

Titre: Synthèse spatialisée de la population pour fins de modélisation des
Title: comportements de mobilité

Auteur: Mohamed Khachman
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Khachman, M. (2023). Synthèse spatialisée de la population pour fins de
Citation: modélisation des comportements de mobilité [Thèse de doctorat, Polytechnique
Montréal]. PolyPublie. <https://publications.polymtl.ca/57064/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/57064/>
PolyPublie URL:

**Directeurs de
recherche:** Catherine Morency, & Francesco Ciari
Advisors:

Programme: Génie civil
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Synthèse spatialisée de la population pour fins de modélisation des
comportements de mobilité**

MOHAMED KHACHMAN

Département des génies civil, géologique et des mines

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie civil

Décembre 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Synthèse spatialisée de la population pour fins de modélisation des comportements de mobilité

présentée par **Mohamed KHACHMAN**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Martin TRÉPANIÉ, président

Catherine MORENCY, membre et directrice de recherche

Francesco CIARI, membre et codirecteur de recherche

Antoine LEGRAIN, membre

Moez KILANI, membre externe

DÉDICACE

À l'Essence de mon existence

À celui que j'attends avec tant d'impatience

À ceux qui ont arrosé les fleurs de nos vies

À celle qui a fait l'homme que je suis [Fadile Fayad]

À celui qui m'a donné tout ce qu'il n'a guère [Hussein Khachman]

À celui qui a toujours couvert mes arrières [Ali Khachman]

À celle qui partage épreuves et bonheurs [Zeinab Khanafer]

À celui pour qui longues étaient les heures [Mohamad Mahdi Khachman]

À celles qui ont toujours supporté leur frère [Dalia et Samar Khachman]

À ceux qui m'ont reçu portes et bras ouverts [Ahmed Hodroge, Bahie Zorkot et Khalil Charara]

REMERCIEMENTS

Je voudrais d’abord remercier ma directrice de recherche Catherine Morency et mon codirecteur de recherche Francesco Ciari dont les contributions étaient essentielles au bon déroulement de ce projet. Grâce à vous, j’ai pu développer des compétences sur les plans scientifique et humain qui m’ont aidé à mener à termes mon parcours doctoral, et continueront à me servir au-delà. Je te remercie Catherine pour m’avoir transmis ta passion et ton énergie, et Francesco pour ton écoute et ton support dans les moments difficiles. Vous m’avez ouvert les portes de la recherche scientifique et j’espère pouvoir continuer à contribuer à l’avancement des connaissances.

Je voudrais aussi remercier la Direction de la Modélisation des Systèmes de Transport (DMST) du ministère des Transports et de la Mobilité Durable (MTMD) qui m’a fourni des outils et des connaissances nécessaires à la réalisation de ce projet de recherche. Merci à Louiselle Sioui et Brigitte Saint-Pierre (DMST) pour leur soutien indéniable.

Merci à Simon Vallée, Vanessa Skelton et Nathalie Schaillée (GHD Consultants Ltée.), et à Dominic Milette et Sylvie Gervais (Kiewit Engineering Group Canada ULC) pour leur flexibilité et leur support qui m’ont permis de concilier famille, études et travail.

Je salue enfin ma famille, mes amis et mes collègues qui ont contribué par une action, un mot, ou une prière à ce que mon passage à Polytechnique Montréal soit couronné de succès.

RÉSUMÉ

De l'identification du besoin à la mise en service d'un projet d'infrastructure de transport, plusieurs années sont souvent requises. De plus, un projet d'infrastructure de transport requiert souvent le déblocage d'investissements considérables. Or, la demande en transport évolue rapidement en nombre comme en nature et les fonds disponibles à investir dans une offre de mobilité sont limités. Une planification préalable menant à une action par anticipation s'impose alors afin de s'assurer de maintenir une offre adaptée et de garantir la pertinence et la rentabilité des investissements. En effet, dans le présent comme dans le futur, la planification de l'offre en transport est basée sur les enjeux de mobilité généralement identifiés par le biais de modèles reflétant le comportement de mobilité de la population.

Les comportements de mobilité des individus sont généralement fortement influencés par leurs caractéristiques sociodémographiques, celles de leurs ménages, et leurs localisations résidentielles. Ainsi, pour alimenter les plateformes de microsimulation à grande échelle basées agents de plus en plus populaires en modélisation des comportements de mobilité, une liste totalement énumérée, caractérisée et spatialisée des ménages et individus composant la population de la zone d'étude est requise. Une telle base de données peut être constituée par compilation des formulaires du recensement complétés par les citoyens de la zone. Or, cette compilation demeure inaccessible aux planificateurs pour des raisons relevant essentiellement de la protection de la vie privée. Les instituts nationaux de la statistique publient généralement deux types de données dérivées des formulaires détaillés qu'ils reçoivent : des tabulations agrégées d'une variété de variables à différents découpages zonaux, aussi appelées résumés du recensement, et un échantillon de microdonnées non géoréférencé dont la taille ne dépasse généralement pas 5% de la population. Les données disséminées par les instituts nationaux de la statistique sont alors, entre autres, utilisées par les planificateurs pour produire une base de données statistiquement équivalente à la compilation confidentielle des formulaires détaillés. Ce processus inversé est ce que l'on appelle la synthèse de population. Une population synthétique est une liste totalement énumérée, caractérisée et spatialisée de ménages et individus non identifiables. Lorsqu'elle est destinée à alimenter une simulation des comportements de mobilité, une population synthétique de qualité montre des caractéristiques sociodémographiques, une distribution spatiale, et par conséquent des comportements de mobilité statistiquement cohérents avec ceux de la population réelle.

Il existe de nombreuses contributions scientifiques visant à améliorer la qualité de la synthèse de population. Or, ces contributions portent essentiellement sur l'amélioration de la représentativité sociodémographique de la population synthétique. L'objectif général de la thèse est alors de proposer un cadre méthodologique permettant d'optimiser la qualité des populations synthétiques en termes de représentativité sociodémographique et de précision spatiale simultanément. Moins discutée dans la littérature, la précision spatiale constitue le sujet principal des contributions de la thèse sans que la représentativité sociodémographique soit pour autant négligée. Afin d'atteindre cet objectif général une méthodologie en trois étapes correspondant chacune à une contribution scientifique originale est établie.

La première étape commence par la définition d'indicateurs de qualité permettant de mesurer la précision spatiale et la qualité globale d'une population synthétique. Ces indicateurs sont ensuite investis dans des analyses de sensibilité menées sur les régions métropolitaines de Montréal, Toronto et Vancouver dans le but de mesurer l'impact du degré d'agrégation de la résolution de référence sur la qualité de la synthèse. Des populations synthétiques sont générées à des résolutions géographiques variées dans chacune des régions métropolitaines et pour chaque population synthétique obtenue, l'ajustement sociodémographique, la précision spatiale et la qualité globale sont calculés. Des mesures potentielles d'amélioration de la qualité de la synthèse telles que l'harmonisation des distributions marginales du recensement et le double contrôle de résolutions géographiques sont aussi évaluées. L'analyse des résultats a montré que lorsque la résolution de référence utilisée est désagrégée, l'ajustement sociodémographique de la population synthétique se dégrade mais sa précision spatiale s'améliore. L'amplitude mesurée de l'amélioration en termes de précision spatiale est toutefois largement plus importante que celle de la dégradation en termes d'ajustement sociodémographique. Il est donc démontré que la désagrégation de la résolution de référence améliore la qualité globale de la synthèse. L'utilisation de distributions marginales harmonisées améliore la qualité de la synthèse. Le double contrôle détériore toutefois la population synthétique produite. La synthèse explicitement spatialisée (i.e., à l'échelle du bâtiment ou de l'unité de logement) en utilisant des distributions marginales harmonisées est alors définie comme meilleure pratique.

Les approches existantes de synthèse explicitement spatialisée sont séquentielles. Elles consistent à d'abord générer une population synthétique zonale pour ensuite distribuer les ménages synthétiques sur des localisations ponctuelles à l'intérieur de leurs zones de synthèse respectives,

i.e., assigner à chaque ménage synthétique des coordonnées (X;Y) dans sa zone de synthèse. Ces approches présentent deux limitations principales. Elles sont d'une part, par leur nature séquentielle, vulnérables aux incohérences entre la population synthétique zonale générée et l'environnement bâti. Un exemple des incohérences possibles est d'avoir dans une zone de synthèse plus de ménages synthétiques que d'unités de logement. D'autre part, ces méthodes présentent des enjeux de transférabilité qui relèvent essentiellement de leur dépendance de la richesse de la zone d'étude en bases de données spatialisées (e.g., échantillon de microdonnées géoréférencé, distributions marginales de variables d'intérêt à l'échelle du bâtiment ou de l'unité de logement, etc.), de la performance de modèles localement calibrés, et de la familiarité des usagers avec la zone d'étude permettant d'émettre certaines hypothèses telles qu'un taux d'inoccupation des unités de logement.

Face à cela, un cadre méthodologique multi-résolution intégré de synthèse explicitement spatialisée est proposé. La conception de ce cadre méthodologique constitue la deuxième étape de la démarche de recherche. L'utilisation du cadre méthodologique intégré proposé permet d'éviter les limitations de l'approche séquentielle puisqu'il permet de générer une population synthétique à l'échelle du bâtiment directement avec un minimum de données. En effet, la distribution marginale d'une seule variable d'intérêt à l'échelle du bâtiment est suffisante pour générer une population synthétique à l'échelle du bâtiment en utilisant le cadre méthodologique proposé. Il inclut une extension d'un algorithme efficace de synthèse par optimisation à la synthèse multi-niveau et multi-résolution où n'importe quel nombre de niveaux et de résolutions peut être contrôlé. Cette flexibilité renforce la transférabilité du cadre méthodologique intégré puisqu'il s'avère adaptable à des besoins variés en termes de contrôle. Un cadre méthodologique séquentiel et le cadre méthodologique intégré proposé sont utilisés pour générer des populations synthétiques spatialisées pour les aires de diffusion de l'île de Montréal, Canada. L'ajustement sociodémographique, la précision spatiale, la qualité globale, et l'ajustement à l'échelle du bâtiment sont calculés pour chacune des populations synthétiques produites afin de comparer les performances des cadres méthodologiques séquentiel et intégré. Malgré une légère perte en termes d'ajustement sociodémographique, les populations synthétiques produites par le cadre méthodologique intégré démontrent une précision spatiale, une qualité globale et un ajustement à l'échelle du bâtiment qui sont meilleurs que ceux des populations synthétiques générées par l'approche séquentielle.

Malgré la transférabilité meilleure et la sensibilité réduite aux incohérences entre la population synthétique et l'environnement bâti dont elle fait montre, l'approche intégrée présente deux limitations principales. En effet, chaque bâtiment constituant en soi une unité géographique de synthèse, la représentativité des populations synthétiques produite pourrait être réduite compte tenu du manque d'intrants de la synthèse (notamment d'un échantillon de microdonnées représentatif) à une résolution aussi désagrégée. De plus, la génération d'une population synthétique à l'échelle du bâtiment pour une ville ou une métropole implique le contrôle simultané de nombreuses contraintes ce qui engendre un enjeu d'évolutivité potentiel compte tenu de la grandeur des dimensions atteintes par la matrice de génération lorsque l'approche intégrée est utilisée. Ainsi, la troisième étape de la démarche de recherche consiste à compléter le choix méthodologique de synthèse spatialisée par un nouveau cadre méthodologique séquentiel de synthèse explicitement spatialisée. L'idée est d'exploiter la richesse de l'information spatiale collectée dans le cadre d'enquêtes par le biais d'algorithmes performants de partitionnement, Clustering Large Applications (CLARA), et de classification, Random Forest, afin de prédire le plus précisément possible les localisations des ménages synthétiques à l'échelle du bâtiment directement. En plus de préserver un ajustement sociodémographique optimal et de réaliser une spatialisation réaliste de la population synthétique, le nouveau cadre méthodologique fait montre d'une transférabilité meilleure que les cadres méthodologiques séquentiels existants grâce à l'efficacité de l'algorithme CLARA. Le nouveau cadre méthodologique séquentiel est appliqué à la génération d'une population synthétique explicitement spatialisée pour l'île de Montréal. Les processus de génération zonale, de partitionnement, de classification, et de distribution spatiale dont est composé le cadre méthodologique proposé montrent tous des performances satisfaisantes.

Basés sur deux approches différentes, les cadres méthodologiques de synthèse spatialisée proposés sont adaptés à des contextes différents. Une démarche de sélection du cadre méthodologique permettant d'optimiser la qualité de la synthèse spatialisée en fonction des caractéristiques de la zone d'étude considérée est alors proposée. En effet, lorsque la zone d'étude dispose de données spatialisées, de modèles locaux de partitionnement et de classification performants, et d'une base de données fiable décrivant l'environnement bâti, l'approche séquentielle est à privilégier. L'approche intégrée permet de réaliser une meilleure qualité de synthèse lorsque l'un ou plusieurs de ces éléments sont manquants dans la zone d'étude considérée.

ABSTRACT

Transportation infrastructure projects generally require several years and considerable investments to see the light of day. However, transportation demand is continuously evolving and the funds available to provide for the population's mobility needs are limited. Prior planning leading to anticipatory action is thus necessary to ensure that a suitable transportation supply is maintained, and relevant and effective investments are made. Indeed, in the present as in the future, the planning of transportation supply is based on mobility issues generally identified through models reflecting the mobility behavior of the population.

Individuals' mobility behaviors are generally strongly influenced by their sociodemographic characteristics, their households' characteristics, and their residential locations. Thus, to feed large-scale agent-based microsimulation platforms, increasingly popular in travel demand modeling, a fully enumerated, characterized and spatialized list of households and individuals making up the population of the study area is required. Such a database can be obtained by compiling census forms completed by the citizens of the study area. However, this compilation remains inaccessible to planners and researchers for privacy reasons mainly. Instead, national statistical agencies generally disseminate two types of data derived from the confidential compilation of detailed forms they receive: aggregate tabulations of a variety of variables at different zonal breakdowns, commonly called census summary files, and a non-georeferenced microdata sample whose size does generally not exceed 5% of the population. The data disseminated by national statistical agencies are then, among others, used by planners to produce a statistical equivalent to the confidential compilation of completed detailed forms. This reversed process is called population synthesis. A synthetic population is a fully enumerated, characterized, and spatialized list of unidentifiable households and individuals. When intended to feed a simulation of mobility behaviors, a reliable synthetic population should show sociodemographic characteristics, spatial distribution, and thus mobility behaviors that are statistically consistent with those of the real population.

Many scientific contributions were intended to improve the quality of population synthesis. However, these contributions essentially focused on improving the sociodemographic representativeness of synthetic populations. The general objective of the thesis is then to introduce a methodological framework that can be used to generate optimal synthetic populations in terms of

both sociodemographic representativeness and spatial precision simultaneously. Less tackled in the literature, improving the synthesis spatial precision is the main concern of the thesis contributions without sociodemographic representativeness being neglected. In order to achieve the general research objective, a three-step methodology, where each step corresponds to an original scientific contribution, is established.

The first step starts with defining goodness-of-fit indicators to measure the spatial precision and the overall quality of a synthetic population. These indicators are then invested in sensitivity analyses carried out on Montreal, Toronto, and Vancouver census metropolitan areas with the aim of measuring the impact of the reference resolution aggregation level on the population synthesis quality. Synthetic populations are generated at various geographic resolutions in each of the three census metropolitan areas and for each synthetic population obtained, its sociodemographic accuracy, spatial precision and overall quality are assessed. Moreover, census targets' harmonization and double geographic resolution control are tested as means of synthesis quality improvement. The results' analyses showed that the disaggregation of the reference resolution leads to worse sociodemographic accuracy but better spatial precision. The gain in terms of spatial precision measured is, however, much greater than the loss in terms of sociodemographic accuracy. Disaggregating the reference resolution is thus proven to improve the synthesis overall quality. Using harmonized marginal distributions is shown to further improve synthesis quality. Double control is however shown to damage synthetic populations. Performing explicitly spatialized (i.e., at the scale of the building or the housing unit) population synthesis using harmonized marginal distributions is hence defined as best practice.

Existing spatialized population synthesis methods follow a sequential framework where generation and spatialization are independently performed. The idea is to generate a synthetic population at a zonal resolution first, then distribute synthetic households on specific locations within their respective synthesis zones (i.e., assign (X;Y) coordinates to each synthetic household within its generation zone). These approaches show two main limitations. On the one hand, they are by their sequential nature vulnerable to inconsistencies between the zonal synthetic population generated and the built environment. An example of possible inconsistencies is having more synthetic households than housing units in a synthesis zone. On the other hand, these methods present major transferability issues which essentially rely in their dependence on the richness of the spatial databases (e.g., georeferenced microdata sample, marginal distributions of variables of interest at

the building or housing unit scale, etc.) available in the study area, the performance of locally calibrated models, and the familiarity of users with the study area allowing them to make certain assumptions such as the dwelling units' vacancy rate.

Therefore, an integrated multiresolution framework for explicitly spatialized population synthesis is proposed. The design of this methodological framework constitutes the second step of the research methodology. The proposed framework makes it possible to avoid the limitations of the sequential approach by generating a synthetic population at the building scale directly with minimal input data. Indeed, the marginal distribution of a single variable of interest at the building scale is sufficient to generate a synthetic population at the building scale using the new framework. It includes an extension of an efficient optimization-based algorithm to multilevel and multiresolution population synthesis where any number of levels and resolutions can be controlled. This flexibility enhances the transferability of the integrated framework since it makes it adaptable to various needs in terms of level and resolution control. A sequential conventional framework and the proposed integrated framework are used to generate spatialized synthetic populations for the dissemination areas of the Montreal Island, Canada. Accuracy, precision, overall, and building-scale errors were calculated for each of the synthetic populations produced in order to compare the performances of the sequential and integrated frameworks. Despite showing a slight loss in terms of sociodemographic accuracy, the synthetic populations produced by the integrated framework demonstrate drastically better spatial precision, overall quality and building-scale accuracy compared to their equivalents generated using the sequential approach.

Despite the better transferability and reduced sensitivity to inconsistencies between the synthetic population and the built environment that it shows, the integrated approach presents two main limitations. Indeed, each building being used as an independent reference geographic unit, the sociodemographic representativeness of the synthetic populations obtained could be damaged given the lack of synthesis inputs (in particular a representative microdata sample) at such a disaggregated resolution. Moreover, the generation of a synthetic population at the building scale for a city or metropolis involves the simultaneous control of numerous constraints, which creates a potential scalability issue when the integrated approach is used. Thus, the third step of the research methodology consists of completing the proposed choice of performant spatialized population synthesis methods with a new sequential framework. The idea is to leverage powerful clustering (Clustering Large Applications (CLARA)) and classification (Random Forest)

algorithms as well as the rich spatial information collected as part of surveys to make accurate predictions of synthetic households' locations at the building scale directly. In addition to preserving an optimal sociodemographic representativeness through zonal generation and achieving a realistic spatialization of the synthetic population, the new sequential framework demonstrates better transferability than existing sequential frameworks thanks to the efficiency of the CLARA algorithm. The new sequential framework is applied to the generation of an explicitly spatialized synthetic population for the Montreal Island. The zonal generation, clustering, classification, and spatial distribution processes making up the proposed framework all show satisfactory performances.

The spatialized population synthesis frameworks introduced are based on different approaches and are thus to be used in different contexts. Therefore, we propose an approach for selecting the most suitable spatialized population synthesis framework among the two proposed in this thesis according to the considered study area characteristics. When spatialized data, efficient local clustering and classification models, and a reliable database describing the built environment are available in the study area, the sequential approach is to be favored. The integrated approach, however, makes it possible to achieve better quality of synthesis when one or more of these elements are missing in the considered study area.

TABLE DES MATIÈRES

DÉDICACE.....	III
REMERCIEMENTS	IV
RÉSUMÉ.....	V
ABSTRACT	IX
TABLE DES MATIÈRES	XIII
LISTE DES TABLEAUX.....	XIX
LISTE DES FIGURES.....	XXI
LISTE DES SIGLES ET ABRÉVIATIONS	XXIV
CHAPITRE 1 INTRODUCTION.....	1
1.1 Contexte	1
1.2 Synthèse de population.....	2
1.3 Enjeux et problématique.....	4
1.4 Objectifs et contributions	5
1.4.1 Objectif 1 : sélectionner la résolution de référence optimale pour la synthèse de population.....	6
1.4.2 Objectif 2 : générer une population synthétique explicitement spatialisée	7
1.4.3 Objectif 3 : exploiter la précision spatiale des enquêtes dans la synthèse explicitement spatialisée de la population	7
1.5 Structure de la thèse	8
CHAPITRE 2 REVUE DE LITTÉRATURE	9
2.1 Modélisation des comportements de mobilité.....	9
2.1.1 Premières tentatives.....	9
2.1.2 Modélisation agrégée	10
2.1.3 Modélisation désagrégée	12

2.2	Synthèse de population – Génération.....	13
2.2.1	Méthodes basées sur l’ajustement.....	15
2.2.2	Méthodes basées sur la simulation.....	37
2.3	Synthèse de population – Spatialisation.....	49
2.3.1	Méthodes de synthèse spatialisée.....	50
2.3.2	Méthodes de microsimulation spatiale.....	54
2.4	Lacunes dans la littérature.....	55
CHAPITRE 3 MÉTHODOLOGIE GÉNÉRALE.....		57
3.1	Étapes préliminaires.....	57
3.1.1	Système de zonage.....	58
3.1.2	Génération de l’environnement bâti.....	63
3.1.3	Variables de contrôle.....	81
3.1.4	Échantillon désagrégué.....	86
3.1.5	Distributions de probabilités marginales.....	92
3.2	Méthodologie de l’activité de recherche.....	95
3.2.1	Étape 1 : établissement d’un diagnostic.....	95
3.2.2	Étape 2 : définition d’un cadre méthodologique universel de synthèse spatialisée ...	96
3.2.3	Étape 3 : exploitation des données spatialisées.....	99
CHAPITRE 4 ARTICLE 1: IMPACT OF THE GEOGRAPHIC RESOLUTION ON POPULATION SYNTHESIS QUALITY		103
4.1	Introduction.....	104
4.1.1	Problem statement.....	105
4.1.2	Contributions.....	108
4.2	Literature review.....	109
4.2.1	Iterative Proportional Fitting (IPF).....	109

4.2.2	Multilevel synthesizers.....	111
4.2.3	Multiresolution synthesizers	115
4.3	Materials and methods	115
4.3.1	Study area.....	115
4.3.2	Control variables	116
4.3.3	Zoning system	117
4.3.4	Datasets	119
4.3.5	Data inconsistencies	122
4.3.6	Scenarios	125
4.3.7	Assessment criteria.....	127
4.4	Results	129
4.4.1	How Do α , β , and γ vary according to the RGR used?	129
4.4.2	How does γ vary according to β ? In other words, how is the precision improved when decreasing accuracy, i.e., when using a less aggregate RGR, and vice-versa?.....	132
4.4.3	What is the impact of census targets' harmonization and additional control at the CMA resolution on the total error?	133
4.4.4	How does the variation of γ according to β change when census targets are harmonized?	134
4.5	Discussion	137
4.6	Conclusions	138
CHAPITRE 5	ARTICLE 2: INTEGRATED MULTIREOLUTION FRAMEWORK FOR SPATIALIZED POPULATION SYNTHESIS.....	140
5.1	Introduction	142
5.2	Literature review	143
5.2.1	Generation methods.....	143

5.2.2	Multiple simultaneous controls	147
5.2.3	Spatialization methods	149
5.2.4	Contributions	151
5.3	Methodology	154
5.3.1	Household marginals synthesis at the building resolution	157
5.3.2	Multiresolution optimization-based population synthesis.....	157
5.3.3	Assessment	159
5.4	Case study	160
5.4.1	Step 0: preliminary harmonization.....	161
5.4.2	Step 1: generation of the built environment	161
5.4.3	Step 2: HH marginals synthesis at the building resolution	163
5.4.4	Step 3: multiresolution optimization-based synthesis.....	163
5.4.5	Spatialized population synthesis using the CF.....	165
5.5	Results	165
5.5.1	Local assessment	166
5.5.2	Global assessment	168
5.6	Conclusion and outlook.....	171
CHAPITRE 6 ARTICLE 3: A NOVEL MACHINE LEARNING-BASED SPATIALIZED POPULATION SYNTHESIS FRAMEWORK		175
6.1	Introduction	177
6.2	Literature review	178
6.2.1	Population synthesis.....	178
6.2.2	Explicit spatialization.....	181
6.2.3	Machine learning algorithms.....	185
6.2.4	Contributions	187

6.3	Methodology	187
6.3.1	Step 1: population synthesis	188
6.3.2	Step 2: spatial data processing	188
6.3.3	Step 3: nearest neighbor join	188
6.3.4	Step 4: $P_{\text{distance}}(HH \cap BLD)$	190
6.3.5	Step 5: $P_{\text{cluster}}(HH \cap BLD)$	190
6.3.6	Step 6: spatial distribution	190
6.3.7	Assessment	190
6.4	Case study	192
6.4.1	Step 1: population synthesis	192
6.4.2	Step 2: spatial data processing	195
6.4.3	Step 3: nearest neighbor join	196
6.4.4	Step 4: $P_{\text{distance}}(HH \cap BLD)$	196
6.4.5	Step 5: $P_{\text{cluster}}(HH \cap BLD)$	197
6.4.6	Step 6: spatial distribution	200
6.5	Results	201
6.5.1	Population synthesis	201
6.5.2	Clustering	201
6.5.3	Classification	203
6.5.4	Spatial Distribution	204
6.6	Conclusion and outlook	205
CHAPITRE 7	DISCUSSION GÉNÉRALE	208
7.1	Synthèse des contributions de recherche	208
7.1.1	Indicateurs de précision spatiale et de qualité globale	208

7.1.2	Impact de la résolution de référence sur la qualité de synthèse	209
7.1.3	Méthode d'harmonisation des distributions marginales.....	210
7.1.4	Cadre méthodologique intégré	210
7.1.5	Cadre méthodologique séquentiel	213
7.2	Discussion des contributions de recherche.....	216
7.2.1	Intérêt de la synthèse explicitement spatialisée.....	217
7.2.2	Choix méthodologique	221
7.3	Limitations et perspectives.....	225
7.3.1	Limitations	225
7.3.2	Perspectives	231
CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS		234
8.1	Contributions.....	236
8.2	Limitations et contraintes	240
8.3	Perspectives de recherche.....	242
BIBLIOGRAPHIE		244

LISTE DES TABLEAUX

Tableau 3.1 Résolutions géographiques utilisées dans le projet de recherche.....	58
Tableau 3.2 Variables incluses dans la base de données des adresses ponctuelles.....	64
Tableau 3.3 Variables incluses dans la base de données des unités d'évaluation foncière.....	66
Tableau 3.4 Variables incluses dans la base de données des bâtiments 2D.....	69
Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière	70
Tableau 3.6 Variables de la base de données des bâtiments	80
Tableau 3.7 Bassin de variables de contrôle du recensement de la population 2016	82
Tableau 3.8 Bassin de variables de contrôle de l'enquête O-D 2018	84
Tableau 3.9 Résolutions des distributions marginales des variables de contrôle.....	92
Table 4.1 Supposed impacts of RGR aggregation on population synthesis.....	107
Table 4.2 Control variables	116
Table 4.3 Example of data inconsistencies (Montreal CMA).....	123
Table 4.4 Scenarios	125
Table 5.1 Existing options for explicitly spatialized population synthesis	152
Table 5.2 Building-resolution attributes.....	162
Table 5.3 Control variables	163
Table 5.4 Attributes and resolutions considered for each indicator	165
Table 6.1 SA and IA limitations.....	185
Table 6.2 Household and people variables.....	193
Table 6.3 Building attributes	195
Table 6.4 CLARA tuning parameters	198
Table 6.5 Random Forest tuning parameters	199
Table 6.6 Retained clustering model characteristics	202

Table 6.7 Buildings' clustering results.....	202
Table 6.8 Retained Random Forest model characteristics	203
Table 6.9 Example records from the spatialized synthetic population	204

LISTE DES FIGURES

Figure 1.1 Des formulaires du recensement à la population synthétique (Khachman et al., 2021a)	3
Figure 2.1 Matrice O-D à l'issue de l'étape de génération	10
Figure 2.2 Matrice O-D à l'issue de l'étape de distribution	11
Figure 2.3 : exemple d'ajustement itératif proportionnel	17
Figure 2.4 Exemple d'application d'une itération d'IPU	22
Figure 3.1 Hiérarchies des résolutions géographiques considérées pour (a) l'ensemble du Canada et (b) l'île de Montréal	60
Figure 3.2 Résolutions géographiques utilisées dans le projet de recherche	62
Figure 3.3 Extrait graphique des adresses ponctuelles de l'île de Montréal (Ville de Montréal, 2023a)	65
Figure 3.4 Extrait graphique des unités d'évaluation foncière de l'île de Montréal (Ville de Montréal, 2023c)	68
Figure 3.5 Extrait graphique des bâtiments 2D de l'île de Montréal (Ville de Montréal, 2023b)	69
Figure 3.6 Extrait graphique du rôle d'évaluation foncière sur l'île de Montréal (MAMH, 2020)	78
Figure 4.1 δ county (a) synthetic population with the county used as RGR and (b) observed population	106
Figure 4.2 Geographic locations of Montreal, Toronto, and Vancouver CMAs	116
Figure 4.3 CMA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver	118
Figure 4.4 CSD boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver	118
Figure 4.5 ADA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver	118
Figure 4.6 CT boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver	119
Figure 4.7 DA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver	119
Figure 4.8 Distributions of ages per sex in Montreal CMA's PUMF and population	121

Figure 4.9 Distributions of ages per sex in Toronto CMA's PUMF and population	121
Figure 4.10 Distributions of ages per sex in Vancouver CMA's PUMF and population	121
Figure 4.11 Households' distributions according to household size in the PUMF and the population of the CMAs of (a) Montreal, (b) Toronto, and (c) Vancouver	122
Figure 4.12 Households' distribution according to household after-tax income in the PUMF and the population of the CMAs of Montreal, Toronto, and Vancouver	122
Figure 4.13 Example of step 3 of the harmonization process applied to a cluster of n DA.....	124
Figure 4.14 Variation of α according to the RGR	129
Figure 4.15 Variation of β according to the RGR	130
Figure 4.16 Variation of γ according to the RGR	130
Figure 4.17 Variation of $\beta + \gamma$ according to the RGR.....	132
Figure 4.18 Variation of γ according to β	133
Figure 4.19 Variation of $\beta + \gamma$ at the DA resolution according to the configuration used.....	134
Figure 4.20 Variation of γ according to β with raw and harmonized data in Montreal	135
Figure 4.21 Variation of γ according to β with raw and harmonized data in Toronto.....	135
Figure 4.22 Variation of γ according to β with raw and harmonized data in Vancouver	136
Figure 4.23 Variation of ΔSlope according to α_{DA}	136
Figure 5.1 Conventional spatialized population synthesis framework (CF)	155
Figure 5.2 Integrated multiresolution spatialized population synthesis framework (IMF).....	156
Figure 5.3 Example of HH marginals synthesis at the building resolution.....	157
Figure 5.4 Optimization problem formulation for simultaneous control of M levels at K resolutions	158
Figure 5.5 Example of optimization problem formulation for two levels at two resolutions	158
Figure 5.6 Montreal Island (a) CD, (b) DAs and (c) DBs.....	161
Figure 5.7 Point addresses and property assessment units.....	162

Figure 5.8 Distributions of a accuracy, b precision, c total and d building-resolution errors	167
Figure 5.9 Histogram of ΔTE	168
Figure 5.10 Accuracy of the global synthetic population generated through a CF and b IMF ..	169
Figure 5.11 Precision of the global synthetic population generated through a CF and b IMF...	170
Figure 5.12 Total quality of the global synthetic population generated through a CF and b IMF	170
Figure 5.13 Building-resolution quality of the global synthetic population generated through a CF and b IMF.....	171
Figure 6.1 Classification of existing explicitly spatialized population synthesis methods.....	182
Figure 6.2 Proposed explicitly spatialized population synthesis framework.....	189
Figure 6.3 ODS area and Montreal Island area delimitations.....	193
Figure 6.4 Montreal Island (a) MBs and (b) DAs.....	193
Figure 6.5 Comparison of synthetic and observed marginal totals.....	201
Figure 6.6 ASW per nbclust.....	202
Figure 6.7 ASWL per nbclust	203
Figure 6.8 Confusion matrix of the retained model applied to the test dataset.....	203
Figure 6.9 Montreal Island explicitly spatialized synthetic population	204
Figure 6.10 Distribution of synthetic households per building cluster probability rank.....	205
Figure 7.1 Limitations potentielles des approches de synthèse explicitement spatialisée de la population.....	222
Figure 7.2 Arbre de décision permettant d'identifier l'approche de synthèse spatialisée de la population la plus appropriée	224

LISTE DES SIGLES ET ABRÉVIATIONS

AD	Aggregate Data (données agrégées)
ADA	Aggregate Dissemination Area (aire de diffusion agrégée)
AP	Autopartage
ARE	Average Relative Error (erreur relative moyenne)
ARTM	Autorité Régionale de Transport Métropolitain
ASW	Average Silhouette Width (largeur moyenne de silhouette)
ASWL	Average Silhouette Width Loss (perte en largeur moyenne de silhouette)
BE	Built Environment (Environnement bâti)
BLD	Building (bâtiment)
CCD	Cyclic Coordinate Descent (descente cyclique de coordonnées)
CD	Census Division (division de recensement)
CF	Conventional Framework (cadre méthodologique conventionnel)
CGAN	Conditional Generative Adversarial Network (réseau antagoniste génératif conditionnel)
CLARA	Clustering Large Applications (partitionnement de grands ensembles de données)
CMA	Census Metropolitan Area (région métropolitaine de recensement)
CO	Combinatorial Optimization (optimisation combinatoire)
CSD	Census Subdivision (subdivision de recensement)
CT	Census Tract (secteur de recensement)
CTGAN	Composite Travel Generative Adversarial Network (réseau antagoniste génératif composite)
CUBF	Code d'Utilisation des Biens-Fonds
CVAE	Conditional Variational Autoencoder (auto-encodeur variationnel)
DA	Dissemination Area (aire de diffusion)

DAG	Directed Acyclic Graph (graphe acyclique orienté)
DATGAN	Directed Acyclic Tabular Generative Adversial Network (réseau antagoniste génératif tabulaire acyclique orienté)
DB	Dissemination Block (îlot de diffusion)
DD	Disaggregate Data (données désagrégées)
DU	Dwelling Unit (unité de logement)
EIPU	Enhanced Iterative Proportional Updating (mise à jour proportionnelle itérative renforcée)
FBS	Fitness-Based Synthesis (synthèse basée sur les valeurs d'ajustement)
FMGD	Fichier de Microdonnées à Grande Diffusion
GR	Generalized Raking (ratissage généralisé)
HH	Household (ménage)
hIPF	Hierarchical IPF (ajustement proportionnel itératif hiérarchique)
hMCMC	Hierarchical Markov Chain Monte Carlo (simulations de Monte Carlo par chaînes de Markov hiérarchiques)
HMM	Hidden Markov Model (modèle de Markov caché)
hPUMF	Public Use Microdata Files – Hierarchical File (fichier de microdonnées à grande diffusion – fichier hiérarchique)
IA	Integrated Approach (approche intégrée)
IMF	Integrated Multiresolution Framework (cadre méthodologique multi-résolution intégré)
IPF	Iterative Proportional Fitting (ajustement proportionnel itératif)
IPU	Iterative Proportional Updating (mise à jour proportionnelle itérative)
iPUMF	Public Use Microdata Files – Individuals File (fichier de microdonnées à grande diffusion – fichier des individus)
MacroF1	Macro-average F1 score (score F1 moyen non pondéré)

MAMH	Ministère des Affaires Municipales et de l’Habitation
MATSim	Multi-Agent Transport Simulation (simulation de transport multi-agent)
MB	Municipal Borough (secteur municipal)
MCMC	Markov Chain Monte Carlo (simulations de Monte Carlo par chaînes de Markov)
N/A	Not Applicable (non applicable)
O-D	Origine-Destination
ODS	Origin-Destination Survey (enquête origine-destination)
OSM	OpenStreetMap
PAM	Partitioning Around Medoids (partitionnement en k-médoïdes)
PE	Precision Error (erreur de précision)
PP	People (personnes)
PUMA	Public Use Microdata Area (zone de microdonnées à grande diffusion)
PUMF	Public Use Microdata File (fichier de microdonnées à grande diffusion)
PUMS	Public Use Microdata Sample (échantillon de microdonnées à grande diffusion)
RGR	Reference Geographic Resolution (résolution géographique de référence)
RGU	Reference Geographic Unit (unité géographique de référence)
RMSE	Root Mean Square Error (erreur quadratique moyenne)
SA	Sequential Approach (approche séquentielle)
SF	Summary Files (résumés du recensement)
SP	Synthetic Population (population synthétique)
SR	Synthetic Reconstruction (reconstruction synthétique)
TASHA	Toronto Area Scheduling Model for Household Agents (modèle de planification de la région de Toronto pour les agents domestiques)
TAZ	Traffic Analysis Zone (zone d’analyse du trafic)
TC	Transport en commun

TRANSIMS	Transportation Analysis Simulation System (Système de simulation d'analyse des transports)
TRS	Truncate, Replicate, Sample (tronquer, répliquer, échantillonner)
UEF	Unité d'évaluation foncière
VAE	Variational Autoencoder (auto-encodeur variationnel)
VP	Vélopartage
WGAN	Wasserstein Generative Adversarial Network (réseau antagoniste génératif de Wasserstein)

CHAPITRE 1 INTRODUCTION

1.1 Contexte

Je me déplace donc je suis

Que l'on veuille étudier, travailler, faire ses courses, rejoindre nos proches ou simplement se divertir, on commence par se déplacer. Or, que ce soit en termes d'infrastructures, de temps ou de confort de déplacement, l'offre en mobilité ne peut répondre parfaitement aux aspirations individuelles de chacun. De la congestion qui nous prend en otage sur les routes, aux émissions de gaz à effet de serre qui menacent notre avenir, sans oublier les enjeux d'accessibilité, d'équité et de sécurité entre autres, les défis liés au transport sont tangibles dans différents aspects de notre vie. Ainsi, avoir accès à une offre de déplacements permettant à la fois de satisfaire la demande et de faire face aux enjeux qui en découlent, demeure une préoccupation de tout citoyen. Une offre optimale en transport consisterait donc en un compromis permettant de minimiser le coût collectif, au sens large non au sens monétaire, des déplacements des individus.

Un modèle, un diagnostic

Relevant de divers domaines incluant sans s'y limiter l'économie, la sociologie et la psychologie, les défis liés au transport sont divers et complexes. Ainsi, pour les appréhender et y définir des solutions appropriées, il faudra commencer par en établir un modèle. Un modèle scientifique est *“une représentation abstraite des relations entre les paramètres caractéristiques d'un phénomène ou d'un processus. Le modèle scientifique vise à simplifier le phénomène étudié, à unifier des phénomènes a priori sans rapport et à prédire des phénomènes nouveaux”* (Larousse en ligne, s.d.). Un modèle est pour les enjeux de mobilité ce qu'est un diagnostic pour une maladie. Ainsi, un modèle fiable et représentatif de la réalité est sine qua non pour la définition de solutions efficaces aux problèmes encourus par la société. Il permet de tester différents scénarios de réponse et d'en mesurer les impacts pour effectuer un choix éclairé.

Mieux vaut prévenir que guérir

La demande en transport évolue très rapidement en nombre comme en nature. Or, de l'identification du besoin à la mise en service d'un projet d'infrastructure de transport, plusieurs années sont souvent requises. Cela pourrait réduire l'efficacité d'une mesure prise a posteriori. De

plus, le coût des investissements en infrastructures de transport est élevé. Ainsi, pour s'assurer de maintenir une offre adaptée et garantir la pertinence et la rentabilité des investissements, une prévision adéquate de la demande, i.e., une modélisation fiable de la demande future, est requise puisqu'elle permettra d'agir par anticipation et optimisera par conséquent les chances d'aboutissement des stratégies entreprises.

Que ce soit pour le présent ou pour un horizon futur, la qualité de la modélisation des comportements de mobilité constitue alors un élément incontournable dans le processus d'identification des besoins et/ou problèmes, et d'établissement de solutions adéquates. Un modèle représentatif des besoins réels en matière de déplacement permettrait de définir des stratégies de réponse et/ou de prévention et d'optimiser par conséquent les investissements.

L'humain au centre du modèle

Les approches modernes de modélisation de la demande en transport sont des approches désagrégées basées sur des processus de prise de décision individuels relevant des caractéristiques des agents modélisés, des contraintes de leur environnement, et des règles régissant les interactions des agents entre eux d'une part, et avec leur environnement d'autre part. Elles permettent ainsi de rouler des simulations microscopiques des comportements de mobilité de la population. Que ce soit pour estimer les chaînes d'activités individuelles ou pour alimenter des microsimulations exposant les comportements de mobilité individuels, une base de données listant et caractérisant l'ensemble des individus de la population est minimalement requise. Une telle base de données peut être obtenue par compilation des formulaires de recensement complétés par la population.

1.2 Synthèse de population

Une population synthétique : un sortilège renversé

Les instituts nationaux de la statistique tels que Statistique Canada, disposent chacun d'une liste caractérisée et spatialisée de ménages et individus établie à partir des données recueillies via les formulaires du recensement. Toutefois, cette liste demeure inaccessible aux chercheurs et planificateurs pour des raisons de confidentialité et de protection de la vie privée. Toutefois, deux produits en sont généralement dérivés puis publiés. Le premier est un échantillon non spatialisé de microdonnées dont la taille ne dépasse généralement pas 5% de la population : au Canada, seule la région métropolitaine du ménage est précisée dans cet échantillon. Le deuxième est constitué

d'agrégations zonales à différentes échelles, e.g., les nombres d'hommes dans les secteurs de recensement. Tel qu'illustré dans la Figure 1.1 pour un territoire imaginaire δ , la synthèse de population revient à inverser la démarche des instituts de la statistique, i.e., utiliser l'échantillon non spatialisé et les totaux zonaux pour reconstruire une liste totalement énumérée, caractérisée et spatialisée de ménages et individus. C'est cette liste que l'on appelle « population synthétique ».

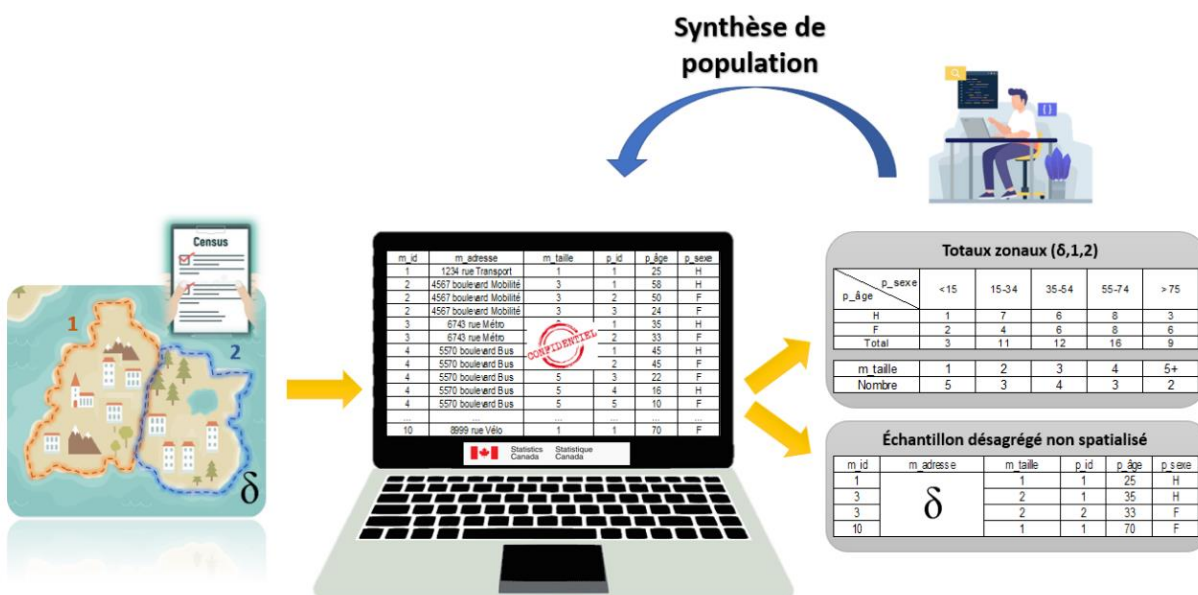


Figure 1.1 Des formulaires du recensement à la population synthétique (Khachman et al., 2021a)

Le but d'une population synthétique n'est pas de reproduire exactement les ménages et individus de la population réelle. Lorsqu'elle est produite pour des fins de modélisation des comportements de mobilité, il s'agit plutôt d'une population de ménages et individus non identifiables qui présente le plus fidèlement possible le comportement de mobilité de la population réelle. L'impossibilité d'identifier les agents synthétiques produits n'est pas un inconvénient mais un avantage de la synthèse de population puisqu'elle permet de publier la population synthétique générée la rendant ainsi accessible aux chercheurs et professionnels sans enjeux de confidentialité.

Variables de contrôle

L'un des choix à effectuer avant de se lancer dans le processus de synthèse est celui des variables de contrôle. Il s'agit des variables selon lesquelles l'on souhaite que la population synthétique générée soit représentative de la population réelle. Autrement dit, il s'agit des caractéristiques de la population réelle dont on souhaite reproduire les distributions dans la population synthétique. Par exemple, si la variable « âge » est contrôlée, une synthèse de qualité mènerait à une population

synthétique où la distribution des âges des individus correspond à celle observée dans la population réelle.

Résolution de référence

Un autre choix à effectuer pour pouvoir démarrer la synthèse de population est celui de la résolution de référence. Il s'agit de la strate ou l'échelle géographique au niveau de laquelle les totaux zonaux du recensement sont extraits. Considérons par exemple que l'on souhaite générer une population synthétique pour le territoire imaginaire δ de la Figure 1.1 composé de deux secteurs (1 et 2) en contrôlant la variable « nombre de personnes ». Supposons que 10 personnes résident dans le territoire δ dont 6 sont dans le secteur 1 et 4 dans le secteur 2. Si la résolution de référence choisie pour la synthèse est celle du territoire, le processus de synthèse devrait idéalement produire une population synthétique composée de 10 individus pour l'ensemble du territoire δ . Toutefois, si la résolution de référence choisie est celle des secteurs, une population synthétique est générée séparément pour chacun des secteurs 1 et 2. Une population synthétique de 6 personnes serait ainsi idéalement obtenue pour le secteur 1 et une autre de 4 personnes serait générée pour le secteur 2.

1.3 Enjeux et problématique

Pour modéliser correctement les comportements de mobilité de la population, il faudrait que les individus synthétiques alimentant le modèle représentent ceux observés dans la population réelle, notamment pour ce qui est des facteurs influençant leurs choix de mobilité. Deux enjeux principaux sont à considérer lors d'une synthèse de population pour fins de modélisation et de prévision de la demande en transport : la représentativité sociodémographique et la précision spatiale.

Dis-moi qui tu es, je te dirai comment tu te déplaces

Un parent ayant de jeunes enfants aurait plus de chances de se diriger vers la garderie à l'heure de sortie qu'un jeune étudiant vivant en colocation. Un jeune étudiant aurait moins de chances d'aller faire ses courses pendant l'horaire des cours qu'un retraité. Un individu de 15 ans aurait moins de chances de conduire sa voiture pour rendre visite à un ami qu'un autre ayant 20 ans. Les caractéristiques sociodémographiques des individus constituent ainsi des facteurs déterminants de leurs comportements de mobilité. Toutefois, certains facteurs régissant les choix de mobilité des individus tels que la motorisation et le motif de déplacement (e.g., accompagnement d'un membre du ménage) relèvent plutôt des caractéristiques sociodémographiques du ménage auquel ils

appartiennent. Ainsi, pour optimiser la vraisemblance des comportements de mobilité modélisés, les modèles utilisés doivent être alimentés par une population synthétique d'individus regroupés en ménages et présentant une distribution de caractéristiques sociodémographiques statistiquement cohérente avec celle de la population réelle. C'est ce que l'on appelle la représentativité sociodémographique de la population synthétique.

Dis-moi où tu habites, je te dirai comment tu te déplaces

La localisation géographique exerce aussi une influence importante sur les choix de mobilité des individus. Par exemple, un individu habitant à 10 km du réseau de transport en commun aurait moins de chances de se déplacer en transport en commun uniquement qu'un autre résidant à 200 m d'une station de métro. Un étudiant résidant à 1 km de son école aurait plus de chances d'y aller en marchant qu'un autre résidant à 50 km de la sienne. Ainsi, les décisions prises par un individu qui s'apprête à effectuer un déplacement, telles que le choix modal et l'heure de départ, sont fortement dépendantes de la localisation du ménage auquel il appartient. Ainsi, pour optimiser la vraisemblance des comportements de mobilité modélisés, les modèles utilisés doivent être alimentés par une population synthétique dont les ménages présentent une distribution spatiale similaire à celle de la population réelle. C'est ce que l'on appelle la précision spatiale de la population synthétique.

Dis-moi qui tu es et où tu habites, je te dirai comment tu te déplaces

En résumé, l'un des intrants requis pour une modélisation fiable des comportements de mobilité de la population réelle consiste en une population synthétique dotée d'une représentativité sociodémographique et d'une précision spatiale optimales. Ainsi, la problématique de recherche régissant les travaux de la présente thèse est la suivante : **comment générer une population synthétique reproduisant le plus fidèlement possible les caractéristiques sociodémographiques et la distribution spatiale de la population réelle ?**

1.4 Objectifs et contributions

L'objectif général du projet de recherche est de définir un cadre méthodologique permettant de produire une population synthétique pour des fins de modélisation des comportements de mobilité, i.e., une liste totalement énumérée et spatialisée d'individus et ménages présentant une cohérence statistique avec la population réelle sur les aspects sociodémographique et spatial. Contrairement

à la précision spatiale, la représentativité sociodémographique des populations synthétiques a fait l'objet de nombreux travaux de recherche. Ainsi, tout en veillant à surveiller les deux aspects et dans la limite où cela contribue à l'amélioration de la qualité globale, cette thèse accorde un intérêt particulier à l'optimisation de la précision spatiale des populations synthétiques.

Les enjeux liés à la représentativité sociodémographique et la précision spatiale des populations synthétiques sont ainsi étudiés et les méthodes de synthèse et de spatialisation existantes sont testées et comparées afin de définir un cadre méthodologique permettant d'optimiser la qualité de la population synthétique générée et par conséquent la fiabilité du modèle alimenté. Assez complexe, l'objectif principal énoncé est décomposé en plusieurs objectifs spécifiques qui sont atteints par le biais de contributions scientifiques originales.

1.4.1 Objectif 1 : sélectionner la résolution de référence optimale pour la synthèse de population

De nombreux travaux visant à améliorer la qualité des populations synthétiques sont continuellement menés. Or dans la littérature, cette qualité est mesurée uniquement par l'ajustement sociodémographique des populations synthétiques, i.e., la correspondance des totaux synthétiques aux totaux zonaux du recensement pour le découpage géographique le plus grossier. Par exemple, si l'on synthétise une population pour le territoire δ (Figure 1.1), la population synthétique est considérée parfaitement ajuster la variable « nombre de personnes » si elle reproduit 10 individus synthétiques pour δ . Toutefois, cela ne garantit aucunement une distribution spatiale adéquate des individus synthétiques sur les secteurs 1 et 2. En effet, une population synthétique avec 6 personnes dans le secteur 1 et 4 personnes dans le secteur 2, et une autre avec 10 personnes dans le secteur 1 et aucune dans le secteur 2 par exemple seront toutes les deux jugées parfaites si seul l'ajustement sociodémographique est considéré comme critère d'évaluation. Ces deux exemples de populations synthétiques, présentant des distributions spatiales complètement différentes, ne produiront toutefois pas les mêmes comportements de mobilité. Ainsi, un indicateur de qualité mesurant la précision spatiale d'une population synthétique est requis pour refléter la différence de qualité entre ces deux cas de figure. Il servira aussi de base à la définition d'un indicateur combinant l'ajustement sociodémographique et la précision spatiale permettant de globalement évaluer la qualité de la population synthétique.

Une fois ces indicateurs définis, l'impact de différents choix de résolution de référence sur l'ajustement sociodémographique, la précision spatiale et la qualité globale de la population synthétique peut être évalué. L'objectif est de sélectionner la résolution de référence permettant d'optimiser la qualité de la population synthétique générée. L'hypothèse émise est que la synthèse aux niveaux les plus désagrégés engendre des gains considérables en termes de précision spatiale améliorant par conséquent la qualité globale de la population synthétique.

