27. Elsevier.

Gill, P. E. and Wong, E. (2015). Methods for convex and general quadratic programming. *Mathematical Programming Computation*, 7(1) :71–112.

Hackl, J. and Dubernet, T. (2019). Epidemic spreading in urban areas using agent-based transportation models. *Future Internet*, 11(4) :92.

Harada, T. and Murata, T. (2017). Projecting households of synthetic population on buildings using fundamental geospatial data. *SICE Journal of Control, Measurement, and System Integration*, 10(6) :505–512.

Ji, Z. and Wan, Y. (2021). A novel method for socioeconomic data spatialization. *Spatial Statistics*, page 100501.

Laborie, P., Rogerie, J., Shaw, P., and Vilím, P. (2018). Ibm ilog cp optimizer for scheduling. *Constraints*, 23(2) :210–250.

Long, Y. and Shen, Z. (2015). Population spatialization and synthesis with open data. In *Geospatial Analysis to Support Urban Planning in Beijing*, pages 115–131. Springer.

Rossi, F., van Beek, P., and Walsh, T. (2008). Chapter 4 constraint programming. In van Harmelen, F., Lifschitz, V., and Porter, B., editors, *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 181–211. Elsevier.

Sallard, A., Balać, M., and Hörl, S. (2020). A synthetic population for the greater são paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, 1545.

Su, M.-D., Lin, M.-C., Hsieh, H.-I., Tsai, B.-W., and Lin, C.-H. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *Science of the Total Environment*, 408(20) :4807–4816.

Thomson, D. R., Kools, L., and Jochem, W. C. (2018). Linking synthetic populations to household geolocations : a demonstration in namibia. *Data*, 3(3) :30.

Zhu, Y. and Ferreira Jr, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record*, 2429(1) :168–177.

Conclusion Générale

La quantification et la réduction des expositions aux nuisances environnementales générées par les systèmes de transport, en particulier le bruit et les polluants atmosphériques sont des questions centrales. L'approche par les systèmes multi-agents, contrairement aux modèles à quatre étapes permet de mesurer une exposition dynamique des individus aux externalités (bruit ou pollution) associées aux moyens de transport, i.e. une exposition qui rend compte des mobilités et des changements de localisation vécus par les individus au cours de la journée.

La modélisation des mobilités est appréhendée par une classe spécifique de modèles multi-agents désignée sous le terme de modèles multi-agents basés sur les activités. Ces modèles analysent le comportement des individus lorsqu'ils interagissent avec d'autres individus dans la réalisation de leurs activités quotidiennes ainsi que les implications de ces interactions dans le réseau de transport. La mesure des externalités environnementales consiste ensuite en un couplage des modèles multi-agents basés sur les activités avec des modèles de bruit ou d'émission qui permettent d'estimer et de représenter la répartition spatio-temporelle des niveaux de bruit ou de concentrations en polluants atmosphériques. Il est par exemple possible de quantifier le nombre d'individus exposés en un endroit donné, à un moment précis de la journée et d'identifier les activités pour lesquelles l'exposition est la plus importante.

Le premier chapitre est consacré à la présentation des modèles multi-agents basés sur les activités et à la description de trois plateformes logiciels issues de cette classe de modèles : MATSim, TRANSIMS et SimMobility. L'association des modèles multi-agents

basés sur les activités avec des modèles de bruit ou d'émission pour l'évaluation de l'exposition dynamique des individus aux externalités environnementales est également mentionnée. Les modèles multi-agents basés sur les activités reposent sur une description assez fine des individus, de leurs ménages et de leurs activités. Cette approche totalement désagrégée nécessite la création d'une population synthétique d'individus et de ménages, représentative de la population réelle. Cette population synthétique est générée à partir d'un ensemble de méthodes qui utilisent les sources de données disponibles (recensement, enquêtes...). Une telle population synthétique est appelée population synthétique à deux niveaux : niveau individuel et niveau ménage. La population synthétique a fait l'objet d'une production scientifique importante avec le développement de méthodes, de logiciels commerciaux ou open source. L'analyse de ces méthodes indiquent plusieurs voies d'amélioration. Il s'agit notamment 1) de la réplicabilité des méthodes, 2) de l'affectation aléatoire de certains attributs sociodémographiques et 3) de l'allocation spatiale des résidences qui s'effectue souvent à une échelle géographique qui n'offre pas une précision spatiale suffisante. Les chapitres de cette thèse contribuent à ces trois voies d'améliorations.