1.4.2 Objectif 2 : générer une population synthétique explicitement spatialisée

Si l'hypothèse précédente est confirmée et la désagrégation de la résolution de référence est prouvée améliorer la qualité globale de la population synthétique, il serait idéal de générer une population synthétique à l'échelle spatiale la plus désagrégée possible, soit à l'échelle du bâtiment ou de l'unité de logement. Il s'agit en d'autres termes de générer une population synthétique explicitement spatialisée où chacun des ménages synthétiques dispose de coordonnées (X;Y) précisant sa localisation exacte. Or, une synthèse de population directement au niveau du bâtiment n'a jamais été réalisée pour différentes raisons dont la disponibilité limitée des intrants de synthèse à une échelle aussi fine et la complexité des calculs résultant de la multiplicité des contraintes. Le deuxième objectif spécifique de cette thèse est alors de définir un cadre méthodologique permettant de générer une population synthétique à l'échelle du bâtiment directement. Outre la faisabilité et la qualité de la synthèse explicitement spatialisée, un intérêt particulier sera accordé à l'optimisation de la transférabilité du cadre méthodologique conçu compte tenu de la variabilité des bases de données décrivant l'environnement bâti d'une localisation géographique à une autre.

1.4.3 Objectif 3 : exploiter la précision spatiale des enquêtes dans la synthèse explicitement spatialisée de la population

Les échantillons de microdonnées publiés par les instituts nationaux de la statistique ne disposent que d'une information spatiale peu précise. Par exemple, seule la région métropolitaine des ménages de l'échantillon de microdonnées publié au Canada est connue. Cependant, il existe des échantillons de microdonnées provenant d'enquêtes où les coordonnées des ménages recensés sont disponibles ainsi que leurs caractéristiques et les caractéristiques des individus dont ils sont composés. Ainsi, le troisième objectif spécifique du projet de recherche est de concevoir un cadre méthodologique permettant d'exploiter la richesse de l'information sociodémographique et spatiale

disponible dans la production d'une population synthétique explicitement spatialisée. L'utilisation de méthodes performantes d'apprentissage machine permettraient une mise à profit optimale des données d'enquêtes.

1.5 Structure de la thèse

La suite de la thèse est organisée comme suit : le Chapitre 2 contient une revue de la littérature relative aux thématiques principales régissant les travaux de recherche. Dans le Chapitre 3, la méthodologie générale pour atteindre les objectifs de la thèse est exposée. S'agissant d'une thèse par articles, ce chapitre expliquera l'articulation des articles autour de l'objectif de la thèse. Le Chapitre 4 est constitué du premier article dont le sujet est l'évaluation de l'impact de la résolution de référence sur la qualité de la synthèse de population. Cet article est consacré à l'atteinte de l'objectif spécifique 1 détaillé dans la section 1.4.1. Dans le Chapitre 5, les travaux de recherche menant à la définition d'un cadre méthodologique multi-résolution intégré pour la génération d'une population synthétique explicitement spatialisée sont présentés. Il s'agit du deuxième article de la thèse répondant au deuxième objectif spécifique détaillé dans la section 1.4.2. Le troisième article de la thèse constitue le Chapitre 6. Il traite de l'application d'algorithmes d'apprentissage machine à la synthèse explicitement spatialisée de la population. Ce chapitre répond au troisième objectif spécifique détaillé dans la section 1.4.3. Une discussion générale des travaux et des résultats obtenus est établie dans le Chapitre 7, et les conclusions et recommandations sont détaillées dans le Chapitre 8.

CHAPITRE 2 REVUE DE LITTÉRATURE

La raison d'être du projet étant d'améliorer la modélisation des comportements de mobilité, un aperçu de l'évolution de la pratique de modélisation de la demande en transport jusqu'aux modèles requérant des populations synthétiques est d'abord établi. Ensuite, les méthodes de génération de populations synthétiques existantes sont détaillées et comparées, et les forces et faiblesses de chacune sont exposées. Enfin, les différentes applications de synthèse spatialisée de la population sont explorées afin d'établir une classification des méthodes, d'en tirer les avantages et d'identifier les lacunes à combler.

2.1 Modélisation des comportements de mobilité

Tel qu'expliqué en introduction, les contributions de ce projet de recherche sont destinées à être investies dans la génération d'une population synthétique optimale qui permettrait d'établir des modèles fiables de la demande en transport. Ainsi, avant d'explorer les avancements techniques dans les domaines de la génération et de la spatialisation des populations synthétiques, il serait utile d'élaborer un historique concis de la modélisation de la demande en transport décrivant l'évolution qui a mené aux modèles exigeant des populations synthétiques.

2.1.1 Premières tentatives

Jusqu'à la Seconde Guerre Mondiale, la modélisation de la demande de transport se focalisait essentiellement sur la résolution a posteriori des problèmes qui survenaient sur le réseau routier. Ainsi, face à un problème en cours, des comptages étaient effectués pour alimenter des modèles permettant de déterminer par exemple la largeur nécessaire de la voie en fonction des volumes comptés (Bonnell, 2002). Après la guerre, les problèmes de congestion rapidement amplifiés par la croissance de l'industrie de l'automobile ont imposé une augmentation massive des investissements dans de nouveaux projets d'infrastructures de transport. Ainsi, dans une perspective de rationalisation des investissements, des projets d'infrastructures routières ont, pour la première fois, été accompagnés d'enquêtes Origine-Destination (O-D) visant à identifier les lignes de désir du trafic (Bonnell, 2002). Cependant, la croissance inédite du secteur de l'automobile a rendu les enquêtes O-D, à leur tour, inefficaces. C'est ainsi que la prévision de la demande a été pratiquée pour la première fois au début des années 1950 aux États-Unis dans le but d'anticiper le

développement automobile (Kane et Behrens, 2002) qui était à l'époque synonyme de développement économique (Bonnell, 2002). Les premiers modèles consistaient en une simple projection du trafic futur à partir du trafic existant par le biais d'un facteur de croissance (Kane et Behrens, 2002). Or, l'incapacité de ces projections à prédire les déplacements impliquant une zone en forte urbanisation a mené à l'utilisation des lois gravitaires et des régressions linéaires pour prédire le futur trafic de manière indépendante du trafic existant (Bonnell, 2002). Les modèles agrégés basés sur la procédure séquentielle classique ont ainsi vu le jour aux États-Unis (Kagho et al., 2020).

2.1.2 Modélisation agrégée

La modélisation agrégée de la demande de transport est basée sur la procédure séquentielle classique. Il s'agit d'une approche de modélisation de la demande de transport qui s'opère en quatre étapes (génération, distribution, répartition modale et affectation) correspondant aux quatre séquences du choix de mobilité de l'individu. À l'étape de génération, les totaux des déplacements produits et attirés sont séparément estimés pour chacune des zones de l'aire d'étude à partir de données d'enquêtes recensant des variables jugées explicatives de la production et de l'attraction de déplacements. Les totaux obtenus serviront à compléter les marges de la matrice O-D tel que montré en bleu sur la Figure 2.1 où D_{ij} réfère aux déplacements dont l'origine est la zone i et la destination est la zone j , P_i correspond au total des déplacements dont l'origine est la zone i et A_j correspond au total des déplacements dont la destination est la zone j .

O \ D	Zone₁	Zone₂	Zone_j	Zone_N	Produits
Zone₁					$P_1 = \sum_{j=1}^N D_{1j}$
Zone₂					$P_2 = \sum_{j=1}^N D_{2j}$
Zone_i					$P_i = \sum_{j=1}^N D_{ij}$
Zone_N					$P_N = \sum_{j=1}^N D_{Nj}$
Attirés	$A_1 = \sum_{i=1}^N D_{i1}$	$A_2 = \sum_{i=1}^N D_{i2}$	$A_j = \sum_{i=1}^N D_{ij}$	$A_N = \sum_{i=1}^N D_{iN}$	

Figure 2.1 Matrice O-D à l'issue de l'étape de génération

À l'étape de distribution, les totaux produits et attirés jusque-là séparément estimés, sont reliés et par conséquent transposés en déplacements entre zones (Bonnell, 2002). La distribution permet de trouver le nombre de déplacements entre chaque couple de zones moyennant une fonction de friction reflétant la pénibilité du déplacement (Trépanier, 1999). Les volumes interzonaux ainsi obtenus permettront de compléter les cases de la matrice O-D tel que montré en bleu sur la Figure 2.2.

O \ D	Zone₁	Zone₂	Zone_j	Zone_N	Produits
Zone₁	D_{11}	D_{12}	D_{1j}	D_{1N}	$P_1 = \sum_{j=1}^N D_{1j}$
Zone₂	D_{21}	D_{22}	D_{2j}	D_{2N}	$P_2 = \sum_{j=1}^N D_{2j}$
Zone_i	D_{i1}	D_{i2}	D_{ij}	D_{iN}	$P_i = \sum_{j=1}^N D_{ij}$
Zone_N	D_{N1}	D_{N2}	D_{Nj}	D_{NN}	$P_N = \sum_{j=1}^N D_{Nj}$
Attirés	$A_1 = \sum_{i=1}^N D_{i1}$	$A_2 = \sum_{i=1}^N D_{i2}$	$A_j = \sum_{i=1}^N D_{ij}$	$A_N = \sum_{i=1}^N D_{iN}$	

Figure 2.2 Matrice O-D à l'issue de l'étape de distribution

Tel que son nom l'indique, à l'étape de répartition modale, le choix du (ou des) mode(s) de transport utilisé(s) pour effectuer les déplacements définis à l'étape de distribution sont modélisés. La matrice O-D est alors scindée en matrices O-D spécifiques à un mode ou une combinaison de modes de déplacement moyennant différents modèles dont les variables explicatives relèvent de trois facteurs principaux : l'offre de transport, les caractéristiques du déplacement et les caractéristiques de la personne et de son ménage (Morency, communication personnelle, 28 septembre 2017). Une fois les paires de déplacement déterminées et les parts modales calculées, les itinéraires sont établis. Ceci correspond à la quatrième étape du modèle qui est celle d'affectation de la demande de déplacements sur le réseau. L'affectation peut se faire selon différents principes tels que la minimisation des coûts individuels de transport (e.g., minimisation du temps de déplacement), ou la minimisation du coût total de déplacements (Morency, communication personnelle, 28 septembre 2017).

Dans les années 1950, les modèles à quatre étapes étaient conçus pour résoudre les problèmes relatifs au réseau routier uniquement et la prévision constituait une forme de garantie pour tout

investissement entrepris dans les infrastructures routières (Kagho et al., 2020). Les déplacements n'impliquant pas ou peu l'automobile étaient alors filtrés lors de la phase de la répartition modale (Bonnel, 2002). À partir de la fin des années 1960, les problèmes de circulation se sont amplifiés. L'impasse dans laquelle se trouvait la politique du « tout automobile » devenait de plus en plus critiquée compte tenu de l'ampleur irréaliste, en termes d'emprise et de prix, des projets d'infrastructures qui demeuraient requis pour répondre aux besoins alors prédits (Bonnel, 2002). Les critiques ont augmenté dans les années 1970 et elles allaient cette fois-ci dans deux sens : d'une part, la performance du modèle agrégé de prévision était remise en question (Douglass et Lee, 1973; Atkins, 1977) et d'autre part, la limitation de la planification des transports à l'automobile devenait de moins en moins acceptée (Bonnel, 2002). En effet, l'agrégation de différents facteurs (temps, espace, individus qui se déplacent, etc.) dans la modélisation faite par le biais de la procédure séquentielle classique, rend celle-ci incapable de refléter les différents comportements de mobilité observés à l'échelle individuelle (Kane et Behrens, 2002). Or, une modélisation microscopique basée sur une théorie reflétant les comportements humains individuels plutôt que centrée sur des déplacements zonaux agrégés commençait toutefois à être requise pour répondre aux problématiques émergentes (Kagho et al., 2020). En effet, à la fin des années 1970, les objectifs de la planification des transports ont évolué et une politique de gestion de la demande commençait à remplacer celle de l'expansion des infrastructures. Ce besoin accru de modélisation des comportements individuels a alors mené à l'émergence des modèles désagrégés (Kagho et al., 2020).

2.1.3 Modélisation désagrégée

Par opposition à l'approche agrégée où l'on considère un « *individu moyen* » (Bonnel, 2002) dont les comportements de mobilité constituent à leur tour une moyenne des choix recensés dans sa zone, l'approche désagrégée se focalise sur la modélisation de la variété des choix de mobilité individuels (Trépanier, 1999). Ainsi, dans une perspective de considération de l'environnement social et spatial des individus dans la prévision de la demande de mobilité quotidienne, les approches en termes d'activités ont été développées (Bonnel, 2002). Au lieu d'être basés sur des déplacements estimés à partir d'agrégations zonales de variables plus ou moins explicatives des comportements de mobilité individuels, les modèles d'activités sont basés sur l'activité qui est à la source du déplacement et par conséquent de l'individu qui agit dans un contexte social et spatial

précis (Raux et Chapleau, 1997). Les modèles basés activités sont alors devenus la tendance dans la pratique de la prévision de la demande de transport. À partir de données relatives à la population, une chaîne d'activités est estimée pour chacun des individus incluant par exemple le lieu, le temps et le mode utilisé de chacune des activités. Un modèle d'affectation est ensuite généralement utilisé pour transposer les chaînes d'activités estimées en itinéraires empruntant les différentes infrastructures de transport (routes, lignes de transport en commun, etc.).

Récemment, les modèles basés agents, eux aussi relevant de la modélisation désagrégée, ont fait irruption dans la pratique de la modélisation de la demande en transport. Les trois composantes d'un modèle basé agents sont : les agents (des entités indépendantes et autonomes dans leurs choix), leur environnement, et les règles régissant les interactions des agents entre eux ainsi que les interactions des agents avec leur environnement. Les chaînes d'activités estimées pour chacun des agents par les modèles d'activités constituent la demande initiale simulée par les modèles basés agents. Celle-ci est ensuite itérativement ajustée selon les résultats de chaque simulation jusqu'à ce qu'un équilibre soit atteint (Kagho et al., 2020). Outre les modèles d'activités, les modèles basés agents sont capables d'accueillir de multiples modules constitués de modèles liés au transport tels que les modèles de choix résidentiels et commerciaux, les modèles d'utilisation du sol, etc. (Kagho et al., 2020).

Afin de rouler des simulations microscopiques des comportements de mobilité de la population, les modèles désagrégés doivent être alimentés, entre autres, par des populations synthétiques représentatives de la population réelle. Tel que détaillé en introduction, la qualité d'une population synthétique se mesure par sa représentativité sociodémographique et sa précision spatiale. Dans une optique d'optimisation de la représentativité sociodémographique des populations synthétiques, une variété de méthodes de génération a été conçue. Celles-ci sont discutées en détail dans la section suivante.

2.2 Synthèse de population – Génération

Les microsimulations couplées aux modèles basés agents, de plus en plus utilisées par les planificateurs pour la gestion des politiques de transport, dont UrbanSim (Waddell, 2002), Toronto Area Scheduling Model for Household Agents (TASHA) (Miller et Roorda, 2003), Multi-Agent Transport Simulation (MATSim) (Balmer et al., 2009), DAYSIM (Bowman et al., 2014), et SimMobility (Adnan et al., 2016), exigent comme données entrantes des données désagrégées

décrivant la population, au niveau des ménages et/ou des individus, ainsi que son activité en matière de mobilité. Or, le processus de récolte de données de ce type pose des enjeux vis-à-vis de la protection de la vie privée et s'avère coûteux, chronophage et présente une certaine complexité (Mohammadian et al., 2010). Face à cela, d'autres solutions permettant d'alimenter ces simulations microscopiques ont été envisagées par le biais de méthodes de synthèse de population.

La synthèse de population est typiquement un processus se servant de données agrégées, provenant généralement de résumés de recensements, et de données partielles désagrégées relatives à un échantillon d'enquête, pour produire une population totalement énumérée d'agents (individus et/ou ménages) non identifiables avec des caractéristiques sociodémographiques assignées. L'objectif de la synthèse est de produire une population synthétique ayant des caractéristiques statistiquement cohérentes avec les données disponibles décrivant la population réelle.

Deux choix doivent être effectués avant de se lancer dans le processus de synthèse : celui des variables de contrôle et celui de la résolution de référence. D'abord, les caractéristiques sociodémographiques de la population réelle dont l'on souhaite que la population synthétique soit représentative doivent être choisies. Il s'agit des variables de contrôle pour le processus de synthèse. Pour alimenter des microsimulations de comportements de mobilité, les caractéristiques ayant une influence prépondérante dans la détermination des comportements de mobilité des individus sont sélectionnées. La résolution de référence est ensuite choisie parmi une variété de découpages de la région d'étude souvent correspondant à des zones administratives plus ou moins fines prédéfinies dans le cadre du recensement. Les divisions de recensement, les subdivisions de recensement, les secteurs de recensement et les aires de diffusion sont des exemples de découpages administratifs définis dans le recensement canadien (Statistique Canada, 2019b).

Ainsi, les données entrantes peuvent être préparées. Les données agrégées correspondant à la résolution de référence et aux variables de contrôle choisies sont extraites des résumés du recensement. Il s'agit de tableaux à une, deux ou n dimensions contenant les totaux de la distribution conjointe des variables de contrôle sélectionnées que ce soit au niveau des individus ou des ménages. Quant aux données désagrégées, elles sont tirées à partir d'un échantillon désagrégé et anonymisé de personnes et/ou de ménages complètement caractérisés dans le cadre d'une enquête. Les données de l'échantillon sont alors filtrées pour garder uniquement celles relatives aux variables de contrôle sélectionnées, puis tabulées pour être utilisées dans la synthèse

de population. Les échantillons de microdonnées utilisés, généralement des échantillons désagrégés de moins de 5% de la population publiés par les instituts nationaux de la statistique, ne sont pas géolocalisés. Par exemple, la seule information disponible pour l'échantillon canadien est la région métropolitaine d'appartenance.

La synthèse de population peut être résumée à un problème d'optimisation où l'on cherche à ajuster l'échantillon utilisé aux totaux observés (i.e., totaux extraits des tableaux résumés du recensement) tout en préservant la corrélation existante, dans l'échantillon, entre les variables de contrôle sélectionnées. Plusieurs approches de synthèse de population basées sur différents principes sont documentées dans la littérature. Elles peuvent être classées dans deux grandes catégories : les méthodes basées sur l'ajustement et les méthodes basées sur la simulation. Un aperçu des méthodes appartenant à chacune de ces deux familles incluant leurs fondements théoriques, leurs forces et leurs faiblesses sera donné dans la suite de cette section.

2.2.1 Méthodes basées sur l'ajustement

Les méthodes basées sur l'ajustement sont des méthodes qui procèdent par ajustement mathématique de l'échantillon de microdonnées aux totaux observés dans le recensement. Il existe deux sous-catégories de méthodes basées sur l'ajustement : les méthodes de reconstruction synthétique ou « Synthetic Reconstruction » (SR) basées sur l'ajustement proportionnel itératif ou « Iterative Proportional Fitting » (IPF), et les méthodes basées sur l'optimisation combinatoire ou « Combinatorial Optimization » (CO).

2.2.1.1 Méthodes de reconstruction synthétique

Les méthodes de reconstruction synthétique sont les méthodes de synthèse de population les plus anciennes et les plus populaires. Dans leur papier, Beckman et al. (1996) étaient les premiers à appliquer l'IPF (Deming et Stephan, 1940) aux problèmes de synthèse de population à partir des données du recensement et d'un échantillon désagrégé connu sous le nom de « Public Use Microdata Sample » (PUMS) aux États-Unis. Depuis, une multitude de papiers a été publiée discutant les points faibles de cette technique et proposant des solutions, des variantes et des extensions à l'algorithme basique alors implémenté dans l'outil « Transportation Analysis Simulation System » (TRANSIMS) (Smith et al., 1995).

2.2.1.1.1 Ajustement proportionnel itératif

L'application de l'ajustement proportionnel itératif – aussi appelé calage sur marge - à un tableau de contingence est une procédure d'ajustement du contenu de ses cellules de manière que les sommes des lignes et des colonnes soient égales aux totaux exogènes fixés dans les marges. Les cellules sont donc initialisées à partir d'une base de données puis ajustées par itérations successives pour converger vers des valeurs dont les sommes en lignes et en colonnes correspondent aux totaux marginaux. Pour un tableau 2D par exemple, chaque itération a lieu en 2 étapes :

1. Ajustement des lignes :

$$c_{ij}^{t+1} = \frac{c_{ij}^t}{\sum_{j=1}^{n_c} c_{ij}^t} \times c_i. \quad (2.1)$$

2. Ajustement des colonnes :

$$c_{ij}^{t+1} = \frac{c_{ij}^t}{\sum_{i=1}^{n_r} c_{ij}^t} \times c_j \quad (2.2)$$

où n_r et n_c sont respectivement le nombre de lignes et le nombre de colonnes dans le tableau, c_{ij}^t est la valeur dans la case de la ligne i ($\forall i = 1 \dots n_r$) et la colonne j ($\forall j = 1 \dots n_c$) à la $t^{\text{ième}}$ itération et c_{ij}^{t+1} la valeur dans la même case à la $(t+1)^{\text{ième}}$ itération. c_j et c_i représentent respectivement les totaux exogènes des lignes et des colonnes avec $i = 1 \dots n_r$ et $j = 1 \dots n_c$. Ces étapes sont répétées itérativement jusqu'à convergence.

La convergence peut être définie selon plusieurs critères : une variation d'une case d'une itération à l'autre inférieure à un certain seuil ($c_{ij}^{t+1} - c_{ij}^t < \epsilon$), une différence inférieure à un seuil déterminé entre les sommes des cases en lignes et en colonnes et les marges préfixées ($\forall j = 1 \dots n_c, \sum_{i=1}^{n_r} c_{ij}^t - c_j < \epsilon$ et $\forall i = 1 \dots n_r, \sum_{j=1}^{n_c} c_{ij}^t - c_i < \epsilon$), ou simplement un nombre maximal d'itérations atteint. Ces critères peuvent éventuellement aussi être combinés. L'algorithme converge pour n'importe quel seuil choisi mais les sommes de cases en lignes et en colonnes ne vont généralement pas atteindre exactement les marges préfixées (Hunsinger, 2008).

La Figure 2.3 ci-dessous présente les étapes d'un exemple d'application de l'ajustement proportionnel itératif à un tableau 2D. Les calculs de l'itération 1 ($t=1$) relatifs à la case c_{11} sont détaillés ci-après à titre d'exemple.

1. Ajustement de ligne (c_{11} pour $t=1$) :

$$c'_{11} = \frac{c_{11}}{c_{11} + c_{12} + c_{13} + c_{14}} \times c_{1.} = \frac{7}{7 + 4 + 2 + 3} \times 35 = 15.31 \quad (2.3)$$

2. Ajustement de colonne (c_{11} pour $t=1$) :

$$c''_{11} = \frac{c'_{11}}{c'_{11} + c'_{21} + c'_{31} + c'_{41}} \times c_{.1} = \frac{15.31}{15.31 + 4.77 + 9.47 + 11.25} \times 30 = 11.26 \quad (2.4)$$

Marges exogènes					Ajustement de lignes					Ajustement de colonnes				
t=0	30.00	40.00	35.00	25.00	t=1	40.81	30.25	28.44	30.50	t=1	30.00	40.00	35.00	25.00
35.00	7.00	4.00	2.00	3.00	35.00	15.31	8.75	4.38	6.56	33.59	11.26	11.57	5.38	5.38
35.00	3.00	7.00	7.00	5.00	35.00	4.77	11.14	11.14	7.95	38.46	3.51	14.73	13.71	6.52
30.00	6.00	3.00	7.00	3.00	30.00	9.47	4.74	11.05	4.74	30.71	6.96	6.26	13.60	3.88
30.00	6.00	3.00	1.00	6.00	30.00	11.25	5.63	1.88	11.25	27.24	8.27	7.44	2.31	9.22
					t=2	30.83	39.77	33.91	25.48	t=2	30.00	40.00	35.00	25.00
					35.00	11.73	12.06	5.61	5.60	34.83	11.41	12.13	5.79	5.50
					35.00	3.19	13.40	12.47	5.93	35.28	3.11	13.48	12.87	5.82
					30.00	6.80	6.12	13.29	3.79	30.21	6.62	6.15	13.71	3.72
					30.00	9.11	8.19	2.54	10.16	29.69	8.86	8.24	2.62	9.96
					t=3	30.08	40.00	34.86	25.06	t=3	30.00	40.00	35.00	25.00
					35.00	11.47	12.19	5.82	5.53	34.98	11.44	12.19	5.84	5.51
					35.00	3.08	13.37	12.77	5.77	35.03	3.07	13.37	12.82	5.76
					30.00	6.57	6.11	13.62	3.69	30.03	6.56	6.11	13.67	3.69
					30.00	8.96	8.33	2.65	10.07	29.96	8.93	8.33	2.66	10.04
					t=4	30.01	40.00	34.98	25.01	Convergence				
					35.00	11.44	12.19	5.85	5.51	t=4	30.00	40.00	35.00	25.00
					35.00	3.07	13.36	12.81	5.76	35.00	11.44	12.19	5.85	5.51
					30.00	6.55	6.11	13.66	3.68	35.00	3.07	13.36	12.82	5.75
					30.00	8.94	8.34	2.66	10.05	30.00	6.55	6.11	13.67	3.68
										30.00	8.94	8.34	2.67	10.05

Figure 2.3 : exemple d'ajustement itératif proportionnel

L'approche est détaillée pour un tableau 2D mais demeure bien entendu facilement adaptable à un tableau nD en effectuant n ajustements – un ajustement pour chaque case par dimension – à chaque itération.

2.2.1.1.2 Approche pionnière

La reconstruction synthétique – aussi appelée approche conventionnelle (Guo et Bhat, 2007) - est l'application de l'IPF à la synthèse de population. Le premier exercice de reconstruction synthétique a été mené par Beckman et al. (1996) pour générer une population synthétique de ménages. L'approche alors proposée comportait deux étapes : l'estimation et la sélection (Templ et al., 2017). L'étape d'estimation commence par le choix des caractéristiques des ménages selon

lesquelles la population synthétique de ménages sera générée. Il s'agit des variables de contrôle du processus de synthèse. La résolution de référence est ensuite choisie et les données agrégées et désagrégées requises sont extraites. L'information spatiale disponible pour chaque ménage échantillonné dans le PUMS est son Public Use Microdata Area (PUMA) d'appartenance. Dans Beckman et al. (1996), une zone de référence correspond à un secteur de recensement et un PUMA couvre plusieurs zones de référence.

Pour n variables de contrôle, une matrice nD est établie : ses marges comportent les sommes des totaux marginaux extraits du recensement pour les zones de référence qui sont inclus dans la PUMA de l'échantillon utilisé, et ses cellules sont initialisées en utilisant la distribution conjointe des variables de contrôle dans l'échantillon. Les valeurs des cases sont alors itérativement ajustées pour correspondre aux marges par IPF. La matrice nD résultante croise les variables de contrôle entre elles à l'échelle du PUMA. Cette matrice et les matrices croisant chacune des n variables de contrôle avec les zones de référence sont utilisées comme marges pour ajuster une nouvelle matrice à $n+1$ dimensions dont les cellules sont initialisées à l'unité puisque la localisation géographique par zone de référence des ménages du PUMS n'est pas connue. Quand l'ajustement de cette deuxième matrice converge, on obtient une matrice $(n+1)D$ contenant la distribution conjointe des variables de contrôle dans la population synthétique au niveau de la résolution de référence. Ainsi, la valeur obtenue dans chacune des cellules de cette matrice correspond au nombre d'agents dans une zone de référence ayant une certaine combinaison de caractéristiques sociodémographiques consistant en une réalisation spécifique de chacune des variables de contrôle. Quand la localisation des agents est connue au niveau de la résolution de référence, seul un ajustement d'une matrice contenant la distribution conjointe des variables de contrôle dans l'échantillon contre toutes les distributions marginales – incluant la distribution spatiale ou zonale – est requis. Beckman et al. (1996) détaillent un exemple de leur procédure pour deux variables de contrôle chacune ayant deux catégories possibles pour un PUMA englobant deux secteurs de recensement.

Une fois l'effectif connu dans la population synthétique pour chaque combinaison de caractéristiques et de zone de référence, l'étape de sélection commence. Pour chacune des cases ajustées – correspondant à une combinaison de caractéristiques dans une zone de référence – chacun des ménages de l'échantillon se voit attribuer une probabilité de sélection proportionnelle à la similarité existante entre ses caractéristiques et celles de la case. Selon les probabilités calculées, des ménages sont sélectionnés de l'échantillon jusqu'à atteindre les effectifs des cellules

ajustées et une population synthétique de ménages est par conséquent obtenue. Les membres assignés à ces ménages dans l'échantillon forment une population synthétique d'individus mais celle-ci ne respecte pas les totaux du recensement pour les individus puisque l'ajustement a été effectué uniquement au niveau des ménages. L'étape de sélection peut être opérée suivant une fonction traduisant le degré de divergence ou de convergence entre les combinaisons cibles et celles de l'échantillon (Beckman et al., 1996) ou aléatoirement par une série de simulations de Monte Carlo. Diverses populations synthétiques peuvent ainsi être produites respectant toutes les contraintes imposées.

2.2.1.1.3 Critiques et variantes

La reconstruction synthétique est de loin l'approche la plus populaire dans l'exercice de la synthèse de population. Cela est essentiellement dû à sa simplicité en termes de principe et d'implémentation mais aussi à sa convergence rapide (Mosteller, 1968; Little et Wu, 1991), à son applicabilité à différents systèmes de zonage et à son aptitude à produire une population synthétique à partir d'un échantillon réduit (Choupani et Mamdoohi, 2016). De plus, l'estimation fournie par reconstruction synthétique de la population réelle est caractérisée par une entropie optimale (Ireland et Kullback, 1968) et un maximum de vraisemblance (Little et Wu, 1991) relativement à l'échantillon utilisé, dans la limite du respect des contraintes imposées. Les rapports de cotes - indicateurs mathématiques de la structure de corrélation des variables sociodémographiques utilisés pour caractériser l'échantillon - sont ainsi conservés au niveau des populations synthétiques générées par reconstruction synthétique (Mosteller, 1968; Little et Wu, 1991).

Toutefois, bien qu'assez efficace et répandue, l'approche implémentée par Beckman et al. (1996) demeure critiquée par plusieurs chercheurs qui y proposent des extensions voire des alternatives. En effet, les autres approches de synthèse de population développées dans les sections suivantes ne sont apparues que pour résoudre un problème diagnostiqué au niveau de l'approche basique de reconstruction synthétique proposée par Beckman et al. (1996). Les limitations de l'IPF soulevées dans la littérature peuvent être regroupées en dix catégories.

2.2.1.1.3.1 Contrôle à un seul niveau

L'approche basique n'est pas capable de contrôler des variables relatives à plusieurs niveaux (i.e., plusieurs types d'agents) dans le même exercice (X. Ye et al., 2009; Guo et Bhat, 2007; Pritchard

et Miller, 2012). En effet, une limitation majeure de l'approche basique de reconstruction synthétique réside dans le fait que l'ajustement puisse uniquement être assuré à un seul niveau : celui des ménages ou celui des individus. Les variables de contrôle au niveau des ménages et celles au niveau des individus donnant deux tableaux de contingence distincts, elles ne peuvent pas être prises en considération simultanément dans le processus itératif. La convergence de la reconstruction synthétique au niveau des ménages, niveau privilégié par Beckman et al. (1996), se fait aux dépens de la convergence au niveau des individus. Ainsi, des ménages ayant les mêmes caractéristiques mais des compositions d'individus à caractéristiques différentes se voient attribuer le même facteur de pondération. La négligence des distributions marginales à l'un des niveaux dans la génération de la population synthétique implique donc des divergences entre celle-ci et la population réelle.

Plusieurs extensions de l'approche basique ont été proposées pour remédier à ce problème. Guo et Bhat (2007) proposent d'intégrer à l'étape de sélection un indicateur déterminant la désirabilité d'un ménage qui elle, dépend des individus qui le forment et des contraintes considérées au niveau des individus. Un ménage est considéré désirable si le nombre de ménages du même type et les totaux d'individus des mêmes types que ceux qui le composent sont inférieurs à un seuil préfixé. Ce seuil est supérieur aux totaux marginaux aux niveaux des individus et des ménages par un pourcentage de déviation prédéterminé (Guo et Bhat, 2007). Les totaux ne sont pas utilisés comme seuils afin de laisser un certain degré de liberté à l'algorithme lui permettant d'atteindre le meilleur équilibre en termes de qualité d'ajustement des distributions d'attributs des ménages et des individus. En effet, forcer l'ajustement parfait à un niveau dégraderait la qualité d'ajustement pour l'autre. Arentze et al. (2007) proposent l'utilisation de matrices relationnelles convertissant les contraintes au niveau des individus en contraintes supplémentaires pour l'ajustement de la distribution de ménages. Pritchard et Miller (2012) proposent une modification de l'algorithme ajustant les attributs partagés entre deux niveaux (e.g., ménages et individus) à l'un, puis appliquant les résultats de l'ajustement comme contraintes pour la synthèse à l'autre. Pour éviter les divergences entre les distributions d'individus et de ménages au niveau de l'étape de sélection stochastique, ils proposent d'échantillonner l'un des deux moyennant des simulations de Monte Carlo basées sur des probabilités conditionnelles à la distribution de l'autre.

Bien qu'elles permettent de corriger l'ajustement aux distributions marginales des attributs des individus, ces extensions n'assurent pas un ajustement intrinsèquement simultané aux distributions

des ménages et des individus et ne garantissent par conséquent pas leur cohérence (Yaméogo et al., 2021). Cependant, des méthodes de synthèse multiniveau basées sur la reconstruction synthétique telles que le ratissage généralisé ou « generalized raking » (GR) (Deville, 1993), la mise à jour proportionnelle itérative ou « Iterative Proportional Updating » (IPU) (X. Ye et al., 2009), la synthèse multiniveau par optimisation de l'entropie (Bar-Gera et al., 2009) et l'ajustement proportionnel itératif hiérarchique (hIPF) (Müller et Axhausen, 2011) ont été introduites.

Ratissage généralisé

Il s'agit d'une méthode classique de pondération d'enquêtes introduite par Deville et al. (1993) qui peut être adaptée à l'exercice de synthèse de population. Assez similaire à l'IPF, le ratissage généralisé procède toutefois au calcul des poids d'une manière similaire à la minimisation d'une erreur de régression plutôt que par calage (Fournier et al., 2021). Pour une synthèse multiniveau, l'erreur est minimisée sur les contraintes correspondant aux attributs des ménages et des individus simultanément (Yaméogo et al., 2021). Une propriété intéressante de cette approche réside dans la minimisation, dans la limite des contraintes à respecter, de la distance entre les poids initiaux et les poids finaux calculés (Yaméogo et al., 2021).

Mise à jour proportionnelle itérative

La méthode IPU introduite par X. Ye et al. (2009) et implémentée dans PopGen (MARG, 2016) est un algorithme pionnier de la synthèse multiniveau. Alors que l'approche de base de reconstruction synthétique se contente d'ajuster la distribution à un seul niveau uniquement, l'IPU ajuste les distributions des ménages et des individus simultanément en recalculant les facteurs de pondération attribués aux ménages selon leur composition en individus. À la manière de l'exemple détaillé par X. Ye et al. (2009), la Figure 2.4 ci-dessous présente les étapes d'une itération d'IPU appliquée à un échantillon fictif de dix ménages.

D'abord, une tabulation d'un échantillon hiérarchique, i.e., ménages-individus, est requise. Les colonnes « Type Ménage » correspondent aux catégories des variables de contrôle au niveau des ménages. Par exemple, « Type Ménage 1 » pourrait correspondre à la catégorie « 1 personne » de la variable « taille du ménage ». De même, pour les types d'individus où la colonne « Type Personne 1 » correspondrait par exemple à la catégorie « 0-25 ans » de la variable « âge » et ainsi de suite. Chaque ménage dispose d'abord d'un poids unitaire inscrit dans les colonnes « Type Ménage » qui lui correspondent. Pour chacun des ménages de l'échantillon, le nombre de personnes

qui le composent appartenant à chacune des colonnes « Type Personne » est noté dans les colonnes correspondantes. La colonne « Poids 0 » contient les poids initiaux unitaires des ménages de l'échantillon. Les contraintes pour chacune des catégories d'attributs contrôlés sont inscrites et la « Somme pondérée 0 » des « Poids 0 » est calculée.

ID Ménage	Type Ménage 1	Type Ménage 2	Type Ménage 3	Type Personne 1	Type Personne 2	Type Personne 3	Poids 0	Poids 1	Poids 2	Poids 3	Poids 4	Poids 5	Poids 6
1	1	0	0	1	3	0	1.00	8.33	8.33	8.33	6.06	4.90	4.90
2	1	0	0	3	1	3	1.00	8.33	8.33	8.33	6.06	4.90	4.35
3	1	0	0	2	0	0	1.00	8.33	8.33	8.33	6.06	6.06	6.06
4	0	1	0	2	2	2	1.00	1.00	5.00	5.00	3.64	2.94	2.61
5	0	1	0	1	0	3	1.00	1.00	5.00	5.00	3.64	3.64	3.23
6	0	1	0	0	3	0	1.00	1.00	5.00	5.00	5.00	4.05	4.05
7	0	0	1	3	1	0	1.00	1.00	1.00	2.50	1.82	1.47	1.47
8	0	0	1	3	1	2	1.00	1.00	1.00	2.50	1.82	1.47	1.31
9	0	0	1	1	3	0	1.00	1.00	1.00	2.50	1.82	1.47	1.47
10	0	0	1	0	0	2	1.00	1.00	1.00	2.50	2.50	2.50	2.22
Contraintes	25	15	10	60	45	35							
Somme pondérée 0	3	3	4	16	14	12							
δ_b	0.88	0.80	0.60	0.73	0.69	0.66							
Somme pondérée 1	25.00	3.00	4.00	60.00	43.33	34.00							
Somme pondérée 2	25.00	15.00	4.00	72.00	63.33	54.00							
Somme pondérée 3	25.00	15.00	10.00	82.50	70.83	60.00							
Somme pondérée 4	18.18	12.27	7.95	60.00	55.61	45.00							
Somme pondérée 5	15.87	10.63	6.91	51.56	45.00	39.45							
Somme pondérée 6	15.32	9.88	6.47	48.33	43.62	35.00							
δ_a	0.39	0.34	0.35	0.19	0.03	0.00							

Figure 2.4 Exemple d'application d'une itération d'IPU

Le processus d'ajustement démarre ensuite. Il comporte deux volets principaux : l'ajustement selon les caractéristiques des ménages et l'ajustement selon les caractéristiques des personnes. L'ajustement pour un type de ménages se fait par la multiplication des poids des ménages appartenant au type considéré par le ratio contrainte-somme pondérée de ce même type. Une fois l'ajustement d'un type terminé, les sommes pondérées de tous les types de ménages et personnes sont mises à jour en fonction des nouveaux poids calculés avant de procéder à l'ajustement du type suivant. Une fois l'ajustement selon les contraintes des ménages terminé (colonne « Poids 3 »), les poids sont ajustés selon les types de personnes constituant chacun des ménages. L'ajustement selon un type de personnes se fait de la même manière que l'ajustement selon un type de ménages mais ne modifie que les poids des ménages incluant au moins un individu du type en cours d'ajustement. Les colonnes « Poids 4 », « Poids 5 » et « Poids 6 » sont ainsi calculées pour compléter la première itération et le processus itératif se poursuit.

À chaque itération, un indicateur de la qualité d'ajustement δ , consistant en la différence relative moyenne des contraintes et des sommes pondérées correspondantes, est calculé. La valeur de δ ne décroît pas constamment au cours des itérations. Ainsi, à chaque itération, les poids correspondant à la valeur minimale de δ sont conservés. La différence en valeur absolue entre les valeurs de

l'indicateur δ avant le début de l'itération (moyenne des δ_a) et après la fin de celle-ci (moyenne des δ_b), indique le gain en termes de qualité d'ajustement pour chacune des itérations. L'algorithme converge lorsque la valeur de Δ est en-dessous d'un seuil ε prédéfini donc lorsque le gain en termes de qualité d'ajustement devient minime.

Quand l'algorithme converge, l'étape de sélection est entamée. La probabilité d'un ménage d'être échantillonné dans la population synthétique est calculée par la division de son poids estimé par IPU par la somme des poids des ménages du même type eux aussi estimés par IPU. Le nombre de clones de ce ménage dans la population synthétique peut ensuite être calculé par la multiplication de la probabilité obtenue par la contrainte d'ajustement du même type de ménages. Les nombres obtenus sont arrondis aux entiers les plus proches. Cela permet de prendre en compte la composition en termes de types d'individus de chaque ménage dans le processus de sélection.

L'IPU montre une performance supérieure à celle de l'IPF (X. Ye et al., 2009). Celle-ci demeure toutefois très dépendante de la qualité des données agrégées et désagrégées fournies. En effet, toute contradiction entre les contraintes aux niveaux des individus et des ménages mènerait à une solution partielle ou « corner solution ». Ainsi, un choix entre une solution alignée sur l'ajustement parfait des ménages en gardant l'ajustement maximal possible au niveau des individus, ou une autre alignée sur l'ajustement parfait des individus en gardant l'ajustement maximal possible au niveau des ménages doit être choisie.

X. Ye et al. (2009) mentionnent une situation où leur algorithme ne converge pas. Cela arrive lorsque tous les individus d'un certain type sont inclus dans des ménages du même type. Cette situation est extrêmement improbable mais il suffira de modifier les variables de contrôle et/ou leurs catégories pour y remédier si jamais ce problème se pose.

Optimisation de l'entropie

Une approche similaire à l'IPU est développée par Bar-Gera et al. (2009) et consiste à sélectionner la solution présentant une entropie optimale et assurant le respect des distributions marginales aux niveaux des individus et des ménages simultanément. La différence essentielle entre cette approche et l'IPU réside dans le fait que l'IPU modifie de la même façon les poids des ménages comportant un certain type d'individus lors de l'ajustement de ce type alors que la méthode proposée répondra de façon plus importante les ménages comportant plus de personnes de ce type (Müller et Axhausen, 2012). Le but est de minimiser l'apport en termes de nouvelles informations et par

conséquent de garder une distribution de poids la plus proche possible de la distribution initiale dans l'échantillon tout au long du processus d'ajustement aux contraintes marginales considérées.

Ajustement proportionnel itératif hiérarchique

L'IPF hiérarchique (Müller et Axhausen, 2011) consiste en un ajustement de manière alternée entre le niveau des ménages et celui des individus. Ainsi, pour chaque itération, un ajustement est effectué au niveau des ménages puis les facteurs de pondération calculés pour chacun des ménages sont attribués aux individus qui y appartiennent. Ensuite, un ajustement est effectué au niveau des individus en utilisant les contraintes marginales des attributs relatifs aux individus et la moyenne des facteurs de pondération attribués aux individus d'un certain ménage est utilisée comme facteur de pondération du ménage pour commencer une deuxième itération d'ajustement au niveau des ménages et ainsi de suite.

Selon les expérimentations de Müller et Axhausen (2011), l'IPF hiérarchique performe mieux que les approches proposées par X. Ye et al. (2009) et Bar-Gera et al. (2009) notamment en termes d'ajustement de la distribution conjointe d'attributs. De plus, il montre une capacité à converger dans des cas où l'IPU ne converge pas (Müller et Axhausen, 2011). Une option supplémentaire pour effectuer une synthèse multiniveau serait de procéder par optimisation combinatoire tel que détaillé dans la section 2.2.1.2.

2.2.1.1.3.2 Contrôle à une seule résolution

L'approche basique ne permet pas de contrôler des variables dont les distributions marginales sont extraites à différentes résolutions géographiques dans le même exercice (Konduri et al., 2016). Or, les distributions des variables d'intérêt pourraient n'être disponibles qu'à des résolutions géographiques différentes. Ainsi, un algorithme pouvant contrôler une seule résolution géographique à la fois imposerait la négligence de variables de contrôle jugées importantes ce qui nuirait par conséquent à la qualité de la population synthétique. De plus, la capacité d'un algorithme à contrôler des variables à différentes résolutions géographiques permettrait de tenir compte des différences entre les sous-zones lorsqu'une population synthétique est générée pour une grande zone (Moreno et Moeckel, 2018). Des méthodes de synthèse multi-résolution basées sur la reconstruction synthétique telles que la mise à jour proportionnelle itérative renforcée ou

« Enhanced Iterative Proportional Updating » (EIPU) (Konduri et al., 2016) et l'IPU tri-résolution (Moreno et Moeckel, 2018) ont ainsi été introduites pour remédier à ce problème.

Mise à jour proportionnelle itérative renforcée

Konduri et al. (2016) ont développé une extension de l'IPU capable de considérer les contraintes aux niveaux des individus et des ménages à plusieurs échelles géographiques simultanément démontrant ainsi une amélioration de la qualité de la population synthétique par rapport à l'IPU de base. L'EIPU est applicable pour une synthèse où des variables de contrôle sont observées à un maximum de deux échelles géographiques. Dans les paramètres de l'algorithme, une région ou « region » correspond à la zone géographique la plus grande et l'unité géographique ou « geo » à la zone la plus petite incluse dans la région. La pondération est basée sur un processus identique à celui de la version basique de l'IPU : un poids unitaire est accordé aux ménages de l'échantillon puis les attributs régionaux des ménages, les attributs régionaux des individus, les attributs locaux des ménages, et les attributs locaux des individus sont respectivement ajustés de la même façon que pour la première version de l'IPU. L'ajustement pour les types de ménages puis les types d'individus observés au niveau de la région, se fait de la même manière décrite pour l'IPU. Les poids obtenus à l'issue de l'ajustement au niveau de la région sont utilisés comme poids de base pour l'ajustement au niveau de chaque unité géographique séparément. L'ajustement au niveau de chaque unité géographique se fait aussi tel que décrit pour l'IPU. L'algorithme de l'EIPU est implémenté dans PopGen 2.0 (MARG, 2016).

IPU tri-résolution

Moreno et Moeckel (2018) ont établi un algorithme permettant de considérer trois résolutions géographiques simultanément. Il s'agit d'une extension de l'EIPU incluant une repondération supplémentaire des poids des ménages selon les attributs observés à une troisième résolution géographique.

2.2.1.1.3.3 Cases nulles incorrectes

Un autre problème majeur de l'approche basique est celui des cases nulles incorrectes (Beckman et al., 1996; Guo et Bhat, 2007; Auld et al., 2009). En effet, l'absence d'un agent ayant une certaine combinaison de caractéristiques dans l'échantillon utilisé implique l'absence de ce type d'agents dans la population synthétique. Autrement dit, une case initialisée à 0 dans le tableau de

contingence est condamnée à garder une valeur nulle tout au long du processus d'ajustement ce qui empêcherait la convergence de l'IPF si la contrainte correspondante n'est pas nulle. Cela constitue une limitation pour ce qui est de la qualité de la population synthétique aussi car l'absence, par exemple, d'un individu de sexe masculin ayant entre 15 et 25 ans dans l'échantillon de microdonnées n'implique pas forcément l'absence d'un homme dont l'âge est inclus dans cette fourchette dans l'ensemble de la population.

Pour faire face à ce problème, plusieurs solutions ont été proposées. Un nombre maximum d'itérations peut simplement être fixé comme critère de convergence de l'IPF (Guo et Bhat, 2007). Beckman et al. (1996) proposent de remplacer les valeurs nulles par une valeur très faible de manière à permettre au processus itératif de se lancer. Toutefois, cette solution a aussi des effets indésirables sur la structure de corrélation des variables de contrôle choisies et peut par conséquent biaiser les résultats (Beckman et al., 1996). Une agrégation manuelle ou automatisée des catégories de variables de contrôle à faibles fréquences permettrait aussi de limiter les risques d'obtention de cases nulles incorrectes (Guo et Bhat, 2007; Auld et al., 2009). En effet, Auld et al. (2009) incluent un algorithme de réduction de catégories automatisant le processus d'agrégation et le rendant plus efficace. Dans cette phase du processus, une catégorie ayant un effectif inférieur à un seuil déterminé en termes de pourcentage de l'ensemble de la population, est automatiquement unie avec la catégorie adjacente ayant l'effectif le plus bas. Cette phase est terminée lorsque toutes les catégories restantes ont des effectifs supérieurs au seuil. X. Ye et al. (2009) proposent de remplacer une case nulle incorrecte par une proportion de l'effectif de la même combinaison de caractéristiques échantillonné à une résolution géographique plus agrégée. Toutefois, pour éviter toute surreprésentation, l'effectif calculé n'est utilisé que s'il est inférieur à un seuil préfixé; sinon ce seuil remplace la valeur nulle concernée.

2.2.1.1.3.4 Dépendance de la qualité des intrants et variabilité limitée

La reconstruction synthétique est une méthode fortement dépendante de la qualité des données fournies en entrée et de leur complétude (Farooq et al., 2013). Ainsi, une mauvaise qualité ou un manque d'intrants agrégés et/ou désagrégés (e.g., biais d'échantillonnage, etc.) nuirait fortement voire empêcherait la synthèse de population par reconstruction synthétique. De plus, lorsque générée par reconstruction synthétique, la variabilité des agents observés dans la population synthétique demeure limitée à celle observée dans l'échantillon de microdonnées utilisé. Ainsi, un

ménage et/ou un individu existant dans la population réelle mais qui ne figure pas dans l'échantillon utilisé ne pourra pas être synthétisé.

Malgré leur popularité limitée dans la pratique de synthèse de la population, les algorithmes ne nécessitant pas un échantillon de microdonnées peuvent être utilisés pour contourner ce problème (Gargiulo et al., 2010; Barthelemy et Toint, 2013). Toutefois, cela revient à sacrifier complètement la préservation de la structure de corrélation observée dans l'échantillon au niveau de la population synthétique.

Une autre solution serait d'utiliser des méthodes de synthèse basées sur la simulation (Farooq et al., 2013). Celles-ci sont capables de tirer profit des données agrégées et désagrégées disponibles sans dépendre de leur disponibilité et/ou de leur qualité. Les méthodes de synthèse par simulation seront détaillées dans la section 2.2.2.

2.2.1.1.3.5 Nombre limité d'attributs contrôlables

Le nombre d'attributs pouvant être ajustés par IPF est limité (Pritchard et Miller, 2012). En effet, la structure matricielle utilisée pour opérer l'algorithme classique implique une croissance exponentielle du nombre de cellules dans la matrice à chaque fois qu'une nouvelle variable de contrôle, donc une nouvelle dimension, est ajoutée. Pour réduire l'ampleur de cette matrice tout en conservant l'information qu'elle contient, Pritchard et Miller (2012) proposent une structure en liste où seules les informations figurant dans les cases non nulles sont retenues. Ainsi, grâce à la technique de stockage de données clairsemées proposée, l'algorithme modifié réussit à intégrer plus de variables sans pertes en termes de fluidité d'exécution. Cette structure de stockage de données réduit l'encombrement de la mémoire de l'ordinateur optimisant par conséquent le temps requis pour effectuer les itérations et aboutir à la convergence.

2.2.1.1.3.6 Distributions multivariées des attributs non optimales

À l'étape d'estimation de la synthèse par reconstruction synthétique, l'objectif d'optimisation est centré sur la reproduction de la distribution marginale des variables de contrôle. Ainsi, il n'y a aucune garantie que la distribution multivariée de ces variables dans la population synthétique reproduira fidèlement celle dans la population réelle (Farooq et al. 2013). Ce problème pourrait être contourné par la définition de variables de contrôle complexes, i.e., croisant plusieurs variables simples. Par exemple, au lieu de chercher à reproduire correctement le nombre de personnes entre

15 et 25 ans d'une part et le nombre d'hommes d'autre part, on pourrait chercher à reproduire le nombre d'hommes entre 15 et 25 ans directement. Cependant, lorsque le nombre des catégories de variables simples contrôlées augmente, le croisement de ces catégories mènerait à un très grand nombre de catégories de variables complexes à contrôler ce qui poserait des défis au niveau de la capacité de calcul de l'IPF. De plus, les croisements de variables de contrôle possibles demeurent limités aux distributions croisées disséminées par les instituts nationaux de la statistique.

Des méthodes de synthèse de population permettant de mieux refléter la corrélation entre les variables de contrôle en optimisant le respect de la distribution multivariée de celles-ci directement ont été développées (Farooq et al., 2013). Il s'agit des méthodes par simulation détaillées dans la section 2.2.2.

2.2.1.1.3.7 Attributs discrets uniquement

La structure matricielle de la reconstruction synthétique la rend limitée en termes de contrôle aux attributs à catégorisation discrète. Or, certaines variables de contrôle sont à la base continues.

Une solution serait de passer par une discrétisation des variables continues. Or, pour chacune des variables de contrôle, différentes catégorisations peuvent être définies menant chacune à un résultat de synthèse possiblement différent. Otani et al. (2012) proposent une méthode de sélection d'une catégorisation optimale. Ils considèrent comme optimale la catégorisation la plus compacte (i.e., qui présente un nombre de catégories minimal) menant à des résultats de simulation similaires à ceux de la catégorisation la plus désagrégée extraite du recensement. Les résultats de simulation sont comparés selon une variable de sortie clé définie en fonction des objectifs expérimentaux.

Une autre solution au problème des attributs discrets serait d'utiliser des méthodes de synthèse par simulation (section 2.2.2) capables de considérer aussi bien des variables discrètes que continues (Farooq et al., 2013).

2.2.1.1.3.8 Sensibilité aux incohérences des sommes des totaux du recensement

Les totaux des catégories des variables de contrôle, notamment à des résolutions spatiales désagrégées, subissent parfois des traitements tels que des arrondis afin d'empêcher toute possibilité d'identification des individus recensés. Ainsi, les sommes des effectifs des catégories de chacune des variables de contrôle ne coïncident souvent pas. Ce problème est aussi amplifié lorsque les variables de contrôle sont croisées et des variables complexes sont contrôlées. Malgré

l'amplitude généralement faible des différences, celles-ci pourraient significativement nuire à la qualité de l'optimisation mathématique par IPF (Moreno et Moeckel, 2018; Konduri et al., 2016; Barthelemy et Toint, 2013; Harland et al., 2012). Pour remédier à cela, Rich et Mulalic (2012) proposent une variante de l'approche basique de reconstruction synthétique où les marginaux sont rangés par ordre de priorité puis linéairement ajustés de manière à respecter les totaux des variables qui les dominent.

2.2.1.1.3.9 Non garantie de la correspondance de la solution à la réalité

À l'étape d'estimation, les méthodes de reconstruction synthétique cherchent à reproduire les totaux observés dans le recensement en calculant un poids pour chacun des agents figurant dans l'échantillon de microdonnées. Toutefois, il est généralement possible de respecter les contraintes du recensement par différents ensembles de poids estimés. Ainsi, un ensemble de poids respectant ces contraintes ne correspond pas forcément à la réalité (Zhuge et al., 2017). Ce problème est aussi rencontré au niveau de l'étape de sélection lorsque celle-ci est effectuée moyennant des simulations de Monte Carlo basées sur les poids estimés. Les erreurs potentielles intrinsèques aux processus d'échantillonnage aléatoire mènent à une variance importante des populations qui en résultent et les rendent ainsi individuellement moins fiables (Huang et Williamson, 2001).

Zhugue et al. (2017) proposent « Pop-H », une méthode heuristique de synthèse de population qui considère la fréquence des agents dans l'échantillon utilisé comme représentative de la réalité et ajuste par conséquent l'échantillon aux contraintes du recensement en appliquant les altérations minimales possibles aux fréquences initiales. Le poids final ajusté de chacun des ménages de l'échantillon demeure ainsi dans les limites d'un intervalle d'ajustement précalculé. De plus, l'approche proposée ne se limite pas à minimiser la moyenne des différences entre les totaux générés et les contraintes marginales imposées mais considère aussi la minimisation de l'écart-type de ces différences dans le processus d'optimisation. Ceci permet d'éviter qu'un écart inacceptablement important pour une contrainte ne soit masqué par des écarts infimes sur les autres contraintes (Zhuge et al., 2017). L'objectif d'optimisation plus complexe limite ainsi le nombre d'ensembles de poids permettant d'atteindre la valeur optimale de la fonction d'optimisation.

2.2.1.1.3.10 Valeurs décimales

Les valeurs dans les cases de la matrice ajustée à l'issue de l'étape d'estimation ne sont pas entières. Par conséquent, un problème d'arrondis pourrait s'imposer avant l'étape de sélection (Frick et Axhausen, 2004). Arrondir simplement à l'entier le plus proche altère la structure de corrélation dans la distribution obtenue et mène à une violation des contraintes marginales utilisées pour l'ajustement. Beckman et al. (1996) proposent d'arrondir les nombres trouvés puis d'ajuster le total au nombre total de ménages par zone en ajoutant ou supprimant quelques ménages. Frick et Axhausen (2004) proposent que l'élimination ou l'ajout des ménages se fasse de manière à réduire l'écart entre les totaux obtenus pour les variables de contrôle et les contraintes marginales préfixées. Pour arrondir les facteurs de pondération, X. Ye et al. (2009) proposent d'établir une comparaison entre les poids des ménages obtenus par IPU et ceux qui seraient obtenus par IPF pour une synthèse au niveau des ménages uniquement. La somme des poids obtenus par IPU étant constamment inférieure au total des ménages dans la zone d'étude, un ménage est ajouté aux cases qui divergent le plus des résultats IPF jusqu'à atteindre le total de ménages dans la zone considérée. Lovelace et Ballas (2013) comparent cinq méthodes de conversion des valeurs décimales estimées par IPF en valeurs entières : l'approche par arrondis, l'approche du seuil, l'approche compteur-poids ou « counter-weight » (Ballas et al., 2005), l'approche des probabilités proportionnelles, et la méthode « Tronquer, Répliquer, Échantillonner » ou « Truncate, Replicate, Sample » (TRS) (Lovelace et Ballas, 2013).

L'approche des arrondis consiste à simplement arrondir les poids estimés pour chacune des combinaisons de caractéristiques à l'entier le plus proche. Ballas et al. (2005) montrent que la distribution de proportions obtenues par IPF est asymétrique et que la majorité des proportions sont généralement inférieures au seuil de 0.5 ce qui est susceptible de fortement nuire aux résultats.

Pour l'approche du seuil (Ballas et al., 2005), un seuil d'inclusion est initialisé à « 1 » et itérativement réduit de « 0.0001 ». À chaque itération, les ménages ayant des proportions dont la partie décimale est supérieure à la valeur atteinte par le seuil sont échantillonnés. Le processus est arrêté lorsque la taille de la population synthétique devient égale ou supérieure à celle de la population réelle. Cette méthode a tendance à générer un surplus d'agents synthétiques et cela principalement à cause des proportions égales ou dont la différence est inférieure au pas de diminution de seuil utilisé (Lovelace et Ballas, 2013).

L'approche compteur-poids (Ballas et al., 2005) consiste à classer les ménages par ordre croissant de leurs proportions estimées par IPF. Ensuite, deux variables sont initialisées à « 0 » : le compteur et le poids. L'algorithme passe la liste des ménages itérativement. À chaque itération le compteur est incrémenté d'une unité et le poids est cumulé. Lorsque le poids cumulé dépasse « 1 », le ménage atteint hérite de la valeur arrondie du poids cumulé atteint et celle-ci est ensuite soustraite du poids cumulé. Le processus continue jusqu'à ce que le compteur atteigne le nombre total de ménages.

L'approche des probabilités proportionnelles (Lovelace et Ballas, 2013) consiste à considérer les proportions estimées des ménages par IPF comme des probabilités d'échantillonner ces ménages jusqu'à ce que la population synthétique atteigne la taille de la population réelle. Le caractère probabiliste de l'approche fait qu'il existe des chances qu'un ménage ayant une proportion très faible à l'issue de l'étape d'estimation se trouve échantillonné plus de fois qu'un ménage ayant une proportion largement plus élevée (Lovelace et Ballas, 2013).

Afin de remédier aux problèmes des autres méthodes, Lovelace et Ballas (2013) introduisent l'approche TRS. Celle-ci consiste d'abord à répliquer les ménages selon la troncature entière de leurs proportions calculées par IPF. La partie décimale correspondant à chacun des ménages est ensuite utilisée comme probabilité d'échantillonner ce ménage. Un échantillonnage probabiliste est ainsi lancé jusqu'à ce que la taille de la population synthétique atteigne la taille de la population réelle. Parmi les cinq approches testées, l'approche TRS a permis de réaliser la conversion en poids entiers la plus optimale (Lovelace et Ballas, 2013).

Une manière d'éviter le problème de la conversion des proportions décimales estimées en entiers est l'utilisation des méthodes basées sur l'optimisation combinatoire détaillées dans la section 2.2.1.2 ou des méthodes par simulation détaillées dans la section 2.2.2.

2.2.1.2 Méthodes d'optimisation combinatoire

Williamson et al. (1998) proposent l'optimisation combinatoire comme alternative à la reconstruction synthétique. Au lieu des tableaux de contingence sur lesquels sont basées les méthodes de reconstruction synthétique, les méthodes par optimisation combinatoire utilisent une structure en liste considérée plus flexible et efficace en termes de stockage et de manipulation.

2.2.1.2.1 Approche pionnière

Pour une population de « n » ménages, le principe de base est de tester toutes les combinaisons possibles de « n » ménages tirés avec répétitions de l'échantillon de microdonnées disponible. Pour chacune des combinaisons, un indicateur de qualité mesurant l'ajustement de celle-ci aux totaux du recensement pour la zone et les variables de contrôle considérées est calculé (Voas et Williamson, 2001). La combinaison présentant le meilleur indicateur d'ajustement constituerait alors la population synthétique retenue. Or, le grand nombre de combinaisons candidates rend une telle recherche exhaustive pratiquement impossible (Williamson et al., 1998; Huang et Williamson, 2001). Ainsi, dans le but d'effectuer une recherche sélective de la combinaison optimale, plusieurs méthodes d'optimisation combinatoire sont proposées dont la méthode de descente ou « hill climbing », le recuit simulé ou « simulated annealing » et les algorithmes génétiques (Williamson et al., 1998).

2.2.1.2.1.1 Méthode de descente ou « hill climbing »

Lorsque la méthode de descente est utilisée pour la génération d'une population synthétique par CO, une liste de « n » ménages est d'abord aléatoirement tirée de l'échantillon de microdonnées avec répétitions, « n » étant le nombre total de ménages dans la population. Un ménage aléatoirement choisi dans la liste initiale établie est ensuite échangé avec un ménage aléatoirement choisi dans l'échantillon de microdonnées et la qualité d'ajustement de la nouvelle combinaison est mesurée. Si l'ajustement est amélioré, le remplacement est retenu, sinon, la liste avant remplacement est gardée (Yaméogo et al., 2021; Voas et Williamson, 2000). Ce processus est itérativement répété jusqu'à l'atteinte d'un critère d'arrêt ou de convergence prédéfini. Il pourrait s'agir simplement d'un nombre maximal d'itérations, d'un seuil minimal d'ajustement ou d'un seuil minimal d'amélioration de l'ajustement d'une itération à l'autre. Williamson et al. (1998) ont testé une variante de la méthode de descente. En effet, au lieu de sélectionner aléatoirement un ménage de la liste pour être remplacé, le ménage qui contribue le plus à la divergence entre la combinaison considérée et les observations du recensement est considéré pour remplacement. Toutefois, la version classique (i.e., avec sélection aléatoire) s'est avérée plus performante (Williamson et al., 1998). Le défaut principal de la méthode de descente est qu'elle peut facilement être piégée dans un minimum local (Williamson et al., 1998) puisqu'un remplacement ne peut être effectué que s'il apporte une amélioration en termes d'ajustement. En effet, une détérioration

temporaire de l'ajustement qui serait nécessaire pour ensuite y apporter de meilleures améliorations n'est pas permise par l'algorithme.

2.2.1.2.1.2 Le recuit simulé ou « simulated annealing »

Contrairement à la méthode de descente, le recuit simulé est une méthode d'optimisation permettant un remplacement menant à une détérioration de l'ajustement dans sa quête de la combinaison optimale (Williamson et al., 1998; Harland et al., 2012). L'amplitude des détériorations acceptées demeure toutefois limitée et elle décroît au cours du processus. Par exemple, un certain pourcentage de réduction peut y être appliqué à chaque fois qu'un certain nombre de remplacements sont réalisés avec succès. Il est intéressant de signaler que plus l'amplitude de la détérioration acceptée est réduite, plus l'algorithme tend à performer de manière similaire à la méthode de descente. Le défi principal rencontré lors de l'utilisation du recuit simulé est le paramétrage de la méthode notamment en termes du seuil de détérioration initialement accepté, de l'amplitude de la réduction graduellement appliquée à celui-ci et du rythme auquel celle-ci est appliquée (Williamson et al., 1998). Ainsi, une analyse de sensibilité pour choisir un paramétrage optimal constitue un prérequis incontournable à l'utilisation du recuit simulé (Williamson et al., 1998).

2.2.1.2.1.3 Les algorithmes génétiques

Une autre technique d'optimisation combinatoire est celle basée sur les algorithmes génétiques (Williamson et al., 1998; Birkin et al., 2006). Dans le problème d'optimisation de la combinaison de ménages retenue, un chromosome correspond à une combinaison de « n » ménages, « n » étant le nombre de ménages dont est constituée la population. L'idée est d'abord d'établir plusieurs chromosomes et de calculer une fonction d'évaluation ou « fitness function » pour chacun des chromosomes afin de les classer par ordre décroissant de leur évaluation. Ensuite, la deuxième génération est établie (Williamson et al., 1998). Elle est composée en partie des « p » meilleurs chromosomes de la génération initiale, où « p » est un pourcentage prédéfini. Le reste de la deuxième génération est dérivé de la génération initiale selon un processus itératif. En effet, à chaque itération, un nombre prédéfini de chromosomes de la génération initiale est aléatoirement sélectionné par simulations de Monte Carlo basées sur les valeurs des fonctions d'évaluation établies (Williamson et al., 1998). Ces chromosomes subissent des opérations de croisement et de mutation pour en dériver un nombre équivalent de chromosomes « enfants » qui sont par

conséquent intégrés à la deuxième génération. Le processus itératif est répété jusqu'à ce que la deuxième génération atteigne la taille de la génération initiale. Une troisième génération est par la suite établie de la même manière à partir de la deuxième génération et ainsi de suite. Lorsqu'un nombre prédéfini de générations est atteint, le meilleur chromosome de la dernière génération est sélectionné. Il constitue la population synthétique. Comme pour le recuit simulé, les algorithmes génétiques présentent des défis de calibration (Williamson et al., 1998).

La comparaison menée par Williamson et al. (1998) de ces trois méthodes d'optimisation combinatoire a montré que le recuit simulé permet généralement de meilleurs ajustements notamment grâce à sa capacité à éviter d'être piégé dans des minima locaux (Williamson et al., 1998). Le produit final des différentes méthodes d'optimisation combinatoire détaillées consiste en une liste ajustée d'agents à poids entiers unitaires directement générée à partir des données entrantes.

2.2.1.2.1.4 La synthèse basée sur les valeurs d'ajustement ou « Fitness-Based Synthesis » (FBS)

Srinivasan et al. (2008) utilisent un algorithme glouton ou « greedy heuristic » pour résoudre le problème d'optimisation combinatoire. Il s'agit d'une approche itérative où à chaque itération la contribution d'un ménage de l'échantillon au respect des contraintes aux niveaux des ménages et des individus simultanément est d'abord mesurée par une valeur unique appelée « valeur d'ajustement » ou « fitness value » (Srinivasan et al., 2008). Le ménage dont la valeur d'ajustement est la plus élevée est ensuite ajouté à la liste de la population synthétique, les valeurs d'ajustement sont recalculées et ainsi de suite. Le processus s'arrête lorsque le nombre total de ménages est atteint ou lorsqu'il ne reste que des ménages à valeurs d'ajustement négatives à sélectionner.

Ma et Srinivasan (2015) introduisent une formulation améliorée de l'algorithme proposé par Srinivasan et al. (2008). En effet, la valeur d'ajustement unique est cette fois-ci remplacée par deux valeurs d'ajustement. L'une quantifie l'apport d'un ajout d'un ménage de l'échantillon de microdonnées à la population synthétique et l'autre quantifie l'apport d'un retrait de ce même ménage de la population synthétique s'il y a été préalablement ajouté bien évidemment. La formulation des deux valeurs d'ajustement est faite de manière qu'elles ne peuvent pas être simultanément positives (Ma et Srinivasan, 2015). Les ménages de l'échantillon de microdonnées dont la valeur d'ajustement d'ajout est positive et ceux dont la valeur d'ajustement de retrait est positive et qui sont déjà intégrés à la population synthétique forment un ensemble de candidats. Un

ménage est aléatoirement sélectionné dans cet ensemble puis ajouté ou retiré de la population synthétique selon ses valeurs d'ajustement. Les valeurs d'ajustement sont ensuite actualisées, la prochaine sélection aléatoire est lancée et ainsi de suite. Le processus est arrêté lorsqu'aucun ménage de l'échantillon de microdonnées ne montre au moins une valeur d'ajustement positive (Ma et Srinivasan, 2015).

2.2.1.2.2 *Critiques et variantes*

La synthèse par optimisation combinatoire présente plusieurs avantages comparativement à celle par reconstruction synthétique. Par sa structure en listes, la synthèse par CO évite le problème des cases nulles incorrectes inhérentes aux tableaux de contingence de la reconstruction synthétique. L'ajustement des populations synthétiques produites par CO aux totaux du recensement est aussi généralement supérieur à celui de leurs homologues produites par reconstruction synthétique (Huang et Williamson, 2001; Ryan et al., 2009). La variance des populations synthétiques produites par CO est plus faible que celles produites par reconstruction synthétique via des simulations Monte Carlo ce qui rend une population produite par CO plus fiable (Huang et Williamson, 2001; Ryan et al., 2009). De plus, contrairement à la reconstruction synthétique nécessitant une étape de conversion des poids décimaux estimés en nombres entiers (section 2.2.1.1.3.10), l'optimisation combinatoire est une méthode de résolution aboutissant directement à des agents à poids entiers. Aucune altération des poids n'est ainsi requise ce qui préserve la qualité de la population synthétique en termes de respect des totaux observés dans le recensement (Williamson et al., 1998). La CO présente aussi une sensibilité moins prononcée aux incohérences des marginaux extraits du recensement (Huang et Williamson, 2001). En effet, une telle incohérence se traduit par l'inexistence d'une solution optimale permettant de respecter toutes les contraintes d'ajustement simultanément. Cela n'empêcherait toutefois pas la convergence de l'algorithme vers la meilleure solution possible. Enfin, alors que l'approche de base de la reconstruction synthétique ne peut contrôler qu'un niveau (i.e., un type d'agents), l'approche de base de la CO peut être exécutée sur une liste des ménages et des individus qui les composent (Huang et Williamson, 2001). Ainsi, la CO est par définition une méthode de synthèse multiniveau.

Malgré ses nombreux avantages, l'optimisation combinatoire présente plusieurs limitations dont certaines communes avec l'approche de base de la reconstruction synthétique. Il s'agit par exemple du problème de la capacité de l'algorithme à contrôler des attributs observés à une seule résolution

géographique à la fois. Ce problème est détaillé dans la section 2.2.1.1.3.2. Abraham et al. (2012) proposent une extension de la méthode de descente permettant de résoudre un problème d'optimisation combinatoire multi-résolution. Un autre exemple de limitation commune à la reconstruction synthétique et la CO est celui de dépendance de la qualité de la population synthétique générée vis-à-vis de la qualité des intrants, et la variabilité des agents synthétiques limitée à celle observée dans l'échantillon. Cette limitation est détaillée dans la section 2.2.1.1.3.4. La nature univariée de l'optimisation qui se fait aux dépens de la distribution multivariée des variables de contrôle (section 2.2.1.1.3.6) et l'impossibilité de contrôler des variables continues sans passer par une discrétisation (section 2.2.1.1.3.7) sont aussi des problèmes communs aux deux familles de méthodes de synthèse par ajustement. Enfin, le nombre d'attributs et le détail des catégories d'attributs pouvant être considérés sont limités pour la CO comme pour la reconstruction synthétique. Les raisons sont toutefois différentes. Alors que le problème de la multiplicité des contraintes limite la reconstruction synthétique sur l'aspect du stockage de la matrice, il entrave la CO sur l'aspect des ressources et du temps de calcul requis. Ce problème d'inefficacité ainsi que la perte de la structure de corrélation des attributs observée au niveau de l'échantillon ont contribué à limiter la popularité de la CO par rapport à celle de la reconstruction synthétique. Ces deux limitations spécifiques à la CO sont détaillées ci-dessous.

2.2.1.2.2.1 Inefficacité

Le problème principal de la CO est son inefficacité. En effet, comparée à la reconstruction synthétique, l'optimisation combinatoire s'avère extrêmement chronophage (Huang et Williamson, 2001; Lee et Fu, 2011; Pritchard et Miller, 2012; Yaméogo et al., 2021). Elle est ainsi généralement appliquée dans le cadre d'un exercice de synthèse de taille limitée (Huang et Williamson, 2001; Ryan et al., 2009 ; Lee et Fu, 2011). De plus, la convergence de l'algorithme n'est souvent pas attendue compte tenu de la durée requise pour l'atteindre, et le processus est par conséquent tronqué au bout d'un nombre prédéterminé d'itérations. Ainsi, ce problème d'inefficacité empêche pratiquement la mise à profit du plein potentiel de l'algorithme, notamment sur un exercice de synthèse à grande échelle, et la solution acceptée est la meilleure pouvant être atteinte dans un certain intervalle de temps et non la solution optimale garantie (Huang et Williamson, 2001).

2.2.1.2.2.2 Structure de corrélation observée non préservée

Contrairement à la reconstruction synthétique, la CO ne permet pas de préserver la structure de corrélation observée entre les attributs contrôlés au niveau de l'échantillon (Pritchard et Miller, 2012). En effet, la structure de corrélation des variables de contrôle au sein de l'échantillon est assez bien préservée au niveau de la liste initiale lorsque le tirage aléatoire effectué est basé sur les poids des ménages dans l'échantillon. Toutefois, les altérations subséquentes de cette liste ne tiennent aucunement compte de cette structure et ne peuvent ainsi pas la préserver. Ma et Srinivasan (2015) affirment que leur approche basée sur les valeurs d'ajustement, où la sélection d'un ménage candidat se fait aléatoirement plutôt que selon la valeur d'ajustement maximale (Srinivasan et al., 2008), permet de mieux reproduire la structure de corrélation observée entre les variables de contrôle au niveau de l'échantillon dans la population synthétique.

2.2.2 Méthodes basées sur la simulation

Compte tenu de la vulnérabilité des approches basées sur l'ajustement vis-à-vis de la disponibilité et de la qualité des données entrantes et de leur faiblesse dans la reproduction de la distribution multivariée des variables de contrôle entre autres, Farooq et al. (2013) proposent une méthode de synthèse de nature différente : il s'agit d'une approche basée sur la simulation ou « simulation-based approach », aussi appelée approche probabiliste.

2.2.2.1 Approche pionnière

L'approche par simulation proposée par Farooq et al. (2013) consiste à aborder les variables de contrôle comme des variables aléatoires et à échantillonner des agents, i.e., des combinaisons de réalisations de ces variables, à partir de la distribution conjointe de celles-ci dans la population réelle. Or, les données désagrégées décrivant l'ensemble de la population réelle n'étant généralement pas disponibles et les données disponibles étant limitées et sujettes à maintes approches d'anonymisation, la distribution de probabilités multivariée des variables de contrôle dans la population réelle demeure non accessible aux chercheurs (Sun et Erath, 2015). Il s'avère ainsi complexe voire impossible d'échantillonner directement à partir de celle-ci (Farooq et al., 2013; Sun et Erath, 2015).

Pour reproduire une estimation de la distribution multivariée des variables de contrôle, Farooq et al. (2013) proposent de passer par l'établissement des distributions de probabilités de chacune de ces variables conditionnelles aux autres variables de contrôle. Ces distributions de probabilités

conditionnelles peuvent généralement être extraites des échantillons désagrégés où les variables d'intérêt sont recensées. Toutefois, la taille des échantillons désagrégés accessibles étant généralement limitée, il s'avère impossible d'en extraire la totalité des probabilités conditionnelles requises. Ce problème s'amplifie lorsque le nombre de catégories des attributs contrôlés augmente. En effet, cette augmentation engendre une augmentation exponentielle du nombre de combinaisons de ces catégories rendant la probabilité de ne pas toutes les observer dans l'échantillon désagrégé plus importante. Farooq et al. (2013) proposent deux solutions à ce problème. La première consiste à établir les probabilités conditionnelles manquantes en utilisant des modèles paramétriques de choix discrets ou continus entraînés sur une combinaison des échantillons désagrégés et des distributions marginales du recensement. Si les distributions de probabilités conditionnelles ne sont toujours pas complètes, la deuxième solution consiste à émettre des hypothèses basées sur des liens logiques tels qu'une probabilité nulle d'avoir un diplôme universitaire sachant qu'un individu a deux ans, ou sur l'indépendance d'une variable par rapport à une autre (Farooq et al., 2013). En effet, procéder avec des hypothèses d'indépendance entre des variables ou des combinaisons de variables de contrôle revient à procéder avec des distributions partielles ou incomplètes de probabilités conditionnelles (Sun et Erath, 2015).

Une fois les probabilités conditionnelles établies, Farooq et al. (2013) proposent d'utiliser un échantillonneur de Gibbs pour générer une population synthétique à partir de celles-ci. Un échantillonneur de Gibbs est une méthode de simulations de Monte Carlo par chaînes de Markov ou « Markov Chain Monte Carlo » (MCMC) qui permet d'échantillonner à partir des probabilités conditionnelles reliant plusieurs variables aléatoires. Il est particulièrement utilisé lorsque l'échantillonnage ne peut se faire directement à partir de la distribution multivariée des variables d'intérêt. Ainsi, une fois les distributions conditionnelles des variables de contrôle prêtes, l'échantillonneur de Gibbs est démarré. Les réalisations de l'échantillonneur de Gibbs dépendent chacune de celle qui la précède. Cette propriété implique deux mesures à prendre lors de la sélection des agents synthétiques parmi ceux échantillonnés par l'échantillonneur de Gibbs : attendre que la période d'échauffement ou « burn-in period » s'écoule et appliquer un amincissement ou « thinning ». La période d'échauffement est la période où les réalisations de l'échantillonneur sont supposées être influencées par la combinaison de départ utilisée. Ces réalisations sont par conséquent ignorées puisque la combinaison de départ pourrait correspondre à un type d'agents très peu probable dans la population. Lorsque l'échantillonneur atteint une distribution stationnaire,

une infinité d'agents synthétiques relevant d'une estimation adéquate de la distribution multivariée de probabilités des variables de contrôle peut être échantillonnée. Différentes populations synthétiques peuvent alors être générées en sélectionnant le nombre requis d'agents parmi ceux échantillonnés (Farooq et al., 2013). Généralement, un certain nombre d'agents échantillonnés est ignoré entre les agents sélectionnés pour éviter une corrélation importante entre ces derniers (Farooq et al., 2013). En effet, la dépendance entre les réalisations de l'échantillonneur de Gibbs implique, après avoir dépassé la période d'échauffement, de définir un saut de réalisations pour récupérer un agent synthétique. Deux agents successifs seraient quasiment une copie l'un de l'autre. Par conséquent, pour garder une certaine hétérogénéité dans la population, un agent est récupéré à chaque fois qu'un certain nombre de réalisations passe. C'est ce que l'on appelle l'amincissement.

2.2.2.2 Critiques et variantes

L'approche MCMC détaillée ci-dessus présente de nombreux avantages par rapport aux méthodes basées sur l'ajustement. D'abord par sa formulation probabiliste, le problème des cases nulles incorrectes inhérentes aux tableaux de contingence de la reconstruction synthétique est évité (Casati et al., 2015). Ensuite, elle requiert peu de ressources en termes de mémoire et de calculs ce qui, outre la rendre pratique à utiliser, lui permet de tirer profit des distributions conditionnelles de probabilités disponibles en plus des distributions marginales (Farooq et al., 2013). Tel qu'expliqué précédemment, les versions de base peu efficaces des méthodes par ajustement sont pratiquement limitées aux distributions marginales puisque l'intégration de distributions conditionnelles est susceptible d'augmenter drastiquement la dimension de l'exercice. La méthode MCMC est ainsi capable de mieux reproduire la distribution multivariée des variables de contrôle (Farooq et al., 2013). Un autre avantage de l'approche probabiliste réside dans sa capacité à contrôler, en plus des variables discrètes, des variables continues (Farooq et al., 2013). Cette approche présente aussi une flexibilité en termes d'intrants. En effet, l'incomplétude des distributions marginales et/ou conditionnelles des attributs contrôlés n'empêchent pas son utilisation. Celles-ci peuvent être extraites ou être établies à partir de modèles de choix discrets et/ou continus en utilisant les différentes bases de données agrégées et désagrégées disponibles. Des hypothèses d'indépendance ou autres peuvent aussi être établies pour compléter les distributions de probabilités conditionnelles (Farooq et al., 2013). Enfin, un avantage majeur de la synthèse par simulation réside dans sa

capacité à échantillonner des agents qui ne figurent pas dans l'échantillon désagrégé utilisé. Les populations synthétiques générées par simulation présentent par conséquent une hétérogénéité supérieure à celle des populations synthétiques générées par ajustement constituées, elles, de clones des agents observés dans l'échantillon de microdonnées.

La méthode de synthèse proposée par Farooq et al. (2013) présente toutefois quelques limitations. En effet, elle présente d'une part des problèmes communs avec d'autres méthodes tels que le contrôle limité à un seul niveau, diagnostiqué aussi pour l'approche de base de reconstruction synthétique, et le contrôle limité à une seule résolution géographique caractérisant les approches de base de la reconstruction synthétique et de la CO. D'autre part, des défis spécifiques à la synthèse par simulation sont identifiés dont les risques de sous-ajustement ou « underfitting » liés aux probabilités conditionnelles modélisées et/ou supposées, les risques de surajustement ou « overfitting » à l'échantillon désagrégé, l'absence de garantie du respect des contraintes marginales, les défis liés à l'interdépendance des réalisations de l'échantillonneur et la possibilité d'échantillonnages incorrects. Ces défis et les variantes et solutions proposées pour y remédier, lorsqu'elles existent, sont détaillés ci-dessous.

2.2.2.2.1.1 Contrôle à un seul niveau

Une limitation majeure de l'approche introduite basée sur la simulation décrite par Farooq et al. (2013) réside dans son incapacité à contrôler les attributs de plusieurs niveaux simultanément (Casati et al., 2015). Cela limite la qualité et par conséquent l'utilité de la population synthétique générée puisque les variables expliquant les comportements de mobilité sont observées à plusieurs niveaux (e.g., ménages, individus, etc.).

Casati et al. (2015) introduisent une extension de la méthode MCMC permettant de contrôler des attributs aux niveaux des ménages et des individus simultanément. Pour ce faire, les agents observés dans l'échantillon désagrégé sont d'abord typifiés selon une hiérarchie qui structure le ménage et l'échantillon désagrégé est segmenté selon ces types. Dans l'exemple traité par Casati et al. (2015), trois types ont été utilisés : les propriétaires ou « owners », les intermédiaires (conjoint, enfant), et les autres. La typification se fait selon des règles logiques prédéfinies comme considérer l'individu le plus âgé et au revenu le plus haut comme propriétaire, l'individu de sexe opposé présentant l'écart d'âge le plus réduit avec le propriétaire comme conjoint, etc. Une autre étape préliminaire consiste à faire hériter les propriétaires des caractéristiques de leurs ménages

respectifs tels que la taille du ménage, le type de l'unité de logement et l'ethnicité des habitants du ménage.

Le processus de synthèse peut alors être lancé. Une approche MCMC classique (Farooq et al., 2013) est utilisée pour échantillonner les propriétaires en considérant l'ensemble de leurs attributs, i.e., leurs attributs individuels et ceux dont ils ont hérité caractérisant leurs ménages. Si le ménage est composé de plus d'un individu, l'échantillonnage des autres types est entamé. Les distributions de probabilités conditionnelles de chaque type d'individus sont calculées en considérant en plus de ses attributs, celles des types supérieurs auxquels il se rattache. Ainsi, les distributions de probabilités conditionnelles des conjoints considèrent leurs caractéristiques individuelles et les caractéristiques des propriétaires de leurs ménages respectifs incluant celles des ménages eux-mêmes. Les distributions de probabilités conditionnelles des enfants considèrent en plus de leurs caractéristiques celles des conjoints et des propriétaires incluant celles des ménages, et ainsi de suite. Des individus hiérarchisés et par conséquent regroupés en ménages sont alors synthétisés.

Sun et Erath (2015) proposent une application hiérarchique similaire du réseau bayésien permettant de générer une population synthétique d'individus regroupés en ménages. Sun et al. (2018) introduisent un modèle hiérarchique comprenant trois composantes : une factorisation tensorielle probabiliste ou « probabilistic tensor factorization » (Sun et Axhausen, 2016), un modèle multiniveau de classes latentes ou « multilevel latent class model » (Vermunt, 2003 ; 2008), et un échantillonnage par la méthode de rejet ou « rejection sampling ». L'idée est de d'abord capturer la distribution multivariée des variables de contrôle à chacun des niveaux (ménages et individus) par la factorisation tensorielle probabiliste, puis de capturer les interactions entre les deux niveaux par le modèle multiniveau de classes latentes (Sun et al., 2018). La distribution multivariée multiniveau de probabilités est alors estimée et un échantillonnage par la méthode de rejet basé sur celle-ci est mené. Cette méthode d'échantillonnage permet de préserver les relations entre les individus formant un ménage dans la population synthétique (Sun et al., 2018). Récemment, Borysov et al. (2019) ont introduit un algorithme de synthèse basé sur un auto-encodeur variationnel ou « Variational Autoencoder » (VAE) qui est capable d'appréhender la distribution multivariée des attributs contrôlés aux niveaux des ménages et des individus simultanément.

2.2.2.2.1.2 Contrôle à une seule résolution

Une autre limitation importante de l'approche introduite basée sur la simulation décrite par Farooq et al. (2013) réside dans son incapacité à exercer un contrôle simultané des attributs observés à différentes résolutions géographiques. Cela limite la qualité de la population synthétique tel que détaillé dans la section 2.2.1.1.3.2.

2.2.2.2.1.3 Risques de sous-ajustement

Comme expliqué précédemment, la quantité limitée de données désagrégées décrivant la population ne permet généralement pas d'en extraire l'ensemble des probabilités conditionnelles requises pour reconstruire la distribution multivariée des variables de contrôle. Cela introduirait des risques de sous-ajustement du modèle. Ainsi, deux solutions sont proposées. La première consiste à utiliser des modèles de choix discrets et/ou continus pour établir les distributions de probabilités conditionnelles manquantes. La seconde serait d'émettre des hypothèses logiques et/ou simplificatrices pour atteindre le même objectif. Cela reviendrait à procéder avec des distributions conditionnelles partielles ou incomplètes. Or, ces deux solutions proposées par Farooq et al. (2013) pour combler le manque en termes de probabilités conditionnelles présentent chacune des défis. En effet l'utilisation des modèles paramétriques pourrait poser des défis en termes de calculs. Il s'agit de réussir à estimer des modèles de qualité notamment lorsque le nombre de catégories de variables de contrôle augmente, que le nombre d'observations est limité et que le nombre de paramètres à estimer est par conséquent trop important (Sun et Erath, 2015). Pour ce qui est de l'utilisation des hypothèses logiques et simplificatrices, cela requiert une connaissance du domaine et des particularités de l'environnement d'application dont tout le monde ne disposerait pas forcément. Par exemple, pour utiliser des probabilités conditionnelles partielles, il faut avoir une bonne idée des interactions entre les variables de contrôle pour éviter de rendre une variable indépendante par rapport à une autre qui en réalité en est fortement déterminante (Sun et Erath, 2015). Le sous-ajustement d'un modèle MCMC pourrait aussi provenir d'une incohérence des distributions de probabilités conditionnelles qui empêcherait la convergence de la chaîne Markov vers une distribution stationnaire de probabilités. Il en résulterait des biais au niveau de la population synthétique (Saadi et al., 2018).

Pour éviter ces problèmes entre autres, Sun et Erath (2015) proposent d'utiliser un réseau bayésien pour établir une estimation de la distribution conjointe des variables de contrôle dans la population réelle. Il s'agit d'un modèle graphique constitué de nœuds, représentant les variables d'intérêt, et

de liens, représentant les interactions entre les nœuds, permettant de coder les relations de causalité ou de dépendance entre les variables de contrôle (Sun et Erath, 2015). L'origine d'un lien représente la variable « parent » et son extrémité la variable « enfant ». Pour chacun des nœuds, une distribution de probabilités conditionnelles est calculée. Les auteurs démontrent que la distribution de probabilités d'un nœud dépend de ce qu'ils s'appellent sa couverture Markov ou « Markov blanket » (Sun et Erath, 2015). Il s'agit de ses nœuds parents, ses nœuds enfants et ses nœuds co-parents, i.e., les autres nœuds parents de ses nœuds enfants. Ainsi, la distribution complète de probabilités conditionnelles d'une variable de contrôle est approximée par sa distribution partielle de probabilités conditionnelles considérant uniquement les autres variables de contrôle formant sa couverture Markov. La définition de la distribution de probabilités conditionnelles d'un nœud n'est donc pas limitée aux parents ce qui soulève une autre limitation des hypothèses simplificatrices émises dans le cadre des méthodes MCMC (Sun et Erath, 2015).

Un réseau bayésien est ainsi composé de deux éléments : sa structure reflétant les interactions entre les variables d'intérêt d'une part, et les distributions de probabilités conditionnelles de chacune des variables par rapport aux variables de sa couverture Markov d'autre part (Zhang et al., 2019). La combinaison de ces deux éléments constitue la distribution de probabilités multivariée des variables incluses dans le modèle (Sun et Erath, 2015). Le réseau bayésien est entraîné en utilisant les données désagrégées disponibles pour définir la combinaison de structure et de distributions de probabilités conditionnelles reproduisant de manière optimale la distribution multivariée de probabilités des variables de contrôle. Un critère permettant de définir la performance des différents réseaux possibles est utilisé et un algorithme de recherche heuristique est lancé pour localiser assez rapidement l'optimum du critère et par conséquent conserver le réseau démontrant la meilleure qualité. Selon les auteurs, un réseau bayésien performant peut être établi à partir de données limitées (Sun et Erath, 2015). Un échantillonnage de réalisations indépendantes est enfin lancé sur la base du réseau bayésien optimal sélectionné.

Ainsi, le réseau bayésien permet d'éviter l'utilisation de modèles paramétriques pour compléter les distributions de probabilités conditionnelles des variables de contrôle en se contentant de distributions partielles/incomplètes de probabilités conditionnelles. De plus, au lieu d'être arbitrairement définies et imposées comme dans le cas de la méthode MCMC, ces distributions partielles de probabilités conditionnelles sont cette fois-ci directement inférées des données désagrégées disponibles (Sun et Erath, 2015). Cela permet de minimiser le risque de perte

d'information lors de l'utilisation de distributions partielles de probabilités conditionnelles (Saadi et al., 2018). Le réseau bayésien s'avère alors, contrairement à la méthode MCMC, capable d'identifier la structure de corrélation entre les variables de contrôle à partir d'un échantillon limité démontrant un ajustement et une hétérogénéité supérieurs à ceux de l'approche MCMC au niveau de la population synthétique (Sun et Erath, 2015). Le réseau bayésien demeure toutefois sujet à un risque de sous-ajustement lorsqu'une ou plusieurs variables qui feraient partie de la couverture Markov d'une autre variable ne font pas partie de l'échantillon sur lequel le réseau est entraîné.

2.2.2.2.1.4 Risques de surajustement

La méthode MCMC présente des risques de surajustement notamment lorsque la taille de l'échantillon désagrégé est limitée et le nombre de catégories des variables de contrôle est important (Sun et Erath, 2015) puisque cela réduit les chances d'avoir des observations suffisantes des différentes combinaisons de ces catégories. Ce risque se présente à plusieurs moments tels que lors de l'extraction des distributions de probabilités directement des fréquences dans l'échantillon et lorsque la totalité des probabilités conditionnelles est utilisée (Sun et Erath, 2015; Saadi et al., 2018). Le surajustement mène à un échantillonnage biaisé où une population synthétique correspondante à l'échantillon, mais pas à l'ensemble de la population réelle, est générée.

Une solution à ce problème réside dans l'utilisation de distributions incomplètes de probabilités conditionnelles. Or, la définition « manuelle » de ces distributions présente des défis (section 2.2.2.2.1.3). Ainsi, un réseau bayésien sélectionné en utilisant un critère pénalisant l'ajout de nœuds permettrait une extraction optimale de ces distributions partielles de probabilités conditionnelles ce qui aboutirait à une appréhension adéquate de la distribution multivariée de probabilités avec un risque minime de surajustement aux données d'entraînement (Sun et Erath, 2015).

Cependant, les critères de sélection de la structure optimale du réseau bayésien doivent être choisis avec prudence afin de ne pas excessivement pénaliser la complexité du réseau et l'empêcher par conséquent d'appréhender correctement la distribution multivariée de probabilités de la population. Ainsi, un compromis entre la robustesse du réseau, dont l'excès mènerait à un problème de sous-ajustement, et sa complexité, dont l'excès mènerait à un problème de surajustement, doit être établi pour lui permettre de reproduire le plus fidèlement possible la distribution conjointe de probabilités des variables de contrôle dans la population.

2.2.2.2.1.5 Respect non garanti des distributions marginales

Malgré l'ajustement satisfaisant des distributions marginales des populations synthétisées par MCMC dans les expérimentations de Farooq et al. (2013), un ajustement parfait tel que celui généralement produit par reconstruction synthétique et CO n'est généralement pas garanti (Casati et al., 2015). En effet, les méthodes basées sur l'ajustement fournissent une meilleure garantie de l'ajustement de la distribution marginale synthétique aux contraintes du recensement puisque cet ajustement constitue leur objectif direct d'optimisation.

Pour remédier à ce problème des méthodes de synthèse mixtes ont été conçues. L'idée est de générer une population synthétique par simulation puis d'y appliquer un algorithme de reconstruction synthétique où les distributions marginales des variables d'intérêt sont contrôlées. Ainsi, ces méthodes permettraient de générer des populations synthétiques optimales en bénéficiant simultanément des points forts des méthodes de synthèse par simulation et par ajustement. Dans la version hiérarchique de la méthode MCMC, Casati et al. (2015) ajoutent une étape d'ajustement a posteriori par ratissage généralisé (Deville et al., 1993) de la population synthétique obtenue par MCMC hiérarchique. Saadi et al. (2018) proposent de générer un bassin d'agents synthétiques en utilisant un modèle de Markov caché ou « Hidden Markov Model » (HMM) puis de l'ajuster par IPF aux totaux du recensement. La synthèse par HMM a d'abord été introduite par Saadi et al. (2016). Il s'agit d'une autre approche basée sur un processus markovien où un HMM est utilisé pour estimer la distribution multivariée des variables de contrôle dans la population. Les attributs contrôlés sont traités comme une séquence d'états cachés commençant par l'attribut ayant le plus de catégories et se terminant par celui présentant le moins de catégories (Yaméogo et al., 2021). Chacune des catégories d'un attribut est connectée à l'ensemble des catégories de l'attribut subséquent. Les matrices de probabilités de transition entre les états cachés et de probabilités d'émission de leurs symboles correspondants sont optimisées produisant ainsi une estimation de la distribution conjointe de probabilités des variables de contrôle. L'échantillonnage peut par conséquent être démarré. Sun et Erath (2015) mentionnent la possibilité d'appliquer un post-traitement similaire à une population générée en utilisant le réseau bayésien pour l'ajuster aux totaux zonaux. Ilahi et Axhausen (2019) et Zhou et al. (2022) ont effectivement utilisé un cadre méthodologique composé d'un réseau bayésien et d'un ratissage généralisé. Fournier et al. (2021) proposent un algorithme de synthèse mixte où la méthode MCMC et le réseau bayésien sont utilisés pour générer des bassins séparés de ménages et d'individus synthétiques qui sont ensuite

conjointement repondérés par optimisation basée sur la descente cyclique de coordonnées ou « Cyclic Coordinate Descent » (CCD). Sun et al. (2018) mentionnent aussi la possibilité d'ajuster les agents synthétiques générés avec leur modèle probabiliste hiérarchique aux distributions zonales en y appliquant un ratissage généralisé.

2.2.2.2.1.6 Interdépendance des réalisations

L'interdépendance des réalisations de l'échantillonneur de Gibbs est source de plusieurs défis. D'une part, elle impose de négliger un certain nombre des premières réalisations de l'échantillonneur (période d'échauffement) pour que l'échantillonnage soit effectué à partir d'une estimation fiable de la distribution multivariée des variables de contrôle, puis de définir un pas de sélection entre les réalisations intégrées à la population synthétique (amincissement) afin d'éviter qu'elles soient fortement corrélées. Ainsi, la détermination de la période d'échauffement et du facteur d'amincissement optimaux constituent un défi de l'utilisation de la méthode MCMC (Farooq et al., 2013).

D'autre part, cette interdépendance des réalisations rend la méthode MCMC vulnérable à une éventuelle multipolarisation de l'échantillon. Un échantillon est multipolarisé lorsque la distribution multivariée des variables d'intérêt dans celui-ci présente une concentration de probabilités élevées autour d'un nombre restreint de combinaisons et des probabilités d'occurrence faibles pour l'ensemble des autres combinaisons. Casati et al. (2015) comparent cette situation à des îlots de probabilités élevées entourés d'un océan de probabilités faibles où la chaîne de Markov sur laquelle est basée l'échantillonneur de Gibbs aura besoin d'un nombre irréaliste de pas pour atteindre tous les îlots, i.e., pour couvrir l'ensemble de la distribution multivariée. Le processus serait donc probablement arrêté avant cela, ce qui mènerait à un échantillonnage biaisé où des combinaisons à probabilités d'occurrence élevées ne figureraient pas dans la population synthétique et d'autres à probabilités d'occurrence faibles seraient excessivement échantillonnées.

Pour remédier aux effets indésirables de l'interdépendance des réalisations de l'échantillonneur de Gibbs, Sun et Erath (2015) proposent un échantillonnage à réalisations indépendantes basées sur un réseau bayésien.

2.2.2.2.1.7 Possibilité d'échantillonnages incorrects

La capacité de la méthode MCMC à générer des agents synthétiques non observés dans l'échantillon est à double tranchant. Il est vrai que cela permettrait de générer des agents présents dans la population réelle qui ne figurent pas dans l'échantillon désagrégé utilisé. Toutefois, la probabilité de générer un agent qui n'existe pas dans la population réelle ou un agent dont la combinaison de caractéristiques est impossible demeure non nulle. Cela arrive notamment lorsque la distribution de probabilités conditionnelles est estimée par le biais d'un modèle ne pouvant pas attribuer des probabilités nulles, mais seulement des probabilités extrêmement faibles, à certaines combinaisons d'attributs.

2.2.2.2.1.8 Problèmes d'évolutivité

Bien qu'acceptablement performants sur des exercices de synthèse de dimension limitée, les méthodes basées sur un processus markovien ainsi que les réseaux bayésiens s'avèrent généralement moins performants et inefficaces lorsque la dimension de l'exercice de synthèse augmente, i.e., lorsque les variables de contrôle et/ou les catégories de variables de contrôle deviennent nombreuses (Saadi et al., 2018; Borysov et al., 2019). En effet, lorsque la dimension de l'exercice de synthèse augmente et que la taille de l'échantillon demeure limitée, les cases nulles se multiplient et entraînent des problèmes de sous-ajustement et de surajustement. Ce problème limite la capacité de ces algorithmes à contrôler des attributs continus imposant une discrétisation en des catégories assez grossières.

Pour remédier à cela, entre autres, plusieurs algorithmes relevant de la modélisation profonde générative ont été récemment adaptés à la pratique de synthèse de la population. Borysov et al. (2019) comparent une approche basée sur un auto-encodeur variationnel à la méthode basée sur l'échantillonneur de Gibbs et celle utilisant le réseau bayésien. Ces deux dernières s'avèrent plus performantes pour un exercice de dimension limitée. Cependant, avec 47 attributs contrôlés, l'auto-encodeur variationnel a donné de meilleurs résultats en termes d'ajustement et d'hétérogénéité en un temps largement inférieur à celui requis par les autres algorithmes (Borysov et al., 2019). Borysov et Rich (2021) mentionnent qu'une version conditionnelle de l'auto-encodeur variationnel ou « Conditional Variational Autoencoder » (CVAE) peut aussi être appliquée à l'appréhension de la distribution multivariée de probabilités de centaines de variables de contrôle. Garrido et al. (2020) comparent les performances de l'auto-encodeur variationnel à ceux d'un autre algorithme génératif profond, le réseau antagoniste génératif de Wasserstein ou « Wasserstein Generative

Adversarial Network » (WGAN), pour un exercice de synthèse à haute dimension. Le WGAN performe mieux que le VAE en termes de qualité de prédiction ou d'ajustement mais génère une population synthétique moins hétérogène que ce dernier (Garrido et al., 2020). En effet, la formulation graphique qu'utilise le VAE pour stocker une distribution de probabilités serait responsable de la dégradation relative de la qualité de la population synthétique qu'il génère puisque l'échantillonnage est fait à partir d'une distribution multivariée de probabilités simplifiée (Badu-Marfo et al., 2022). Badu-Marfo et al. (2022) introduisent une variante composite du réseau antagoniste génératif ou « Composite Travel Generative Adversarial Network » (CTGAN) qui surpasse le VAE. Johnsen et al. (2022) montrent que la version conditionnelle du VAE réalise des prédictions plus précises que celles du réseau antagoniste génératif conditionnel ou « Conditional Generative Adversarial Network » (CGAN). Lederrey et al. (2021) proposent une variante du réseau antagoniste génératif, le réseau antagoniste génératif tabulaire acyclique orienté ou « Directed Acyclic Tabular Generative Adversarial Network » (DATGAN). Alors que les autres variantes du réseau antagoniste génératif adaptées à la synthèse de population sont totalement axées sur les données, cette variante permet l'intégration des connaissances disponibles dans la définition du modèle via un graphe acyclique orienté ou « Directed Acyclic Graph » (DAG) (Lederrey et al., 2021).

L'adaptation des modèles génératifs profonds à l'exercice de synthèse de population a permis de s'affranchir du problème d'évolutivité et du problème des cases nulles incorrectes qu'elle engendre. Cependant, un effet indésirable de cela est la génération de combinaisons de catégories d'attributs correspondant à des cases nulles structurelles, i.e., des combinaisons de catégories d'attributs impossibles ou non observées dans la population (Kim et Bansal, 2023). Kim et Bansal (2023) proposent ainsi deux fonctions de coût qu'ils appliquent à l'auto-encodeur variationnel et au réseau antagoniste génératif afin de maximiser la diversité de la population synthétique (i.e., la génération des combinaisons correspondant aux cases nulles incorrectes) tout en limitant la génération de combinaisons correspondant à des cases nulles structurelles. Les modèles génératifs profonds demeurent aussi difficiles à interpréter et sujets à un surajustement potentiel aux données d'entraînement (Borysov et Rich, 2021; Lederrey et al., 2021; Basu et Ferreira, 2020).

2.3 Synthèse de population – Spatialisation

Cette section est basée sur les revues des méthodes de synthèse spatialisée incluses dans Khachman et al. (2022) constituant le Chapitre 5 et Khachman et al. (2023) constituant le Chapitre 6 de cette thèse.

Alors que l'optimisation de la représentativité sociodémographique des populations synthétiques a été, et continue à être, le sujet de nombreux travaux de recherche, l'amélioration de la distribution spatiale des ménages synthétiques demeure relativement moins discutée dans la littérature. Cependant, la corrélation entre les attributs sociodémographiques d'intérêt varie d'un milieu à un autre (Zhou et al., 2022) impliquant une variation spatiale des comportements de mobilité de la population (Zhu et al., 2018). Ainsi, la fiabilité des simulations des comportements de mobilité dépendent autant de la précision spatiale que de la représentativité sociodémographique globale des populations synthétiques (Yaméogo et al., 2023; Chapuis et al., 2018).

La majorité des applications de synthèse de population recensées dans la littérature sont des applications de synthèse zonales où les ménages synthétiques sont assignés à des surfaces correspondant à des zones administratives plus ou moins désagrégées (Yaméogo et al., 2023; Nejad et al., 2021; Long et Shen, 2015, p. 115–131; Anderson et al., 2014) ou des cellules d'une grille prédéfinie (Ji et Wan, 2021; Su et al., 2010). Alors qu'une spatialisation zonale de la population synthétique demeure suffisante pour de nombreuses applications, l'émergence de la pratique de la planification urbaine microscopique et à long terme a instauré un besoin croissant de populations synthétiques géoréférencées ou explicitement spatialisées (i.e., où chaque ménage synthétique est assigné à un bâtiment ou une unité de logement). Cela inclut la planification des transports actifs, les modèles de choix résidentiels et professionnels, la planification énergétique urbaine, la gestion des catastrophes, la sécurité intérieure, l'aménagement de services, et la santé publique (Chen et al., 2021; Liu et al., 2021; Pajares et al., 2021; Chapuis et al., 2018; Thomson et al., 2018; Zhu et al., 2018; Bast et al., 2015; Muñoz and Peters, 2014). Des méthodes de spatialisation explicite ont par conséquent été introduites pour assigner des localisations ponctuelles, i.e., des coordonnées (X ;Y), aux ménages synthétiques générés.

2.3.1 Méthodes de synthèse spatialisée

Les méthodes de synthèse explicitement spatialisées existantes sont toutes basées sur une approche séquentielle où la spatialisation de la population synthétique constitue une étape subséquente à sa génération. Une population synthétique est ainsi d’abord générée à une résolution géographique plus ou moins grossière et les ménages synthétiques générés sont ensuite assignés à des localisations ponctuelles à l’intérieur de leur zone de synthèse.