Le deuxième chapitre a consisté en une revue et une comparaison des principales méthodes de génération de population synthétique à deux niveaux. La littérature scientifique recense trois principales catégories de méthodes de génération. Ces catégories sont la reconstruction synthétique (SR), l'optimisation combinatoire (CO) et l'apprentissage statistique (SL). Il est alors important de savoir quelle méthode privilégier en fonction des informations que l'utilisatrice ou l'utilisateur possède. La comparaison à partir d'un certain nombre de critères (caractéristiques de l'algorithme, données d'entrée nécessaires, sorties obtenues) a permis d'illustrer les avantages et les inconvénients d'utilisation de chaque méthode. La principale contribution de ce chapitre est l'élaboration d'un arbre de décision. A travers cet arbre, les chercheurs et les professionnels ont désormais accès à un outil de décision standardisé et complet pour choisir la méthode de génération appropriée à leurs données et leurs objectifs de modélisation. Ainsi, ce chapitre améliore la réplicabilité des méthodes existantes.

La plupart des instituts nationaux de statistique diffusent à la fois des données désagrégées qui représentent un échantillon de la population totale avec des attributs individuels et de ménages et des données agrégées qui traduisent les distributions marginales de ces attributs individuels et de ménages. Les données françaises de ce type (échantillon et données agrégées) diffusées par l'Insee proviennent du recensement de la population. Une particularité de la France est la mise à disposition d'un échantillon d'individus et de ménages qui correspond à 30% de la population totale des individus et des ménages contre souvent moins de 5% dans les autres pays. Ces sources de données sont diffusées à l'échelle des communes ou des Ilôts Regroupés pour l'Information Statistique (IRIS). Il est ainsi possible de générer une population synthétique d'individus et de ménages à ces deux niveaux.

En prenant en compte la configuration des données françaises et l'arbre de décision, les méthodes de la catégorie SR étaient les plus adaptées à la génération d'une population synthétique à deux niveaux. Dans le troisième chapitre j'ai comparé et testé quatre algorithmes de reconstruction synthétique : l'Iterative Proportional Update (IPU), le Hierarchical Iterative Proportional Fitting (HIPF), la minimisation de l'entropie (ent) et le calage sur marges (GR). Ces quatre algorithmes ont été associés à deux méthodes de conversion des poids décimaux en poids entiers à savoir l'approche des probabilités proportionnelles (PP) et l'approche TRS (Truncate Replicate Sample). La comparaison des algorithmes et des méthodes s'est effectuée sur la base d'un cadre conceptuel commun avec une harmonisation des notations et une description détaillée de chaque algorithme.

A partir de données agrégées et d'un échantillon constitué d'environ 294 000 individus et 136 000 ménages répartis dans 307 communes et IRIS, j'ai généré (pour chaque algorithme associé à chaque méthode de conversion) la population de l'aire urbaine estimée à environ 949 000 individus et 418 000 ménages. Les populations synthétiques obtenues ont été validées à l'aide de différents indicateurs : R^2, TAE, SAE, SRMSE et approche de Bland-Altman. Les résultats de la validation indiquent que toutes les méthodes considérées génèrent une population synthétique à deux niveaux dont les caractéristiques sont proches des données agrégées du recensement. Sur la base de ces indicateurs, la minimi-

sation de l'entropie et le Hierarchical Iterative Proportional Fitting se sont avérées être les méthodes les plus efficaces ; également l'approche TRS est plus performante que celle des probabilités proportionnelles.