2.3.1.1 Classification

Les méthodes de synthèse explicitement spatialisée recensées dans la littérature peuvent être classées en quatre catégories selon le degré de considération de l’environnement bâti et de ses caractéristiques dans le processus de spatialisation : les distributions supposées ou hypothétiques, les distributions basées sur les adresses, les distributions basées sur les capacités, et les distributions par correspondance totale.

2.3.1.1.1 Distributions supposées

Cette catégorie inclut les méthodes de spatialisation où l’environnement bâti est totalement négligé. Il s’agit d’appliquer des distributions géographiques simples telles qu’une distribution aléatoire, uniforme ou pondérée dans un polygone (Zhang et al., 2019; Gallagher et al., 2018; Swarup et Marathe, 2016) ou tout au long du réseau routier (Jiang et al., 2022; Gallagher et al., 2018) pour assigner à chaque ménage synthétique des coordonnées (X;Y) à l’intérieur de son unité géographique de synthèse.

2.3.1.1.2 Distributions basées sur les adresses

Cette catégorie inclut les méthodes de spatialisation explicite considérant uniquement les adresses des bâtiments. Les ménages synthétiques sont aléatoirement distribués sur les bâtiments situés dans leur unité géographique de synthèse sans considérer les caractéristiques de chacun des bâtiments telles que sa capacité, son type, son usage, etc. (Zhang et al., 2019; Sallard et al., 2021; Thomson et al., 2018).

2.3.1.1.3 Distributions basées sur les capacités

Cette catégorie inclut les méthodes de spatialisation explicite considérant uniquement les capacités des bâtiments ou leurs caractéristiques permettant d'en estimer les capacités. Chaque ménage synthétique est alors aléatoirement assigné à un bâtiment situé dans son unité géographique de synthèse dont la capacité n'est pas atteinte (Chapuis et al., 2018; Zhu et Ferreira, 2014).

Zhu et Ferreira (2014) utilisent un double ajustement proportionnel itératif pour générer une population synthétique pour Singapour. Le premier ajustement permet de générer une population synthétique multiniveau à l'échelle de la zone d'analyse du trafic ou « Traffic Analysis Zone » (TAZ). Un deuxième IPF est ensuite appliqué à l'échelle du bâtiment où la capacité des bâtiments, estimée par fusion des données spatialisées disponibles, est minimalement contrôlée. Des modèles de choix discret peuvent aussi être utilisés pour enrichir les distributions marginales contrôlées à l'échelle du bâtiment. Chapuis et al. (2018) proposent un processus de spatialisation en deux étapes : l'interpolation surfacique et la localisation explicite. L'idée est d'abord d'interpoler la population à une échelle zonale plus fine que celle à laquelle elle a été générée (e.g., les cellules d'une grille prédéfinie) pour ensuite assigner des localisations ponctuelles aux ménages synthétiques à l'intérieur de leurs cellules respectives. L'interpolation est réalisée selon une distribution statistique (une régression linéaire par exemple) calibrée en utilisant des données auxiliaires disponibles telles qu'un plan d'utilisation du sol, une carte du réseau routier, un fichier de rôle foncier, etc. Les ménages synthétiques sont ensuite imbriqués dans des objets spatiaux (e.g., zones, bâtiments, etc.) situés à l'intérieur de l'unité surfacique dans laquelle ils ont été transférés à l'issue de l'interpolation surfacique. L'imbrication dans un objet spatial peut être effectuée selon une distribution uniforme, proportionnelle à l'aire de l'objet spatial, ou proportionnelle à la capacité de cet objet. Des contraintes géométriques, de contingence et de densité peuvent également être définies lors de l'imbrication pour guider le processus de spatialisation. Une fois imbriqué dans un objet spatial, le ménage synthétique peut être attribué soit à son centroïde, soit à un emplacement aléatoire à l'intérieur de celui-ci. La méthodologie proposée permet de tirer profit des différentes couches disponibles dans la définition des distributions et des contraintes requises. Les ménages synthétiques pourraient ainsi par exemple être distribués aléatoirement, régulièrement à une certaine distance du réseau routier, ou ponctuellement sur les bâtiments dans les limites de leurs unités surfaciques respectives.

2.3.1.1.4 Distributions par correspondance totale

Cette catégorie inclut les méthodes où les caractéristiques des bâtiments et des ménages synthétiques disponibles sont considérées lors de la spatialisation des ménages synthétiques. Ces caractéristiques peuvent être investies dans une variété de processus tels qu'une optimisation combinatoire (Muñoz and Peters, 2014), une heuristique de correspondance (Zhou et al., 2022), une modélisation probabiliste (Liu et al., 2021), etc. pour réaliser une spatialisation explicite par correspondance totale des caractéristiques des ménages synthétiques avec ceux des bâtiments.

Pour générer une population synthétique pour Singapour, Zhou et al. (2022) commencent par générer une population de ménages et individus à l'échelle de la zone de planification ou « planning area » en utilisant un algorithme de synthèse mixte composé d'un réseau bayésien et d'un ratissage généralisé. L'environnement bâti est ensuite séparément synthétisé par fusion de plusieurs bases de données et en considérant un ensemble d'hypothèses (e.g., le taux d'inoccupation) basées sur la familiarité des auteurs avec le contexte local. Une heuristique de correspondance est enfin appliquée pour assigner une unité de logement synthétique à chaque ménage synthétique par correspondance de l'unité géographique de synthèse (i.e., de zone de planification) et du type de logement. Liu et al. (2021) introduisent une application web géospatiale permettant de visualiser des données sociodémographiques spatialement désagrégées. Pour générer les données alimentant l'interface web, les données relatives à l'environnement bâti sont d'abord fusionnées et un modèle de capacité résidentielle est estimé puis appliqué pour estimer le nombre d'unités de logement dans chaque projet résidentiel où l'information n'est pas disponible. Une population synthétique est ensuite générée en utilisant un algorithme de reconstruction synthétique (Zhu et Ferreira, 2015). Chaque ménage synthétique est enfin associé à une unité de logement moyennant un modèle de choix discret sous les contraintes de capacité (De Palma et al. 2007). Muñoz et Peters (2014) associent les ménages synthétiques aux unités de logement par optimisation combinatoire basée sur la méthode de descente.

2.3.1.2 Critiques

Les méthodes de synthèse spatialisée existantes présentent trois types d'enjeux : la non-considération de l'environnement bâti, les incohérences potentielles entre la population synthétique générée et l'environnement bâti, et les problèmes de transférabilité.

2.3.1.2.1 Non-considération de l'environnement bâti

Moins l'environnement bâti est considéré dans le processus de spatialisation, moins la spatialisation effectuée est fiable. Ainsi, l'utilisation de distributions supposées n'aboutit évidemment pas à une spatialisation réaliste des ménages synthétiques puisqu'ils ne sont pas assignés à des bâtiments et peuvent même se trouver dans des endroits irréalistes tels que des zones non résidentielles ou au bord d'un pont ou d'un tunnel (Gallagher et al. 2018). Lorsqu'une distribution basée sur les adresses est utilisée, les capacités des bâtiments ne sont pas prises en compte. De plus, ces méthodes pourraient assigner des ménages synthétiques à des composantes non résidentielles de l'environnement bâti telles qu'un bâtiment non résidentiel, un garage détaché, un chalet de jardin ou une unité commerciale au rez-de-chaussée d'un bâtiment. Les distributions basées sur les capacités permettent de réaliser une spatialisation plus réaliste que celles réalisées par distributions supposées ou basées sur les adresses. Elle demeure toutefois aléatoire à part pour l'unique contrainte de capacité considérée. Les distributions par correspondance totale permettent de réaliser les spatialisations explicites les plus réalistes des quatre types de méthodes identifiées dans la littérature. Leur qualité demeure toutefois dépendante de l'abondance et de la pertinence des caractéristiques accessibles de la population synthétique et de l'environnement bâti.

2.3.1.2.2 Incohérences entre la population synthétique et l'environnement bâti

Aborder la génération et la spatialisation comme des processus indépendants mènerait potentiellement à des incohérences entre la population synthétique générée et l'environnement bâti supposé l'accueillir. Par exemple, si l'on considère la contrainte de capacité, il pourrait y avoir plus de ménages synthétiques que d'unités de logement dans une zone. Ainsi, certains ménages synthétiques pourraient ne pas trouver de bâtiment correspondant. Pour garantir la complétion du processus de spatialisation, des mesures de mitigation artificielle ou forcée telles que le relâchement de règles et contraintes (Zhou et al., 2022; Chapuis et al., 2018), la correction de valeurs négatives (Chapuis et al., 2018), et l'utilisation d'algorithmes intégrant des contraintes de capacité (Liu et al., 2021; De Palma et al., 2007) sont requises. Zhu et Ferreira (2014) mentionnent que l'intégration de la génération et de la spatialisation dans un seul IPF est théoriquement possible. Elle n'a cependant pas été réalisée puisque la dimension de la matrice devient intraitable pour un algorithme de reconstruction synthétique lorsque des contraintes à l'échelle du bâtiment y sont intégrées.

2.3.1.2.3 Problèmes de transférabilité

Contrairement aux données du recensement, la disponibilité, le format, et la qualité des données spatialisées (e.g., enquêtes où les ménages sont géoréférencés, données décrivant l'environnement bâti, etc.) varient fortement d'une zone d'étude à une autre. Cela a rendu la définition d'un cadre méthodologique universel de synthèse spatialisée assez complexe. Ainsi, les méthodes de synthèse explicitement spatialisées existantes présentent des défis de transférabilité. Il s'agit principalement d'applications ad hoc fortement dépendantes de la richesse de la zone d'étude en bases de données spatiales décrivant la population et l'environnement bâti et d'une connaissance approfondie du contexte local ce qui les rend difficilement adaptables à des contextes différents de ceux où ils ont été conçus (Liu et al., 2021). De plus, même lorsque ces facteurs sont réunis dans une zone d'étude, la transférabilité des méthodes incluant des modèles localement calibrés demeure problématique puisque la performance d'un modèle varie selon la base de données sur laquelle il est entraînée et pourrait par conséquent être dégradée dans une autre zone d'étude.

2.3.2 Méthodes de microsimulation spatiale

La microsimulation spatiale est une pratique qui a pour objectif de désagréger spatialement des données sociodémographiques (Zhou et al. 2022). Bien qu'une telle désagrégation des distributions marginales demeure différente d'une synthèse de population à une échelle désagrégée, plusieurs méthodes de synthèse de population telles que l'IPF (Panori et al., 2017 ; Anderson, 2012), la CO (Farrell et al., 2012; Hynes et al., 2009) et le ratissage généralisé (Vidyattama et al. 2013; Harding et al. 2011; Tanton et al. 2011; Ballas et al. 2007) ont aussi été appliquées à la microsimulation spatiale. La majorité des applications de microsimulation spatiale sont limitées à une désagrégation zonale des données sociodémographiques. Toutefois, certaines applications ad hoc de microsimulation spatiale à l'échelle du bâtiment sont trouvées dans la littérature. Pajares et al. (2021) ont désagrégué la population à l'échelle du bâtiment moyennant des estimations de superficies brutes de plancher. Calka et al. (2016) se sont basés sur les caractéristiques des bâtiments pour en estimer la population. Bast et al. (2015) ont utilisé l'apprentissage automatique pour désagréger la population à l'échelle du bâtiment en se basant sur OpenStreetMap (OSM) et les données du recensement à l'échelle de la municipalité. Chen et al. (2021) et Xie et al. (2015) ont utilisé des données LiDAR pour atteindre des objectifs similaires. Une revue plus complète et détaillée des méthodes de microsimulation spatiale est réalisée par Tanton et al. (2014).

2.4 Lacunes dans la littérature

Plusieurs lacunes identifiées dans la revue de littérature sont à l'origine des contributions de cette thèse. D'abord, dans les travaux existants, la qualité de la population synthétique est résumée à sa représentativité sociodémographique. Celle-ci se traduit par le degré auquel les distributions de probabilités des variables de contrôle dans la population synthétique correspondent à leurs distributions de probabilités observées dans la population réelle. Or, tel qu'expliqué précédemment, la localisation résidentielle des individus est aussi un facteur déterminant de leurs comportements de mobilité. Ainsi, il est aussi important qu'une population synthétique destinée à alimenter des microsimulations des comportements de mobilité reproduise la variation spatiale des distributions de probabilités des variables de contrôle. Le jugement de la qualité d'une population synthétique par sa représentativité sociodémographique uniquement demeure alors partiel. Un indicateur de précision spatiale est par conséquent requis pour avoir une idée plus complète de la qualité d'une population synthétique générée.

Ensuite, parmi les différents choix à effectuer avant le lancement d'un processus de synthèse de population, seulement ceux qui sont jugés impacter la représentativité sociodémographique de la population synthétique sont discutés. Il s'agit par exemple du choix des variables de contrôle, de la catégorisation des variables contrôlées, de l'algorithme de synthèse, etc. Le choix de la résolution de référence, souvent implicitement effectué lors de l'extraction des distributions de probabilités marginales des variables de contrôle, n'est pas abordé. Or, ce choix est susceptible d'impacter la qualité de la synthèse, notamment en termes de précision spatiale. Il est ainsi important, dans une optique d'optimisation de la qualité de la synthèse, d'analyser l'impact de la résolution de référence sur la qualité de la synthèse afin d'effectuer un choix éclairé.

Enfin, les méthodes existantes de synthèse spatialisée les plus fiables suivent une approche séquentielle basée sur des distributions par correspondance totale (section 2.3.1.1.4). Lorsqu'utilisées, ces méthodes présentent deux limitations principales pouvant essentiellement altérer la précision spatiale de la population synthétique produite. D'une part, l'indépendance des processus de génération et de spatialisation pourrait être à l'origine d'incohérences entre la population synthétique et l'environnement bâti (section 2.3.1.2.2) qui compromettraient la qualité de la synthèse. D'autre part, les méthodes de synthèse spatialisée existantes présentent un problème majeur de transférabilité. Il s'agit d'applications ad hoc dépendantes du format et de l'abondance

locaux de données spatialisées ainsi que d'une connaissance approfondie de la zone d'étude considérée. En effet, contrairement aux données sociodémographiques ayant généralement des formats similaires dans différentes zones d'étude, les données spatialisées décrivant l'environnement bâti entre autres varient fortement en quantité, en nature et en format d'une zone d'étude à une autre. Cela limite la transférabilité et par conséquent l'utilité des méthodes de synthèse spatialisée existantes.

CHAPITRE 3 MÉTHODOLOGIE GÉNÉRALE

L'objectif général du projet de recherche est de définir une approche permettant de générer une population synthétique explicitement spatialisée de qualité optimale. La qualité d'une population synthétique ayant longtemps été résumée à sa représentativité sociodémographique mesurée à des résolutions géographiques plus ou moins grossières, les travaux menés dans le cadre de cette thèse ont particulièrement été axés sur la précision spatiale des populations synthétiques tout en veillant à ne pas compromettre la qualité globale de synthèse. La démarche générale suivie pour atteindre l'objectif de recherche est détaillée dans ce chapitre. L'étude de cas principale choisie pour le développement et l'application de l'ensemble de la méthodologie de recherche est l'île de Montréal, Canada.

En plus des données agrégées et désagrégées du recensement, ce territoire dispose d'un échantillon d'enquête O-D présentant des caractéristiques intéressantes en termes de taille, de variables enquêtées et de détail spatial, ainsi que de données permettant de reconstruire l'environnement bâti. Cette richesse en données accessibles aux chercheurs constitue un atout considérable pour le développement de méthodes de synthèse spatialisée présentant des besoins différents en termes d'intrants. Le développement d'un choix de méthodes plus ou moins exigeantes en termes d'intrants spatialisés permettrait de disposer d'une solution adaptée à toute zone d'étude selon la richesse des intrants qui y sont disponibles. Étant les plus récentes dans la zone d'étude au moment du début des travaux, les données du recensement de la population 2016 et de l'enquête O-D 2018 sont utilisées pour le développement et l'application de la méthodologie de recherche.

3.1 Étapes préliminaires

Cette section est dédiée à la description des étapes préliminaires à l'exécution de la méthodologie de recherche. Il s'agit de l'établissement du système de zonage, de la génération de l'environnement bâti, du choix des variables de contrôle, et de l'extraction et du traitement de l'échantillon désagrégé ainsi que des distributions de probabilités marginales nécessaires pour alimenter l'algorithme de synthèse.

3.1.1 Système de zonage

Une variété de résolutions géographiques est considérée dans les expérimentations menées dans le cadre de ce projet de recherche. Il s'agit des différentes résolutions géographiques où les résumés du recensement sont publiés, du découpage en secteurs municipaux utilisé pour la pondération de l'enquête O-D et des bâtiments dérivés de la base de données des adresses ponctuelles (Ville de Montréal, 2023a). Les résolutions géographiques considérées dans le cadre des travaux menés sont décrites dans le Tableau 3.1. Les définitions des résolutions géographiques du recensement sont tirées du dictionnaire du recensement (Statistique Canada, 2018) et celle du secteur municipal provient du site de l'Autorité Régionale de Transport Métropolitain (ARTM) (ARTM, 2023). La hiérarchie des résolutions considérées pour l'ensemble du Canada est illustrée sur la Figure 3.1(a). Une flèche liant deux résolutions géographiques indique que le découpage de la résolution vers laquelle elle pointe respecte les limites du découpage de la résolution constituant son origine. Une hiérarchie plus verticale (i.e., avec plus de correspondances entre les découpages des différentes résolutions) est observée au niveau de l'île de Montréal constituant une division de recensement. Celle-ci est illustrée sur la Figure 3.1(b). Les découpages correspondant aux résolutions géographiques utilisées dans le projet de recherche sont illustrés sur les cartes de la Figure 3.2.

Tableau 3.1 Résolutions géographiques utilisées dans le projet de recherche

Résolution géographique	Source de données	Définition
Région Métropolitaine de Recensement ou « Census Metropolitan Area » (CMA)	Recensement	Une ou plusieurs municipalités adjacentes situées autour d'un centre de population (aussi appelé le noyau). Elle doit avoir une population totale d'au moins 100 000 habitants, et son noyau doit compter au moins 50 000 habitants.
Division de Recensement ou « Census Division » (CD)	Recensement	Région géographique intermédiaire entre la municipalité (subdivision de recensement) et la province/territoire.

Tableau 3.1 Résolutions géographiques utilisées dans le projet de recherche (suite)

Résolution géographique	Source de données	Définition
Subdivision de Recensement ou « Census Subdivision » (CSD)	Recensement	Une municipalité ou un territoire considéré comme un équivalent municipal à des fins statistiques (p. ex. une réserve indienne, un établissement indien et un territoire non organisé).
Secteur Municipal ou « Municipal Borough » (MB)	Enquête O-D	Un regroupement de secteurs ou de subdivisions de recensement.
Aire de Diffusion Agrégée ou « Aggregate Dissemination Area » (ADA)	Recensement	Aire ayant une population de 5 000 à 15 000 habitants.
Secteur de Recensement ou « Census Tract » (CT)	Recensement	Petite région géographique relativement stable qui compte habituellement une population de 2 500 à 8 000 habitants.
Aire de Diffusion ou « Dissemination Area » (DA)	Recensement	Plus petite région géographique normalisée pour laquelle toutes les données du recensement sont diffusées. Elle compte de 400 à 700 habitants.
Îlot de diffusion ou « Dissemination Block » (DB)	Recensement	Territoire dont tous les côtés sont délimités par des rues et/ou des limites de régions géographiques normalisées. L'îlot de diffusion est la plus petite unité géographique pour laquelle les chiffres de population et des logements sont diffusés.

Tableau 3.1 Résolutions géographiques utilisées dans le projet de recherche (suite)

Résolution géographique	Source de données	Définition
Bâtiment ou « Building » (BLD)	Adresses ponctuelles	Il s'agit du regroupement d'adresses ponctuelles partageant la même localisation (X;Y).

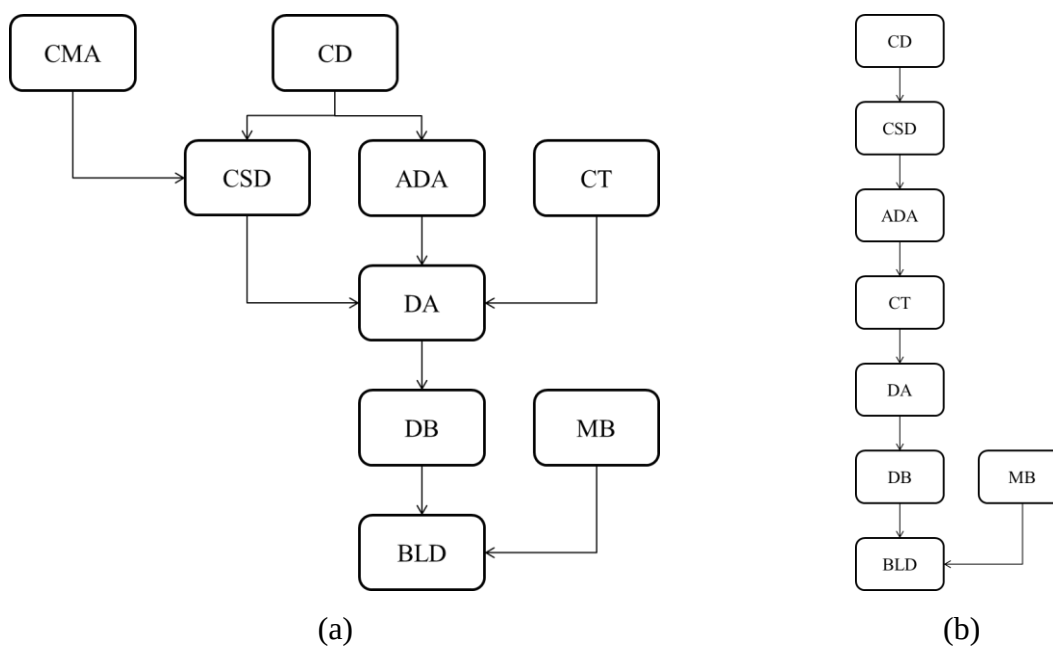
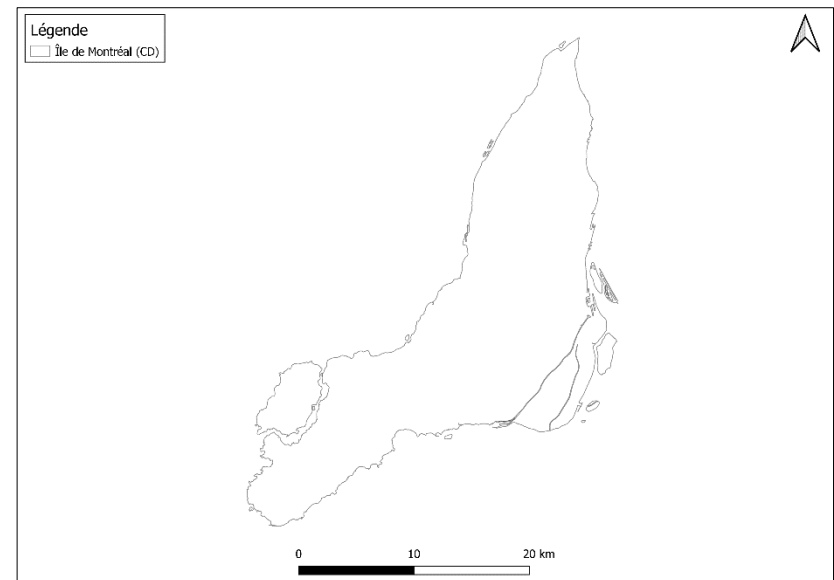
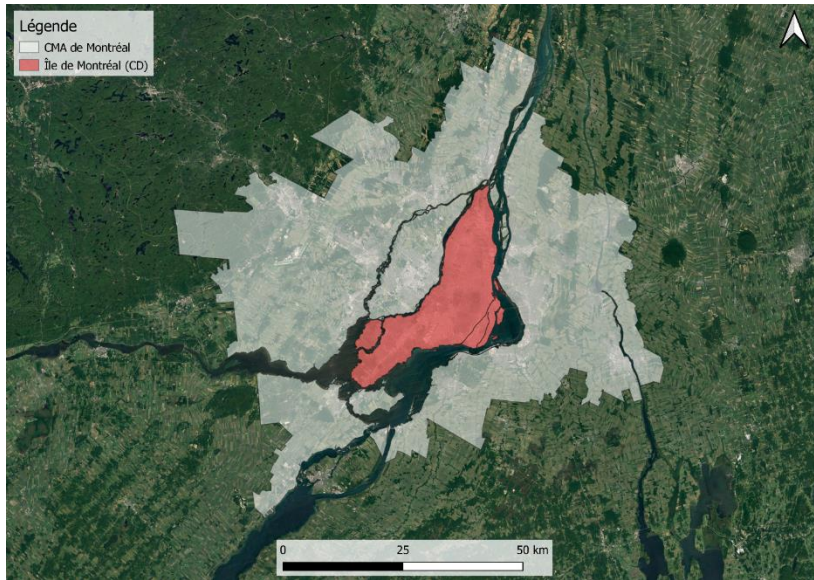


Figure 3.1 Hiérarchies des résolutions géographiques considérées pour (a) l'ensemble du Canada et (b) l'île de Montréal



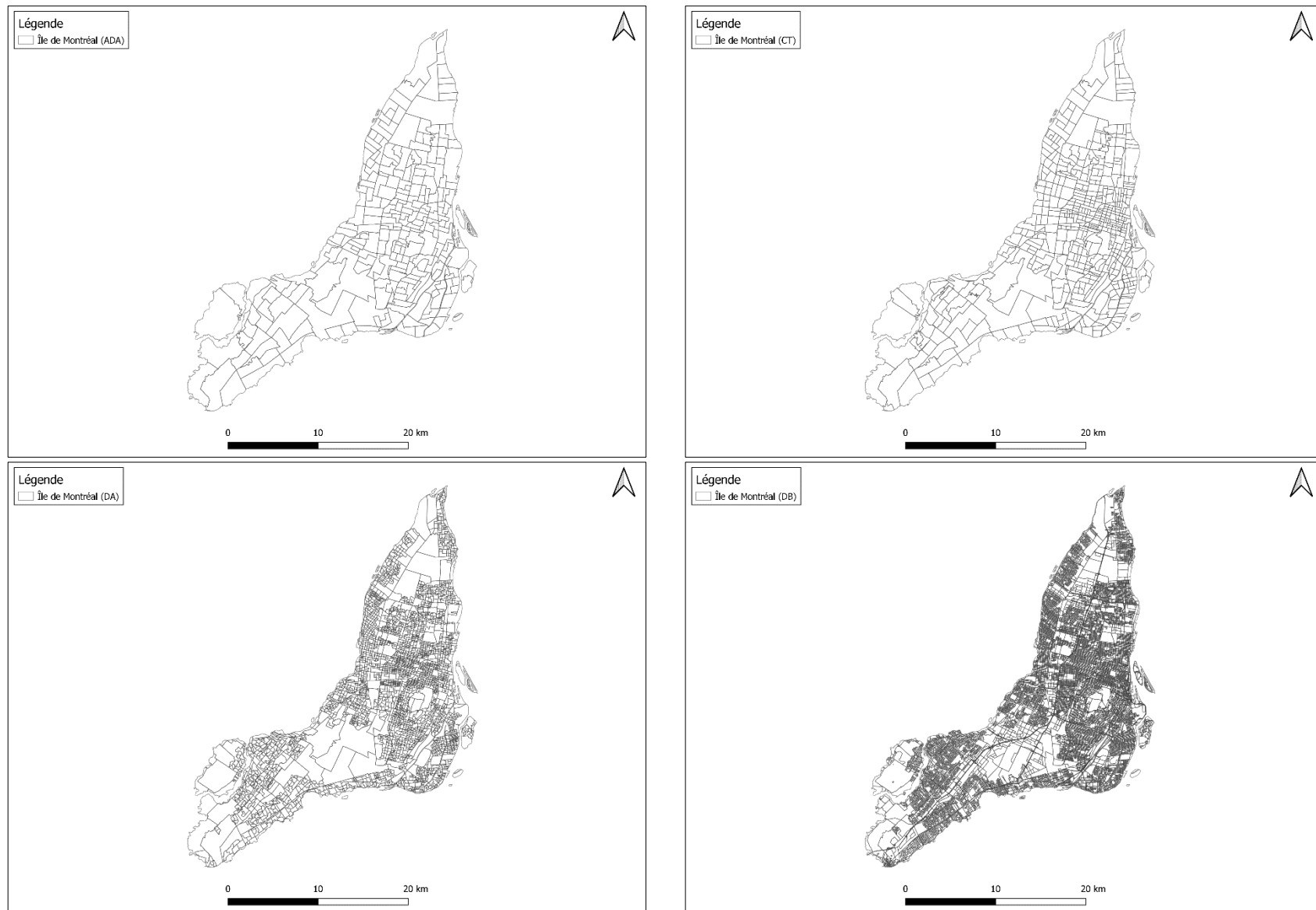


Figure 3.2 Résolutions géographiques utilisées dans le projet de recherche

3.1.2 Génération de l'environnement bâti

Cette étape de la méthodologie générale n'est requise que dans une zone d'étude ne disposant pas d'une base de données listant et caractérisant ses bâtiments et/ou ses unités de logement, ce qui est le cas de l'île de Montréal. Ainsi, l'environnement bâti de l'île de Montréal a été généré à partir de plusieurs bases de données accessibles dont :

- La base de données des adresses ponctuelles publiée par la Ville de Montréal (2023a);
- La base de données des unités d'évaluation foncière (UEF) publiée par la Ville de Montréal (2023c);
- La base de données des bâtiments 2D publiée par la Ville de Montréal (2023b);
- La base de données du rôle d'évaluation foncière dont des variables sont fournies sur demande par le ministère des Affaires Municipales et de l'Habitation (MAMH) pour des fins de recherche (MAMH, 2020)¹.

3.1.2.1 Adresses ponctuelles

Cette base de données consiste en une liste localisée des adresses répertoriées sur l'île de Montréal. Chaque point correspond à une adresse ou à un ensemble d'adresses partageant les mêmes localisation, type de voie et nom de voie. Cette base de données n'inclut pas de variables caractérisant les bâtiments telles que le nombre d'unités de logement, le type de bâtiment, l'année de construction, etc. Les variables comprises dans la base de données des adresses ponctuelles sont résumées dans le Tableau 3.2. Les définitions proviennent du site de la Ville de Montréal (2023a). Un extrait graphique d'adresses ponctuelles est montré sur la Figure 3.3.

¹ Un ensemble plus large de variables du rôle d'évaluation foncière que celui fourni pour les travaux de cette thèse est récemment devenu accessible au public.

Tableau 3.2 Variables incluses dans la base de données des adresses ponctuelles

Variable	Définition	Exemple
ID_ADRESSE	Identifiant unique	72
TEXTE	Texte pour annotation	#5664
SPECIFIQUE	Nom de la voie (toponymie)	1 ^{re}
ORIENTATION	Orientation de la voie (toponymie)	X
LIEN	Particule de la voie (toponymie)	-
HAUTEUR	Hauteur du texte de l'annotation	2
GENERIQUE	Type de voie (toponymie)	Avenue
ANGLE	Angle de l'annotation	22.60579999999999856186
ADDR_DE	Adresse minimale sur le bâtiment	5664
ADDR_A	Adresse maximale sur le bâtiment	5664
X	Coordonnée X dans la projection NAD83 MTM8 (EPSG 32188)	298491.8800000000046566
Y	Coordonnée Y dans la projection NAD83 MTM8 (EPSG 32188)	5045167.129999999888241
LONGITUDE	Positionnement géographique WGS84	73.58078399999999419379
LATITUDE	Positionnement géographique WGS84	45.54638899999999779311

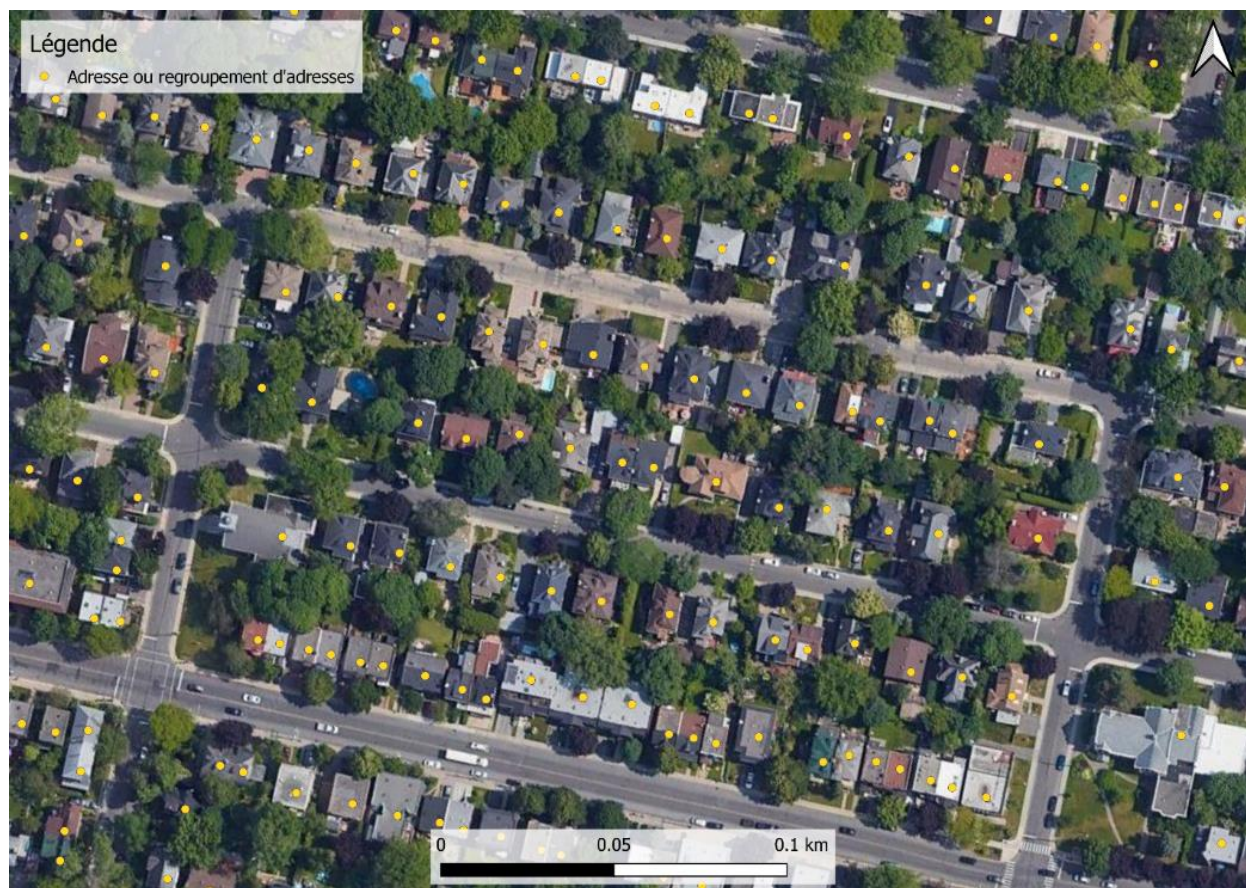


Figure 3.3 Extrait graphique des adresses ponctuelles de l'île de Montréal (Ville de Montréal, 2023a)

3.1.2.2 Unités d'évaluation foncière

Cette base de données consiste en un découpage vectoriel des unités d'évaluation foncière situées sur l'île de Montréal. Une unité d'évaluation foncière est le « *regroupement d'immeubles adjacents, appartenant à un même propriétaire (ou groupe de propriétaires par indivis), utilisés à une même fin prédominante et n'étant cessibles que globalement compte tenu de l'utilisation la plus probable qui peut en être faite* » (MAMH, 2010). Chaque UEF est représentée par un polygone (Figure 3.4) disposant des attributs listés dans le Tableau 3.3. Les définitions du Tableau 3.3 sont tirées du site de la Ville de Montréal (2023c).

Tableau 3.3 Variables incluses dans la base de données des unités d'évaluation foncière

Variable	Définition	Exemple
ID_UEV	Identifiant unique système	1000702
CIVIQUE_DEBUT	Numéro civique (plage - début)	3483
CIVIQUE_FIN	Numéro civique (plage - fin)	3483
NOM_RUE	Nom de rue	Rue Jeanne-Mance (MTL)
SUITE_DEBUT	Numéro unité (appartement ou local)	-
ETAGE_HORS_SOL	Nombre d'étages maximal : • Si l'UEF comprend un seul bâtiment : nombre d'étages du bâtiment • Si l'UEF comprend plusieurs bâtiments : nombre d'étages du bâtiment comptant le plus d'étages (maximum)	50
NOMBRE_LOGEMENT	Nombre de logements	3
ANNEE_CONSTRUCTION	Année de construction	2
CODE_UTILISATION	Codification [selon le] code d'utilisation des biens-fonds (CUBF) du MAMH	1885
LETTRE_DEBUT	Première lettre de l'appartement	1000
LETTRE_FIN	Dernière lettre de l'appartement	-
LIBELLE_UTILISATION	Descriptif [selon le] CUBF	-

Tableau 3.3 Variables incluses dans la base de données des unités d'évaluation foncière (suite)

Variable	Définition	Exemple
CATEGORIE_UEF	Catégorie de l'unité d'évaluation (Régulier ou Condominium)	Logement
MATRICULE83	Matricule au rôle foncier (système géographique NAD83 MT8)	Régulier
SUPERFICIE_TERRAIN	Superficie du terrain pour fin d'évaluation foncière (mètres carrés)	9941-11-2750-1-000-0000
SUPERFICIE_BATIMENT	Aire d'étages du bâtiment, soit surface brute correspondant à la somme des aires de chacun des étages entiers du bâtiment principal et, le cas échéant, de celles de l'attique, du garage intégré et de la verrière intégrée (mètres carrés)	208
NO_ARROND_ILE_CUM	Identifiant de l'arrondissement (Référence identifiant MAMH)	288
MUNICIPALITE	Identifiant interne de la municipalité	REM21



Figure 3.4 Extrait graphique des unités d'évaluation foncière de l'île de Montréal (Ville de Montréal, 2023c)

3.1.2.3 Bâtiments 2D

Cette base de données consiste en une représentation 2.5D des bâtiments montréalais. Il s'agit d'une représentation 2D à laquelle des informations altimétriques ponctuelles sont ajoutées. Cette base de données inclut les cinq calques suivants (Ville de Montréal, 2023b) :

- CARTO-BAT-TOIT : bordure du toit (polygones) ;
- CARTO-BAT-CONSTRUCTION : bâtiment en construction (polygones) ;
- CARTO-BAT-RUINE : bâtiment en ruine (polygones) ;
- CARTO-BAT-DETAIL : détail des bâtiments (polylignes) ;
- CARTO-BAT-SUPER-STRUCTURE : structure extérieure d'un bâtiment ayant un impact visuel important (polylignes) ;

- CARTO-BAT-COTE : cote altimétrique du toit (points).

Les calques CARTO-BAT-TOIT et CARTO-BAT-COTE (Figure 3.5) sont utilisés dans la génération de l’environnement bâti. Les attributs d’intérêt inclus dans ces calques sont décrits dans le Tableau 3.4.

Tableau 3.4 Variables incluses dans la base de données des bâtiments 2D

Variable	Définition	Exemple
Superficie	Superficie du toit du bâtiment (m ²)	36.0
Élévation	Élévation du bâtiment au point considéré (m)	192.6

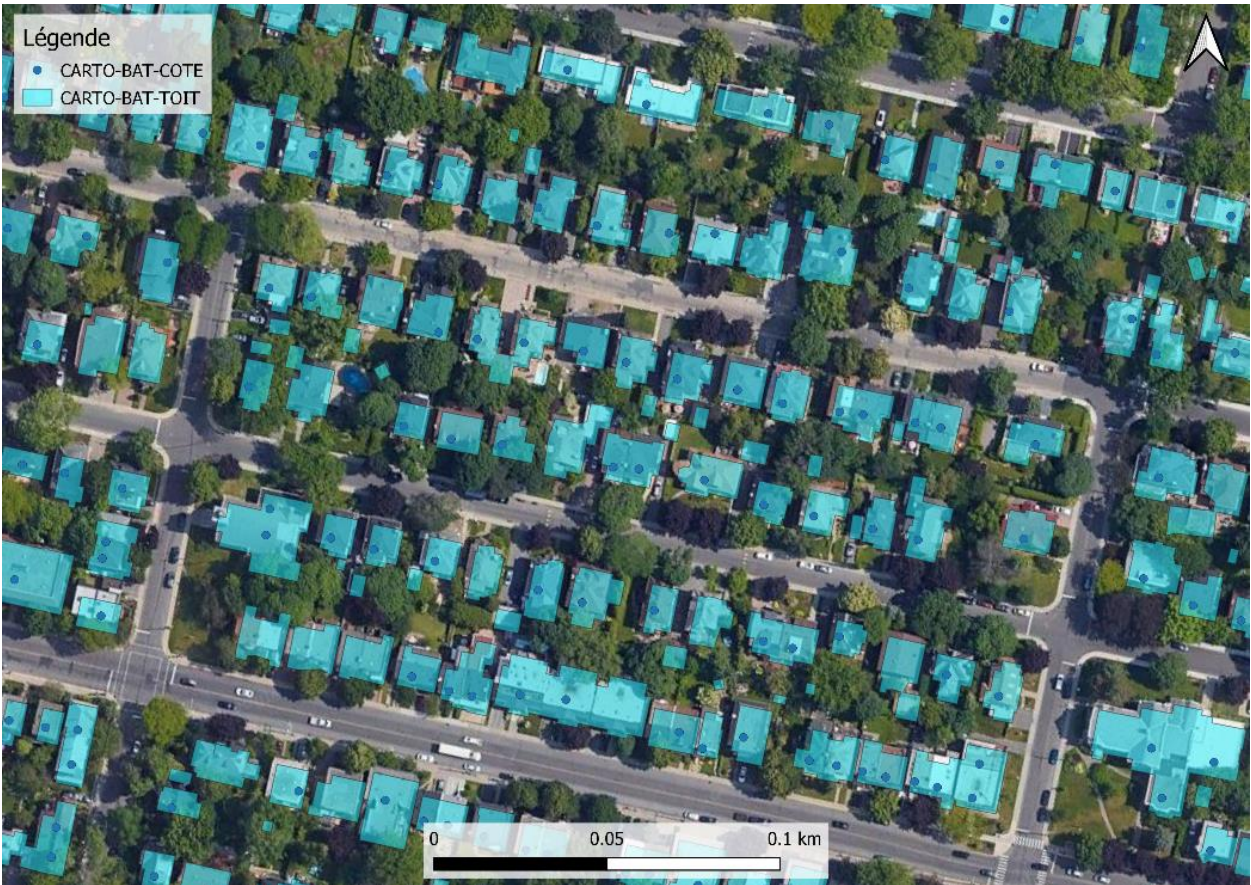


Figure 3.5 Extrait graphique des bâtiments 2D de l’île de Montréal (Ville de Montréal, 2023b)

3.1.2.4 Rôle d'évaluation foncière

Le rôle d'évaluation foncière du MAMH est un résumé de l'inventaire des unités d'évaluation foncières situées sur les territoires des municipalités du Québec. Ce fichier contient les tables d'attributs suivantes :

- Le rôle d'évaluation foncière ponctuel (rol_unite_p) ;
- La table des caractéristiques de l'immeuble (B05V_UNITE_EVALN) ;
- La table des adresses de l'immeuble (B05V_ADR_UNITE_EVALN) ;
- La table des répartitions du cadastre (B05V_REPAR_FISC) ;
- La table des données du cadastre (B05V_LOT_CADST) ;
- La table des adresses des propriétaires (B05V_ADR_PROPRIO).

Pour les travaux de cette thèse, seulement certaines variables des tables rol_unite_p et B05V_UNITE_EVALN ont été fournies par le MAMH (MAMH, 2020). Récemment, l'ensemble des variables des tables rol_unite_p, B05V_UNITE_EVALN, B05V_ADR_UNITE_EVALN et B05V_REPAR_FISC est devenu accessible au public (MAMH, 2023). Les variables des autres tables demeurent toujours caviardées. Le Tableau 3.5 liste l'ensemble des variables des six tables du rôle d'évaluation foncière. Les lignes à fond gris indiquent les variables qui étaient fournies pour les travaux de cette thèse. Les définitions proviennent du document décrivant la structure des attributs du rôle d'évaluation foncière (MAMH, 2022). Un aperçu graphique du rôle d'évaluation foncière montréalais est montré sur la Figure 3.6.

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière

Variable	Définition	Exemple
Rôle d'évaluation foncière ponctuel (rol_unite_p)		
X	Latitude (WGS 84)	- 73.64982670 35

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
Y	Longitude (WGS 84)	45.48628757 42
OBJECTID	Identifiant unique de l'enregistrement	501332
ID_PROVINC	Identifiant composé du code géographique de la municipalité à cinq chiffres et du MAT18. Renseignement servant à faire le lien entre les différentes tables	6606293380 4879390000 000
CODE_MUN	Code géographique de la municipalité	66062
ARROND (RL0102A)	Le numéro d'arrondissement de la municipalité lorsque présent	-
REJET	Code identifiant les unités d'évaluation foncière qui sont à l'extérieur de la municipalité ou qui présentent d'autres particularités pour fin de production	0
DATE_ENTREE	Année de l'activation de l'unité d'évaluation foncière	2006
MAT18	Matricule de l'unité d'évaluation foncière à 18 chiffres (division-section-emplacement-chiffre autovérificateur-bâtiment-local)	9338048793 90000000
Table des caractéristiques de l'immeuble (B05V_UNITE_EVALN)		
ANROLE	Année de l'exercice financier du rôle	2020
RL0104G	Fuseau de projection MTM	08

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0104H	Discriminant qui s'applique aux municipalités dont l'étendue en longitude et en latitude est plus grande que 100 KM	-
RL0105A	Code d'utilisation prédominant de l'unité d'évaluation	1000
RL0106A	Numéro distinctif permanent attribué à l'unité d'évaluation par l'évaluateur ou par l'organisme qui a le mandat	7-F04000400
RL0107A	Numéro de l'unité de voisinage dont fait partie l'unité d'évaluation	0908
RL0301A	Dimension linéaire du terrain en front sur la voie publique	21.42
RL0302A	Superficie totale du terrain porté au rôle	610.80
RL0303A	Code indiquant comment le terrain est assujetti au zonage agricole (loi sur la protection du territoire agricole)	0
RL0304A	Superficie totale de l'exploitation agricole enregistrée	0
RL0305A	Superficie zonée agricole de l'exploitation agricole enregistrée	0
RL0306A	Nombre maximal d'étages des bâtiments de l'unité d'évaluation	2

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0307A	Année de construction originelle du bâtiment principal, s'il n'en existe qu'un seul	1963
RL0307B	Mention du fait que l'année de construction originelle indiquée est réelle ou estimée	R
RL0308A	Aire d'étages du bâtiment principal, s'il n'en existe qu'un seul	295.80
RL0309A	Code du lien physique de l'unité d'évaluation lorsqu'il existe un seul bâtiment principal	1
RL0311A	Nombre total de logements de l'unité d'évaluation	1
RL0312A	Nombre total de chambres locatives de l'unité d'évaluation, soit la somme de celles situées dans tous les bâtiments compris dans cette unité d'évaluation	0
RL0313A	Nombre total de locaux non résidentiels de l'unité d'évaluation	0
RL0402A	Valeur du terrain inscrite au rôle d'évaluation en vigueur	629100
RL0403A	Valeur du ou des bâtiments inscrite au rôle d'évaluation foncière en vigueur	740500
RL0404A	Valeur de l'immeuble inscrite au rôle d'évaluation en vigueur	1369600

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0405A	Valeur du même immeuble inscrite au rôle d'évaluation antérieur	1104500
RL0501A	Indicateur d'unité d'évaluation considérée comme étant un terrain vague desservi	0
RL0502A	Code de la classe de la catégorie des immeubles non-résidentiels	-
RL0503A	Code de la classe de la catégorie des immeubles industriels	-
RL0505A	Valeur totale imposable pour fins scolaires d'une exploitation agricole enregistrée	0
IND_NOUV_CTRID	Indicateur d'un nouveau centroïde	Non
DAT_COND_MRCHE	Date à laquelle les conditions du marché ont été considérées	7/1/2018
Table des adresses de l'immeuble (B05V_ADR_UNITE_EVALN)		
NO_SEQ_ADR	Numéro de la séquence de l'adresse de l'unité d'évaluation foncière	0
RL0101A	Numéro inférieur de l'adresse civique de l'immeuble	2020
RL0101B	Inscription de la fraction ou de la lettre du numéro inférieur de l'adresse (1/2, A, B, inf.)	-
RL0101C	Numéro supérieur de l'adresse civique de l'immeuble lorsque présent	1

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0101D	Inscription de la fraction ou de la lettre du numéro supérieur de l'adresse (1/2, A, B, sup.)	5701
RL0101E	Code générique désignant la voie publique pour uniformiser leur inscription	-
RL0101F	Code de lien entre le code générique et le nom de la voie publique afin d'uniformiser les inscriptions d'adresses	-
RL0101G	Nom de la voie publique	-
RL0101H	Code de direction (orientation) de la voie publique sur laquelle est située l'unité d'évaluation	RU
RL0101I	Numéro d'appartement ou du local	-
RL0101J	Partie complémentaire du numéro d'appartement ou de local de l'adresse de l'unité d'évaluation	FERNCROF T
Table des répartitions fiscales (B05V_REPAR_FISC)		
RL0504A	Numéro de la loi accordant une exemption fiscale	-
RL0504B	Numéro de l'article de la loi accordant une exemption fiscale	-
RL0504C	Numéro ou lettre de l'alinéa ou du paragraphe de l'article de la loi accordant une exemption fiscale	-
RL0504D	Montant résultant de la répartition des valeurs du bloc 61 selon le terrain, le bâtiment ou l'immeuble	-

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0504E	Partie d'immeuble correspondante	-
RL0504F	Imposabilité de chacun des montants inscrits	-
Table des données du cadastre (B05V_LOT_CADST)		
RL0103A	Le numéro de lot du cadastre québécois, rénové ou non	NA
RL0103B	Suffixe du numéro de lot du cadastre du Québec, rénové	NA
RL0103C	Nom du cadastre non rénové	NA
RL0103D	Désignation secondaire du cadastre non rénové	NA
RL0103E	Numéro de lot non rénové	NA
RL0103F	Indicateur de partie non subdivisée du lot non rénové	NA
Table des adresses des propriétaires (B05V_ADR_PROPRIO)		
RL0201AX	Nom légal du propriétaire	NA
RL0201BX	Prénom du propriétaire	NA
RL0201U	Code de type de possession (code des conditions d'inscription)	NA
NO_COPROP	No de copropriétaire	NA
RL0201CX	Adresse postale non structurée	NA

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0201DX	Nom de la municipalité de l'adresse postale	NA
RL0201EX	Code postale de l'adresse postale du propriétaire	NA
RL0201FX	Complément d'adresse relatif à l'adresse postale du propriétaire	NA
RL020GX	Date de l'inscription au rôle	NA
RL0201HX	Statut du propriétaire aux fins d'imposition scolaire	NA
RL0201IX	Numéro civique de l'adresse postale	NA
RL0201JX	Fraction ou partie de l'adresse postale du propriétaire	NA
RL0201KX	Code de lien de l'adresse postale	NA
RL0201LX	Code de générique de l'adresse postale	NA
RL0201MX	Nom de la voie publique de l'adresse postale du propriétaire	NA
RL0201NX	Code de l'orientation de l'adresse postale du propriétaire	NA
RL0201OX	Numéro d'appartement ou de local de l'adresse postale du propriétaire	NA
RL0201PX	Fraction ou partie d'adresse du numéro d'appartement ou de locale de l'adresse postale du propriétaire	NA
RL0201QX	Province ou état de l'adresse postale du propriétaire	NA

Tableau 3.5 Variables incluses dans la base de données du rôle d'évaluation foncière (suite)

Variable	Définition	Exemple
RL0201RX	Nom du pays	NA
RL0201TX	Nom de la succursale postale	NA
RL0201SX	Numéro du casier postal	NA

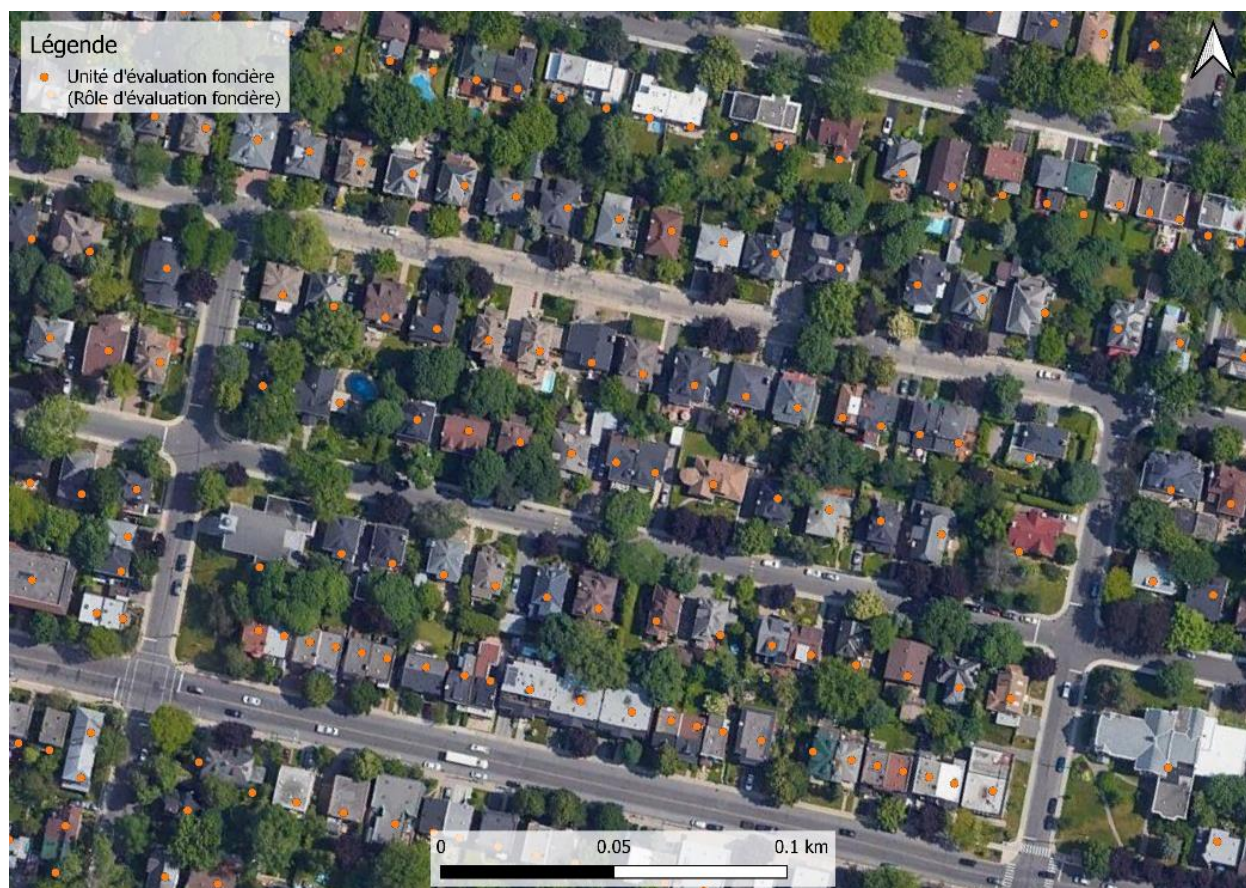


Figure 3.6 Extrait graphique du rôle d'évaluation foncière sur l'île de Montréal (MAMH, 2020)

3.1.2.5 Démarche

Pour générer l'environnement bâti, les bases de données sont d'abord filtrées à l'année de synthèse. Il s'agit de l'année 2016 pour une population synthétique générée à partir des données du recensement de la population de 2016 (Statistique Canada, 2019b) et de l'année 2018 pour une synthèse de population basée sur l'enquête O-D 2018 (ARTM, 2018). Une démarche en trois étapes

est ensuite suivie pour dériver une base de données listant et caractérisant les bâtiments de l'île de Montréal à partir des bases de données disponibles. Il s'agit d'établir dans un premier temps une liste des bâtiments, puis de s'assurer que chaque UEF polygonale (Ville de Montréal, 2023c) et ponctuelle (MAMH, 2020) est associée à un seul bâtiment, et enfin d'agrégier les valeurs des attributs des UEFs par bâtiment.

3.1.2.5.1 Établir les bâtiments

La première étape consiste à établir les bâtiments à partir de la base de données des adresses ponctuelles. En effet, les adresses ayant la même localisation, i.e., les mêmes coordonnées (X;Y), sont considérées constituer un bâtiment unique. Une liste des bâtiments de l'île de Montréal et des adresses qui y sont associées est alors établie. Ces bâtiments demeurent cependant non caractérisés puisque les variables caractérisant l'environnement bâti sont diffusées dans la base de données portant sur les unités d'évaluation foncières de la Ville de Montréal (2023c) et dans le rôle d'évaluation foncière du MAMH (2020). Or, la correspondance entre une UEF et un bâtiment n'est pas toujours évidente. En effet, une UEF peut autant correspondre à un bâtiment qu'à une partie d'un bâtiment ou à plusieurs bâtiments.

3.1.2.5.2 Association de chaque UEF à un seul bâtiment

La deuxième étape consiste alors à déterminer, pour chaque UEF son bâtiment d'appartenance parmi la liste établie à la première étape. Cette association est possible par comparaison des localisations des bâtiments et des UEFs et/ou par correspondance des adresses des UEFs aux adresses ponctuelles composant les bâtiments. L'association des UEFs ponctuelles du rôle d'évaluation foncière aux adresses ponctuelles s'avère compliquée pour deux raisons. D'une part, il est rare que le point représentant une adresse et celui représentant une UEF qui y est située soient superposés ce qui complique l'association par localisation. Une association d'une UEF ponctuelle à l'adresse ponctuelle voisine la plus proche n'a pas permis de régler ce problème puisque dans le rôle d'évaluation foncière utilisé, une UEF ponctuelle a des chances non minimales d'être plus proche de l'adresse d'une UEF voisine que de sa propre adresse. D'autre part, les données du MAMH (2020) fournies pour les travaux de cette thèse ne contenaient pas la table des adresses limitant les moyens d'association d'une UEF à son bâtiment à la correspondance de leurs localisations spatiales qui, comme expliqué précédemment, demeure non fiable notamment à cause de la géométrie ponctuelle des deux bases de données. Pour remédier à cela, les UEFs polygonales

de la Ville de Montréal (2023c) sont d’abord associées aux adresses correspondantes. La géométrie polygonale d’une UEF et la disponibilité de son adresse dans la base de données diffusée par la Ville de Montréal (2023c) permettent d’optimiser les chances d’une association correcte. Une UEF qui selon tous les critères d’association établis correspond à plusieurs bâtiments est associée aléatoirement à un bâtiment parmi ceux-ci.

Ensuite, chaque UEF ponctuelle du rôle d’évaluation foncière est associée à l’UEF polygonale correspondante par correspondance spatiale et par correspondance des caractéristiques de l’UEF communes aux deux bases de données. Chaque UEF ponctuelle est ainsi associée à un bâtiment par le biais d’une UEF polygonale. L’association des UEFs ponctuelles aux bâtiments demeure optionnelle. Elle permet toutefois d’une part l’inclusion des caractéristiques des UEFs uniquement disponibles dans le rôle d’évaluation foncière du MAMH (2020) telles que les valeurs du terrain et du bâtiment dans la base de données des bâtiments générée, et d’autre part l’imputation des valeurs manquantes de certaines variables pour certaines UEFs dans la base de données de la Ville de Montréal (2023c) lorsque la variable concernée figure aussi dans la base de données du MAMH (2020) et que sa valeur pour les UEFs concernées n’y sont pas manquantes. À l’issue de cette deuxième étape, chaque UEF polygonale et ponctuelle est associée à un seul bâtiment.

3.1.2.5.3 Agrégation des caractéristiques des UEFs par bâtiment

La dernière étape consiste à agréger les caractéristiques des unités d’évaluation foncière associées à chaque bâtiment de manière à obtenir les caractéristiques globales de ce dernier. Les valeurs qui demeurent manquantes pour certains bâtiments dans les colonnes de certains attributs (e.g., aire du bâtiment, nombre d’étages du bâtiment, etc.) pourraient être imputées par correspondance des localisations spatiales des bâtiments ponctuels aux polygones des bâtiments 2D et aux points d’élévation y appartenant. Les variables sélectionnées pour constituer la base de données des bâtiments sont détaillées dans le Tableau 3.6.

Tableau 3.6 Variables de la base de données des bâtiments

Variable	Définition
bld_id	Identifiant unique du bâtiment

Tableau 3.6 Variables de la base de données des bâtiments (suite)

Variable	Définition
bld_lon	Longitude à laquelle est situé le bâtiment
bld_lat	Latitude à laquelle est situé le bâtiment
bld_flats	Nombre d'unités de logement dans le bâtiment
bld_nonflats	Nombre d'unités non résidentielles dans le bâtiment
bld_area	Aire cumulée des étages du bâtiment
bld_value	Valeur du bâtiment
bld_landvalue	Valeur du terrain sur lequel est situé le bâtiment
bld_consy	Année de construction du bâtiment
bld_dutype	Type des unités de logement constituant le bâtiment

L'algorithme établi pour générer l'environnement bâti est inclus dans un fichier supplémentaire joint à l'article constituant le Chapitre 6 de la présente thèse (Khachman et al., 2023).

3.1.3 Variables de contrôle

Les variables disponibles dans le recensement de la population 2016 et les variables incluses dans l'enquête O-D 2018 ne sont pas identiques. Les variables communes à ces deux bases de données présentent aussi parfois des catégorisations différentes. Ainsi, deux bassins de variables de contrôle sont définis pour les expérimentations menées dans le cadre de cette thèse : le bassin des variables de contrôle provenant du recensement de la population 2016, et le bassin des variables de contrôle de l'enquête O-D 2018.

3.1.3.1 Recensement de la population 2016

Le Tableau 3.7 liste et décrit le bassin de variables de contrôle du recensement et leurs catégories utilisées dans le cadre du projet de recherche. Ce bassin est partiellement ou complètement

considéré à chacune des étapes de la démarche de recherche où une population synthétique est générée en utilisant les données du recensement.

Tableau 3.7 Bassin de variables de contrôle du recensement de la population 2016

Variable	Définition	Catégorie	Description
hh_count	Nombre de ménages	1	1 ménage
hh_size	Taille du ménage	1	1 individu
		2	2 individus
		3	3 individus
		4	4 individus
		5	5 individus ou plus
hh_type	Type du ménage	1	Sans famille de recensement ²
		2	1 famille de recensement sans enfants
		3	1 famille de recensement avec enfants
		4	Plusieurs familles de recensement
hh_income	Revenu net du ménage	1	Moins que 30k\$
		2	[30k\$; 60k\$[
		3	[60k\$; 100k\$[
		4	100k\$ et plus
hh_dtype	Type de l'unité de logement	1	Condominium
		2	Autre

² « Un couple marié et les enfants, le cas échéant, du couple et/ou de l'un ou l'autre des conjoints; un couple en union libre et les enfants, le cas échéant, du couple et/ou de l'un ou l'autre des conjoints; ou un parent dans une famille monoparentale, peu importe son état matrimonial, habitant avec au moins un enfant dans le même logement et cet ou ces enfants. » (Statistique Canada, 2018)

Tableau 3.7 Bassin de variables de contrôle du recensement de la population 2016 (suite)

Variable	Définition	Catégorie	Description
hh_dbuilt	Année de construction de l'unité de logement	1	Avant 1960
		2	[1961 ; 1980[
		3	[1981 ; 2000[
		4	[2001 ; 2016[
pp_count	Nombre d'individus	1	1 individu
pp_age	Catégorie d'âge de l'individu	1	[0 ; 15[
		2	[15 ; 65[
		3	65 ou plus
pp_sex	Sexe de l'individu	1	Homme
		2	Femme
pp_marst	État matrimonial de l'individu	1	Jamais marié (et ne vivant pas en union libre)
		2	Légalement marié (et non séparé)
		3	Vivant en union libre
		4	Séparé, divorcé, ou veuf (et ne vivant pas en union libre)
pp_lfst	Statut d'activité de l'individu	1	Employé
		2	Sans emploi
		3	Pas dans la population active

3.1.3.2 Enquête O-D 2018

Le Tableau 3.8 liste et décrit le bassin de variables de contrôle de l'enquête O-D 2018 et leurs catégories utilisées dans le cadre du projet de recherche. L'enquête O-D 2018 inclut des variables plus directement liées aux comportements de mobilité des individus.

Tableau 3.8 Bassin de variables de contrôle de l'enquête O-D 2018

Variable	Définition	Catégorie	Description
hh_count	Nombre de ménages	1	1 ménage
hh_size	Taille du ménage	1	1 individu
		2	2 individus
		3	3 individus
		4	4 individus
		5	5 individus ou plus
hh_income	Revenu net du ménage	1	Moins que 30k\$
		2	[30k\$; 60k\$[
		3	[60k\$; 90k\$[
		4	[90k\$; 120k\$[
		5	[120k\$; 150k\$[
		6	[150k\$; 180k\$[
		7	[180k\$; 210k\$[
		8	210k\$ et plus
hh_motorization	Motorisation du ménage	1	1 véhicule
		2	2 véhicules
		3	3 véhicules
		4	4 véhicules
		5	5 véhicules ou plus
pp_count	Nombre d'individus	1	1 individu
pp_age	Catégorie d'âge de l'individu	1	[0;5[
		2	[5;10[

Tableau 3.8 Bassin de variables de contrôle de l'enquête O-D 2018 (suite)

Variable	Définition	Catégorie	Description
		3	[10;15[
		4	[15;20[
		5	[20;25[
		6	[25;30[
		7	[30;35[
		8	[35;40[
		9	[40;45[
		10	[45;50[
		11	[50;55[
		12	[55;60[
		13	[60;65[
		14	[65;70[
		15	[70;75[
		16	75 ou plus
pp_sex	Sexe de l'individu	1	Homme
		2	Femme
pp_occupation	Occupation principale de l'individu	1	Travailleur à temps plein
		2	Travailleur à temps partiel
		3	Étudiant
		4	Retraité
		5	Autre

Tableau 3.8 Bassin de variables de contrôle de l'enquête O-D 2018 (suite)

Variable	Définition	Catégorie	Description
pp_drlicense	Permis de conduire de l'individu	0	Sans permis de conduire
		1	Détient un permis de conduire
pp_transpass	Abonnement de transport en commun (TC) de l'individu	0	Sans abonnement TC
		1	Détient un abonnement TC
pp_carshsub	Abonnement d'autopartage (AP) de l'individu	0	Sans abonnement d'AP
		1	Détient un abonnement d'AP
pp_bikeshsub	Abonnement de vélopartage (VP) de l'individu	0	Sans abonnement de VP
		1	Détient un abonnement de VP

3.1.4 Échantillon désagrégé

Dans la zone d'étude choisie, deux échantillons désagrégés sont considérés : le Fichier de Microdonnées à Grande Diffusion (FMGD) ou « Public Use Microdata File » (PUMF) du recensement de la population 2016 et l'échantillon de l'enquête O-D 2018. Présentant des caractéristiques différentes, ces deux échantillons sont traités pour être investis dans les expérimentations menées.

3.1.4.1 Fichier de microdonnées à grande diffusion

Le PUMF du recensement est publié en deux versions pour l'ensemble du Canada : une version portant uniquement sur les individus et leurs caractéristiques (iPUMF) et une version hiérarchique (hPUMF) associant chaque individu à son ménage et caractérisant à la fois les individus et les ménages. Malgré le fait que le iPUMF est plus riche en attributs individuels que le hPUMF, le hPUMF du recensement de la population 2016 était un choix naturel compte tenu de l'importance de l'objet ménage dans la détermination des comportements de mobilité. La seule information spatiale incluse dans le PUMF du recensement est la CMA d'appartenance des agents. Afin de le rendre utilisable pour la synthèse de population, les modifications suivantes au hPUMF brut doivent être effectuées :

1. Filtrer aux enregistrements appartenant à la CMA de Montréal ou à une autre CMA le cas échéant.
2. Dériver des variables d'intérêt existantes dans les résumés du recensement mais inexistantes dans le hPUMF, et ce à partir des variables disponibles dans le hPUMF brut.
3. Recatégoriser certaines variables. Cette opération est généralement requise lorsque les catégories d'une variable dans le hPUMF et dans les résumés agrégés du recensement ne sont pas identiques. De plus, une variable peut être recatégorisée de manière à obtenir des segments de population plus pertinents pour les objectifs ultimes de la synthèse. Par exemple, pour une population synthétique destinée à alimenter une microsimulation des comportements de mobilité de la population, une variable peut être recatégorisée de manière que les catégories reflètent a priori des comportements de mobilité différents. Enfin, lorsqu'un problème de dimensionnalité se pose, une recatégorisation des variables permettrait d'y remédier par fusion de catégories.
4. Imputer les valeurs manquantes lorsque possible puis éliminer les ménages où la valeur d'un attribut du ménage et/ou d'un individu qu'il inclut n'est pas connue et ne peut être logiquement imputée.

Les variables du hPUMF conservées pour les expérimentations menées dans le cadre de ce projet de recherche sont résumées dans le Tableau 3.7. Les opérations exécutées sur le hPUMF brut dans le but d'obtenir un hPUMF incluant ces variables et prêt à être importé par un algorithme de synthèse sont détaillées ci-dessous.

Nombre de ménages

La variable « hhcount » n'existe pas dans le hPUMF brut. Il s'agit d'une variable ajoutée qui associe une valeur unitaire « 1 » à chaque identifiant unique de ménage « HH_ID ».

Taille du ménage

La variable « hh_size » n'existe pas dans le hPUMF brut. Elle est calculée en comptant le nombre d'identifiants uniques d'individus « PP_ID » dans chacun des ménages. Elle est ensuite catégorisée de manière à respecter les catégories de taille des ménages dans les résumés du recensement tel que décrit dans le Tableau 3.7.

Type du ménage

La variable « hh_type » n'existe pas dans le hPUMF brut. Pour estimer le type d'un ménage, deux variables temporaires, elles aussi inexistantes dans le hPUMF brut, sont calculées : le nombre de familles de recensement dans le ménage et le nombre d'enfants dans le ménage.

Le nombre de familles de recensement dans un ménage est calculé en comptant les identifiants uniques de familles de recensement « CF_ID » appartenant à ce ménage. Une catégorisation en trois catégories est ensuite utilisée : ménage sans famille de recensement, ménage constitué d'une famille de recensement, et ménage à multiples familles de recensement. Le nombre d'enfants dans un ménage est calculé à partir de la variable qualifiant la situation des particuliers dans le ménage « CFSTAT ». La somme du nombre d'enfants de couples (CFSTAT=4) et du nombre d'enfants de familles monoparentales (CFSTAT=5) est calculée pour chacun des ménages. Une catégorisation en deux catégories est considérée pour ce qui est du nombre d'enfants dans le ménage : ménage avec enfants et ménage sans enfants. Chacun des ménages se voit ainsi attribuer un type parmi quatre selon le nombre de familles de recensement et le nombre d'enfants qu'il inclut tel que décrit dans le Tableau 3.7.

Revenu net du ménage

La variable « hh_income » n'existe pas de la hPUMF brut. Elle est calculée en sommant les revenus nets individuels dans chacun des ménages. Avant de calculer la somme par ménage des revenus nets individuels, ceux dont la valeur est « 99999999 » sont remplacés par la valeur « 0 » puisqu'il s'agit d'individus sans revenus. Les ménages sont distribués dans quatre catégories selon leurs revenus nets tel que décrit dans le Tableau 3.7.

Type de l'unité de logement du ménage

La variable « hh_dtype » correspond à la variable « CONDO » du hPUMF brut. Les étiquettes des catégories de cette variable sont modifiées tel que décrit dans le Tableau 3.7.

Année de construction de l'unité de logement du ménage

La variable « hh_dbuilt » correspond à la variable « BUILT » du hPUMF brut. Cette variable est recatégorisée en quatre catégories tel que décrit dans le Tableau 3.7.

Nombre d'individus

La variable « pp_count » n'existe pas dans le hPUMF brut. Il s'agit d'une variable ajoutée qui associe une valeur unitaire « 1 » à chaque identifiant unique d'individu « PP_ID ».

Catégorie d'âge de l'individu

La variable « pp_age » correspond à la variable « AGEGRP » du hPUMF brut. Une reclassification en trois catégories est effectuée tel que décrit dans le Tableau 3.7.

Sexe de l'individu

La variable « pp_sex » correspond à la variable « SEX » du hPUMF brut. Les étiquettes des catégories de cette variable sont modifiées tel que décrit dans le Tableau 3.7.

État matrimonial de l'individu

La variable « pp_marst » correspond à la variable « MarStH » du hPUMF brut.

Statut d'activité de l'individu

La variable « pp_lfst » correspond à la variable « LFTAG » du hPUMF brut. Une recatégorisation en trois classes est effectuée tel que décrit dans le Tableau 3.7.

Les valeurs manquantes de toutes les variables considérées sont remplacées par « 88 » et les ménages correspondant sont complètement éliminés de l'échantillon. Le hPUMF brut de la CMA de Montréal compte 17 272 ménages composés de 40 003 individus. L'élimination des ménages incluant des valeurs manquantes a entraîné une perte de 2.6% des ménages et de 5.7% des individus. L'algorithme permettant de préparer le hPUMF pour la synthèse de population tel que décrit dans cette section est inclus dans un fichier supplémentaire joint à l'article constituant le Chapitre 5 de la présente thèse (Khachman et al., 2022).

3.1.4.2 Échantillon de l'enquête O-D 2018

L'échantillon de l'enquête O-D 2018 est un échantillon hiérarchique explicitement spatialisé, i.e., où les coordonnées de chaque ménage enquêté sont disponibles. Cet échantillon n'est pas accessible au public mais peut être fourni sur demande pour des fins de recherche entre autres. Afin de le rendre utilisable pour la synthèse de population, les modifications suivantes à l'échantillon O-D brut doivent être effectuées :

1. Filtrer aux enregistrements appartenant à l'île de Montréal.
2. Dériver des variables d'intérêt à partir des variables disponibles dans l'échantillon O-D brut.

3. Recatégoriser certaines variables. Dans le cas de l'échantillon O-D, cette opération est généralement effectuée pour obtenir des segments de population plus pertinents pour les objectifs ultimes de la synthèse ou lorsqu'un problème de dimensionnalité se pose.
4. Imputer les valeurs manquantes lorsque possible puis éliminer les ménages où la valeur d'un attribut du ménage et/ou d'un individu qu'il inclut n'est pas connue et ne peut être logiquement imputée.

Les variables de l'échantillon de l'enquête O-D 2018 conservées pour les expérimentations menées dans le cadre de ce projet de recherche sont résumées dans le Tableau 3.8. Les opérations exécutées sur l'échantillon O-D brut dans le but d'obtenir un échantillon O-D incluant ces variables et prêt à être importé par un algorithme de synthèse sont détaillées ci-dessous.

Nombre de ménages

La variable « hh_count » n'existe pas dans l'échantillon O-D brut. Il s'agit d'une variable ajoutée qui associe une valeur unitaire « 1 » à chaque identifiant unique de ménage « FEUILLET ».

Taille du ménage

La variable « hh_size » correspond à la variable « PERSLOGI » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8.

Revenu net du ménage

La variable « hh_income » correspond à la variable « REVENU » de l'échantillon O-D brut. Environ 22.7% des ménages de l'échantillon O-D ont une valeur de « REVENU » parmi les trois suivantes : « 9 » (Refus), « 10 » (Ne sait pas) ou « 0 » (Non définie dans le guide de l'enquête). Afin d'éviter la suppression de ces enregistrements et par conséquent une réduction importante de la taille de l'échantillon, ces valeurs ont été remplacées par une valeur entre « 1 » et « 8 » (définies dans le Tableau 3.8) par échantillonnage aléatoire proportionnel à la prévalence de ces dernières dans l'échantillon.

Motorisation du ménage

La variable « hh_motorization » correspond à la variable « AUTOLOGI » de l'échantillon O-D brut catégorisée tel que décrit dans le Tableau 3.8.

Nombre d'individus

La variable « pp_count » n'existe pas dans l'échantillon O-D brut. Il s'agit d'une variable ajoutée qui associe une valeur unitaire « 1 » à chaque identifiant unique d'individu composé de la concaténation des variables « FEUILLET » et « RANG » de l'échantillon O-D brut.

Catégorie d'âge de l'individu

La variable « pp_age » correspond à la variable « COHORTE » de l'échantillon O-D brut.

Sexe de l'individu

La variable « pp_sex » correspond à la variable « SEXE » de l'échantillon O-D brut.

Occupation principale de l'individu

La variable « pp_occupation » correspond à la variable « P_STATUT » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8. Les valeurs « 6 » (N/A : enfant de 4 ans et moins), « 7 » (À la maison) et « 8 » (Refus) sont remplacées par la valeur « 5 » (Autre).

Permis de conduire de l'individu

La variable « pp_drlicense » correspond à la variable « PERMIS » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8. Les valeurs « 2 » (Non), « 3 » (Ne sait pas), « 4 » (Refus) et « 5 » (Non-applicable – moins de 16 ans) sont remplacées par la valeur « 0 » (Sans permis de conduire).

Abonnement TC de l'individu

La variable « pp_transpass » correspond à la variable « TITRE » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8. Les valeurs « 1 » à « 6 » correspondant à différents titres TC sont remplacées par la valeur « 1 » (Détient un abonnement TC), et les valeurs « 8 » (Enfant de 5 ans et moins), « 9 » (Aucun titre mensuel) et « 10 » (NSP / Refus) sont remplacées par la valeur « 0 » (Sans abonnement TC).

Abonnement AP de l'individu

La variable « pp_carshsub » correspond à la variable « ABON_AP » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8. Les valeurs « 2 » (Non), « 3 » (Refus / Ne sait pas) et « 4 » (N/A) sont remplacées par la valeur « 0 » (Sans abonnement d'AP).

Abonnement VP de l'individu

La variable « pp_bikeshsub » correspond à la variable « ABON_VLS » de l'échantillon O-D brut recatégorisée tel que décrit dans le Tableau 3.8. Les valeurs « 2 » (Non), « 3 » (Refus / Ne sait pas) et « 4 » (N/A) sont remplacées par la valeur « 0 » (Sans abonnement de VP).

L'échantillon O-D brut de l'île de Montréal compte 33,202 ménages composés de 71,914 individus. L'algorithme permettant de préparer l'échantillon O-D pour la synthèse de population tel que décrit dans cette section est inclus dans un fichier supplémentaire joint à l'article constituant le Chapitre 6 de la présente thèse (Khachman et al., 2023).

3.1.5 Distributions de probabilités marginales

Les distributions marginales des variables de contrôle listées dans la section 3.1.3 sont considérées aux résolutions géographiques décrites dans la section 3.1.1 où il est possible soit de les extraire directement soit de les dériver à partir d'autres données disponibles. Le Tableau 3.9 résume pour chacune des variables de contrôle les résolutions où leurs distributions de probabilités marginales sont disponibles. Pour chacune des variables, une résolution en italique indique que la distribution marginale de la variable correspondante n'y est pas directement disponible mais peut être dérivée d'autres données disponibles.

Tableau 3.9 Résolutions des distributions marginales des variables de contrôle

Variable	Définition	Résolution(s) géographique(s)
hh_count	Nombre de ménages	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA, DB, <i>BLD</i>
hh_size	Taille du ménage	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA
hh_type	Type du ménage	CMA, CD, CSD, ADA, CT, DA
hh_income	Revenu net du ménage	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA
hh_dtype	Type de l'unité de logement	CMA, CD, CSD, ADA, CT, DA, BLD
hh_dbuilt	Année de construction de l'unité de logement	CMA, CD, CSD, ADA, CT, DA, BLD
hh_motorization	Motorisation du ménage	<i>MB</i>

Tableau 3.9 Résolutions des distributions marginales des variables de contrôle (suite)

Variable	Définition	Résolution(s) géographique(s)
pp_count	Nombre d'individus	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA, DB
pp_age	Catégorie d'âge de l'individu	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA
pp_sex	Sexe de l'individu	CMA, CD, <i>MB</i> , CSD, ADA, CT, DA
pp_marst	État matrimonial de l'individu	CMA, CD, CSD, ADA, CT, DA
pp_lfst	Statut d'activité de l'individu	CMA, CD, CSD, ADA, CT, DA
pp_occupation	Occupation principale de l'individu	<i>MB</i>
pp_drlicense	Permis de conduire de l'individu	<i>MB</i>
pp_transpass	Abonnement de transport en commun (TC) de l'individu	<i>MB</i>
pp_carshsub	Abonnement d'autopartage (AP) de l'individu	<i>MB</i>
pp_bikeshsub	Abonnement de vélopartage (VP) de l'individu	<i>MB</i>

Les démarches d'extraction et/ou de calcul des distributions de probabilités marginales des variables de contrôle du recensement de la population 2016 et de celles de l'enquête O-D 2018 sont détaillées ci-dessous.

3.1.5.1 Recensement de la population 2016

Les distributions de probabilités marginales des variables de contrôle provenant du recensement (Tableau 3.7) sont extraites des résumés du recensement de la population 2016 (Statistique Canada, 2019b) pour les résolutions CMA, CD, CSD, ADA, CT et DA. En effet, l'aire de diffusion est l'échelle la plus fine où les distributions marginales de l'ensemble des variables du recensement sont disséminées. Ainsi, pour les résolutions DB et BLD, certaines hypothèses et/ou calculs sont nécessaires pour dériver des variables de contrôle à partir des variables disponibles.

À l'échelle du DB, les distributions marginales des trois variables « ppcount » (nombre de personnes), « dcount » (nombre de logements privés) et « urdcount » (nombre de logements privés occupés par des résidents habituels) sont disponibles. La variable « urdcount » est assimilée à la variable « hhcount » à l'échelle du DB. Ainsi, les distributions marginales des variables « hhcount » et « ppcount » sont disponibles à cette résolution. À l'échelle du bâtiment, les variables « bld_consy » et « bld_dutype » sont respectivement équivalentes aux variables « hhdbuilt » et « hhdtype » du recensement. Le nombre de ménages dans chacun des bâtiments est équivalent au nombre de logements occupés. Le taux d'occupation des bâtiments d'un DB est supposé être égal à la division du nombre de logements privés occupés par des résidents habituels « urdcount » par le nombre de logements privés « dcount » dans le DB. Ainsi, le nombre de ménages « hhcount » pour chacun des bâtiments est obtenu en multipliant le nombre de logements dans le bâtiment « bld_flats » par le taux d'occupation du DB dans lequel il est situé. Les distributions marginales des variables « hhcount », « hhdbuilt » et « hhdtype » sont alors disponibles à l'échelle du bâtiment.

L'algorithme permettant d'établir les distributions de probabilités marginales des variables de contrôle du recensement de la population 2016 est inclus dans un fichier supplémentaire joint à l'article constituant le Chapitre 5 de la présente thèse (Khachman et al., 2022).

3.1.5.2 Enquête O-D 2018

Lorsque l'échantillon de l'enquête O-D 2018 est utilisé comme intrant désagrégé à la synthèse, l'utilisation des distributions marginales du recensement de la population 2016 réduirait drastiquement le nombre de variables de contrôle potentielles puisque le nombre de variables communes entre le recensement de la population 2016 et l'enquête O-D 2018 et dont une catégorisation commune peut être établie est limité. Ainsi, pour les expérimentations de synthèse de population menées en utilisant l'enquête O-D 2018, les distributions marginales des variables de contrôle considérées sont calculées à l'échelle du MB par la somme des facteurs de pondération des individus qui y résident pour les variables relatives aux individus, et la somme des facteurs de pondération des ménages qui y sont localisés pour les variables relatives aux ménages. Les distributions marginales des variables de contrôle sont ainsi produites à l'échelle du MB puisque le MB constitue l'unité géographique de pondération de l'enquête O-D 2018.

L'algorithme permettant d'établir les distributions de probabilités marginales des variables de contrôle de l'enquête O-D 2018 est inclus dans un fichier supplémentaire joint à l'article constituant le Chapitre 6 de la présente thèse.

3.2 Méthodologie de l'activité de recherche

Cette section donne un aperçu des contributions des articles constituant les Chapitre 4 à Chapitre 6, de leur articulation, ainsi que des méthodologies utilisées pour les atteindre. Les aspects discutés dans cette section sont détaillés dans leur chapitre respectif.

3.2.1 Étape 1 : établissement d'un diagnostic

Avant de lancer un processus de synthèse de population, plusieurs choix doivent être effectués dont les variables de contrôle, la catégorisation des variables contrôlées, l'algorithme de synthèse mais aussi la résolution de référence. En effet, ce dernier choix est souvent implicitement effectué lors de l'extraction des distributions de probabilités marginales des attributs sociodémographiques du recensement. Alors que des analyses de sensibilité ont été menées pour analyser l'impact de la variation des autres choix sur la qualité de synthèse, celle-ci ayant été longtemps résumée à l'ajustement sociodémographique, la question de l'impact de l'utilisation de différentes résolutions de référence sur la précision spatiale de la population synthétique n'a pas été abordée.

Ainsi, la première étape de la méthodologie générale suivie consiste à établir un diagnostic. Il s'agit d'étudier l'impact du choix de la résolution géographique de référence sur la qualité de la synthèse. Pour ce faire, deux sous-étapes sont naturellement requises : la définition d'indicateurs de qualité et les analyses de sensibilité.

3.2.1.1 Définition d'indicateurs de qualité

Pour définir un cadre méthodologique permettant d'optimiser la précision spatiale d'une population synthétique en particulier et sa qualité globale en général, il faut d'abord définir ce qu'est la précision spatiale, comment la mesurer, et comment l'intégrer dans un indicateur de qualité globale incluant à la fois les aspects d'ajustement sociodémographique et de précision spatiale. Ainsi, dans un premier temps, un indicateur de précision spatiale et un indicateur de qualité globale sont définis.

3.2.1.2 Analyses de sensibilité

Les distributions de probabilités marginales des variables de contrôle sont généralement disséminées à plusieurs résolutions géographiques par les instituts nationaux de la statistique. L'idée est alors de fixer le choix des variables de contrôle, de leur catégorisation, et de l'algorithme de synthèse, puis de générer plusieurs populations synthétiques pour le même territoire basées chacune sur la distribution marginale des attributs contrôlés à une résolution différente. Pour chacune des populations synthétiques générées, trois indicateurs de qualité sont calculés : l'ajustement sociodémographique, la résolution de référence et la qualité globale incluant les deux indicateurs précédents. Les valeurs que prend chaque indicateur pour les différentes résolutions de référence sont comparées pour en tirer des conclusions quant à l'impact de la résolution de référence sur la qualité de la synthèse. Cette analyse est menée sur les trois régions métropolitaines les plus grandes du Canada (Toronto, Montréal et Vancouver) afin de vérifier la possibilité de généralisation des conclusions tirées. L'hypothèse testée est que la désagrégation de la résolution de référence permettrait d'améliorer la qualité de synthèse. Celle-ci est confirmée par les résultats des expérimentations menées. La meilleure pratique consisterait alors à générer une population synthétique à l'échelle du bâtiment ou de l'unité de logement, autrement dit, une population synthétique explicitement spatialisée.

Les travaux effectués dans le cadre de cette première étape de la démarche de recherche sont détaillés dans l'article constituant le Chapitre 4 de cette thèse intitulé « Impact of the Geographic Resolution on Population Synthesis Quality » et publié dans le journal « ISPRS International Journal of Geo-Information » en 2021.

3.2.2 Étape 2 : définition d'un cadre méthodologique universel de synthèse spatialisée

Les méthodes conventionnelles de synthèse explicitement spatialisée les plus fiables suivent une approche séquentielle basée sur des distributions par correspondance totale (section 2.3.1.1.4). Ces méthodes présentent deux limitations principales. D'une part, tel que détaillé précédemment (section 2.3.1.2.2), des problèmes d'incohérence entre la population synthétique et l'environnement bâti pourraient émerger compte tenu de l'indépendance des processus de génération et de spatialisation de la population synthétique caractéristique de l'approche

séquentielle. D'autre part, il s'agit essentiellement d'applications ad hoc dépendantes de l'abondance de données spatialisées et d'une connaissance approfondie de la zone d'étude considérée et des données locales ce qui soulève des problèmes de transférabilité. En effet, contrairement aux données sociodémographiques ayant généralement des formats similaires dans différentes zones d'étude, les données spatialisées décrivant l'environnement bâti entre autres varient fortement en quantité, en nature et en format d'un endroit à un autre. Cela rend les méthodes de spatialisation explicite existantes difficilement transférables limitant par conséquent leur utilité. La transférabilité est en effet un élément clé d'une contribution méthodologique.

Pour remédier à ce problème, la deuxième étape de la méthodologie de recherche suivie consiste à proposer un cadre méthodologique universel de synthèse explicitement spatialisée de la population. L'idée est de proposer une méthodologie générique (i.e., ne comportant ni modèles localement calibrés ni paramètres tirés de caractéristiques spécifiques de la zone d'étude), basée sur une approche intégrée (i.e., permettant d'intégrer les étapes de génération et de spatialisation en générant une population synthétique à l'échelle du bâtiment directement), et présentant des exigences minimales en termes d'intrants et une qualité optimale de synthèse.

Parmi les algorithmes de synthèse existants, les algorithmes multi-résolutions auraient théoriquement pu être utilisés pour une synthèse explicitement spatialisée basée sur une approche intégrée, i.e., pour générer une population synthétique à l'échelle du bâtiment directement. Un contrôle multi-résolution est essentiel puisque les distributions marginales des variables d'intérêt ne sont généralement pas toutes disponibles à une échelle aussi fine que celle du bâtiment. Il s'agit essentiellement d'extensions de l'IPF telles que l'IPU renforcé (Konduri et al., 2016) permettant de contrôler des attributs observés à une résolution en plus de ceux dont les marginaux sont disponibles à la résolution de référence, et l'IPU tri-résolution (Moreno et Moeckel, 2018) permettant de contrôler des attributs observés à deux résolutions en plus de ceux dont les marginaux sont disponibles à la résolution de référence. Zhu et Ferreira (2014) ont aussi mentionné la possibilité d'intégrer la génération et la spatialisation d'une population synthétique dans un IPF unique. Ces algorithmes n'ont toutefois jamais été investis dans un exercice de synthèse explicitement spatialisée de la population probablement à cause de leur incapacité à traiter la dimension d'une matrice intégrant des contraintes à l'échelle du bâtiment.

Un nouveau cadre méthodologique intégré basé sur un algorithme efficace de synthèse par optimisation est alors introduit. Cet algorithme est défini de manière flexible permettant d'intégrer des contraintes à n'importe quels nombres de niveaux et de résolutions géographiques ce qui lui permet de répondre aux besoins variés d'une zone d'étude à l'autre. En effet, cette flexibilité permet de contrôler les variables d'intérêt aux échelles les plus désagrégées où leurs distributions marginales sont observées, peu importe leur nombre. Ainsi, grâce à son efficacité et sa flexibilité, le cadre méthodologique proposé permet d'inclure dans la matrice à optimiser des contraintes à l'échelle du bâtiment, alors adoptée comme résolution de référence. Une population synthétique est ainsi générée directement à l'échelle du bâtiment ce qui permet d'éviter toute incohérence entre les contraintes sociodémographiques et les caractéristiques de l'environnement bâti. Le cadre méthodologique proposé présente enfin une transférabilité optimale puisque ses exigences en termes d'intrants dans le cadre d'un exercice de synthèse explicitement spatialisée se limitent à une seule distribution de probabilités marginales d'une caractéristique des unités de logement à l'échelle du bâtiment (e.g., le nombre d'unités de logement dans chaque bâtiment).

Pour comparer la performance du cadre méthodologique multi-résolution intégré proposé à celle du cadre méthodologique conventionnel de synthèse explicitement spatialisée, chacun a été utilisé pour générer des populations synthétiques spatialisées pour l'ensemble des aires de diffusion de l'île de Montréal. La qualité des populations synthétiques générées est mesurée par les trois indicateurs de qualité définis à l'étape précédente en plus d'un indicateur isolant l'erreur à l'échelle du bâtiment. Malgré un ajustement sociodémographique légèrement inférieur essentiellement expliqué par la multiplicité des contraintes, le nouveau cadre méthodologique réalise des performances drastiquement meilleures en termes de précision spatiale améliorant par conséquent la qualité globale des populations synthétiques. Le cadre méthodologique intégré génère aussi moins d'erreurs à l'échelle du bâtiment que le cadre conventionnel.

Les travaux effectués dans le cadre de cette deuxième étape de la démarche de recherche sont détaillés dans l'article constituant le Chapitre 5 de cette thèse intitulé « Integrated Multiresolution Framework for Spatialized Population Synthesis » et publié dans le journal « Transportation » en 2022.

3.2.3 Étape 3 : exploitation des données spatialisées

Malgré ses avantages, notamment en termes de transférabilité et de cohérence entre la population synthétique et l'environnement bâti, la synthèse explicitement spatialisée de la population basée sur l'approche intégrée présente deux limitations principales : la représentativité sociodémographique potentiellement détériorée à cause de la disponibilité extrêmement limitée d'intrants à l'échelle du bâtiment, et la dimension potentiellement intraitable d'un exercice de synthèse incluant un contrôle d'attributs à une résolution aussi fine.

3.2.3.1 Problème de représentativité sociodémographique

En effet, les intrants généralement requis pour les algorithmes de synthèse tels que les distributions marginales d'attributs sociodémographiques et l'échantillon désagrégé, sont rarement disponibles à une échelle aussi fine que celle du bâtiment pour des raisons relevant essentiellement de la protection de la vie privée. Ainsi, la solution adoptée consiste à contrôler les variables non observées à l'échelle du bâtiment ou de l'unité de logement aux résolutions où elles sont observées moyennant un algorithme multi-résolution. Cependant, l'indisponibilité d'un échantillon désagrégé à cette résolution demeure problématique. En effet, l'information spatiale incluse dans les échantillons de microdonnées publiés par les instituts nationaux de la statistique est généralement très peu précise. Par exemple, la seule information incluse dans le hPUMF canadien est la Région Métropolitaine de Recensement d'appartenance d'un ménage. Face à cela, une solution serait d'utiliser l'ensemble du hPUMF pour générer une population synthétique pour chacun des bâtiments. Une telle manœuvre ferait toutefois émerger des problèmes au niveau de la représentativité sociodémographique de la population synthétique puisque l'échantillon de la CMA ne serait pas représentatif de la population de chacune des unités géographiques de synthèse considérées (Yaméogo et al., 2023; Nejad et al., 2021). En effet, plus l'écart d'agrégation entre la résolution de synthèse et la résolution de l'échantillon utilisé est important, plus ce problème de non-représentativité potentielle serait amplifié. Il constitue par conséquent un défi considérable de la synthèse à l'échelle du bâtiment.

Lorsque disponibles, les échantillons désagrégés explicitement spatialisés (e.g., échantillon de l'enquête O-D) constitueraient une option alternative à l'emprunt d'un échantillon relatif à une zone plus agrégée. En effet, il est possible d'en extraire un échantillon pour n'importe quelle zone de synthèse en sélectionnant les enregistrements localisés à l'intérieur de celle-ci. Toutefois cela

ne résout pas complètement le problème de synthèse à une résolution désagrégée car il n'est pas garanti de disposer d'enregistrements dans toutes les zones dont est constituée une telle résolution. En l'occurrence, lorsqu'il s'agit d'une synthèse à l'échelle du bâtiment, de nombreux bâtiments de la région d'étude n'incluront pas des ménages enquêtés. De plus, la taille des échantillons désagrégés pour de petites zones de synthèse, lorsque disponibles, est extrêmement limitée. Ils incluent par conséquent un nombre limité de combinaisons de catégories d'attributs sociodémographiques ce qui ferait émerger, lorsqu'utilisés dans un exercice de synthèse, des problèmes au niveau de la convergence de l'algorithme de synthèse et de l'ajustement sociodémographique de la population synthétique générée par conséquent (Guo et Bhat, 2007).

3.2.3.2 Problème d'évolutivité

Un autre défi rencontré lors de l'utilisation d'une approche intégrée pour la synthèse explicitement spatialisée de la population réside dans la dimension immense en termes de stockage de données, de manipulation de données et/ou de calculs que prendrait la matrice à optimiser lorsque les contraintes à une résolution de référence aussi fine y sont intégrées. Des machines virtuelles et/ou des structures particulières de données pourraient alors être requises.

3.2.3.3 Nouveau cadre méthodologique

Ainsi, l'objectif de cette troisième étape de la méthodologie de recherche est d'établir un cadre méthodologique basé sur l'approche séquentielle permettant à la fois d'éviter les limitations de l'approche intégrée décrites précédemment et d'exploiter la richesse des données spatialisées et la puissance des algorithmes d'apprentissage machine dans la performance d'une synthèse de population explicitement spatialisée de qualité. Le nouveau cadre méthodologique est développé de manière à limiter autant que possible les enjeux de transférabilité et d'incohérences entre la population synthétique et l'environnement bâti généralement rencontrées lors de l'utilisation d'une méthode de synthèse spatialisée basée sur l'approche séquentielle.

La méthode proposée consiste d'abord à générer une population synthétique zonale moyennant un algorithme de synthèse par ajustement. Ensuite, les bâtiments de la zone d'étude sont partitionnés en utilisant la méthode de partitionnement d'un grand ensemble de données ou « Clustering Large Applications » (CLARA) (Kaufman and Rousseeuw, 1990). Cette méthode particulièrement efficace et robuste face aux valeurs aberrantes permet d'optimiser d'une part l'applicabilité du

cadre méthodologique proposé à des zones d'études comptant de nombreux bâtiments telles que l'île de Montréal et d'autre part la qualité du partitionnement effectué. La méthode performante de classification par forêt aléatoire ou « Random Forest » (Breiman, 1996) est par la suite utilisée pour estimer la probabilité de chaque ménage synthétique d'habiter dans chaque bâtiment de sa zone de synthèse en fonction du groupe auquel ce dernier appartient, des caractéristiques du ménage et de la distance entre le bâtiment et le clone du ménage synthétique dans l'échantillon désagrégé spatialisé utilisé. Les paires ménages-bâtiments dans chacune des zones de synthèse sont enfin sélectionnées par ordre décroissant de leurs probabilités d'occurrence estimées, dans les limites de la fréquence des ménages synthétiques et de la capacité des bâtiments.

Ce cadre méthodologique est le premier de son type (partitionnement + classification) qui exploite des algorithmes d'apprentissage non supervisé « CLARA » et supervisé « Random Forest » performants pour prédire avec précision l'emplacement des ménages synthétiques à l'échelle du bâtiment directement. Il est appliqué pour générer une population explicitement spatialisée pour l'ensemble des secteurs municipaux de l'île de Montréal. Différents indicateurs sont calculés pour mesurer la performance des composantes du cadre méthodologique proposé soit la génération de la population synthétique zonale, le partitionnement des bâtiments, la classification des ménages synthétiques selon le groupe du bâtiment d'accueil et la distribution spatiale. Les valeurs obtenues sont toutes satisfaisantes.

Outre éviter les problèmes de l'approche intégrée tels que les problèmes de représentativité et d'évolutivité, la méthode basée sur l'approche séquentielle proposée permet de mitiger, sans totalement s'en débarrasser, les problèmes d'incohérences potentielles entre la population synthétique et l'environnement bâti et de transférabilité inhérents à cette famille de méthodes. Elle fait ainsi montre d'une certaine robustesse face aux potentielles incohérences entre la population synthétique et l'environnement bâti et demeure relativement transférable. En effet, l'utilisation d'une approche probabiliste d'association des ménages synthétiques et des bâtiments fournit une certaine flexibilité permettant d'intégrer les contraintes de fréquences des ménages synthétiques et de capacités des bâtiments dans la distribution spatiale effectuée. Les résultats montrent que 78.9% des ménages synthétiques ont été associés à un bâtiment du groupe le plus probable pour eux. La divergence des résultats obtenus avec ceux d'une approche similaire mais sans contrainte demeure alors limitée. De plus, aucune relaxation de contrainte (e.g., assigner un ménage synthétique à un bâtiment à l'extérieur de sa zone de synthèse) n'a été requise pour spatialiser l'ensemble des

ménages synthétiques. Sur le plan de la transférabilité, l'efficacité et la performance des algorithmes « CLARA » et « Random Forest » utilisés, et la disponibilité de plus en plus courante de données spatialisées, telles que les échantillons désagrégés explicitement spatialisés et les données relatives à l'environnement bâti, confèrent au cadre méthodologique proposé une transférabilité acceptable comparée aux méthodes existantes de synthèse explicitement spatialisée basées sur l'approche séquentielle. Cette transférabilité demeure toutefois limitée comparée à celle d'une méthode basée sur l'approche intégrée puisqu'elle est dépendante de la disponibilité de telles bases de données dans la zone d'étude considérée et de la performance des modèles entraînés sur celles-ci.

Les travaux effectués dans le cadre de cette troisième étape de la démarche de recherche sont détaillés dans l'article constituant le Chapitre 6 de cette thèse intitulé « A Novel Machine Learning-Based Spatialized Population Synthesis Framework » et soumis dans le journal « Transportation » en 2023.

CHAPITRE 4 ARTICLE 1: IMPACT OF THE GEOGRAPHIC RESOLUTION ON POPULATION SYNTHESIS QUALITY

Rédigé par :

Mohamed Khachman, Catherine Morency et Francesco Ciari

Publié le 19 novembre 2021 dans :

International Society for Photogrammetry and Remote Sensing (ISPRS) - International Journal of
Geo-Information,

No. 10(11):790, DOI: 10.3390/ijgi10110790

Abstract

Microsimulation-based models, increasingly used in the transportation domain, require richer datasets than traditional models. Precisely enumerated population data being usually unavailable, transportation researchers generate their statistical equivalent through population synthesis. While various synthesizers are proposed to optimize the accuracy of synthetic populations, no insight is given regarding the impact of the geographic resolution on population synthesis quality. In this paper, we synthesize populations for the Census Metropolitan Areas of Montreal, Toronto, and Vancouver at various geographic resolutions using the enhanced iterative proportional updating algorithm. We define accuracy (representativeness of the sociodemographic characteristics of the entire population) and precision (representativeness of the real population's spatial heterogeneity) as metrics of synthetic populations' quality and measure the impact of the reference resolution on them. Moreover, we assess census targets' harmonization and double geographic resolution control as means of quality improvement. We find that with a less aggregate reference resolution, the gain in precision is higher than the loss in accuracy. The most disaggregate resolution is thus found to be the best choice. Harmonization proves to further optimize synthetic populations while double control harms their quality. Hence, synthesizing at the Dissemination Area resolution using harmonized census targets is found to yield optimal synthetic populations.

Keywords: population synthesis; travel demand modelling; iterative proportional fitting; iterative proportional updating; enhanced iterative proportional updating; geographic resolution.

4.1 Introduction

Microsimulation-based models performed by transportation planners and engineers in the context of travel demand forecasting require complete disaggregate datasets describing a population of agents (households and/or individuals) as input. Collecting this type of data is costly, time-consuming, and complex (Mohammadian et al., 2010); thus, synthesis of the required datasets is the typical solution. Population synthesis is a process using aggregate and partially disaggregate data to list a fully enumerated population of agents (individuals and/or households) with sociodemographic characteristics. The goal is to generate a synthetic population that is statistically consistent with the real population as described by aggregate data (usually from the censuses).

The population synthesis process starts with the selection of sociodemographic characteristics according to which the synthetic population will be generated. When the synthetic population is intended to feed microsimulations of mobility behaviors, the characteristics having the most important behavioral effects in terms of transportation habits are used as control variables. Then, aggregate data (AD) at a chosen geographic resolution are extracted from census summary tables (e.g., Summary Files (SF) in the U.S.) which consist of one-, two-, or multiway tables containing the total marginals of the joint distribution of people and households' most important characteristics. Disaggregate datasets (DD) are drawn from a representative microdata sample of households and people with full sociodemographic characteristics detailed for each anonymized agent (e.g., Public Use Microdata Sample (PUMS) in the U.S. and Public Use Microdata Files (PUMF) in Canada). Entire multiway cross tabulations of control variables are drawn from the 5% - or less - disaggregate sample to be used in the population synthesis process. The correlation structure existing among sociodemographic variables in the microdata sample should be preserved in the synthetic population while fitting the totals of different combinations of sociodemographic characteristics to those observed in the census.

Fitting-based approaches, specifically synthetic reconstruction techniques, are the oldest and the most frequently used population synthesizing methods. In their paper, Beckman et al. (1996) were the first to apply the iterative proportional fitting (IPF) technique (Deming & Stephan, 1940) to synthesize a population of households using census and PUMS data. Since then, many papers addressing weaknesses of this technique have been published suggesting alternatives to the basic

algorithm implemented by Beckman et al. (1996) in the Transportation Analysis and Simulation System (TRANSIMS).

The IPF basic method is unable to concurrently account for individual and household control variables. Hence, synthetic populations obtained using this technique can match either individual-level or household-level constraints, but not both. X. Ye et al. (2009) made a major advancement in the field (Fournier et al., 2021) proposing an algorithm known as iterative proportional updating (IPU) that allows the synthetic population to match individual and household joint distributions. Hence, different weights are assigned to households that are identical with respect to household attributes but have different compositions of individuals. More details about IPF and IPU algorithms are provided in Section 4.2. Considering that control variables may sometimes be available at different geographic levels, Konduri et al. (2016) introduced an enhanced version of the IPU algorithm generating a synthetic population at two geographic resolutions simultaneously.