La diffusion des fichiers du recensement français suit un format standard à l'échelle de la France métropolitaine. Bien qu'appliquée à l'aire urbaine de Nantes, la méthodologie développée dans ce chapitre (formalisation mathématique, implémentation et validation des algorithmes de reconstruction statistique dans un cadre d'analyse harmonisé) peut être généralisée et adaptée sans grand efforts à d'autres zones géographiques du territoire français. Le cadre d'analyse proposé permet également de mieux positionner les algorithmes utilisés par rapport aux autres techniques existantes et peut éventuellement stimuler le développement de nouvelles techniques.

Une procédure habituelle consiste à rajouter de nouveaux attributs aux individus ou aux ménages, une fois la population synthétique générée. Dans la modélisation multi-agents basée sur les activités, les attributs rajoutés concernent par exemple des plans d'activités et de déplacements. Ces attributs supplémentaires proviennent de sources externes qui sont dans leur grande majorité des échantillons d'enquêtes. Il arrive souvent que les attributs supplémentaires à rajouter ne soient disponibles que sous la forme de distributions agrégées. Les méthodologies d'enrichissement de la population synthétique d'un attribut provenant d'une source externe de données agrégées sont peu présentes dans la littérature.

Dans le quatrième chapitre, j'ai développé une méthodologie qui offre la possibilité de rajouter des attributs à une population synthétique à partir de données agrégées, de telle sorte que les associations observées entre les attributs de la population synthétique (par exemple, la taille et le type de ménage) et les nouveaux attributs soient conservés. Dans un souci de validation de la démarche méthodologique, j'ai affecté un revenu (plus précisément un niveau de vie) aux ménages synthétiques de la commune de Nantes en utilisant une source de données agrégées dénommée « Fichier Localisé Social et Fiscal » (FiLoSoFi). Le choix du revenu se justifie en raison de l'importance de cet attribut dans la prise en compte de nombreux aspects sociaux et économiques (pouvoir d'achat

des ménages, politique de redistribution, politique fiscale...).

La méthodologie intègre trois étapes distinctes. La première étape a consisté à modéliser théoriquement l'ajout du niveau de vie comme un problème de maximisation de l'entropie en exploitant les attributs disponibles dans le recensement et dans le dispositif FiLoSoFi. La résolution de ce problème (dans ce cas d'étude spécifique) n'était pas possible en raison du grand nombre de contraintes impliquées. Dans la deuxième étape, une heuristique qui combine le théorème de Bayes avec l'algorithme de minimisation de l'entropie croisée est proposée. En raison du nombre élevé de paramètres, certains des résultats obtenus se sont avérés non cohérents. Pour remédier à ce problème, une méthode de post-traitement est appliquée au cours d'une troisième étape pour garantir la cohérence des résultats.

Les résultats obtenus suggèrent que la méthodologie développée produit des résultats proches des données agrégées du dispositif FiLoSoFi. Toutefois, des différences plus importantes ont été constatées pour certaines catégories de ménage ainsi que pour les déciles les plus élevés (huitième et neuvième déciles). Ces écarts sont principalement dus aux différences méthodologiques entre le recensement et le dispositif FiLoSoFi et aux hypothèses que nous avons formulées. La méthodologie formalisée reste générale et peut être modifiée pour intégrer différents types de données agrégées. Ce chapitre répond ainsi à l'amélioration de l'ajout d'attribut à une population synthétique existante.

La localisation des agents synthétiques (individus ou ménages) à une échelle spatiale fine est importante dans une simulation multi-agents basée sur les activités. Dans de nombreuses situations, les lieux de résidence ou de travail des agents synthétiques sont tirés au sort parmi tous les emplacements disponibles. Dans le cinquième et dernier chapitre de la thèse, une méthodologie générique d'affectation d'une population synthétique d'une échelle spatiale supérieure (désignée par le terme de zone statistique conteneur -ZSC-) à une échelle spatiale plus fine (désignée par zone statistique élémentaire -ZSE-) est proposée. Dans l'exemple d'application, des ménages synthétiques du niveau IRIS ou communes (pour les communes non irisées) sont affectés à des carreaux. Les IRIS correspondaient aux ZSC et les carreaux représentaient les ZSE. Les données carroyées

françaises sont diffusées par l'Insee, à partir du dispositif FiLoSoFi. Ces données permettent de disposer de données marginales sur les ménages à des échelles fines (carreaux de 200m de côté par exemple).