4.1.1 Problem statement

To ease the understanding of the paper, it is helpful at this point to clarify the terminology used. In this paper, a reference geographic resolution (RGR) refers to the type of census standard geographic areas at which the population synthesis is performed, i.e., for which the target AD are extracted. Each geographic resolution is made of geographic units. For instance, if we are synthesizing a population for all the census tracts of a city, the geographic division of the whole city into census tracts is the RGR, and each census tract is a reference geographic unit (RGU).

The choice of the RGR has an important impact on the synthetic population and the microsimulation it feeds. The more aggregate the RGR, the more likely spatialization errors will occur. This is because when an RGR is used for population synthesis, the population segments of less aggregate geographic resolutions are implicitly assumed to be homogeneous, i.e., uniformly distributed across each RGU. In other words, the population is assumed to be uniformly distributed on the less aggregate geographic units comprised in each RGU. A simple example would help to clarify this point. In Figure 4.1, a county δ comprised of two municipalities (orange and blue) is depicted. If a population is synthesized for δ considering the county as the reference geographic resolution, the synthetic population is assumed to be uniformly distributed on δ - as per Figure 4.1a - which means that the two municipalities' populations are assumed to be homogeneous. However, in reality, the orange municipality would account for more young men and the old women would

be more prevalent in the blue municipality as per Figure 4.1b. The mobility behaviors in such two municipalities would be drastically different due to the sociodemographic differences of their populations even though they are included in the same RGU δ . Hence, synthesizing a population at an aggregate level would lead to spatialization errors, thus altering the simulations of mobility behaviors fed by such a synthetic population.



Figure 4.1 δ county (a) synthetic population with the county used as RGR and (b) observed population

Except for a truly homogeneous population, the more aggregate the RGR used, the stronger is the homogeneity (spatial uniformity) assumption, and the more altered the simulated mobility behaviors will be. Thus, choosing a less aggregate RGR allows for more heterogeneity, in terms of sociodemographic characteristics and mobility behaviors, to be considered. It follows that the quality of spatialization of the synthetic population can better be assessed at the most disaggregate geographic resolution available. Furthermore, when the synthetic population is intended to be spatialized at the building scale (fully disaggregate from a spatial point of view), performing population synthesis at the least aggregate geographic resolution available in the census can ease the further spatialization by reducing the plausible locations for each synthetic household.

However, one assumption is that using census totals at the least aggregate geographic resolution may severely harm the performance of a fitting-based synthesizer. This is because lacking combinations of attributes and rounded zero marginals for privacy issues are more likely to occur at a less aggregate resolution. In fact, the more aggregate a geographic resolution is, the more its census marginals are expected to reflect reality. This is mainly due to a lower necessity to pre-

process the data (namely, round small values up or down) to preserve privacy. It follows that the quality of fit of the synthetic population can better be assessed at the most aggregate geographic resolution. Another drawback of using a less aggregate RGR is an increase in the synthesis complexity. In fact, a less aggregate RGR implies more RGUs, and thus more targets for the synthesizer to fit. For example, if a population is synthesized at the county resolution for δ , the synthesizer tries only to fit to the targets at the county level, e.g., the number of men in δ . However, for population synthesis at the municipality level, census targets for both the blue and orange municipalities need to be well fitted, e.g., the number of men in the blue municipality and the number of men in the orange municipality, etc. The synthesis targets and thus the potential fitting errors are doubled when shifting from the county to the municipality as an RGR. Hence, fitting errors become more numerous when using a less aggregate RGR, which means that the sociodemographic characteristics of the synthetic population will deviate more from those of the real population, and thus the simulation of mobility behaviors it feeds will become less accurate. The supposed impacts of different RGR aggregations on synthetic populations are summarized in Table 4.1.

Table 4.1 Supposed impacts of RGR aggregation on population synthesis

Reference Resolution Aggregation	Benefits	Drawbacks	Impact on Synthetic Population
More aggregate	- Fewer combinations of attributes missing - Fewer rounded zero marginals - Fewer targets to fit	- Stronger homogeneity (uniform spatial distribution) assumption	- Fewer potential fitting errors - More potential spatialization errors
Less aggregate	- Weaker homogeneity (uniform spatial distribution) assumption	- More combinations of attributes missing - More rounded zero marginals - More targets to fit	- More potential fitting errors - Less potential spatialization errors

As increasing and decreasing the RGR can both have benefits and drawbacks, synthesizing a population at two resolutions simultaneously would help take the best of both worlds. Multi-resolution population synthesis would allow the synthesizer to account for the heterogeneity of the population at the less aggregate geographic resolution while fitting to the more reliable marginal totals at the more aggregate geographic resolution. An ideal synthetic population is thus a

population which perfectly fits the households and individuals' constraints at both the least and the most aggregate geographic resolutions among the census standard geographic areas. However, the perfect fit of households and people distributions at two geographic resolutions is unlikely to occur. As for the IPU algorithm, the enhanced IPU solution for a simultaneous perfect fit of household and people distributions at two resolutions would probably involve negative weights due to the multiplicity of constraints (X. Ye et al., 2009). In this case, a corner solution (X. Ye et al., 2009), i.e., a solution prioritizing the perfect fit of one constraint over the others or averaging the fit of multiple constraints, is considered. Moreover, the data processing applied to census totals introduces inconsistencies of totals between different geographic resolutions, making the perfect fit of all the constraints even less likely. Hence, accounting for two resolutions may further damage the quality of the generated synthetic population. The choice of the RGR and whether to apply multiple geographic resolution controls or not, should thus be done cautiously to reach the best compromise between the spatial precision of the synthetic population (representativeness of the real population's spatial heterogeneity) and its accuracy (representativeness of the sociodemographic characteristics of the entire population). Accuracy and precision are thoroughly defined in Section 4.3.6.

4.1.2 Contributions

To improve the quality of the synthetic population, its accuracy and precision should be optimized. Optimizing the accuracy amounts to minimizing fitting errors and optimizing precision to minimizing spatialization errors. Since a more aggregate RGR would potentially lead to more spatialization errors and a less aggregate RGR to more fitting errors, the magnitudes of both types of errors at various geographic resolutions should be assessed to determine the geographic resolution yielding the best trade-off.

The main objective of this paper is to assess the impact of the RGR on the quality of the synthetic population, thus suggesting means of minimizing fitting and spatialization errors. Specifically, fitting and spatialization errors are measured for various RGRs with a focus on the impact on the errors from (1) the aggregation of the RGR, (2) data inconsistencies between census geographic resolutions, and (3) multiple geographic resolution controls.

The enhanced IPU algorithm is used in this paper to generate various synthetic populations for the Census Metropolitan Areas (CMAs) of Montreal, Toronto, and Vancouver. To the best of our

knowledge, the two most recent population synthesizers handling multiple geographic resolutions are the one introduced by Moreno and Moeckel (2018) and the enhanced IPU (Konduri et al., 2016). Moreno and Moeckel's algorithm (Moreno & Moeckel, 2018) can handle three resolutions simultaneously. However, as our need is limited to retrieving the best fit at two geographic resolutions (i.e., the most and the least aggregate geographic resolutions), an enhanced IPU-based algorithm is used.

The remainder of this paper is organized as follows. In Section 4.2, we discuss properties and variants of IPF and IPU-based population synthesis techniques as well as their advantages and limitations as exposed in the literature. Other multilevel and multiresolution population synthesis approaches are also briefly mentioned in this section. Section 4.3 is devoted to describing the methodology we have developed to assess the impact of various RGRs on enhanced IPU-based synthetic populations (Konduri et al., 2016). The comparison of results is then performed in Section 4.4 and discussed in Section 4.5. Section 4.6 concludes the paper and proposes some research perspectives.

4.2 Literature review

In this section, IPF, multilevel, and multiresolution synthesizers are briefly described based on the literature. A special focus is given to the evolution of population synthesis approaches from IPF to IPU (and enhanced IPU) as the latter is the main algorithm used in this paper. Their inputs, outputs, advantages, and limitations are also detailed.

4.2.1 Iterative Proportional Fitting (IPF)

Iterative proportional fitting (Deming & Stephan, 1940) is an algorithm that generally adjusts the cells of a table to pre-determined marginal totals. Cells' values are initialized and modified iteratively to fit with margins. The fitting process continues until convergence (tolerance must be set beforehand) or until a maximum number of iterations is reached (which also must be set beforehand). The algorithm converges for any convergence threshold chosen generally without exactly summing up to all the predetermined marginal totals (Hunsinger, 2008).

IPF-based population synthesis

IPF - also referred to as the conventional approach (Guo & Bhat, 2007) - has already been used in transportation modeling to synthesize populations of households (Beckman et al., 1996; Auld et al., 2009). For population synthesis purposes, a multiway table is seeded using the frequencies in the disaggregate sample of different types of agents with respect to chosen control variables, and the target marginals are extracted from the census summary files. Each dimension corresponds to a variable; thus, each cell of the table represents a unique type of agent, i.e., a unique combination of the control variables' categories. For instance, if a household's size and income are controlled, a cell would refer to the frequency of one-person households earning between 50 k\$ and 60 k\$. The standard IPF-based procedure (Beckman et al., 1996) takes place in two steps: fitting and allocation (Müller & Axhausen, 2010). The fitting step's objective is to make the frequencies in the table cells fit with the marginal targets. The sample frequencies are iteratively expanded and at the convergence, the frequency of each type of agent in the entire population is obtained. Then, the allocation process begins. Households are drawn from the microdata sample to match the expanded frequencies using Monte Carlo simulations and a synthetic population of households is thereby obtained. People belonging to selected households make up the synthetic population of individuals, but fitting is done at the household level only. A detailed example of this procedure is developed in the paper of Beckman et al. (1996).

IPF is a simple approach of population synthesis that has been proven to provide constrained maximum entropy estimates of the true population (Beckman et al., 1996). It also maintains the correlation structure of sociodemographic variables in the sample while fitting frequencies to census totals (Müller & Axhausen, 2010; Farooq et al., 2013). The IPF also converges quickly: a sufficient convergence is generally reached in about 10 to 20 iterations (Beckman et al., 1996).

Despite its attractive features, many modifications to the basic IPF algorithm have been proposed since the procedure also shows significant weaknesses. First, two types of zero cells are likely to be found in the initial matrix. Some zero cells are structural (Fournier et al., 2021) and consist of combinations of characteristics that do not exist in the sample or in the real population. The rest of the zeros are incorrect zero cells which consist of combinations of characteristics which do not exist in the sample but do exist in the real population. These incorrect zero frequencies prevent the IPF convergence. This issue was addressed using different techniques such as tweaking (Beckman et al., 1996), limiting the number of iterations (Guo & Bhat, 2007), or aggregating the most infrequent categories (Auld et al., 2009).

Integerization consists of converting the proportions of types of households obtained at the estimation step to integers representing the number of households of this type in the synthetic population (Guo & Bhat, 2007). This is another limitation of the IPF, as rounding inevitably alters the correlation structure of the multiway table and leads to unbalance of the total marginals against which the seed matrix has been fitted. Williamson et al. (1998) used a computationally expensive alternative method known as combinatorial optimization where integerization is avoided by drawing zone-by-zone agents from the DD into the zone list and iteratively assessing the contribution of the drawn agent to the goodness of fit of the distribution contained in the list.

The most important problem encountered when using IPF is that the basic procedure allows either household-level variables or person-level variables to be considered, but not both. Controlling only household-level variables leads the IPF to assign equal weights to households of the same type without considering their compositions in terms of the types of individuals. In this way, the joint distribution of person-level variables in the synthetic population could significantly diverge from marginals that appear in the census since it has not been fitted to them. Many modified IPF algorithms that overcome this issue were proposed. Guo and Bhat (2007) proposed checking for “household desirability” before drawing a household from the microdata sample to feed the synthetic population. Arentze et al. (2007) used relation matrices to convert marginal constraints at the person level to additional household-level constraints before using the IPF basic procedure to estimate household joint distributions. However, these methods do not fit households and people distributions simultaneously, and thus do not warrant their consistency (Yaméogo et al., 2021).

4.2.2 Multilevel synthesizers

As the mobility behaviors are determined both by people and households’ characteristics (Fabre & Morency, 2019; Loo & Lam, 2013; Kalter & Geurs, 2016), multilevel synthesizers are proposed. The multilevel synthesizers try to fit both households and people distributions by reweighting households according to their compositions of individuals (Yaméogo et al., 2021). The multilevel synthesizers can be divided into three categories: synthetic reconstruction, combinatorial optimization, and statistical learning (Yaméogo et al., 2021; Sun et al., 2018).

4.2.2.1 Synthetic reconstruction

The multilevel synthesizers that fall under the synthetic reconstruction category are an extension of the IPF fitting both households and people distributions mainly by reweighting households according to their composition of individuals. Müller and Axhausen (2011) suggest the hierarchical IPF as a multilevel synthesizer. At each iteration, the algorithm first fits the households' distribution, and each individual inherits its corresponding household's weight. Then, people's distribution is fitted, and each household's weight is calculated as the average of its people weights and so on. Bar-Gera et al. (2009) used entropy optimization to fit households and people distributions simultaneously while minimally altering initial households' weights (Müller & Axhausen, 2012). Generalized raking (Deville et al., 1993) can also be used for multilevel fit by distance functions minimization (Yaméogo et al., 2021). Fournier et al. (2021) tried to achieve multilevel fit using optimization-based reweighting approaches, such as non-negative least squares, non-negative least deviation, and cyclical coordinate descent.

Iterative proportional updating (IPU) (X. Ye et al., 2009) is a multilevel synthesizer used in this paper. The algorithm calculates a single weight for each household in the disaggregate sample that allows households and people distributions to be fitted simultaneously. Households of the same type with regard to the households' attributes but comprising individuals of different types thus get different weights. The weighting process starts with assigning a unit weight to each household in the disaggregate sample (X. Ye et al., 2009). The weights are then progressively updated so that the weighted sum of each household type meets its corresponding constraint. When the weighting according to the households' attributes is done, the weighting according to people's attributes begins. For each person type, the weights of the households that contain at least one individual of that type are updated so that the weighted sum of each person type meets its corresponding constraint (X. Ye et al., 2009). A complete set of adjustments to all households and people attributes constitutes a single iteration. At the end of each iteration, the gap between the constraints and the updated weighted sums (Δ) is calculated (X. Ye et al., 2009). The process is repeated iteratively until the Δ reduction is less than a pre-set tolerance. If a solution where household and person-level total values are simultaneously perfectly matched is impossible to find, the algorithm yields a corner solution (X. Ye et al., 2009), which usually consists of a perfect match of household-level totals, thus compromising the quality of fit at the person level. Even with a corner solution, the algorithm is found to considerably improve the fitting of person-level marginals compared to IPF.

A detailed example illustrating how the IPU algorithm operates is developed in the paper of X. Ye et al. (2009).

In addition to allowing the fit at person and household levels simultaneously, IPU has many other important features. First, unlike many population synthesis algorithms, IPU is adaptable to different situations, i.e., different control variables and categories. Second, IPU tackles the incorrect zero-cell problem and proposes a new solution that consists of borrowing the value from the microdata sample of the entire region when the considered type of households and/or people is missing from the sample of a smaller zone. To avoid side effects of this method, such as over-representing a character more frequently in the entire region than in the zone, a threshold value is pre-specified so that frequencies are borrowed only if they are below this value, which is otherwise used to fill a zero-cell. Once all zero cells have been modified, all non-zero cells are decreased by the sum of borrowed values divided by the number of non-zero cells, thus keeping the marginal sums unchanged (X. Ye et al., 2009). Finally, when generating a synthetic population for a small area, the zero marginals problem could occur, preventing the algorithm from converging. X. Ye et al. (2009) proposed assigning 0.01 values to zero-marginal cells, claiming that the effect of such a measure on the results is negligible.

At the selection step, the probability of a household being drawn from the microdata sample is calculated by dividing its weight by the total weight of households of the same type (X. Ye et al., 2009). The value obtained when this probability is multiplied by the total number of households in the considered area represents the number of households of the same type and with the same composition to be drawn and used in the synthetic population. Hence, an integerization problem occurs and the total number of households synthesized is generally inferior to the real number of households. X. Ye et al. (2009) proposed a new way of dealing with the integerization problem: a household is added to the cells of the household-level attributes' joint distribution where frequencies diverge the most from those estimated by IPF when household-level attributes are controlled for. Since the selection is based on Monte Carlo simulations, several synthetic populations should be drawn - at least 13, according to X. Ye et al. (2009) - before choosing the best one among them. When a population is generated while controlling household-level attributes, IPU outperforms IPF in terms of fit of person-level attributes (X. Ye et al., 2009). An improvement of the IPU algorithm allowing more social organization types to be synthesized and ensuring convergence to an optimal solution is proposed (P. Ye et al., 2022). Balakrishna et al. (2020)

suggest a simpler and faster population synthesis approach feeding the selection step with a household fitted distribution and IPU-adjusted household initial weights.

4.2.2.2 Combinatorial optimization

Williamson et al. (1998) developed a combinatorial optimization-based method using a conditional Monte Carlo drawing procedure that simultaneously maintains the match to household distribution and improves the quality of the fit at the person level. A set of households is initially arbitrarily drawn to reach the total number of households in the study area. Then, addition, removal, or replacement trials using the sample's households are carried out. For each trial, a goodness-of-fit indicator is calculated (Voas & Williamson, 2001). If the fit is improved, the new household is kept; otherwise, it is disregarded (Yaméogo et al., 2021).

Ma and Srinivasan (2015) used a combinatorial-optimization-based approach where the contribution of a sample household to satisfying constraints at all levels simultaneously is measured by an indicator called the “fitness value”. Households are selected in a decreasing order of assigned fitness values. The drawing process stops if the target total number of households is reached, or only households with negative fitness values remain. Abraham et al. (2012) used a hill-climbing-based solution to achieve the multilevel fit. Other algorithms have been used to resolve the combinatory optimization problem, such as simulated annealing (Harland et al., 2012), genetic algorithm (Birkin et al., 2006), and greedy heuristic (Srinivasan et al., 2008).

4.2.2.3 Statistical learning

Population synthesis algorithms falling in the statistical learning category are generally comprised of two steps: (1) estimating the joint distribution of control variables in the population and (2) sampling from the joint distribution estimated (Farooq et al., 2013; Saadi et al., 2016). Within this framework, the hierarchical Markov chain Monte Carlo (hMCMC) method (Casati et al., 2015) first defines a typology of people living within a household. It relies on conditional probabilities comprising variables relating to certain types of agents to estimate the multilevel joint distribution. The Bayesian network (Sun & Erath, 2015; Zhang et al., 2019) use graphical means to capture the joint distribution of households and people characteristics. Recently, deep generative modeling was applied to generate synthetic populations according to the learned joint distribution (Borysov et al., 2019; Garrido et al., 2020).

4.2.3 Multiresolution synthesizers

As control variables can be available at different spatial resolutions, multiresolution synthesizers are proposed to allow a simultaneous control of variables at two (Konduri et al., 2016) or three (Moreno & Moeckel, 2018) geographic scales. Konduri et al. (2016) developed an enhanced IPU algorithm that can consider person and household-level constraints at two geographic resolutions simultaneously. The weighting process is based on the same principle as the basic version of IPU. Sample households' weights, initially equal to 1, undergo multiple iterations of four fitting steps where they are sequentially modified to fit household attributes at the REGION level, then person attributes at the REGION level, then household attributes at the GEO level, then person attributes at the GEO level. Here, the REGION refers to the more aggregate and the GEO to the less aggregate geographic resolution. During the fitting sequence, a household's weight is updated only if, at the geographic resolution considered, (1) it belongs to the household type being fitted or (2) it comprises the type of people being fitted. The authors demonstrate that doing so improves the fit of the generated synthetic population at the more aggregate geographic resolution, i.e., at the REGION level, especially when various control variables are available at different geographic resolutions.

Moreno and Moeckel (2018) developed a population synthesis algorithm that can handle three geographic resolutions simultaneously. However, as stated in the Introduction, we aim to minimize errors at two geographic resolutions: the most aggregate (fitting errors) and the most disaggregate (spatialization errors) ones. Hence, controlling more than two geographic resolutions simultaneously does not help answer this paper's research questions, especially as the control variables we use are available at all the geographic resolutions considered. This algorithm is thus not used in this paper.

4.3 Materials and methods

4.3.1 Study area

In this paper, an enhanced-IPU based algorithm was used to generate synthetic populations for the CMAs of Montreal, Toronto, and Vancouver, Canada. These three CMAs were chosen since they are the three largest Canadian CMAs in terms of population. The geographic locations of the three CMAs are shown in Figure 4.2.



Figure 4.2 Geographic locations of Montreal, Toronto, and Vancouver CMAs

4.3.2 Control variables

A preliminary step to launching the algorithm is making the choice of variables that will be controlled along the population synthesis process. Some people and households' attributes that are typically included in travel studies were selected. For instance, age, sex, and marital status were controlled for people, and size, type, and net income were controlled for households. The total number of people and the total number of households were also controlled for. The categories of control variables were chosen to minimize incorrect zero values as well as because of their relevance for travel studies. Table 4.2 summarizes the control variables and associated categories.

Table 4.2 Control variables

Variable	Definition	Categories	Categories' Description
ppcount	People count	1	1 person
ppage	People age	1	[0, 15]
		2	[15, 65]
		3	65+
ppsex	People sex	1	Men
		2	Women
ppmarst	People marital status	1	Never legally married (and not living in common law)
		2	Legally married (and not separated)
		3	Living common law
		4	Separated, divorced, or widowed (and not living in common law)

Table 4.2 Control variables (suite)

Variable	Definition	Categories	Categories' Description
hhcount	Household count	1	1 household
hhsize	Household size	1	1 person
		2	2 people
		3	3 people
		4	4 people
		5	5+ people
hhtype	Household type	1	Non census family ¹
		2	1 census family without children
		3	1 census family with children
		4	Multiple census families
hhnetinc	Household net income	1	Less than 30k\$
		2	[30k\$, 60k\$]
		3	[60k\$, 100k\$]
		4	More than 100k\$

¹ “Census family is defined as a married couple and the children, if any, of either and/or both spouses; a couple living common law and the children, if any, of either and/or both partners; or a lone parent of any marital status with at least one child living in the same dwelling and that child or those children.” (Statistics Canada, 2018).

4.3.3 Zoning system

The RGRs used were selected among five standard geographic areas defined by Statistics Canada (2019b). The standard geographic areas for the three CMAs are shown in Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7 with the corresponding number of geographic units. They are defined - in decreasing order of aggregation - as follows:

- Census Metropolitan Area (CMA): Area with a total population of 100,000 where at least 50,000 are concentrated in a population centre (Statistics Canada, 2018);
- Census Subdivision (CSD): Generally equivalent to a municipality (Statistics Canada, 2018);
- Aggregate Dissemination Area (ADA): Area gathering 5000 to 15,000 people according to the previous census counts (Statistics Canada, 2018);
- Census Tract (CT): Area gathering between 2500 to 8000 people (Statistics Canada, 2018);
- Dissemination Area (DA): A subdivision of the CT with a population of 400 to 700 people (Statistics Canada, 2018).

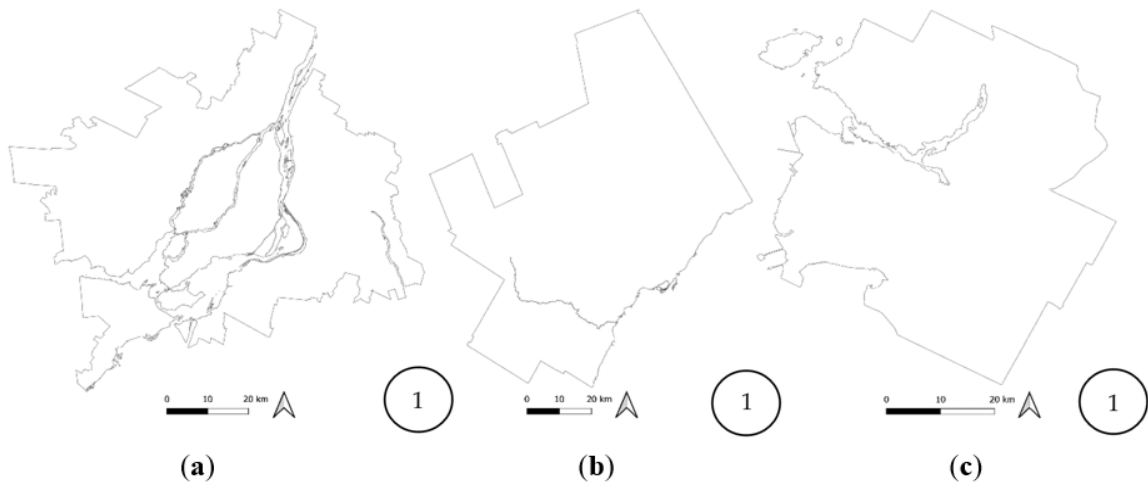


Figure 4.3 CMA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver

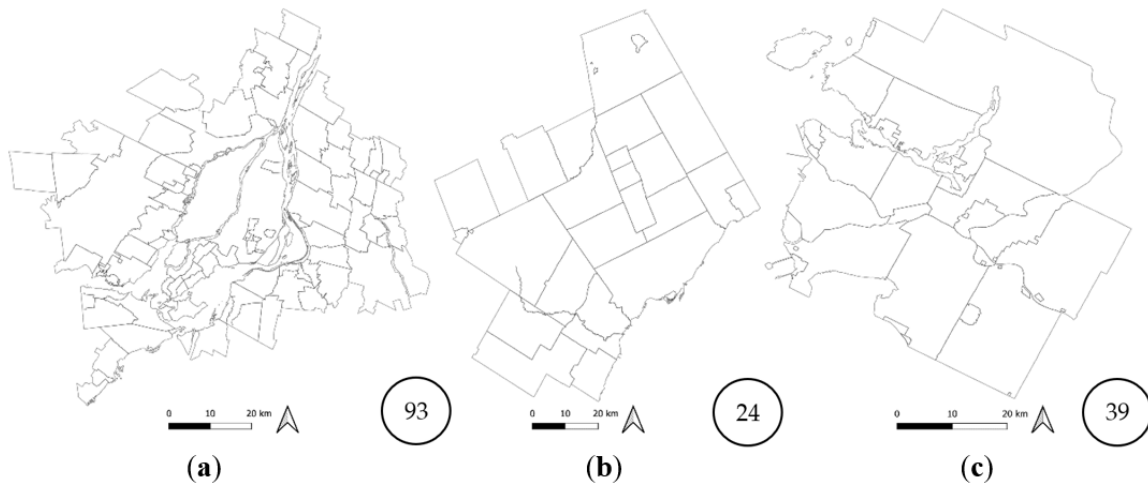


Figure 4.4 CSD boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver

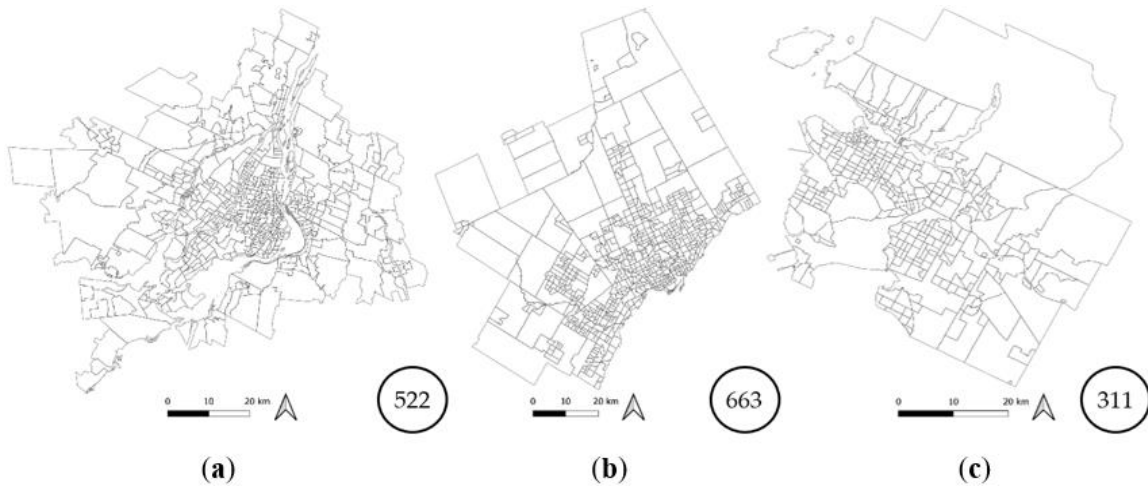


Figure 4.5 ADA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver



Figure 4.6 CT boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver



Figure 4.7 DA boundaries for (a) Montreal, (b) Toronto, and (c) Vancouver

It is important to mention that the variations of population and area between geographic resolutions are proportional. For example, each CT which is a more aggregate geographic unit than the DA in our definition, has at the same time a larger population and a larger area than each DA it contains. It is also worth mentioning that the zoning system defined is consistent, i.e., each geographic unit is included in only one geographic unit of the more aggregate geographic resolutions. For example, a DA belongs to only one CT, one ADA, one CSD, and one CMA.

4.3.4 Datasets

Although some sample-free population synthesizers were conceived (Barthelemy & Toint, 2013; Gargiulo et al., 2010), the majority of population synthesizers still require both aggregate (AD) and

disaggregate (DD) data to be used as inputs. To perform our tests, we used datasets at different geographic resolutions from the 2016 Canadian census summary tables and PUMF. The AD used come from the 2016 Canadian census summary tables (Statistics Canada, 2019b) while the DD were extracted from the hierarchical PUMF 2016 (Statistics Canada, 2019a) for each CMA. The hierarchical PUMF includes a full set of demographic and socioeconomic characteristics for each people and household. Hence, an initial frequency matrix, to be expanded by enhanced IPU later (Konduri et al., 2016), can easily be derived from the sample. All the data were filtered to the selected CMAs. Census summary tables were extracted for CMA, CSD, ADA, CT, and DA resolutions. However, PUMF are only available at the CMA resolution. Hence, the disaggregate sample at the CMA resolution was used to synthesize populations at less aggregate geographic resolutions. An underlying assumption of such a practice is that the correlation structure among control variables is constant across the considered geographic units (Beckman et al., 1996). However, a larger sample, i.e., having a richer pool of agents, especially when synthesizing at the finest geographic scales, helps avoid poor variance among synthesized agents.

The Montreal CMA has around 4.1 million people grouped into roughly 1.7 million households. The Toronto CMA comprises around 5.9 million people in nearly 2.1 million households. The Vancouver CMA is the smallest of the three considered areas with around 2.5 million people in a little less than 1 million households. For the three CMAs, the PUMF size is about 1% of the population. A brief portrait of these areas is drawn in Figure 4.8-Figure 4.12. The distributions of ages per sex in Figure 4.8-Figure 4.10 were extracted from the 2016 Census PUMF and summary files. The three CMAs have similar distributions, showing peaks for people who are less than 10 years old and those who are between 55 and 65 years old. The distributions of men and women per age group are fairly similar. The PUMF sampling generally respects the distribution in the population. People between 10 and 30 years old were found to be slightly undersampled in the PUMF while people between 30 and 45 were slightly oversampled for the three CMAs.

The distributions of households according to their size in the PUMF and the population of the three CMAs are shown in Figure 4.11. The distributions of households according to their size in the PUMF fairly followed the distributions reported in their respective populations, except for households of 5 people or more, who are noticeably underrepresented. Figure 4.12 exhibits the distributions of households according to their net income in the PUMF and the population of the three CMAs. No explicit divergence was detected. The distributions of the other control variables



Figure 4.8 Distributions of ages per sex in Montreal CMA's PUMF and population



Figure 4.9 Distributions of ages per sex in Toronto CMA's PUMF and population



Figure 4.10 Distributions of ages per sex in Vancouver CMA's PUMF and population

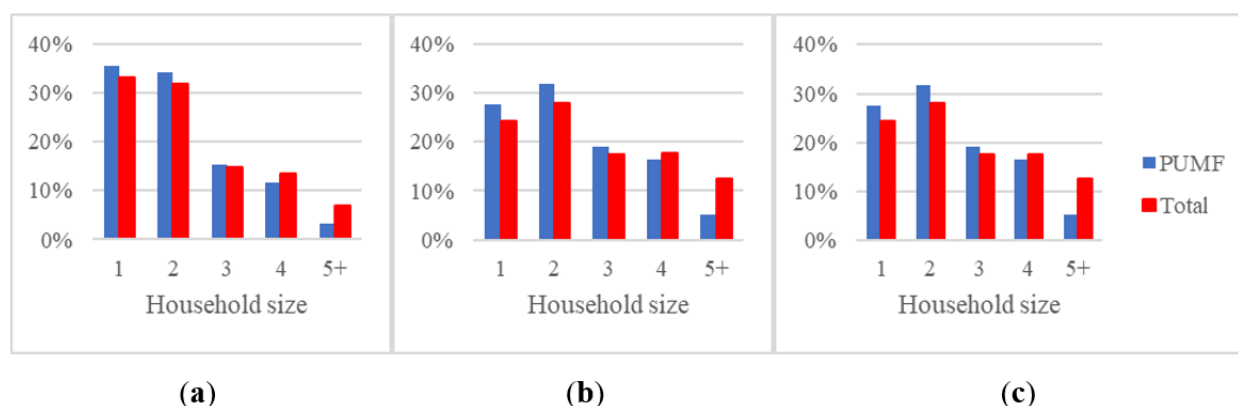


Figure 4.11 Households' distributions according to household size in the PUMF and the population of the CMAs of (a) Montreal, (b) Toronto, and (c) Vancouver

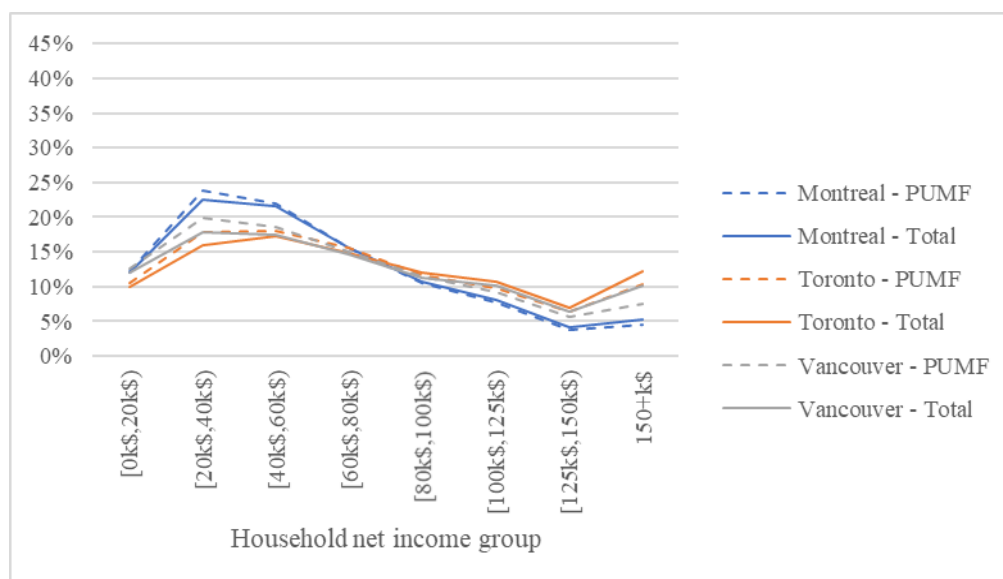


Figure 4.12 Households' distribution according to household after-tax income in the PUMF and the population of the CMAs of Montreal, Toronto, and Vancouver

are not exhibited in this paper for the sake of brevity. However, the PUMFs were found to reflect well the distributions of the control variables in their corresponding populations.

4.3.5 Data inconsistencies

It is important to note that some inconsistencies exist among the aggregate data. The census data present intra- and inter-resolution inconsistencies. The intra-resolution inconsistencies are the differences between the variables totals, e.g., the sum of men and women for the CMA being different than the sum of people belonging to all the age groups at the CMA resolution. The inter-

resolution inconsistencies are the differences between the frequency of a variable category for a geographic unit and the sum of the frequencies of the same variable category for the less aggregate geographic units within it. An example of intra- and inter-resolution data inconsistencies in Montreal CMA is shown in Table 4.3.

Table 4.3 Example of data inconsistencies (Montreal CMA)

Variable	Category	CMA	$\sum DA$	Inter-resolution Inconsistencies $ CMA - \sum DA $
ppcount	-	4,098,927	4,096,356	2571
ppsex	Men	2,000,935	1,999,445	1490
	Women	2,097,990	2,096,820	1170
Intra-resolution inconsistencies $ ppcount - \sum ppsex $ (total)	-	2	91	

Harmonization process

To isolate the effect of these inconsistencies on the enhanced IPU-based synthesizer, harmonized census totals, i.e., census totals without intra- and inter-resolution inconsistencies, were calculated. Before describing the harmonization process, a cluster of geographic units, intra-resolution adjustments, and inter-resolution adjustments should be defined. A cluster of geographic units is comprised of all geographic units belonging to the same more aggregate geographic unit. For example, a cluster of DAs is all the DAs belonging to the same CT. An intra-resolution adjustment refers to the adjustment of categories' frequencies of a control variable in a geographic unit so that their sum meets the total number of corresponding agents in the same geographic unit. For example, the frequencies of men and women in a CSD are proportionally adjusted so that their sum meets the total number of people in the same CSD. An inter-resolution adjustment refers to the adjustment of the frequencies of a control variable's category in all the geographic units of a cluster, so that their sum meets the frequency of the same variable's category at the more aggregate geographic unit they belong to. For example, the frequencies of households of three people in all the DAs of a DA's cluster are proportionally adjusted so that their sum meets the frequency of households of three people at the corresponding CT.

The harmonization process is comprised of three steps:

1. Intra-resolution adjustment of all control variables at the CMA resolution;
2. Inter-resolution adjustment of hhcount and ppcount for all geographic units' clusters at CSD, ADA, CT and DA resolutions;
3. Iterative application of intra-resolution adjustment for all control variables, and inter-resolution adjustment for all geographic units' clusters at CSD, ADA, CT, and DA resolutions, respectively. When convergence is reached at the CSD resolution, the iterative adjustment is launched at the ADA resolution and so on until the algorithm converges at the DA resolution. A convergence threshold of 10^{-5} is considered.

Step 3 is indeed an IPF applied to the frequencies of each variable's categories in all the geographic units of each cluster. The goal is to fit them first, to the total of their corresponding agent within each geographic unit, then to their corresponding frequencies at the more aggregate geographic unit they belong to. For example, men frequencies at the DA resolutions are adjusted as follows (Figure 4.13): first, the frequencies of men and women in each DA are proportionally adjusted so that their sum meets the total number of people in the same DA. Then, the men's frequency in each DA undergoes a proportional adjustment so that the sum of men's frequencies in each cluster of DAs meets the men's frequency at the corresponding CT. The process is then iteratively repeated until the frequencies of all variable categories at the DA resolution meet both constraints within a convergence threshold. Harmonized datasets, i.e., datasets without intra- and inter-resolution inconsistencies, are thus obtained for the five geographic resolutions considered.

Multiway table	DA1	DA2	...	DAn	Constraint 2
Men	x	x	x	x	Men in correspond-ing CT
Women	x	x	x	x	Women in correspond-ing CT

Constraint 1	People in DA1	People in DA2	People in DA3	People in DA4
---------------------	---------------	---------------	---------------	---------------

Figure 4.13 Example of step 3 of the harmonization process applied to a cluster of n DA

4.3.6 Scenarios

As mentioned previously, the enhanced IPU algorithm (Konduri et al., 2016) can simultaneously take into consideration two geographic resolutions for households and people attributes. In this paper, the REGION was always set to the CMA when a double control was applied. This is because the fitting errors are better assessed at the CMA resolution, as explained in the Introduction; thus, adding controls at the CMA resolution would be the best way to reduce fitting errors. Using the enhanced IPU algorithm implemented in PopGen2.0 (MARG, 2016), 18 synthetic populations were generated for each CMA according to the scenarios enumerated in Table 4.4. Scenarios with harmonized data were tested to assess the effect of intra- and inter-resolution inconsistencies. Scenarios with two controlled resolutions were compared to scenarios with a single controlled resolution to show the impact of the additional control at the CMA resolution.

Table 4.4 Scenarios

Scenario	Data Type	Controlled Levels	REGION	GEO
1	Raw	1	-	CMA
2	Raw	1	-	CSD
3	Raw	2	CMA	CSD
4	Raw	1	-	ADA
5	Raw	2	CMA	ADA
6	Raw	1	-	CT
7	Raw	2	CMA	CT
8	Raw	1	-	DA
9	Raw	2	CMA	DA
10	Harmonized	1	-	CMA
11	Harmonized	1	-	CSD
12	Harmonized	2	CMA	CSD
13	Harmonized	1	-	ADA
14	Harmonized	2	CMA	ADA
15	Harmonized	1	-	CT
16	Harmonized	2	CMA	CT
17	Harmonized	1	-	DA
18	Harmonized	2	CMA	DA

Accuracy and precision

For each synthetic population generated, the accuracy and the precision were assessed. The accuracy reflects the representativeness of the sociodemographic characteristics of the entire population and is measured by the fit of the total synthetic population to the targets at the CMA resolution. Hence, the sum of estimated frequencies of each variable's category across the RGUs

was calculated and compared to the observed frequency of the same variable's category at the CMA resolution. For example, the sum of synthetic men across DAs was calculated and compared to the frequency of men at the CMA level.

The precision reflects the representativeness of the real population's spatial heterogeneity. Precision assessment requires prior data processing. The frequencies of variables' categories were first interpolated from each RGU to the DAs within it. The interpolation was done proportionally to the distribution of the RGU's households on the DAs within it. This is because the household is the main synthesis agent for the enhanced IPU algorithm. The calculations were performed according to the following formula:

$$im_{i,DA_j} = m_{i,RGU} \times \frac{hhcount_{DA_j}}{hhcount_{RGU}} \quad (4.1)$$

where

- i denotes the i th variable category;
- j denotes the j th DA within the RGU;
- RGU refers to a reference geographic unit;
- $m_{i,RGU}$ refers to the frequency of the i th variable category in an RGU, as estimated by the enhanced IPU;
- im_{i,DA_j} refers to the interpolated frequency of the i th variable category in the j th DA;
- $hhcount_{DA_j}$ refers to the households' count in the j th DA;
- $hhcount_{RGU}$ refers to the households' count in the RGU.

Then, a synthetic population was drawn for each DA using the interpolated frequencies. This can be done using the “synthesize” function in the ipfr R package (Ward, 2020). The frequencies of variables' categories were compared to the census targets for each DA, and the more similar they were, the more precise was the synthetic population. The mathematical formulas for fitting and spatialization errors calculations are detailed in Section 4.3.7. The higher the fitting errors, the less accurate the synthetic population was, and the higher the spatialization errors, the less precise was the synthetic population.

As stated previously, the main objective was to assess the impact of the RGR's choice on the enhanced IPU algorithm performance to provide insights into the best compromise between

precision and accuracy. The variables were controlled in the same order for all scenarios with the people counts being the last variable to be fitted. The 18 scenarios were run for the CMAs of Montreal, Toronto, and Vancouver.

4.3.7 Assessment criteria

Three indicators were calculated to assess the various scenarios: census inter-resolution inconsistencies, fitting errors, and spatialization errors. These criteria are described below. As we were interested in a good fit of households and people, errors on both types of agents were integrated in the formulas. Moreover, the indicators were calculated per 1000 agents. They were made relative to the number of agents in order to keep the results of the three CMAs comparable, as they have drastically different sizes of populations. Finally, as the selection step was not deterministic (X. Ye et al., 2009), i.e., the synthetic population differed for each simulation, the indicators were calculated on synthetic populations that were averaged across 50 simulations to allow for general conclusions.

4.3.7.1 Census inter-resolution inconsistencies (α)

As described previously, the census inter-resolution inconsistencies indicator was calculated as the sum of the absolute values of the differences between the observed frequency of each variable's category in a geographic unit and the sum of the observed frequencies of the same variable's category in the less aggregate geographic units within it. As the census targets at the CMA resolution were the most reliable, the inter-resolution inconsistencies were calculated relative to the CMA resolution as per the following formula:

$$\alpha(\text{‰}) = \frac{\sum_{i=1}^m \left| \sum_{j=1}^n M_{i,\text{RGU}_j} - M_{i,\text{CMA}} \right|}{\text{hhcount}_{\text{CMA}} + \text{ppcount}_{\text{CMA}}} \times 1000 \quad (4.2)$$

where

- i denotes the i th variable category $\forall i = 1 \dots m$;
- j denotes the j th RGU $\forall j = 1 \dots n$;
- M_{i,RGU_j} refers to the observed frequency of the i th variable category in the j th RGU;
- $M_{i,\text{CMA}}$ refers to the observed frequency of the i th variable category in the CMA;
- $\text{hhcount}_{\text{CMA}}$ refers to the households' count in the CMA;

- $ppcount_{CMA}$ refers to the people's count in the CMA.

4.3.7.2 Fitting errors (β)

The fitting errors formula is similar to the census inter-resolution inconsistencies one, except for the differences being calculated between the observed frequencies of each variable's category in a geographic unit, and the sum of the simulated frequencies of the same variable's category in the less aggregate geographic units within it. The fitting errors were calculated as follows:

$$\beta(\%) = \frac{\sum_{i=1}^m \left| \sum_{j=1}^n m_{i,RGU_j} - M_{i,CMA} \right|}{hhcount_{CMA} + ppcount_{CMA}} \times 1000 \quad (4.3)$$

where

- i denotes the i th variable category $\forall i = 1 \dots m$;
- j denotes the j th RGU $\forall j = 1 \dots n$;
- m_{i,RGU_j} refers to the simulated frequency of the i th variable category in the j th RGU;
- $M_{i,CMA}$ refers to the observed frequency of the i th variable category in the CMA;
- $hhcount_{CMA}$ refers to the households' count in the CMA;
- $ppcount_{CMA}$ refers to the people's count in the CMA.

4.3.7.3 Spatialization errors (γ)

To measure the spatialization errors, the absolute differences between the simulated frequencies of the variable categories interpolated at the DA resolution (im_{i,DA_j}) and the observed frequencies of the variable categories at the DA resolution were calculated, then summed over all the variable categories and the DAs as follows:

$$\gamma(\%) = \frac{\sum_{i=1}^m \sum_{j=1}^n \left| im_{i,DA_j} - M_{i,DA_j} \right|}{hhcount_{CMA} + ppcount_{CMA}} \times 1000 \quad (4.4)$$

where

- i denotes the i th variable category $\forall i = 1 \dots m$;
- j denotes the j th RGU $\forall j = 1 \dots n$;
- im_{i,DA_j} refers to the interpolated simulated frequency of the i th variable category in the j th DA;

- M_{i,DA_j} refers to the observed frequency of the i th variable category in the j th DA;
- $hhcount_{CMA}$ refers to the households' count in the CMA;
- $ppcount_{CMA}$ refers to the people's count in the CMA.

4.4 Results

In this section, different groups of the 18 generated synthetic populations were compared. This provides insights into how the accuracy and the precision of the synthetic population were impacted by (1) the RGR's aggregation, (2) the data inconsistencies between census geographic resolutions, and (3) the multiple geographic resolutions control. Means to improve both accuracy and precision of synthetic populations are then suggested.

4.4.1 How Do α , β , and γ vary according to the RGR used?

The variations of α , β , and γ according to the RGR are depicted in Figure 4.14, Figure 4.15 and Figure 4.16. The indicators generally show the expected trends, with the three CMAs showing similar results for each indicator.

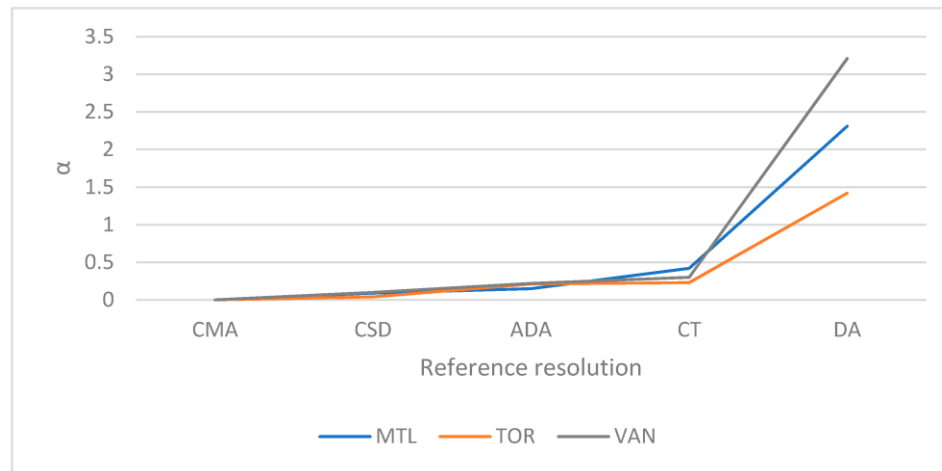


Figure 4.14 Variation of α according to the RGR

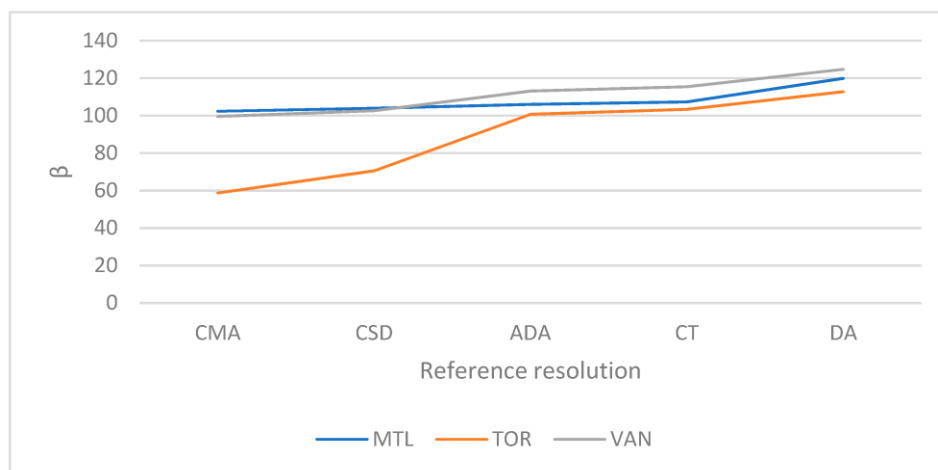


Figure 4.15 Variation of β according to the RGR

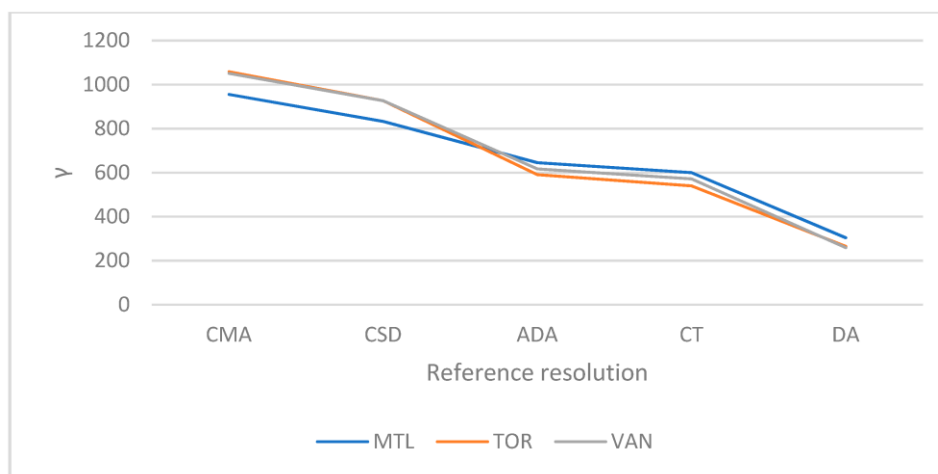


Figure 4.16 Variation of γ according to the RGR

α increased when the RGR became less aggregate, i.e., when the number of RGUs increased. In fact, the less aggregate the geographic unit is, the more rounded frequencies for privacy issues are likely to occur in the census data, thus yielding higher inter-resolution inconsistencies. α was, however, found to be of relatively low magnitude ranging from 0 at the CMA resolution to less than 3.5 at the DA resolution.

β also increased when the RGR became less aggregate for the three CMAs. This is also expected since both the inter-resolution inconsistencies (α) and the synthesis exercise complexity increased when the RGR became less aggregate, thus impacting the accuracy of the synthetic population. When synthesizing at the CMA resolution, only one set of census targets has to be met, compared to 6469 sets of targets at the DA level for Montreal, more than 7525 for Toronto, and more than

3450 for Vancouver. This makes the synthesis exercise more complex at this level and thus potential fitting errors multiply. It is important to mention that the problem of incorrect zero cells, more important at less aggregate resolutions, does not explain the variation of fitting errors in our case since the PUMF was always taken at the CMA level. However, the zero marginals problem, mainly due to rounding for privacy issues, is in itself a challenge for fitting-based synthesizers convergence, apart from increasing α , and thus is damaging the synthetic population's accuracy. β was found to range from a minimum of around 60 (Toronto) at the CMA level to around 120 at the DA level. A lower β was observed for Toronto at the CMA and CSD levels when compared to Montreal and Vancouver. The fluctuation of the corner solution found by the algorithm is a plausible explanation. However, quantifying the impact of the CMA's structure in terms of types of households and people on the corner solution is beyond the scope of this paper. However, the β maximum values at the DA resolution, being fairly similar for the three CMAs, provide insight on the cost in terms of fitting errors of synthesizing at the least aggregate geographic resolution.