Le problème d'affectation a été modélisé dans une première approche, comme un problème de programmation quadratique mixte en nombres entiers (MIQP) mono-objectif. Il s'agit d'affecter les ménages synthétiques dans des carreaux, en minimisant l'écart entre la valeur agrégée du niveau de vie (attribut d'intérêt) estimé au niveau de l'IRIS et la valeur agrégée réelle du niveau de vie diffusé dans le carreau. Les attributs présents à la fois dans l'IRIS et dans le carreau (attributs communs) sont utilisés comme contraintes du MIQP problème. Un solveur de MIQP a été appliqué sur un jeu de données simulées représentatif de la configuration des données réelles : 6 000 ménages synthétiques d'un IRIS de la commune de Nantes à allouer à 6 carreaux. Cette méthode de résolution ne pouvait être utilisée pour affecter un nombre important de ménages en un temps de calcul raisonnable. Le problème a alors été modélisé comme un problème de programmation quadratique (QP) dans une deuxième approche. Cela a permis d'obtenir pour la totalité des ménages du jeu de données, une résolution efficace avec un temps de calcul rapide et des résultats proches des données réelles. La méthodologie et l'heuristique proposées restent génériques et peuvent être utilisées pour allouer une population synthétique dans un large éventail de cas.

L'ensemble des données utilisées dans le cadre de la thèse sont librement accessibles. Les méthodologies mises en œuvre restent génériques et transposables à des sources de données non françaises. Un soin particulier a donc été appliqué au traitement des données afin de faciliter l'appropriation et l'adaptation des méthodologies. Les algorithmes développés dans la thèse ont été implémentés avec le langage R (logiciel libre) et les différents scripts associés seront mis à la disposition des utilisateurs. Certains aspects abordés dans cette thèse méritent toutefois d'être complétées par de nouvelles analyses et hypothèses.

Combiner différentes sources de données peut s'avérer délicat notamment s'il existe des différences méthodologiques entre les sources. C'est le cas par exemple du recense-

ment français et du dispositif FiLoSoFi avec une définition du ménage qui diffère. Dans FiLoSoFi, un ménage (fiscal) regroupe tous les individus figurant sur une même déclaration de revenus. Une personne peut être affiliée à un ménage du point de vue fiscal sans nécessairement résider dans ce ménage. En revanche, dans le recensement, seuls les résidents d'un ménage donné sont pris en compte. Le nombre, la taille et le type de ménages peuvent ainsi diverger dans les deux sources. La méthodologie d'ajout du revenu que nous avons développé intègre ces différences méthodologiques.

Dans le dispositif FiLoSoFi, les distributions du niveau de vie sont données en déciles (du premier au neuvième). Aucun revenu minimal ni maximal n'est mentionné ; également, il n'existait aucune information entre deux déciles de niveau de vie. Certaines hypothèses ont alors été formulées :

— $R_{min} = 0$ (le niveau de vie minimal a une valeur nulle) ;

— $R_{max} = D9 \times 1,5$ (pour chaque catégorie, le niveau de vie maximal est égal à 1,5 le neuvième décile) ;

— la distribution des niveaux de vie est linéaire ;

Ces hypothèses pourraient être améliorées. Il serait avantageux pour de prochaines études d'acquérir une meilleure connaissance a priori de la distribution des niveaux de vie et, en particulier, des niveaux de vie extrêmes en se référant notamment à la littérature économique ou à des personnes expertes du domaine. Cela permettrait d'affiner les analyses et de trouver une meilleure discrétisation des niveaux de vie afin de produire des intervalles de revenus plus larges. Ce changement entraînerait également moins de probabilités invalides.

Les différentes recherches menées dans ce travail de thèse laissent entrevoir des perspectives de recherche intéressantes.

La configuration des données du recensement français (échantillon de taille importante, présence de données agrégées) favorise l'utilisation des méthodes de reconstruction synthétique. Il est important de tester d'autres types de méthodes de génération en particulier les méthodes d'apprentissage statistique sur des échantillons de taille réduite

comme par exemple des données d'enquêtes. Les méthodes SL peuvent s'avérer utiles dans l'utilisation de différents types de données (enquêtes classiques, données GPS) pour générer une population synthétique à deux niveaux.

Plusieurs améliorations de la méthodologie de spatialisation sont envisagées dans des études futures. Une extension de l'heuristique proposée serait de considérer le cas où plusieurs attributs d'intérêt peuvent être calculés et utilisés pour l'allocation des ménages. Cela donnerait lieu à un problème d'optimisation à objectifs multiples.

Une autre piste intéressante à explorer est le cas où une Zone Statistique Elémentaire (ZSE) se retrouve sur plusieurs Zones Statistiques Conteneurs (ZSC). Dans ce cas de figure, les ZSE pourraient être divisées en plusieurs sous-unités correspondant aux ZSC dans lesquelles elle sont incluses. Ce processus impliquerait de calculer de nouvelles données agrégées pour ces sous-unités de ZSE. Il serait également intéressant d'appliquer la méthodologie de spatialisation à une commune de grande taille, comme par exemple la commune de Nantes et d'analyser les résultats obtenus.

A terme, le processus de génération de plans d'activités individuels, deuxième étape de l'implémentation d'un modèle multi-agents basé sur les activités sera analysé.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE NANTES

Titre : Méthodologie de calibration d'un modèle multimodal des déplacements pour l'évaluation des externalités environnementales à partir de données ouvertes (open data) : le cas de l'aire urbaine de Nantes

Mots-clés : modélisation multi-agents, population synthétique, microsimulation, entropie, spatialisation, MIQP

Résumé :

Ce travail s'articule autour de la thématique de la population synthétique, pierre angulaire de tout modèle multi-agents basé sur les activités. L'objectif de la thèse est de proposer des solutions permettant de pallier certaines insuffisances relevées dans la littérature sur ce sujet. Dans un premier temps, les principales méthodes de génération de population synthétique à deux niveaux (individus et ménages) sont présentées de façon détaillée. La recension de ces méthodes s'achève sur la proposition d'un arbre de décision qui permet de faire un choix raisonné entre les méthodes présentées. Une population synthétique d'individus répartis dans des ménages est générée dans un second temps en comparant différents algorithmes. Après génération de la population synthétique, une méthodologie d'attribution de caractéristiques supplémentaires à partir de données agrégées est développée. L'ajout d'une nouvelle caractéristique est formulé comme une maximisation de l'entropie en associant les attributs disponibles dans la population synthétique et les données agrégées. La validation de cette méthodologie a consisté à l'affectation d'un revenu à un nombre conséquent de ménages synthétiques. Enfin, une approche innovante de spatialisation d'une population synthétique à une échelle géographique plus fine est proposée. Le processus d'affectation est modélisé comme un problème de programmation quadratique mixte en nombres entiers. Bien qu'appliqués à des données françaises librement accessibles, la totalité des algorithmes implémentés restent génériques et transposables à d'autres sources de données.

Title: Calibration methodology for a multimodal travel model for the assement of environmental externalities based on open data: application in the urban area of Nantes

Keywords: agent-based modeling, synthetic population, microsimulation, entropy, spatialization, MIQP

Abstract:

This work focuses on the topic of synthetic population, which serves as the cornerstone of any activity-based model. The objective of this thesis is to propose solutions that overcome some of the shortcomings found in the literature on this subject. First of all, the main methods for generating a two-layered synthetic population (individuals and households) are presented in detail. The review of these methods is completed by proposing a decision tree that allows for a reasoned choice between the various methods presented. A synthetic population of individuals distributed in households is generated during a second step by comparing the different algorithms. Afterwards, a methodology is developed for assigning additional characteristics from the aggregated data. The addition of a new feature is formulated as an entropy maximization problem by combining the attributes available in the synthetic population with the aggregated data. The methodology validation step consists of assigning an income to a large number of synthetic households. Lastly, an innovative approach to spatialize a synthetic population at a finer geographical scale is proposed. This assignment process is modeled as a mixed-integer quadratic programming problem. Although applied here to freely available French data, all the algorithms implemented remain generic and transposable to other data sources.