γ was found to decrease with a less aggregate RGR. This is expected since higher aggregation results in a stronger spatial homogeneity assumption and thus a less precise synthetic population. The spatial homogeneity assumption became stronger with more aggregate RGRs because a large population over a wide area is more likely to be spatially heterogeneous than a small population over a compact area. The three CMAs showed similar trends and values with the maximum γ being around 1000 at the CMA resolution and the minimum around 300 at the DA resolution. Two observations are worth mentioning: first, the spatialization errors' magnitude was higher than the fitting errors' magnitude for the same synthesis area (the CMA). Second, the ratio of the highest error to the lowest error was higher than 3 for spatialization errors, while it remained lower than 2 for fitting errors (1.2 if calculated for Montreal and Vancouver). This shows that a synthetic population is generally susceptible to more spatialization errors than fitting errors. Hence, for the same synthesis area, perfect precision is more difficult to achieve than perfect accuracy. Moreover, it shows that the gain in terms of precision when synthesizing at the least aggregate RGR is more important than the loss in terms of accuracy and vice-versa.

As we were interested in optimizing both accuracy and precision, i.e., minimizing both fitting and spatialization errors, the variation of the total error ($\beta + \gamma$) according to the RGR used was calculated as depicted in Figure 4.17

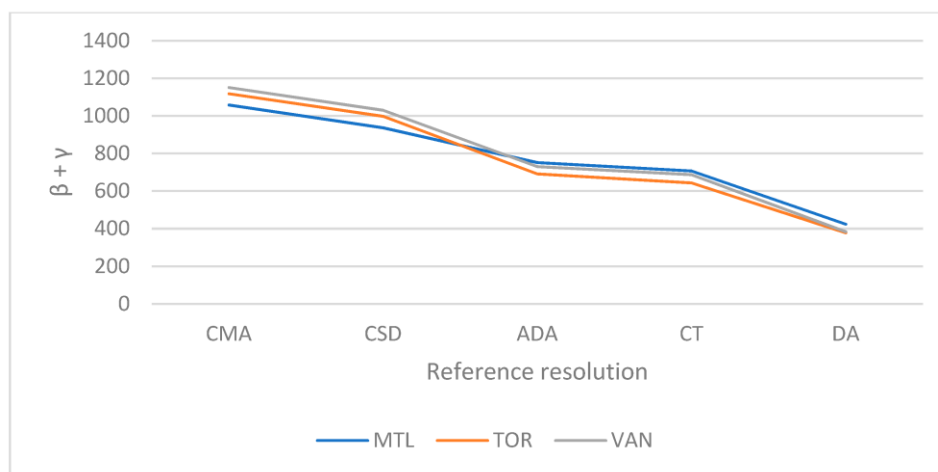


Figure 4.17 Variation of $\beta + \gamma$ according to the RGR

The synthetic populations at the DA resolution showed around 400 total errors per 1000 agents, while at the CMA resolution around 1100 errors per 1000 agents were observed. The total error was reduced by nearly 64% at the DA resolution. Hence, using the DA as the RGR was shown to be the best compromise between fitting and spatialization errors. In other words, using the DA as the RGR allows the quality, i.e., the combination of accuracy and precision, of the synthetic population to be optimized.

4.4.2 How does γ vary according to β ? In other words, how is the precision improved when decreasing accuracy, i.e., when using a less aggregate RGR, and vice-versa?

β was found to increase and γ to decrease when the RGR became less aggregate. The variation of γ according to β was then further investigated in the three CMAs (Figure 4.18). The relation between γ and β could be fitted well by a decreasing linear trend as evidenced by the high R^2 calculated. For Montreal and Vancouver, when the RGR decreased, each additional fitting error per 1000 agents gave in return around 30 less spatialization errors per 1000 agents. For Toronto, each additional fitting error per 1000 agents was found to give in return around 13 spatialization errors per 1000 agents. However, Toronto's divergence from Montreal and Vancouver was mainly due to its low β errors at the CMA and CSD resolutions as shown in Figure 4.15. If the errors at CMA and CSD resolutions for Toronto are neglected, the slope of the trendline fitting the variation of γ according to β (black trendline) becomes similar to that of Montreal and Vancouver.

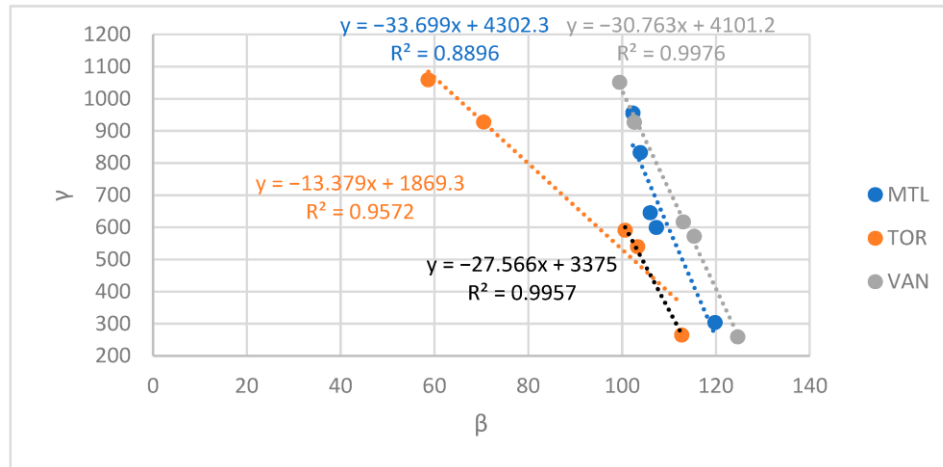


Figure 4.18 Variation of γ according to β

The key takeaway of this analysis is the decreasing linear variation of γ according to β with a single fitting error being equivalent to at least 13 spatialization errors per 1000 agents. The minimal total error was thus found at the DA resolution because the decrease of spatialization errors when the RGR became less aggregate was much higher than the increase of fitting errors. In other words, γ errors are more sensitive to the RGR aggregation than β errors.

4.4.3 What is the impact of census targets' harmonization and additional control at the CMA resolution on the total error?

In the quest for an optimal synthetic population, i.e., a synthetic population with minimum $\beta + \gamma$ error, the following four configurations were tested:

- 1R: Raw data with single control;
- 2R: Raw data with double control;
- 1H: Harmonized data with single control;
- 2H: Harmonized data with double control.

The variation of the total error according to each configuration was assessed. As the DA resolution was found to minimize the total error, the different configurations' results at the DA resolution are shown in Figure 4.19. The goal was to detect which configuration would help to further reduce the total error, i.e., further optimize the synthetic population.

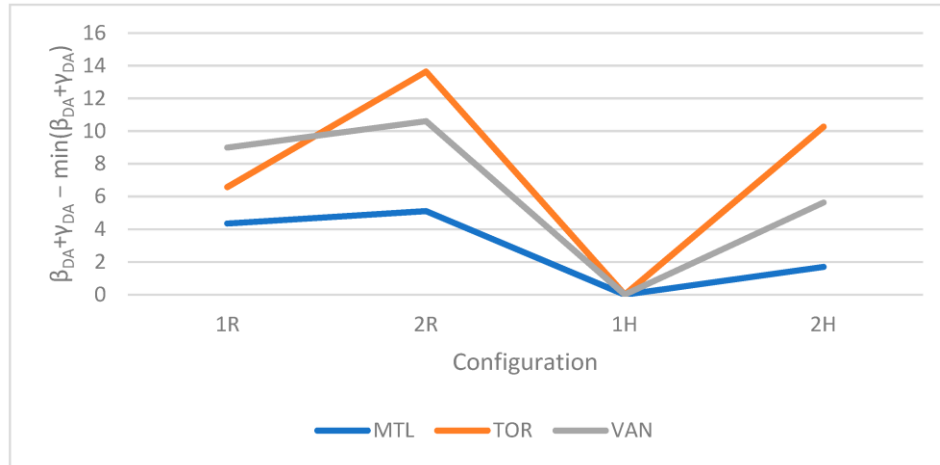


Figure 4.19 Variation of $\beta + \gamma$ at the DA resolution according to the configuration used

The total errors were lower for the 1H and 2H scenarios compared to those of 1R and 2R. This shows that harmonizing census targets helps in improving the quality of the synthetic population. This is mainly due to the intra- and inter-resolution inconsistencies between all geographic resolutions being reduced to zero. However, 2R and 2H showed higher total errors than 1R and 1H, respectively. In fact, the double control improved the synthetic population's fit at the REGION resolution, thus damaging its fit at the GEO resolution. As shown in Figure 4.18, when β varies, γ undergoes an opposite, more important, variation. Hence, as the double control applied improves the fit at the CMA resolution, i.e., reduces β errors, γ errors undergo an increase the magnitude of which is more important than β errors reduction, and the total error thus increases. In summary, harmonizing census targets helps improve the synthetic population's quality. The total error reduction obtained ranged from ~ 4 to ~ 9 errors per 1000 agents in the study area's CMAs.

4.4.4 How does the variation of γ according to β change when census targets are harmonized?

Figure 4.20, Figure 4.21 and Figure 4.22 show the variation of γ according to β for the 1R and 1H configurations in Montreal, Toronto, and Vancouver CMAs, respectively. The goal was to assess how this variation is altered when harmonized census targets are used. For the three CMAs, the absolute value of the slope of the 1H trendline was higher than the absolute value of the slope of the 1R trendline. This means that when census targets are harmonized, for an additional fitting error we get a higher reduction of spatialization errors. Hence, synthesizing at a less aggregate

RGR becomes more beneficial with harmonized census targets as the γ error, and thus the total error, are further reduced for the same β error increase.

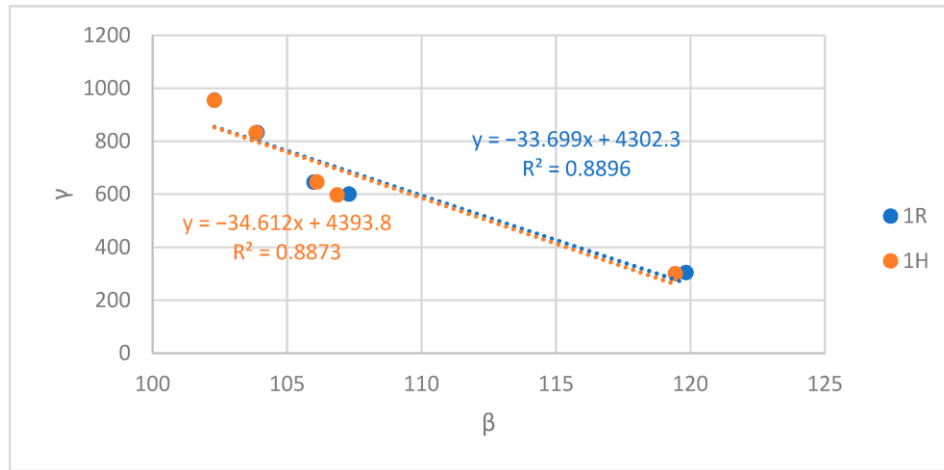


Figure 4.20 Variation of γ according to β with raw and harmonized data in Montreal

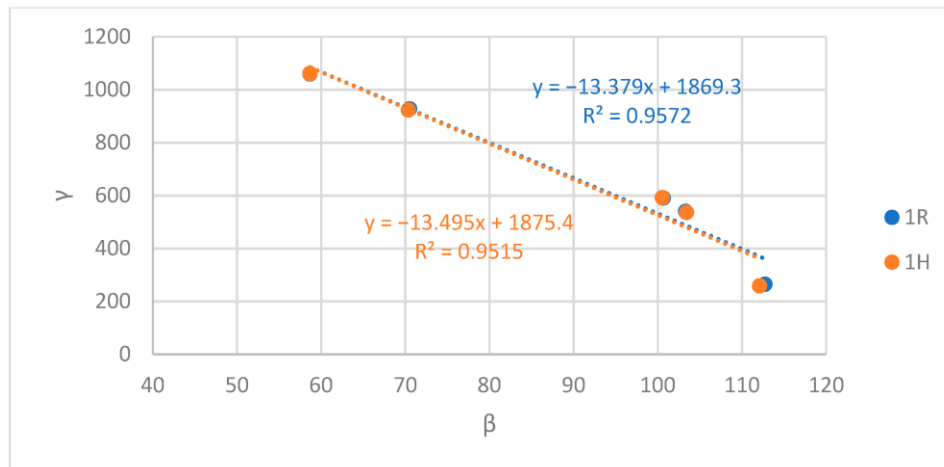


Figure 4.21 Variation of γ according to β with raw and harmonized data in Toronto

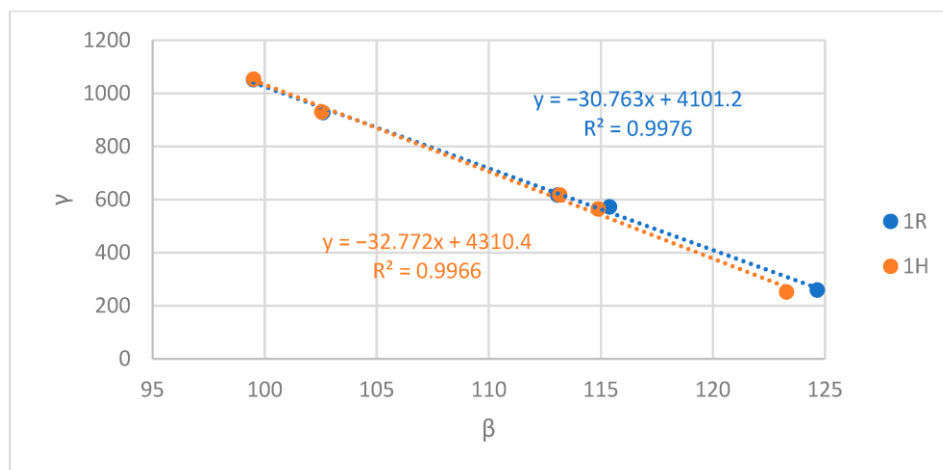


Figure 4.22 Variation of γ according to β with raw and harmonized data in Vancouver

For each CMA, the difference between the 1R trendline slope and the 1H trendline slope (ΔSlope) was calculated, and its variation according to the inter-resolution inconsistencies at the DA resolution (α_{DA}) is depicted in Figure 4.23. The variation was found to be linear, with the ΔSlope increasing when the α_{DA} increases. This means that the more inter-resolution inconsistencies a CMA shows, the more harmonizing its census targets allows to save spatialization errors for each additional fitting error, i.e., when decreasing the RGR. In other words, the more inter-resolution inconsistencies a CMA shows, the more harmonization helps to increase its synthetic population's accuracy for the same cost in terms of precision.

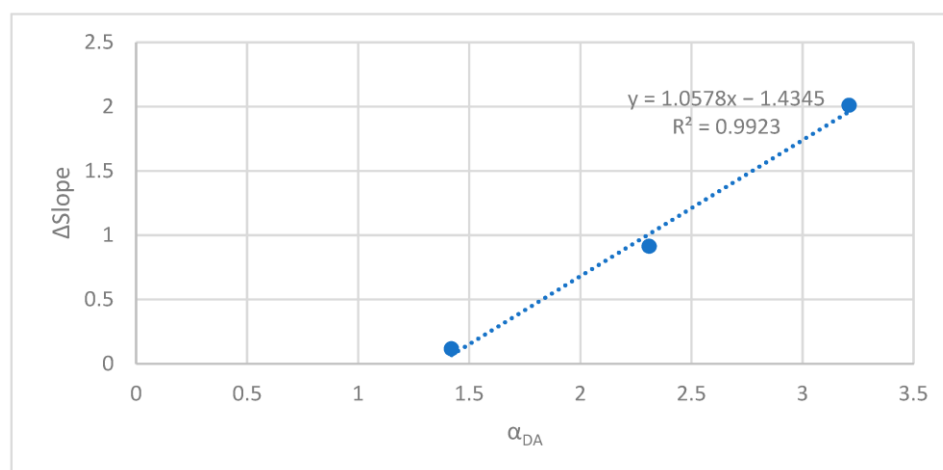


Figure 4.23 Variation of ΔSlope according to α_{DA}

4.5 Discussion

A synthetic population's quality is usually assessed by its fit to the census targets at the RGR (Beckman et al., 1996; X. Ye et al., 2009; Konduri et al., 2016; Moreno & Moeckel, 2018; Guo & Bhat, 2007; Auld et al., 2009). This paper introduced the concepts of sociodemographic accuracy and spatial precision as components of the quality measure. The accuracy of the synthetic population (representativeness of the sociodemographic characteristics of the entire population) was measured by its fit to the census targets at the CMA resolution. This is because the census targets at the CMA resolution are considered as the ground truth due to the rare or non-existent need to adjust census targets to preserve privacy at this level. To measure the precision (representativeness of the real population's spatial heterogeneity), the synthetic population was first interpolated from its RGR to the DA resolution. The precision was measured by the fit of the interpolated population to the census targets at the DA resolution. Fitting errors (β) and spatialization errors (γ) were thus calculated to assess accuracy and precision, respectively. An optimal synthetic population is a synthetic population showing minimal total ($\beta + \gamma$) error.

While population synthesizers are continually conceived to improve the quality of population synthesis, the choice of the RGR and its impact on the quality of synthetic populations have not been studied to the best of our knowledge. Hence, another contribution of this paper was its assessment of the impact of the RGR characteristics on the quality of population synthesis. The main characteristics of the RGR considered are its aggregation and the inter-resolution inconsistencies (α) it shows. β was found to increase and γ to decrease with less aggregate geographic resolutions, with γ magnitude being generally more important than β . γ was also found to be more sensitive to the RGR's aggregation than β , thus yielding a minimal total error at the least aggregate RGR.

An additional contribution of this paper was its testing of the impact of double control and census targets' harmonization on the quality of the synthetic population. While double control is originally introduced to control variables that are not available at the same geographic resolution (Konduri et al., 2016), we tested it as a plausible means of reducing the total error since it makes it possible to control the least and the most aggregate geographic resolutions simultaneously, i.e., the resolutions where γ and β are respectively measured. In their paper, Konduri et al. (2016) showed that double control improves the quality of the synthetic population. The quality indicator they used comprises

only the fit at the more aggregate level (Konduri et al., 2016), which corresponds to accuracy in this paper. However, we found that the double control damages the quality of the synthetic population when both accuracy and spatial precision are considered as components of the quality indicator. In fact, when correcting the fit to the most aggregate geographic resolution (reducing β), the fit to the least aggregate resolution was compromised (γ increased). As γ variation is more important than β variation, the total error increased, hence damaging the synthetic population's quality. On the contrary, census targets' harmonization, i.e., applying proportional adjustments to census targets to reduce intra- and inter-resolution inconsistencies to zero, was found to improve the quality of the synthetic population. Moreover, for the same β , a lower γ was obtained when population synthesis was performed using harmonized census targets. It is worth mentioning that despite the importance of intra- (Balakrishna et al., 2020) and inter-resolution (Konduri et al., 2016) consistency for population synthesizers being mentioned in the literature, quantifying the impact of such inconsistencies, both in terms of accuracy and precision, and suggesting a complete framework to harmonize census data within and between resolutions simultaneously were, to the best of our knowledge, done for the first time in this paper.

4.6 Conclusions

In this paper, the enhanced IPU algorithm (Konduri et al., 2016) was used to synthesize populations at five RGRs (CMA, CSD, ADA, CT, and DA) for the CMAs of Montreal, Toronto, and Vancouver. Eighteen scenarios involving different RGRs, data types (raw or harmonized), and control types (single or double) were compared to assess the impact of the RGR characteristics on the quality of population synthesis. Specifically, the impact of the following factors on the synthetic populations' accuracy and precision was assessed: (1) the aggregation of the RGR, (2) data inconsistencies between census geographic resolutions, and (3) multiple geographic resolutions control.

Three indicators were calculated to compare the synthetic populations generated: census inter-resolution inconsistencies (α), fitting errors (β), and spatialization errors (γ). The impact on the three indicators of the RGR was first investigated. α and β were found to increase, and γ to decrease when the RGR became less aggregate. The total error was found to decrease when the RGR became less aggregate yielding, an optimal synthetic population at the DA resolution. Then, the variation of γ according to β was investigated. A decreasing linear trend was observed with an important

slope, meaning that spatialization errors are more sensitive to the RGR than fitting errors. Finally, the impacts of harmonization and double control were assessed. The double control was found to damage the quality of the synthetic population while the harmonization was found to reduce the total error. The harmonization was also found to be more effective when the census inconsistencies were higher. In summary, synthesizing a population at the DA resolution using harmonized census targets was found to be the best practice.

The main limitation of the conclusions of this paper lies in the number of CMAs investigated. Although finding similar trends for three large and different CMAs allows for conclusions to be made, they would be better founded if based on a larger number of CMAs. Next steps involve comparing the impacts of the RGR choice on various population synthesizers, considering the spatial precision as a component of the synthetic population's quality measure. Various spatialization approaches to allocate the synthetic households to individual dwellings should then be tested on a synthetic population with the best accuracy and precision possible.

Author contributions

Study conception and design: Mohamed Khachman, Catherine Morency, and Francesco Ciari; analysis and interpretation of results : Mohamed Khachman; draft manuscript preparation : Mohamed Khachman, Catherine Morency, and Francesco Ciari. All authors have read and agreed to the published version of the manuscript.

Funding

The authors wish to acknowledge the contribution and financial support of the Mobilité research chair partners: Ministère des Transports du Québec (MTQ), Société de Transport de Montréal (STM), Autorité Régionale de Transport Métropolitain, exo, and Ville de Montréal.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm> (accessed on 12 November 2021).

Conflicts of interest

The authors declare no conflict of interest.

CHAPITRE 5 ARTICLE 2: INTEGRATED MULTIREOLUTION FRAMEWORK FOR SPATIALIZED POPULATION SYNTHESIS

Rédigé par :

Mohamed Khachman, Catherine Morency et Francesco Ciari

Publié le 23 novembre 2022 dans :

Transportation,

DOI : 10.1007/s11116-022-10358-w

Abstract

Large-scale agent-based microsimulation platforms, increasingly used in transportation demand modelling, require fully enumerated and spatialized lists of the population and its sociodemographic characteristics as input. The quality of the synthetic population, measured as its ability to reproduce the sociodemographic characteristics of the real population and their spatial distributions, is thus a determinant factor of the model reliability. While many efforts were devoted to improving the sociodemographic accuracy of synthetic populations, less attention was paid to perfecting their spatial precision. Conventional spatialized population synthesis methods, where the generation and spatialization processes are separated, are vulnerable to inconsistencies between zonal synthetic populations, and the built environments on which they are then distributed. These methods also present transferability issues that lie in their high reliance on rich spatialized datasets and knowledge of the local context. Hence, we propose an integrated multiresolution framework (IMF) that overcomes the limitations of the conventional framework (CF) by its ability to directly generate synthetic populations at the building resolution with minimal data requirements. The IMF includes an extension of an optimization-based method to multiresolution applications where any number and aggregation of spatial resolutions can efficiently be controlled. The CF and the IMF are applied to generate synthetic populations for Montreal, Canada. We define and measure sociodemographic accuracy, spatial precision, overall quality, and building-resolution fit of the synthetic populations to compare the frameworks' performances. Despite a small loss in accuracy, the IMF achieves drastically better spatial precision, overall quality and building-resolution fit of synthetic populations, compared to the CF.

Keywords: population synthesis; spatial disaggregation; integrated framework; multiresolution control; optimization-based method; travel demand modelling.

5.1 Introduction

The classic four-step model having shown its limits (Kitamura et al., 2000), large-scale simulation platforms are taking the upper hand in the field of transportation modelling and demand forecasting where an important shift from classical trip-based models to activity-based models is witnessed. Various examples of these large-scale platforms used for transportation and land-use planning purposes were introduced in the last two decades including UrbanSim (Waddell, 2002), TASHA (Miller & Roorda, 2003), MATSim (Balmer et al., 2009), DAYSIM (Bowman et al., 2014), and SimMobility (Adnan et al., 2016). This evolution in practice induced a growing need of key inputs to such models consisting of spatialized, fully enumerated and characterized populations of agents (e.g., households, individuals, etc.). Despite such datasets being periodically collected by national statistical offices (e.g., U.S. Census Bureau, Statistics Canada, etc.), they remain unavailable for modelers and practitioners, mainly for privacy issues. However, aggregations of these datasets at various geographic resolutions, as well as a microdata sample - generally 5% or less of the population - are often made public. Hence, various methods were conceived to reverse the process, i.e., use zonal aggregations of the population's characteristics and/or a public microdata sample to generate a fully enumerated list of synthetic households and individuals with a set of chosen attributes. This reversed data processing is known as population synthesis.

While many efforts were made to introduce and improve synthetic populations' generation methods, the optimization of synthetic agents' spatial distribution did not garner as much attention in the literature. However, sociodemographic correlations show important variations across spatial locations (Zhou et al., 2022), which exert a significant influence on the spatial variation of mobility behaviors (Zhu et al., 2018). Moreover, with fine-resolution urban planning becoming popular, reliable explicit spatial distribution of sociodemographic characteristics of the population is increasingly required (Liu et al., 2021). For instance, data at the building scale is usually required for transportation planning studies related to non-motorized modes such as walking and cycling (Pajares et al., 2021) and to model long-term urban decisions such as residential and workplace choices (Zhu et al., 2018). This paper is meant to contribute to the improvement of synthetic populations' spatialization through an efficient framework that is: (1) integrated (i.e., integrates the population generation and spatialization processes), (2) flexible (i.e., can handle constraints for any

number of agents and geographic resolutions), and (3) transferable (i.e., does not depend on locally available rich spatial datasets).

The remainder of this paper is organized as follows: “Literature review” section comprises a review of the state of the art about synthetic populations’ generation and spatialization techniques. “Methodology” section describes the methodology i.e., the proposed framework and the assessment criteria. “Case study” section is dedicated to the application of the suggested methodology for spatialized population synthesis in the city of Montreal, Canada. Results are shown and discussed in “Results” section. Conclusions, limitations, and potential extensions of this research are finally included in “Conclusion and outlook” section.

5.2 Literature review

Generation and spatialization of synthetic populations have always been tackled as complementary yet independent processes. In this section, we briefly showcase the different methods used for each process as well as their advantages, limitations, and improvements over time.

5.2.1 Generation methods

Population synthesizers can be grouped into three categories: fitting-based synthesizers, simulation-based synthesizers, and mixed synthesizers.

5.2.1.1 Fitting-based synthesizers

Fitting-based population synthesis is the oldest and the most widespread practice in the field. Fitting-based synthesizers generally aim to expand a disaggregate microdata sample so that it fits to zonal totals for a selection of variables of interest, also called control variables. The idea is to preserve the correlation structure of population’s attributes in the available disaggregate sample, while expanding it to the total population. Synthetic Reconstruction and Combinatorial Optimization are the most widely used fitting methods.

Synthetic Reconstruction techniques comprise two steps: estimation and selection. The first step consists of estimating the joint distribution of chosen control variables in the population. Synthetic agents are then selected from the microdata sample (e.g., Public Use Microdata Sample (PUMS) in the United States, Public Use Microdata Files (PUMF) in Canada, etc.) according to the estimated distribution. The estimation step is usually based on the iterative proportional fitting

(IPF) algorithm initially proposed by Deming and Stephan (1940). The IPF, also known as matrix balancing method (Moreno and Moeckel 2018), is an algorithm iteratively modifying the cells of an n-dimensional table to make them fit with pre-determined margins. The iterative process stops when the algorithm converges (i.e., the error is reduced to a pre-set tolerated level), or when a pre-set number of iterations is reached. Beckman et al. (1996) implemented the first Synthetic Reconstruction method in TRANSIMS (Smith et al. 1995). A contingency table is seeded with frequencies of chosen households' attributes in the PUMS. IPF is then applied to the seeded table to make it fit with corresponding totals at the census tract resolution. The fitted table hence constitutes the estimated distribution of controlled attributes in the population with each cell containing the number or the proportion of households having a determined combination of characteristics. Monte Carlo simulations are then used to draw households from the microdata sample according to the estimated distribution.

Despite its various advantages such as quick convergence (Beckman et al. 1996), providing constrained maximum entropy estimates of the distribution in the real population (Ireland and Kullback 1968), and preserving the correlation structure of controlled attributes in the microdata sample (Müller and Axhausen 2011), the synthesizer suggested by Beckman et al. (1996) shows numerous weaknesses. First, only one level of constraints can be handled i.e., the algorithm allows to control attributes for a single type of agents (households in this case) at a time. This leads to non-fitted synthetic populations of other agents considered (e.g., individuals). Second, constraints at only one geographic resolution (census tracts in this case) can be handled which implies that some variables of interest must be neglected if the totals are only available at distinct geographic resolutions. Third, due to the limited size of the sample, there is high probability that observations for some combinations of characteristics of the population are missing. The algorithm convergence at the estimation step is prevented by such incorrect zero cells (Fournier et al. 2021). Fourth, even if convergence is reached at the estimation step, the variability of agents drawn in the synthetic population remains limited to that of the microdata sample i.e., a type of households that exists in the population but is not observed in the PUMS cannot be found in the synthetic population. Finally, the integerization process i.e., converting decimal proportions obtained at the end of the estimation step to integer numbers of synthetic agents to be drawn (Guo and Bhat, 2007), damages the quality of fit.

These weaknesses were extensively addressed in the literature and improved population synthesizers were introduced. Multilevel and multiresolution synthesizers allow to handle constraints at multiple levels and geographic resolutions simultaneously. The incorrect zero cells problem was tackled through tweaking, aggregation of the categories with lowest frequencies, and proportional usage of frequencies observed at more aggregate geographic resolutions. Simulation-based synthesizers generating non-observed agents were conceived. Some papers introduced population synthesizers not requiring a disaggregate sample to be used as input (Gargiulo et al. 2010; Barthelemy and Toint 2013). Finally, various integerization techniques were suggested to limit the fitting deterioration.

A means of avoiding the integerization problem is using Combinatorial Optimization-based synthesizers. Combinatorial Optimization-based synthesis starts with arbitrarily selecting with repetitions n households from the microdata sample, with n being the total number of households in the population. Each household in the initial list is then iteratively replaced with the microdata sample households and the combination providing the best fit to census totals is kept. The process is repeated until convergence or until a pre-set number of iterations is reached. Various optimization techniques such as simulated annealing (Openshaw and Rao 1995; Williamson et al. 1998; Voas and Williamson 2000), genetic algorithm (Williamson et al. 1998; Birkin et al. 2006), hill climbing (Williamson et al. 1998; Abraham et al. 2012) and greedy heuristic (Srinivasan et al. 2008) were used for population synthesis purposes.

Despite being found more accurate than IPF for the generation of a small synthetic population (Ryan et al. 2009), Combinatorial Optimization methods remain less popular in the field mainly because of their computational complexity and inability to systematically generate optimally fitted large synthetic populations (Lee and Fu 2011).

5.2.1.2 Simulation-based synthesizers

Statistical learning or simulation-based population synthesizers establish an estimation of the joint distribution of control variables using available sociodemographic datasets. Synthetic agents are then sampled based on the estimated distribution. Farooq et al. (2013) introduced a Markov Chain Monte Carlo simulation-based population synthesis approach fed with conditional probability distributions of control variables. Saadi et al. (2016) suggested a Hidden Markov Model-based population synthesis method that treats the attributes as a sequence of connected states. As using

Markov Chains requires prior knowledge of the attributes' interdependencies (Zhou et al. 2022), an alternative data-driven approach based on Bayesian Networks was applied to generate synthetic populations (Sun and Erath 2015; Zhang et al. 2019). Bayesian Networks use a graphical representation to efficiently estimate the joint distribution of control variables. Sun et al. (2018) generated a simulation-based synthetic population using a hierarchical mixture modeling framework including three components: probabilistic tensor factorization, multilevel latent class model, and rejection sampling. More recently, synthesizers leveraging deep generative modeling methods such as Variational Auto-Encoder (Borysov et al. 2019), Wasserstein Generative Adversarial Network (Garrido et al. 2020), Composite Travel Generative Adversarial Network (Badu-Marfo et al. 2022), Conditional Variational Auto-Encoder (Borysov and Rich 2021) and Conditional Generative Adversarial Networks (Johnsen et al. 2022) were introduced. Models' interpretability and overfitting to training data are important challenges faced when deep learning methods are used (Basu and Ferreira 2020).

Simulation-based population synthesis methods present a greater flexibility than fitting-based methods in terms of input data requirements. For instance, simulation-based methods can be used even if no marginal data is available (Yaméogo et al. 2021). They also show a better performance than Synthetic Reconstruction and Combinatorial Optimization when dealing with non-comprehensive, non-representative and/or small sample sizes as they can synthesize non-observed agents which is impossible using Synthetic Reconstruction and Combinatorial Optimization (Farooq et al. 2013). However, while multivariate joint distributions are better reproduced by simulation-based approaches, synthesizers belonging to this category generally fail to respect the observed marginal distributions (Yaméogo et al. 2021). Mixed methods were recommended to overcome this issue (Sun and Erath 2015; Sun et al. 2018). The idea is to post-process a simulation-based synthetic population using a Synthetic Reconstruction reweighting algorithm to ensure that the final synthetic population fits to observed marginal totals.

5.2.1.3 Mixed methods

Mixed methods generating an optimal fit to both multivariate and marginal distributions were proposed. Casati et al. (2015) reweighted a hierarchical Markov Chain Monte Carlo-based synthetic population using Generalized Raking. Saadi et al. (2018) applied a Hidden Markov Model + IPF framework to population synthesis. Fournier et al. (2021) used Markov Chain Monte

Carlo and Bayesian Networks to synthesize separate households and people populations which they then jointly reweighted using an optimization-based algorithm. Ilahi and Axhausen (2019) and Zhou et al. (2022) used a combination of Bayesian Networks and Generalized Raking.

5.2.2 Multiple simultaneous controls

The basic versions of the population synthesizers described previously could handle constraints at a single level (i.e., households or individuals) and a single geographic resolution. However, as households and individuals' attributes both influence mobility behaviors (Fabre and Morency 2019; Kalter and Geurs 2016; Loo and Lam 2013), multilevel population synthesizers fitting the attributes distributions of both types of agents simultaneously were conceived. Similarly, multiresolution synthesizers were used to control variables of interest which marginal totals are not available at the same geographic resolution.

5.2.2.1 Multilevel synthesis

To ensure a simultaneous fit of households and individuals, multilevel Synthetic Reconstruction algorithms such as Iterative Proportional Updating (IPU) (X. Ye et al. 2009), hierarchical IPF (Müller and Axhausen 2011), entropy optimization (Bar-Gera et al. 2009), and applications of Generalized Raking (Deville et al. 1993) were introduced. While IPU reweights each household according to its composition in terms of individuals, hierarchical IPF alternatively fits households and individuals until convergence. When the population synthesis is transposed to a constrained optimization problem, entropy optimization calculates weights that minimize an information-based similarity metric while Generalized Raking yields weights minimizing a distance metric (Yaméogo et al. 2021). Recently, Fournier et al. (2021) presented an optimization-based reweighting method that used Cyclic Coordinate Descent to reach a combination of weights that achieves optimal fit at both households and individuals' levels. They prove Cyclic Coordinate Descent to provide a slightly better fit than IPU in much less computation time.

Multilevel synthesis methods are not limited to the Synthetic Reconstruction category. Combinatorial Optimization methods were also adapted to meet targets at households and individuals' levels simultaneously. For instance, Abraham et al. (2012) used a hill-climbing algorithm to solve the combinatorial problem achieving a multilevel fit. Ma and Srinivasan (2015) presented an alternative solution based on fitness values that measure the fit of each sample

household to all levels' targets simultaneously. A multilevel fit is generally achieved by all simulation-based synthesizers except Markov-process based ones (i.e., Markov Chain Monte Carlo and Hidden Markov Model). However, Casati et al. (2015) presented a hierarchical Markov Chain Monte Carlo method that leverages a typology of people within households to extend the Markov Chain Monte Carlo-based synthesizer (Farooq et al. 2013) to multilevel fitting. For a thorough review and comparison of multilevel population synthesis methods, we refer the reader to Yaméogo et al. (2021)'s paper.

5.2.2.2 Multiresolution synthesis

Multiresolution synthesizers are used when the marginals of variables of interest are not available at a single geographic resolution. They are also beneficial to account for differences between sub-areas when a synthetic population is generated for a large area (Moreno and Moeckel 2018). Khachman et al. (2021b) show that to optimize the quality of synthetic populations defined by the combination of their sociodemographic accuracy and spatial precision, variables of interest should be controlled at the least aggregate resolutions where their marginal distributions are available. Konduri et al. (2016) introduced a Synthetic Reconstruction multiresolution synthesizer called enhanced IPU. It consists of an extension of the IPU algorithm to two geographic resolutions. Moreno and Moeckel (2018) proposed an extension of the same algorithm to three geographic resolutions. Zhu and Ferreira (2014) controlled marginal distributions at two geographic resolutions using a two-step IPF procedure. Abraham et al. (2012) presented a Combinatorial Optimization-based population synthesizer that can fit to marginal totals at multiple geographic resolutions.

The needs in terms of number of resolutions to be controlled vary from a location to another according to data availability and formatting. Hence, a main shortcoming of existing multiresolution synthesizers lies in their inflexibility as they are limited to a pre-determined maximum number of resolutions. Another limitation of these algorithms is their limited efficiency. This makes them impractical for high dimensional exercises such as those involving marginal controls for a very high number of attributes and resolutions, and/or those controlling marginals at extremely disaggregate resolutions such as building or DU resolutions.

5.2.3 Spatialization methods

While many research works focused on improving the marginal and joint distributions of the synthetic populations' attributes to best reflect the population's characteristics, little attention was given to accurately spatialize the synthetic households. Most population synthesis applications remain limited to assigning synthetic households to the least aggregate zones where census totals could be found, e.g., Traffic Analysis Zones, counties, municipalities, etc. However, few, mainly ad-hoc, attempts of explicitly spatialized synthesis could be found in the literature. Zhou et al. (2022) provided a methodology to get a synthetic population for Singapore spatialized at the building scale. A synthetic population of households and individuals was first generated at the planning area level using a Bayesian Networks + Generalized Raking framework. The built environment was then separately synthesized by data fusion of rich spatialized datasets and considering a set of assumptions (e.g., vacancy rate) derived from first-hand knowledge of the local context. A rule-based matching heuristic was finally applied to assign a synthetic dwelling to each synthetic household by correspondence of planning area and dwelling type. The two-step IPF procedure proposed by Zhu and Ferreira (2014) is also meant to generate a spatialized synthetic population for Singapore. A first IPF is applied to synthesize a multilevel population at the Traffic Analysis Zone resolution. Then, a second IPF is applied at the building resolution, controlling at least the building's capacity that can be obtained from available data fusion. Discrete choice models could also be used to enrich controlled marginal distributions at the building resolution. Liu et al. (2021) introduced a web-based geospatial application allowing to visualize disaggregate sociodemographic data. To obtain the background data, available spatial datasets are first joined, and a residential capacity estimation model is applied to estimate the number of residential units for each residential project where the information is lacking. Then, a synthetic population is generated using an IPF-based algorithm (Zhu and Ferreira 2015). Each synthetic household is finally assigned to a residential unit via a capacity-constrained discrete choice model algorithm (De Palma et al. 2007). Chapuis et al. (2018) suggested a two-step spatialization process. The idea is to first interpolate the population data at a finer scale, e.g., cells of a raster grid, then explicitly localize synthetic households. The interpolation is done according to a specified distribution, e.g., linear regression. Synthetic households are then nested to spatial objects, e.g., zones, buildings, etc., using a uniform, area, or capacity distribution function. Geometric, contingency and density nesting constraints can also be defined to guide the spatialization process. Once nested to a spatial object,

the synthetic household can be either assigned to its centroid, or to a random location within it. The proposed methodology allows to take advantage of various available layers in the definition of the distributions and constraints required. Multiple papers suggested to assign specific locations to households within their generation zone by means of simple geographic distributions, e.g., uniform, or weighted distribution within a polygon or along the road network (Jiang et al. 2022; Zhang et al. 2019; Gallagher et al. 2018; Swarup and Marathe 2016).

5.2.3.1 Spatial microsimulation

Spatial disaggregation of sociodemographic data is a focus of another research line englobing works on spatial microsimulation (Zhou et al. 2022). Various population synthesis methods are also used in the spatial microsimulation field such as IPF (Panori et al. 2017; Anderson 2012), Combinatorial Optimization (Farrell et al. 2012; Hynes et al. 2009) and Generalized Raking (Vidyattama et al. 2013; Harding et al. 2011; Tanton et al. 2011; Ballas et al. 2007). For a thorough review of spatial microsimulation methods, we refer the reader to Tanton (2014)'s paper. As for population synthesis methods, many spatial microsimulation applications are limited to calculations of populations' characteristics at less aggregate zones such as pre-defined administrative zones or user-defined cells of raster grids. However, some papers introduced, mainly ad-hoc, methodologies for spatial microsimulation at the building scale. Pajares et al. (2021) estimated population counts at the building access level based on computed residential gross floor areas. Calka et al. (2016) geolocated population counts in buildings considering buildings' characteristics. Bast et al. (2015) estimated population counts at the building resolution based on OpenStreetMap (OSM) and census data at the municipality resolution using machine learning. Chen et al. (2021) and Xie et al. (2015) used LiDAR data to achieve similar goals. Muñoz and Peters (2014) assigned synthetic households to estimated dwelling units using hill climbing.

The explicit spatialization methods suggested in the literature show various limitations summarized in Table 5.1 (bold). First, performing generation and spatialization separately leads to errors mainly due to inconsistencies between synthetic populations and built environments' marginals. Artificial post-processing of results, e.g., relaxing rules and constraints (Zhou et al. 2022; Chapuis et al. 2018), using capacity-constrained approaches (Liu et al. 2021), handling negative values (Chapuis et al. 2018), etc., is thus required to successfully spatialize all synthetic households. Zhu and Ferreira (2014) mentioned that theoretically the integration of generation and spatialization in a

single IPF would be possible. However, this was not possible from a computational point of view because of the huge matrix obtained when margins at the building resolution are integrated.

Second, we understand that the variations in data availability, format, and quality across locations prevent the definition of a fully replicable spatialization framework. However, the suggested methods show high reliance on rich spatialized datasets and knowledge of the local context making them hardly adaptable to other locations (Liu et al. 2021). Moreover, even if a rich building-household joint sample is available at another location, the transfer of the model-based methodologies remains problematic. That is because the replication of the model's goodness of fit to a different dataset cannot be guaranteed, thus the model may perform poorly.

Finally, it is worth mentioning that the approaches relying on assumed geographic distributions (Jiang et al. 2022; Zhang et al. 2019; Gallagher et al. 2018; Swarup and Marathe 2016) do obviously not produce a realistic spatial distribution as the built environment is totally neglected. Such approaches may even assign synthetic households to non-residential areas, water surfaces, tunnels, bridges, etc. (Gallagher et al. 2018). Table 5.1 summarizes various existing options for explicitly spatialized population synthesis, their principles, and potential challenges that can be encountered when using them.

5.2.4 Contributions

The main motivation of this paper is to provide a spatialized population synthesis framework that shows three main advantages compared to existing approaches. (1) Generation and spatialization are integrated in the proposed framework. It thus directly produces a synthetic population at the building resolution. (2) It is a flexible and efficient multiresolution framework. We extended the application of an efficient optimization-based multilevel synthesis algorithm introduced by Fournier et al. (2021) to unlimited multilevel and multiresolution control. (3) It is a transferable framework with minimal data requirements. The proposed framework can be applied whenever the marginal distribution of a single dwelling unit characteristic is available at the building resolution. Finally, a new measure of synthetic populations' spatial precision is introduced in the assessment of the proposed methodology.

Table 5.1 Existing options for explicitly spatialized population synthesis

Approach	Principle	Identified limitations		
		Household-built environment inconsistency	Unreliability	Non transferability
Assumed distributions	- Application of assumed spatial distributions within generation zones - Pre-interpolation of synthetic populations to finer zones also sometimes performed using such distributions	N/A - Built environment not considered	Yes - Synthetic households' locations not necessarily matching buildings' locations	No
Matching heuristics	- Application of rule-based algorithms considering the characteristics of the synthetic population and the built environment within generation zones and capacity limits	Yes - BE synthesis independent from population synthesis - Number of synthetic HHs not necessarily matching the number of occupied DUs	Possible - Factors assumed (e.g., occupancy rates, etc.) - Rules and constraints often relaxed for convergence	Yes - Rich spatial datasets required (some are from private providers and/or unavailable elsewhere) - Knowledge of local context required for some assumptions
Probabilistic methods	- Application of capacity-constrained discrete choice models - Estimation of probabilities for various synthetic household types to live in various dwelling unit types	Yes - BE synthesis independent from population synthesis	Possible - Factors assumed (e.g., occupancy rates, etc.) - Rules and constraints often relaxed for convergence	Yes - Rich spatial datasets required (some are from private providers and/or unavailable elsewhere). Often unavailable explicitly spatialized disaggregate sample particularly required for models' calibration

Table 5.1 Existing options for explicitly spatialized population synthesis (suite)

Approach	Principle	Identified limitations		
		Household-built environment inconsistency	Unreliability	Non transferability
	- Distribution done accordingly	- Number of synthetic HHs not necessarily matching the number of occupied DUs	- Low-quality models	- Knowledge of local context required for some assumptions - Local models' calibration required
Sequential multiresolution fitting	- Additional IPF controlling building-resolution attributes applied	Yes - BE synthesis independent from population synthesis - Number of synthetic HHs not necessarily matching the number of occupied DUs	Possible - Removal of control variables due to inefficiency - Additional rounding implied by additional IPF	Possible - High-dimensional applications (numerous attributes, numerous resolutions, and/or extremely disaggregate resolutions) intractable
Simultaneous multiresolution fitting^a	- Application of multiresolution fitting-based algorithms with the building resolution controlled	Possible - Totals of households and occupied dwelling units to be pre-harmonized	Possible - Removal of control variables due to inefficiency	Yes - Predetermined maximum number of resolutions - High-dimensional applications intractable

^aNever applied to explicitly spatialized population synthesis but can theoretically be used

5.3 Methodology

Despite some differences, spatialized population synthesis methods generally fit within the same framework that we call “conventional” spatialized population synthesis framework (CF) in this paper. The CF comprising three main steps is presented in Figure 5.1. First, available spatial datasets are processed using appropriate data fusion techniques to generate the built environment, generally consisting of dwelling unit (DU) marginals at the building resolution (e.g., number of DUs in each building) (1). Then, a multilevel synthetic population (SP) is generated using people (PP) marginals, household (HH) marginals, and a joint household-people microdata sample (2). Ad-hoc heuristics and/or models, relying on knowledge of the local context and rich spatial datasets, are applied afterwards to assign each synthetic household to a building or a DU (3). An explicitly spatialized synthetic population is then created. The proposed methods that fit within this framework and their weaknesses are covered in previous sections.

To overcome the shortcomings of the CF, we suggest a new “integrated multiresolution” spatialized population synthesis framework (IMF) that is presented in Figure 5.2. The IMF is also comprised of three main steps. As for the CF, DU marginals at the building resolution are first calculated using appropriate data fusion techniques (1). Then, the harmonization technique proposed by Khachman et al. (2021b) is applied to synthesize occupied DU marginals at the building resolution using DU marginals calculated at the first step, and census zonal HH marginals (2). As a household and an occupied DU hold a one-to-one correspondence, the occupied DU marginals correspond to HH marginals at the building resolution. Finally, an optimization-based multilevel and multiresolution population synthesis algorithm is applied to simultaneously control census HH marginals, census PP marginals, and synthetic building-resolution HH marginals obtained at the second step (3).

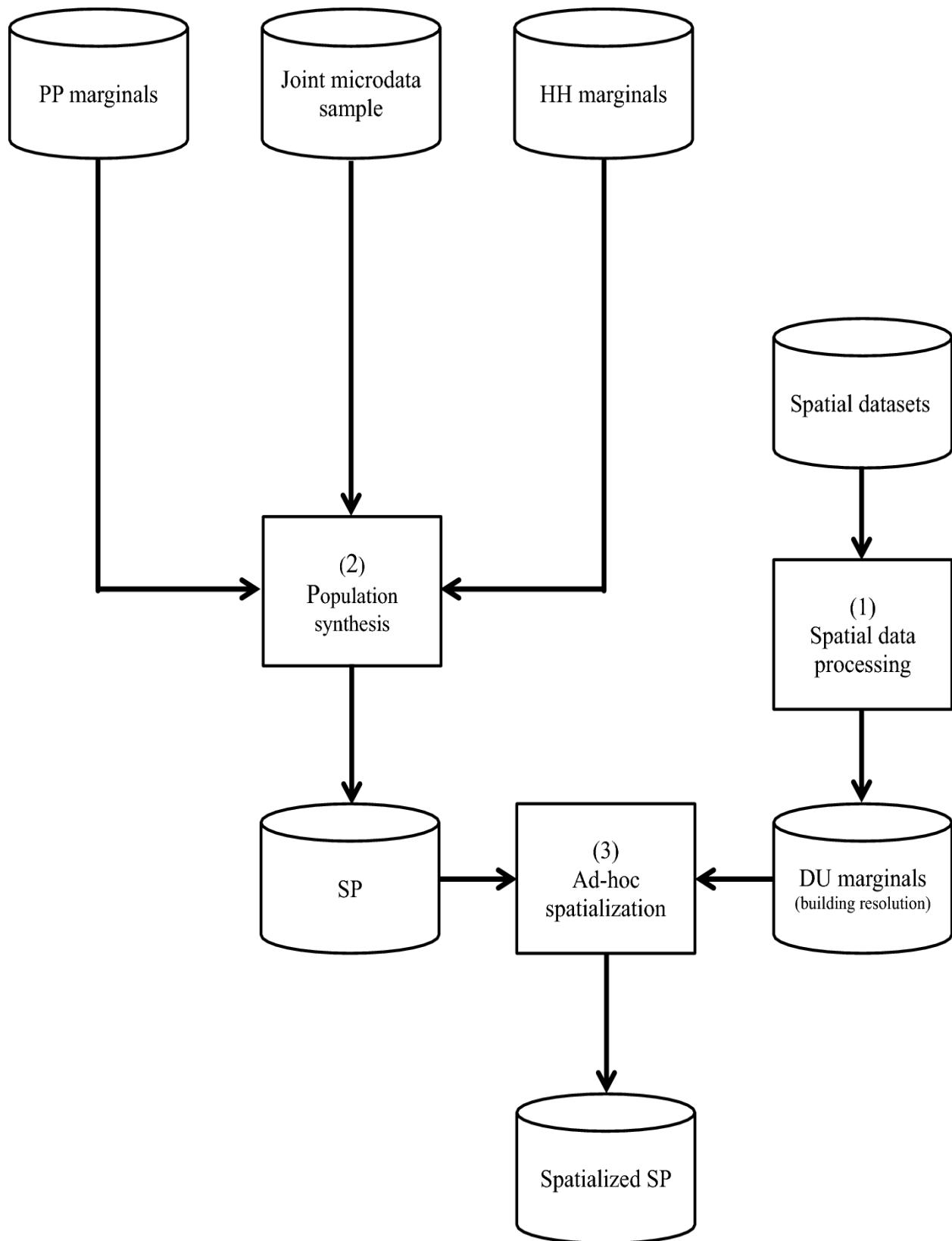


Figure 5.1 Conventional spatialized population synthesis framework (CF)

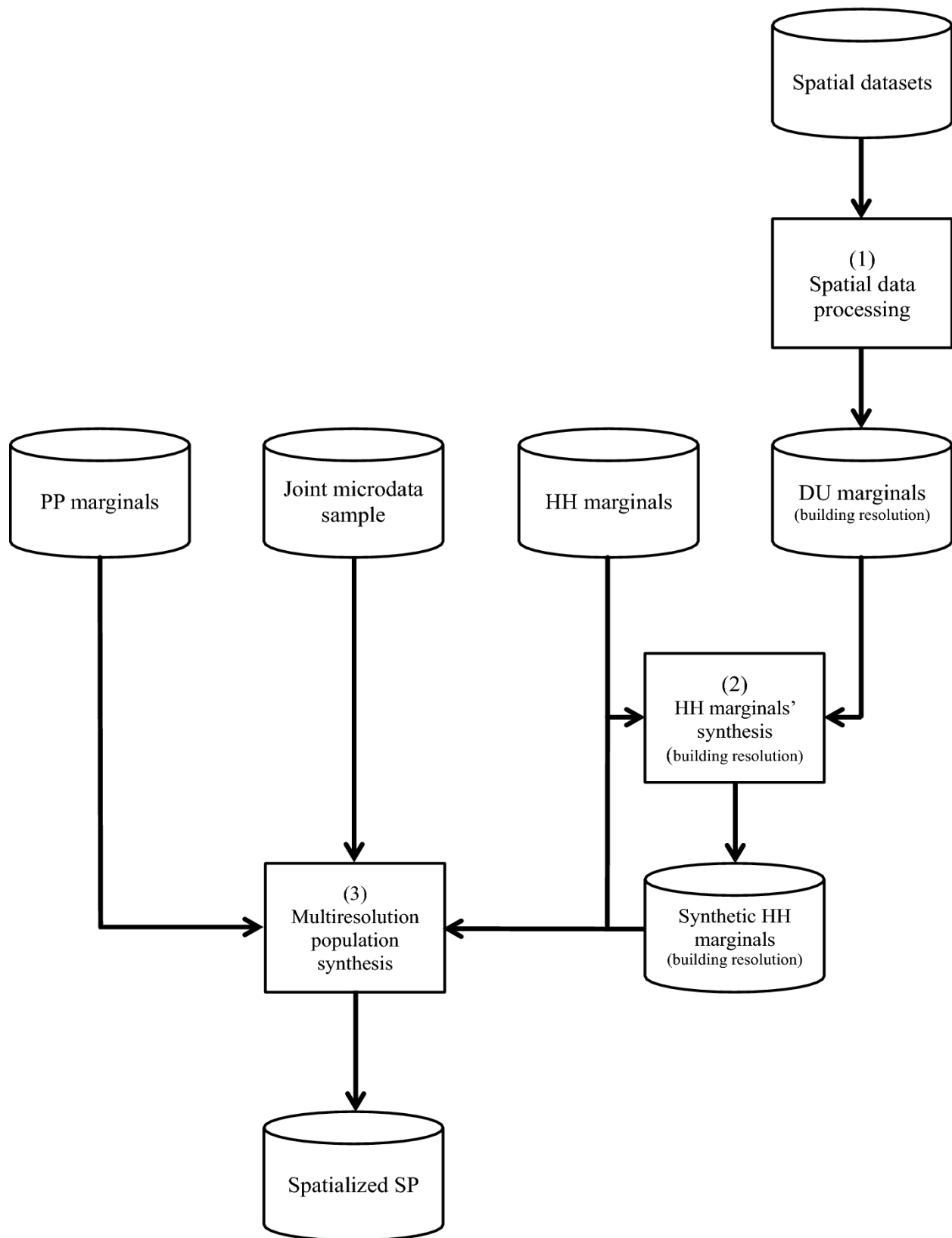


Figure 5.2 Integrated multiresolution spatialized population synthesis framework (IMF)

5.3.1 Household marginals synthesis at the building resolution

Household marginals at the building resolution are synthesized by harmonization of DU marginals at the building resolution to zonal HH marginals extracted from the census. Let the zonal HH marginals be collected at the dissemination area (DA) resolution. The idea is first to apply an intra-resolution adjustment to DA-resolution marginals. That means that categories' totals at the DA resolution are proportionally adjusted so that their sums per variable become equal to the corresponding agent's count. For example, the sum of men and women in a DA are adjusted to meet the count of people in the same DA. Second, an inter-resolution adjustment of DU counts is applied. The counts of DUs for all the buildings belonging to the same DA are proportionally adjusted to meet the total count of HHs in the DA. Synthetic HH counts are thus obtained at the building resolution. Finally, intra-resolution and inter-resolution adjustments are iteratively applied to the characteristics of buildings belonging to the same DA (Figure 5.3) until a pre-set convergence threshold is reached. At convergence, synthetic HH marginals at the building resolution are obtained. The harmonization process applied is extensively described in Khachman et al. (2021b)'s paper.

Multiway table	Building ₁	Building ₂	...	Building _n	Constraint 2
DUs built before 1960	$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$	HHs living in DUs built before 1960 in corresponding DA
DUs built between 1961 and 1980	$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$	HHs living in DUs built between 1961 and 1980 in corresponding DA
DUs built between 1981 and 2000	$x_{3,1}$	$x_{3,2}$	...	$x_{3,n}$	HHs living in DUs built between 1981 and 2000 in corresponding DA
DUs built between 2001 and 2016	$x_{4,1}$	$x_{4,2}$	...	$x_{4,n}$	HHs living in DUs built between 2001 and 2016 in corresponding DA

Constraint 1	Synthetic HH count in Building ₁	Synthetic HH count in Building ₂	...	Synthetic HH count in Building _n
---------------------	---	---	-----	---

Figure 5.3 Example of HH marginals synthesis at the building resolution

5.3.2 Multiresolution optimization-based population synthesis

We adapt the optimization-based reweighting algorithm introduced by Fournier et al. (2021) to simultaneously control M levels at K geographic resolutions as shown in Figure 5.4. To provide a general framework, we use the letters k, l, m, n and i to index resolutions, zones, levels, types, and

surveyed households, respectively. To make it easier to understand, Figure 5.5 presents an application to two levels (HH and PP) at two resolutions (area and sub-area) for a microdata sample of five households.

Resolution	Zone	Level	Type	Level 1							
				1	2	3	...	I			
1	1	1	1	...	...	...	...	...	x_i	$=$	$F(r_k, z_i, l_m, t_n)$
1	1	1	...	...	...	...	...	...			
1	1	1	$N(r_1, l_1)$	...	...	...	...	...			
1	1	...	$1 \dots N(r_1, l_m)$	...	...	...	...	...			
1	1	$M(r_1)$	$1 \dots N(r_1, l_M)$	...	...	$f_i(r_k, z_i, l_m, t_n)$	...	...			
1	...	$1 \dots M(r_1)$	$1 \dots N(r_1, l_{1 \dots M})$	...	...	...	...	...			
1	$L(r_1)$	$1 \dots M(r_1)$	$1 \dots N(r_1, l_{1 \dots M})$	...	...	...	...	...			
...	$1 \dots L(r_k)$	$1 \dots M(r_k)$	$1 \dots N(r_k, l_{1 \dots M})$	...	...	...	...	...			
K	$1 \dots L(r_K)$	$1 \dots M(r_K)$	$1 \dots N(r_K, l_{1 \dots M})$	...	...	...	...	...			

Figure 5.4 Optimization problem formulation for simultaneous control of M levels at K resolutions

Resolution	Zone	Level	Type	HH							
				1	2	3	4	5			
Area	δ	HH	1 person	1	0	0	0	0	x_1	$=$	$\begin{bmatrix} 5 \\ 8 \\ 3 \\ 25 \\ 10 \\ 5 \\ 4 \\ 22 \\ 2 \\ 5 \\ 13 \end{bmatrix}$
Area	δ	HH	2 people	0	1	0	0	0			
Area	δ	HH	3 people and more	0	0	1	1	1			
Area	δ	PP	Men	1	1	1	2	4			
Area	δ	PP	Women	0	1	2	2	2			
Sub-area	1	HH	With children	0	0	1	0	0			
Sub-area	1	HH	Without children	1	0	0	0	1			
Sub-area	1	PP	Count	1	0	3	0	6			
Sub-area	2	HH	With children	0	0	0	1	0			
Sub-area	2	HH	Without children	0	1	0	0	0			
Sub-area	2	PP	Count	0	2	0	4	0			

Figure 5.5 Example of optimization problem formulation for two levels at two resolutions

A type refers to a control variable's category (e.g., the category "1 person" of the variable "household size"). $N(r_k, l_m)$ is the total number of types of the m^{th} level controlled in the k^{th} resolution. For example, in Figure 5.5, for the household level (l_1) at the area resolution (r_1), $N(r_1, l_1)$ is equal to 3 as three types are controlled (1 person, 2 people, and 3 people and more). A level refers to an agent type (e.g., household, person, etc.). The example population of households and people in Figure 5.5 is thus a bilevel population. $M(r_k)$ refers to the number of levels controlled at the k^{th} geographic resolution. In the given example, $M(r_k)$ is equal to 2 for both area (r_1) and sub-area (r_2) resolutions as households and people types are controlled at both resolutions. $L(r_k)$ is the number of geographic zones constituting the k^{th} geographic resolution. In Figure 5.5, $L(r_2)$ is equal to 2 for the sub-area resolution as it includes two sub-areas (1 and 2).

The problem takes the well-known $AX = B$ format. A is a matrix where each column corresponds to a base-level observation in the microdata sample. The base level (l_1) is the only level having a one-to-many relation with the other controlled levels. For example, the household is the base level for a population of households and people. The i^{th} column of A corresponds to the i^{th} household surveyed in the joint microdata sample. Each row of A corresponds to a single resolution-zone-level-type combination. Hence, each cell contains the observed frequency in household i , for type n of level m belonging to the zone l of resolution k . For example, the cell $A(4;3)$ in Figure 5.5 refers to $f_3(r_1, z_1, l_2, t_1)$, i.e., the frequency of men (l_2, t_1) in the third household ($i = 3$) of the microdata sample belonging to the $\delta(z_1)$ area (r_1). X is the vector of weights to be estimated for microdata households included in A to fit to the target vector B . Each element of B refers to $F(r_k, z_l, l_m, t_n)$ which is the target frequency for type n of level m belonging to zone l of resolution k . In Figure 5.5, the first element of B is $F(r_1, z_1, l_1, t_1)$ referring to the total number of 1-person (t_1) households (l_1) in the $\delta(z_1)$ area (r_1).

As it was found to be the most efficient method (Fournier et al. 2021), the Cyclic Coordinate Descent approach implemented in the R package “glmnet” (Friedman et al. 2010) is used to solve the optimization problem. It is proved to provide better results than IPU in much less time (Fournier et al. 2021). The decimal results yielded at both building-resolution HH marginals’ synthesis, and multiresolution optimization processes, are converted to integers using the Truncate, Replicate, Sample method proposed by Lovelace and Ballas (2013). The method first duplicates agents according to the integer truncation of their weights. Then, the lacking number of agents is probabilistically selected from the sample according to the decimal part of agents’ weights. As the Truncate, Replicate, Sample is a probabilistic algorithm, a thousand iterations were run and the one with minimal errors is kept.

5.3.3 Assessment

To assess the proposed framework, we compare synthetic populations generated using both the CF and the IMF. The Root Mean Square Error (RMSE) chosen by Fournier et al. (2021) is used for assessment purposes. It is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{b}_i - b_i)^2}{n}} \quad (5.1)$$

where $\hat{b}_i$ is the i^{th} estimated value, b_i is the i^{th} observed value, and n is the total number of pairs of values compared. While other error measures can be used, the RMSE is preferred because of the model optimization formulation of the population synthesis problem in this paper. Another interesting property of the RMSE is that it increasingly penalizes errors as they become larger while showing the same dimensionality as the measured items. However, using any other indicator will lead to the same conclusions. The quality of the synthetic population is usually measured by its sociodemographic accuracy, i.e., its fit to marginals observed at the most aggregate resolution. However, as suggested by Khachman et al. (2021b), the quality of a synthetic population is also reflected through its spatial precision, i.e., how well the spatial distributions of sociodemographic characteristics controlled is reproduced. Hence, to draw a complete picture of the performance of each framework regarding both, sociodemographic accuracy, and spatial precision, we define four different types of RMSE measures: (1) accuracy error, (2) precision error, (3) total error, and (4) building-resolution error. The accuracy error is measured by calculating the RMSE for all variables at the most aggregate resolution controlled. The precision error is measured by calculating the RMSE for all variables at the least aggregate resolutions where their marginals are available. A RMSE combining both errors is then calculated (total error) to assess the overall quality including sociodemographic accuracy and spatial precision. Finally, a focused assessment of building-resolution error is made through a RMSE that includes variables controlled at the building resolution only. The smaller the RMSE, the closer the estimated values to the observed values, hence the better the fit.

5.4 Case study

We consider the Montreal Island, which includes the city of Montreal together with other municipalities, as our study area. In the 2016 census Standard Geographical Classification, Montreal Island is classified as a Census Division (CD) (Figure 5.6a) which is comprised of 3202 DAs (Figure 5.6b) or 13,808 Dissemination Blocks (DBs) (Figure 5.6c). A CD is a provincially legislated area which size is between the province and municipalities (Statistics Canada 2018). The DA is an area comprising 400 to 700 people (Statistics Canada 2018). DBs are subdivisions of the DAs that are delimited by roads and/or DAs' boundaries (Statistics Canada 2018). The IMF is used to generate a synthetic population for the city of Montreal following the steps detailed hereafter.

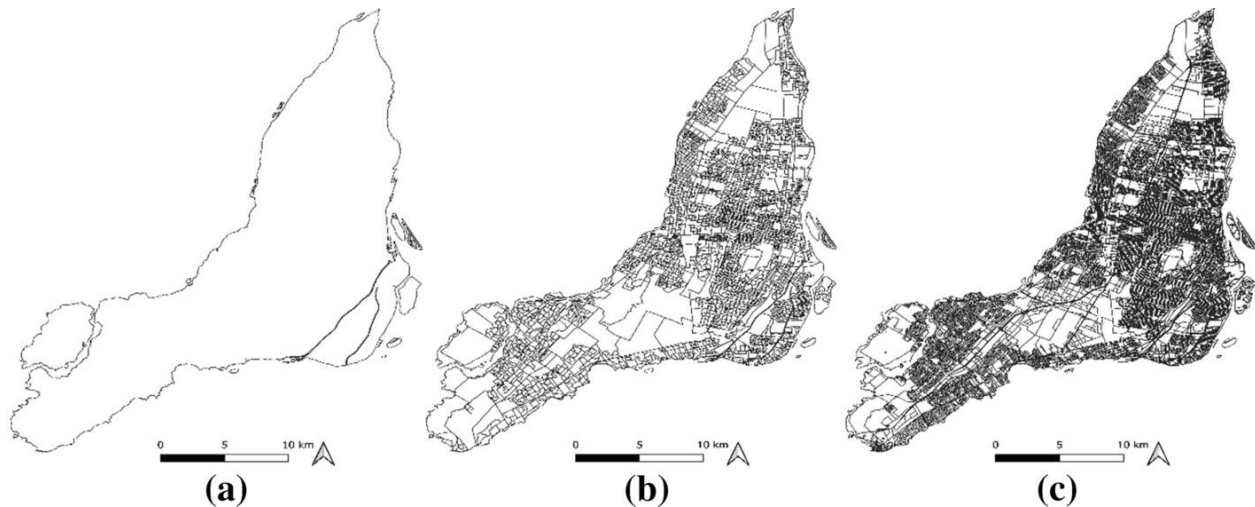


Figure 5.6 Montreal Island (a) CD, (b) DAs and (c) DBs

5.4.1 Step 0: preliminary harmonization

Multiple papers have mentioned the problem of census data inconsistencies (Khachman et al. 2021b; Moreno and Moeckel 2018; Konduri et al. 2016; Barthelemy and Toint 2013; Harland et al. 2012). Khachman et al. (2021b) also measured the harm to the synthetic population's quality caused by using raw census data. Hence, to ensure that optimal synthetic populations are generated, we perform a preliminary harmonization of census data extracted using the harmonization method proposed by Khachman et al. (2021b) and briefly described above.

5.4.2 Step 1: generation of the built environment

To generate the built environment, we fused two public datasets: addresses and property assessment units (Figure 5.7) both publicly disseminated on the City of Montreal's website (Ville de Montréal 2023a; Ville de Montréal 2023b).

The addresses database consists of points corresponding each to a unique address. The complete address and the point's coordinates are specified. However, no information about the characteristics of the built environment is included in this database. The property assessment units' database consists of polygons corresponding each to a property assessment unit. For each polygon, the corresponding address or range of addresses, area, number of floors, number of DUs, construction year and type (condominiums or not) are specified, among others. We used the correspondence of geographic locations and common address information between the two databases to associate to

each point address the following attributes: number of DUs, construction year and condominium status. Characteristics of superposed point addresses are finally summed to produce totals per building (i.e., per unique couple of coordinates). Table 5.2 shows the building-resolution attributes thus obtained. As the 2016 Canadian census observations are considered in the next steps, the built environment is filtered to 2016.



Figure 5.7 Point addresses and property assessment units

Table 5.2 Building-resolution attributes

Variable	Definition	Categories	Description
id	Building unique identifier	—	—
lat	Building latitude	—	—
lon	Building longitude	—	—
ducount	Number of HUs per building	—	—
dutype	Condominium status of buildings' HUs	1	Condominium
		2	Not condominium
dubuilt	Building's construction period	1	Before 1960
		2	1961–1980
		3	1981–2000
		4	2001–2016

5.4.3 Step 2: HH marginals synthesis at the building resolution

As a preliminary optional step, occupancy rates are calculated at the DB resolution where the number of HHs in each DB is divided by the number of DUs in the same DB. The numbers of HHs and DUs at the DB resolution are disseminated by Statistics Canada (Statistics Canada, 2019b). The count of DUs in each building (*ducount*) is then multiplied by the occupancy rate of the DB it belongs to. The application of observed occupancy rates allows to take into consideration the vacancy discrepancies between DBs generating thus a more realistically spatialized synthetic population.

DU counts of buildings located in each DB are first proportionally adjusted to fit to the HH count in the same DB. Synthetic HH counts (*hhcount*) at the building resolution are thus obtained. The iterative intra and inter-resolution adjustments of the harmonization process for *dutype* and *dubuilt* require two sets of constraints. Synthetic HH counts at the building resolution are used for intra-resolution adjustments. For inter-resolution adjustments, the HH's DU type (*hhdtype*) and the HH's DU construction period (*hhdbuilt*) distributions at the DA resolution are respectively used. The example in Figure 5.3 corresponds to the iterative adjustment applied to the *dubuilt* variable. The harmonization process converges to synthetic *hhcount*, *hhdtype* and *hhdbuilt* distributions at the building resolution. These are thus integrated into the multiresolution control.

5.4.4 Step 3: multiresolution optimization-based synthesis

The multiresolution synthesis requires marginal distributions of variables of interest and a disaggregate microdata sample. Control variables, their categorizations, and the resolutions their distributions are available at, are summarized in Table 5.3.

Table 5.3 Control variables

Variable	Definition	Categories	Description	Resolutions
ppcount	PP count	1	1 person	CD, DA, <u>DB</u>
ppage	PP age	1	[0, 15]	CD, <u>DA</u>
		2	[15, 65]	
		3	65+	
ppsex	PP sex	1	Men	CD, <u>DA</u>
		2	Women	
ppmarst	PP marital status	1	Never legally married (and not living in common law)	CD, <u>DA</u>

Table 5.3 Control variables (suite)

Variable	Definition	Categories	Description	Resolutions
		2	Legally married (and not separated)	
		3	Living common law	
		4	Separated, divorced, or widowed (and not living in common law)	
pplfst	PP labour force status	1	Employed	CD, <u>DA</u>
		2	Unemployed	
		3	Not in the labour force	
hhcount	HH count	1	1 Household	CD, DA, DB, <u>Building</u>
hhsz	HH size	1	1 Person	CD, <u>DA</u>
		2	2 People	
		3	3 People	
		4	4 People	
		5	5 + People	
hhtype ^a	HH type	1	Non census family	CD, <u>DA</u>
		2	1 Census family without children	
		3	1 Census family with children	
		4	Multiple census families	
hhnetinc	HH net income	1	Less than 30 k\$	CD, <u>DA</u>
		2	[30 k\$, 60 k\$]	
		3	[60 k\$, 100 k\$]	
		4	100 k\$ and more	
hhdtype	HH's DU type	1	Condominium	CD, DA, <u>Building</u>
		2	Not condominium	
hhdbuilt	HH's DU construction period	1	Before 1960	CD, DA, <u>Building</u>
		2	1961–1980	
		3	1981–2000	
		4	2001–2016	

^a“Census family is defined as a married couple and the children, if any, of either and/or both spouses; a couple living common law and the children, if any, of either and/or both partners; or a lone parent of any marital status with at least one child living in the same dwelling and that child or those children.” (Statistics Canada 2018)

To optimize the synthesis quality, each variable is controlled at the least aggregate resolution its marginal distribution is available at (Khachman et al. 2021b). Synthetic populations are hence generated for each DA separately controlling the resolutions underlined in Table 5.3. The joint microdata sample used is the hierarchical PUMF 2016 for the Montreal Census Metropolitan Area.

5.4.5 Spatialized population synthesis using the CF

The CF is also used to generate a spatialized synthetic population for each DA. The single-resolution multilevel optimization method proposed by Fournier et al. (2021) is first used to generate a synthetic population for each DA with all variables listed in Table 5.3 controlled at the DA resolution. To keep IMF and CF results comparable, the same harmonized marginals at the DA resolution, built environment, and DB-resolution occupancy rates are used.

A heuristic is then defined to distribute synthetic households generated for each DA on buildings located within it. The idea is to assign each household to a building having the same type and construction period as its DU without exceeding the estimated number of HHs in the building. The three constraints (i.e., construction period correspondence, type correspondence, and respect of the number of HHs) are progressively relaxed when needed to assign all synthetic HHs to buildings. Even the number of DUs in buildings (including the ones initially estimated to be vacant) is sometimes exceeded for spatial assignment of all synthetic households to be possible.

5.5 Results

The spatialized synthetic populations generated by both the CF and the IMF are compared using the four errors previously defined (accuracy, precision, total, and building-resolution errors). For each indicator, Table 5.4 summarizes the attributes and the resolution(s) for which the differences between synthetic and observed totals are considered in the RMSE calculation.

Table 5.4 Attributes and resolutions considered for each indicator

Variable	Accuracy error	Precision error	Total error	Building-resolution error
ppcount	DA	DB	DA and DB	–
ppage	DA	DA	DA (2x)	–
ppsex	DA	DA	DA (2x)	–
ppmarst	DA	DA	DA (2x)	–
pplfst	DA	DA	DA (2x)	–
hhcount	DA	BLD	DA and BLD	BLD
hhsize	DA	DA	DA (2x)	–
hhtype	DA	DA	DA (2x)	–
hhnetinc	DA	DA	DA (2x)	–
hhdtype	DA	BLD	DA and BLD	BLD
hhdbuilt	DA	BLD	DA and BLD	BLD

As synthetic populations are generated for each DA separately, a local assessment is first carried out by calculating and comparing the four indicators for the CF-based and the IMF-based populations synthesized in each DA independently. Then, the four error types are calculated for the whole Montreal Island's synthetic populations generated through the CF and the IMF, respectively, to globally assess and compare their performances.

5.5.1 Local assessment

The boxplots in Figure 5.8 show the distributions of (a) accuracy, (b) precision, (c) total, and (d) building-resolution errors for the populations generated through the CF (blue) and the IMF (orange), respectively, in each of the study area's 3202 DAs. The CF generally achieves a better accuracy than the IMF with a median accuracy error of 2.05 and a mean (cross symbol "x") accuracy error of 3.23 (compared to 2.94 and 4.14 respectively for the IMF). This is expected since for the CF all variables are controlled at the DA resolution only, while the IMF controls some variables at more disaggregate resolutions including the building resolution. Hence, the optimization problem's complexity is drastically increased when the IMF is used which explains a worse fit at the DA resolution compared to that achieved by the simpler CF optimization problem. However, the IMF achieves a better spatial precision than the CF. The maximum PE produced by the IMF (1.95) is lower than the median PE produced by the CF (2.19). The IMF median and mean PE are around 1/3 the CF median and mean PE, respectively. Two features of the IMF make the synthetic populations it generates more spatially precise: 1) its ability to address inconsistencies between the built environment and zonal census marginals through the HH marginals synthesis (step 2), and 2) its ability to simultaneously control variables of interest each at the least aggregate resolution where its marginals are available (step 3).

While the Interquartile Ranges of the accuracy errors produced by the CF and the IMF show an important overlap ([1.06; 3.88] for the CF, and [2.27; 4.43] for the IMF), no overlap of precision error Interquartile Ranges ([1.33; 3.68] for the CF, and [0.5; 1.14] for the IMF) could be observed. This would suggest that the IMF generates more precision gain than accuracy loss compared to the CF. The distributions of total errors confirm this observation with the IMF's total error distribution showing a 0.96 median, 1.47 mean, and 2.74 maximum compared to 2.25 median, 3.5 mean, and 7.2 maximum for the CF. Hence, the IMF generates better synthetic populations showing lower total error.

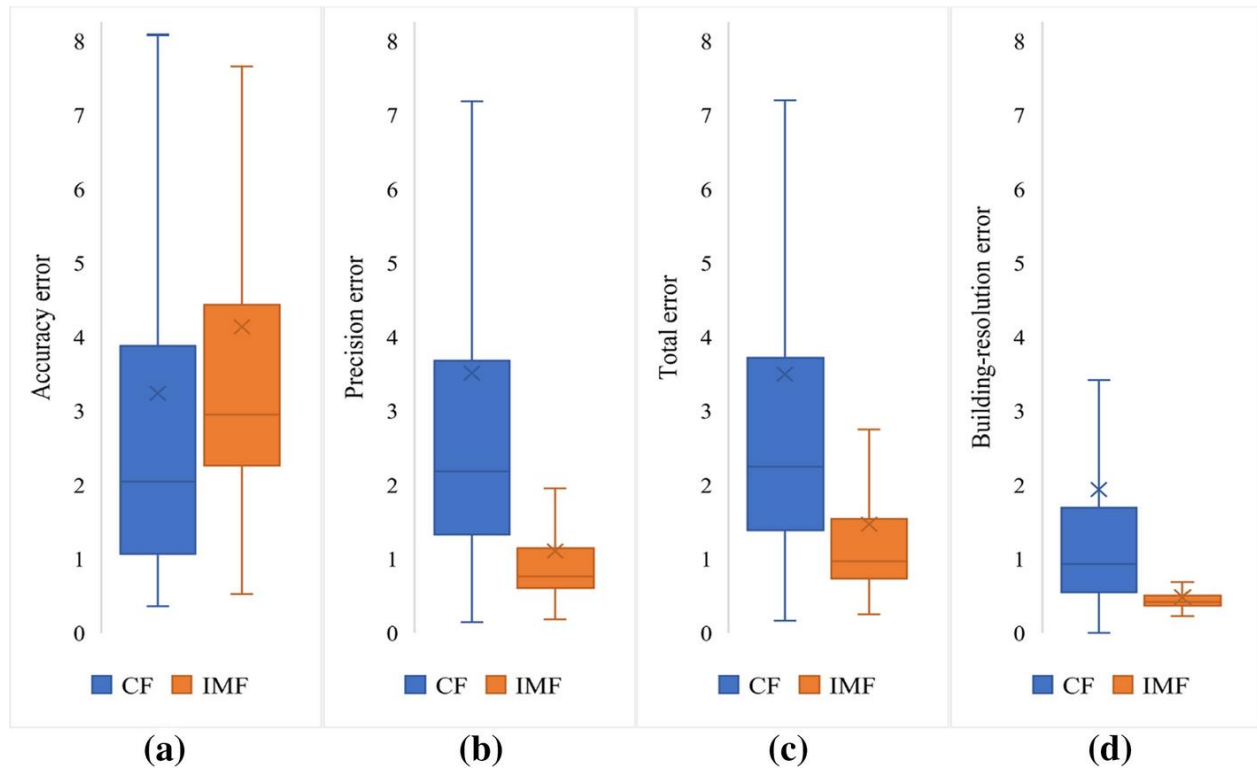


Figure 5.8 Distributions of **a** accuracy, **b** precision, **c** total and **d** building-resolution errors

Finally, the distributions of building-resolution errors clearly show the superiority of the IMF for explicit spatialization tasks with a maximum building-resolution error (0.68) that remains lower than the median building-resolution error produced by the CF (0.92). It is worth mentioning that the distributions of the four indicators show lower ranges and Interquartile Ranges for IMF-based synthetic populations compared to CF-based ones. The differences are particularly noticeable for precision, building-resolution, and total errors. The IMF thus shows a drastically better and more stable performance than the CF regardless of the initial consistency of the built environment data available with census marginals. The IMF proves to be more robust than the CF with a better performance that fluctuates less from a DA to another.

To further investigate the efficiency of the IMF in terms of total error reduction, the total error difference between the IMF-based and CF-based synthetic populations (ΔTE) is calculated for each DA. A positive ΔTE for a DA thus means that the IMF performed better than the CF in that DA, and vice versa. The histogram of ΔTE (Figure 5.9) shows that the IMF generates improved synthetic populations, compared to those generated by the CF, for 95.5% of the DAs. The total error reductions are mainly included in the (0;5] interval with a maximum reduction of 203. Two

factors would explain the low percentage of DAs showing quasi-null total error increases: (1) the probabilistic components of the synthesis frameworks such as decimals to integers conversions occurring more in the IMF; (2) the drastically higher complexity of the IMF's optimization problem. However, the overwhelming majority of DAs where the IMF generated better synthetic populations confirms the general better performance of the IMF. In summary, the observations of Figure 5.8 and Figure 5.9 both confirm that the IMF proposed regularly performs better than the CF, in addition to being more easily transferable.

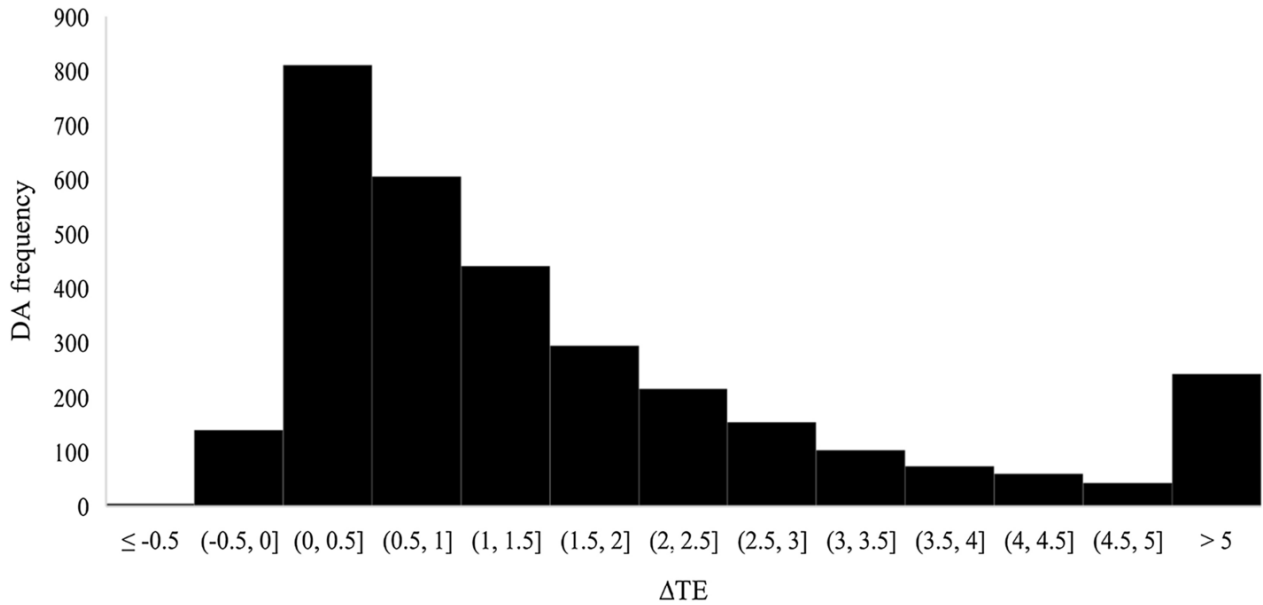


Figure 5.9 Histogram of ΔTE

5.5.2 Global assessment

The synthetic totals of the whole Montreal Island's populations, generated by the CF and the IMF respectively, are compared to their corresponding marginal totals. The accuracy, precision, total, and building-resolution errors are also calculated and compared. Figure 5.10, Figure 5.11, Figure 5.12 and Figure 5.13 show global comparisons of fit reflecting respectively the accuracy, precision, total quality, and building-resolution quality of synthetic populations generated through (a) CF and (b) IMF. The global observations confirm the local observations previously made. A better global accuracy is found to be achieved through the CF (Figure 5.10), and a better global precision is provided by the IMF (Figure 5.11). However, while the difference in terms of global accuracy remains hardly detectable when comparing CF and IMF graphs in Figure 5.10, the precision

improvement generated by the IMF is easily noticeable in Figure 5.11. When compared to the CF, the IMF generates a 15.8% accuracy error increase and a 74.6% precision error decrease. Hence, the gain in precision generated by the IMF exceeds the loss in accuracy yielding a 62.9% decrease of total errors when IMF is used. The overall quality of the global synthetic population is thus drastically improved when generated through the IMF which is reflected by a better diagonal alignment in Figure 5.12b when compared to Figure 5.12a. Finally, the most important performance difference between the IMF and the CF is measured at the building resolution where the IMF achieved a 79% reduction of building-resolution errors compared to the CF (Figure 5.13).

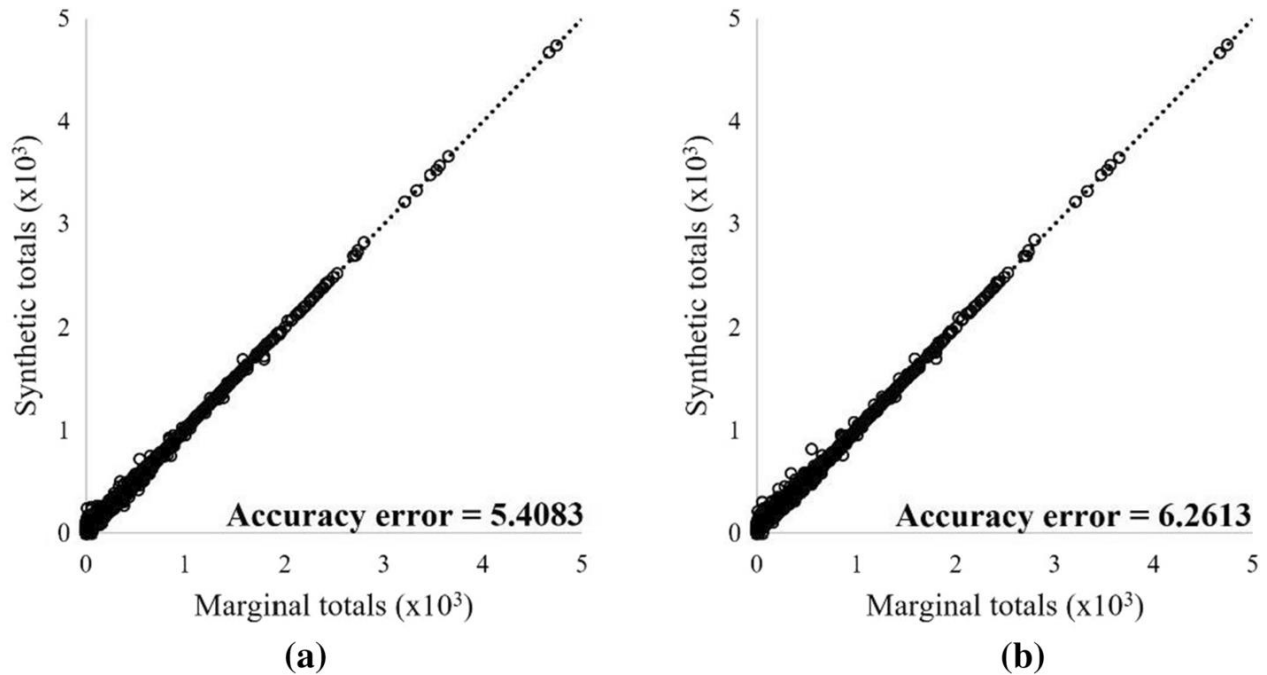


Figure 5.10 Accuracy of the global synthetic population generated through **a** CF and **b** IMF

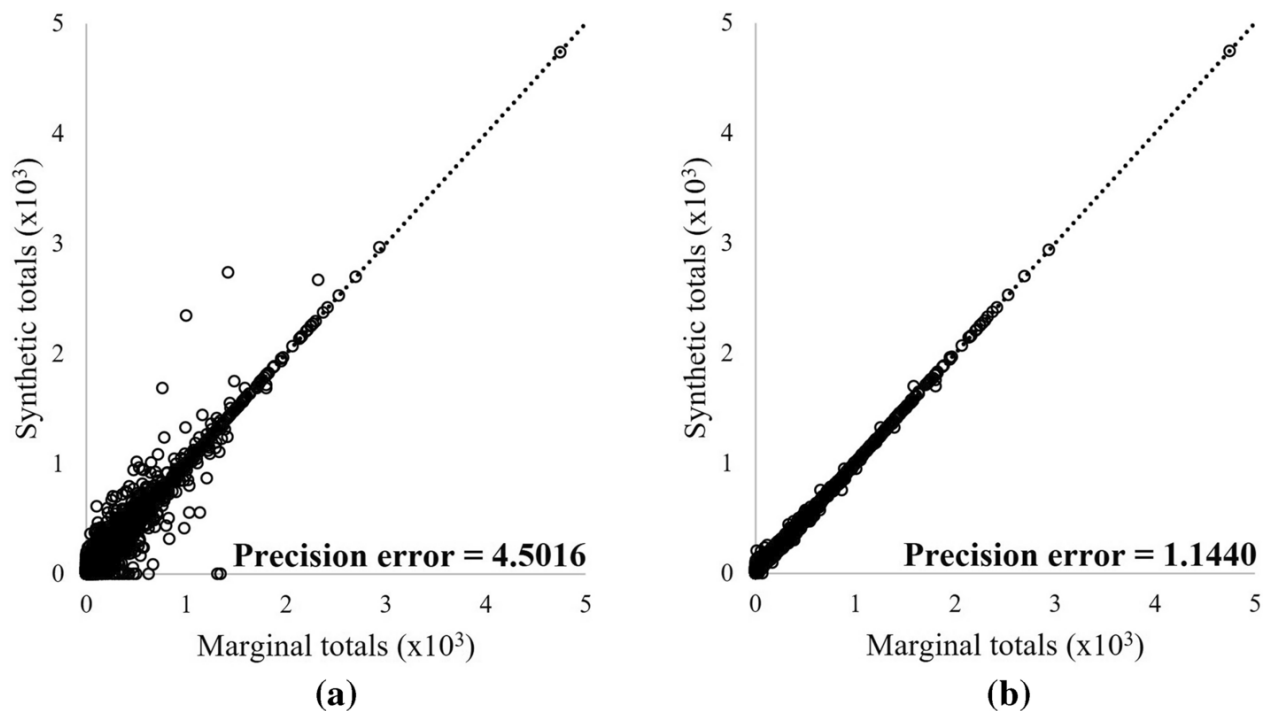


Figure 5.11 Precision of the global synthetic population generated through **a** CF and **b** IMF

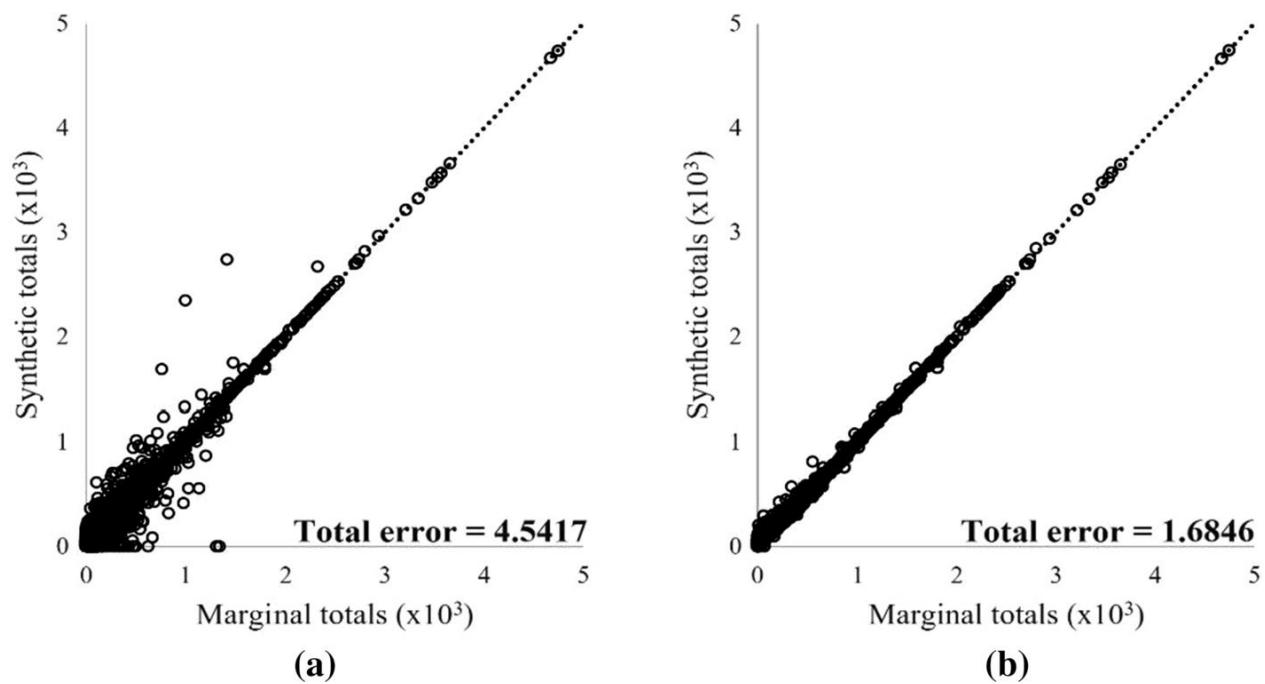


Figure 5.12 Total quality of the global synthetic population generated through **a** CF and **b** IMF

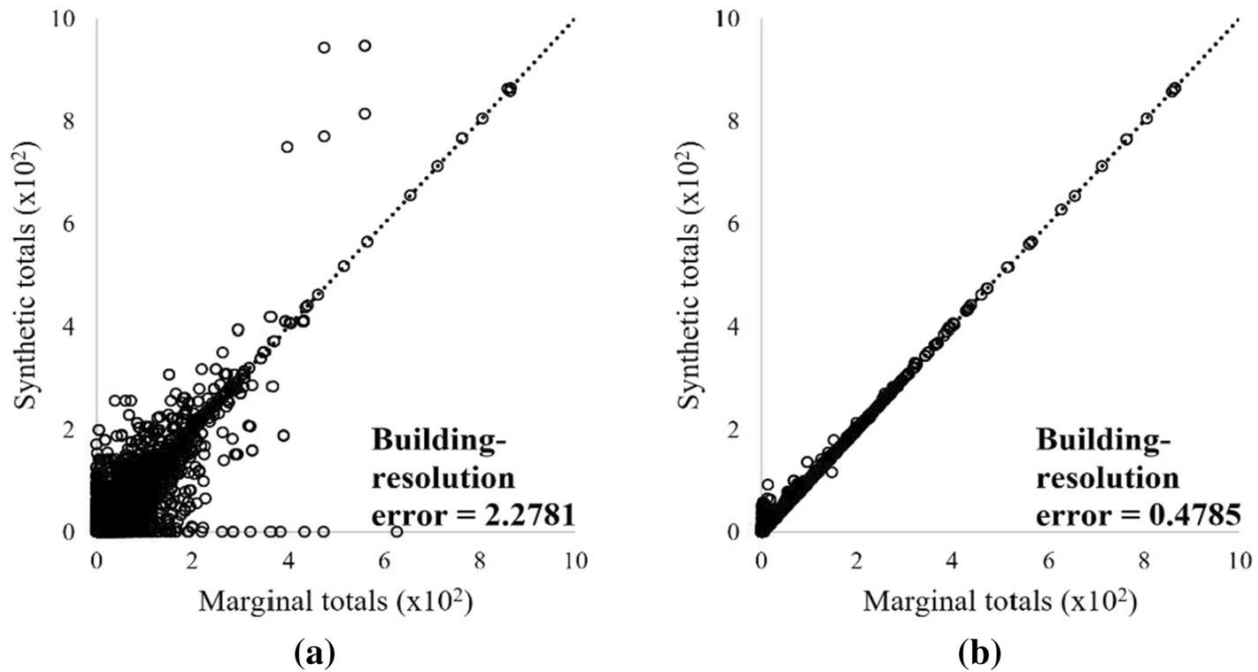


Figure 5.13 Building-resolution quality of the global synthetic population generated through **a** CF and **b** IMF

5.6 Conclusion and outlook

The sociodemographic characteristics of an individual and his household, as well as his residential location are determinant factors of his mobility behavior. Hence, to realistically model the mobility behavior of the population, a spatialized synthetic population reliably reflecting its sociodemographic characteristics and their spatial distribution is needed. While much research focused on improving the sociodemographic accuracy of synthetic populations, little attention was given to enhancing the spatialization process.

The conventional spatialized population synthesis approaches present numerous shortcomings. The separation of generation and spatialization processes makes these methods vulnerable to data inconsistencies between the sociodemographic observations and the built environment. These approaches are also highly reliant on the availability of rich spatial datasets and knowledge of the local context, which raises transferability issues. Some spatialization methods are also based on fully assumed distributions which does not provide a realistic spatial distribution of synthetic households. In this paper, we introduce a new integrated multiresolution framework for spatialized population synthesis. The proposed framework directly generates synthetic populations at the

building resolution avoiding thus data inconsistency issues. The transferability of the suggested framework is also optimized by limiting its data requirements. Only a single marginal distribution of a dwelling unit characteristic at the building resolution (e.g., dwelling unit count in each building) is required. The proposed approach operates using a multiresolution algorithm that is better than its existing pairs in terms of flexibility and efficiency. It can efficiently handle constraints at any desired number of resolutions including the most disaggregate ones such as the building resolution.

Both the conventional and the integrated multiresolution frameworks were used to generate explicitly spatialized synthetic populations for the DAs of the Montreal Island. Four types of RMSE were calculated to assess the generated synthetic populations: sociodemographic accuracy error, spatial precision error, total error, and building-resolution error. Despite generating slightly more accuracy errors than the conventional framework, mainly due to the multiplicity of optimization constraints, the integrated multiresolution framework drastically reduced precision errors providing thus synthetic populations with better overall quality. The integrated multiresolution framework was also found to particularly reduce building-resolution errors, thus showing significantly better performances than the conventional framework for explicitly spatialized population synthesis.

A main challenge when using the proposed framework is storing huge matrices. This problem is particularly encountered when the microdata sample lacks precise spatial information. In our case, the only spatial information available was that the sample belongs to the Montreal Census Metropolitan Area. Hence, the sample was duplicated for each building to maximize variance and avoid zero-cell problems. The largest A matrix handled for a DA of Montreal Island reached a dimension of about 10^{10} . The optimization algorithm used can easily and quickly deal with such dimensions. However, storing them in R while keeping them quickly modifiable is a real challenge. As R by default stores data in the virtual memory, around 100 GB RAM are required to store the largest A matrix we had if defined using R classic matrix structure. This implies a need for a virtual machine when a non-spatialized sample is used. To avoid that, R sparse matrix structures can be used since the A matrices are extremely sparse. These drastically reduce the A matrices' sizes thus making them easily handleable with a 16-GB-RAM machine. This solution should be cautiously used because modifying sparse matrix structures in R is slower than modifying R classic matrix structures. However, this does not impact the synthesis speed.

Research works are in progress to test applications of machine learning methods to explicitly spatialized population synthesis in Montreal. The goal is to explore the potential of these methods in the spatialized synthesis exercise and compare their performances with those of the integrated multiresolution framework proposed. The possibility to upgrade the integrated multiresolution framework by including such methods is also being investigated.

Acknowledgements

The authors wish to acknowledge the contribution and financial support of the Mobilité research chair partners: Ministère des Transports du Québec (MTQ), Société de Transport de Montréal (STM), Autorité Régionale de Transport Métropolitain, exo, and Ville de Montréal.

Authors and affiliations

Mohamed Khachman, Catherine Morency & Francesco Ciari

Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, C.P. 6079, Station Centre-Ville, Montreal, QC, H3C 3A7, Canada.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Mohamed Khachman, Catherine Morency, Francesco Ciari; analysis and interpretation of results: Mohamed Khachman; draft manuscript preparation: Mohamed Khachman, Catherine Morency, Francesco Ciari. All authors reviewed the results and approved the final version of the manuscript.

Corresponding author

Correspondence to Mohamed Khachman.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Below is the link to the electronic supplementary material.

https://static-content.springer.com/esm/art%3A10.1007%2Fs11116-022-10358-w/MediaObjects/11116_2022_10358_MOESM1_ESM.zip

https://static-content.springer.com/esm/art%3A10.1007%2Fs11116-022-10358-w/MediaObjects/11116_2022_10358_MOESM2_ESM.zip

Rights and permissions

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

CHAPITRE 6 ARTICLE 3: A NOVEL MACHINE LEARNING-BASED SPATIALIZED POPULATION SYNTHESIS FRAMEWORK

Rédigé par :

Mohamed Khachman, Catherine Morency et Francesco Ciari

Soumis le 30 novembre 2023 dans :

Transportation

Abstract

Synthetic populations are increasingly required in transportation demand modelling practice to feed the large-scale agent-based microsimulation platforms gaining in popularity. The quality of the synthetic population, i.e., its representativeness of the sociodemographic and the spatial distribution of the real population, is a determinant factor of the reliability of the microsimulation it feeds. While many research works focused on improving the sociodemographic accuracy of synthetic populations, the quality of their spatial distribution remained less covered. Explicitly spatialized population synthesis methods are based on one of two approaches: the Integrated Approach (IA) where the synthetic population is directly generated at the building/dwelling unit scale, or the Sequential Approach (SA) where a synthetic population generated at a coarse resolution is then distributed on buildings/dwelling units. While IA-based methods show better transferability and lower sensitivity to inconsistencies between synthetic populations and built environments, they may suffer from damaged sociodemographic representativeness due to lack of inputs at disaggregate resolutions, and computational issues due to the multiplicity of constraints. This paper suggests a new SA-based explicitly spatialized population synthesis framework. It leverages the performant Clustering Large Applications and Random Forest algorithms as well as rich spatial information collected as part of surveys to make accurate predictions of synthetic households' locations at the building scale directly. In addition to preserving optimal sociodemographic accuracy and achieving realistic explicit spatialization, the new framework shows acceptable transferability thanks to CLARA's efficiency. An explicitly spatialized synthetic population for Montreal Island is generated using the proposed clustering+classification framework. Zonal population synthesis, clustering, classification, and spatial distribution processes making up the proposed framework have all generated satisfactory results.

Keywords: Travel Demand Modelling, Population Synthesis, Explicit Spatialization, Machine Learning, Clustering Large Applications, Random Forest.

6.1 Introduction

Simulating microscopic mobility behaviors using agent-based models is increasingly popular in transportation demand modelling and forecasting. These simulations need to be fed with rich datasets fully enumerating, characterizing, and locating the population of households and individuals. Such datasets are collected by national statistical agencies but generally remain unavailable for research and planning purposes due to privacy issues. Synthetic populations are thus used instead. A reliable synthetic population is a population of non-identifiable agents with a distribution of sociodemographic attributes and a spatial distribution that are similar to those of the real population (Chapuis et al., 2018).

To generate a synthetic population, the first step is to choose the control variables, i.e., variables whose distributions in the real population are to be reproduced. Control variables' choice mainly depends on the purpose of the simulation the synthetic population will be feeding. For most synthesizers, two inputs are then generally sought: zonal aggregations of the control variables published by national statistical agencies, and a disaggregate sample of the population. The synthesis process generally aims at expanding the sample to the population's size while preserving the correlation between sociodemographic attributes in it. From the first Iterative Proportional Fitting (IPF)-based synthesizer (Beckman et al., 1996) to the latest synthesizers using deep generative learning algorithms (Kim & Bansal, 2023) much effort was focused on refining synthetic populations' generation methods to improve the reliability of the control variables' distribution in the synthetic population mainly.

The optimization of synthetic households' spatial distribution is less frequently tackled in the literature. However, an optimal spatial distribution is as important as an optimal distribution of sociodemographic attributes for a reliable mobility behaviors' model (Yaméogo et al., 2023; Chapuis et al., 2018). That is because correlations between sociodemographic attributes vary across spatial locations (Zhou et al., 2022) implying a spatial variation of mobility behaviors (Zhu et al., 2018). Therefore, increasingly popular fine-grained urban planning applications including active transportation planning studies (Pajares et al., 2021), and long-term residential and workplace choices modelling (Zhu et al., 2018), require accurately and explicitly located (i.e., at the building or the dwelling unit scale) synthetic populations (Liu et al., 2021).

The paper’s research objective is to contribute to improving spatial distributions of synthetic populations. It introduces a new spatialization method allowing each synthetic household to be assigned an explicit location, i.e., (X;Y) coordinates. This method leverages machine learning algorithms and spatial information usually collected as part of surveys to improve the quality of the explicit spatialization of synthetic agents.

The remainder of this paper is organized as follows: first, we describe the most popular synthetic population generation and spatialization methods and give an overview of the machine learning algorithms applied to spatialized population synthesis in this paper. Then, a machine-learning-based explicit spatialization method is proposed. After that, an application of the explicit spatialization method introduced to spatialized population synthesis in the city of Montreal, Canada is detailed. Results are subsequently reported and analyzed. Conclusions, limitations, and research outlooks are finally encompassed in the last section.

6.2 Literature review

In this section, we briefly showcase population synthesizers and explicit spatialization attempts. We also give an overview of machine learning algorithms used to achieve explicit spatialization of synthetic households in this paper.

6.2.1 Population synthesis

Synthetic populations generation methods can be grouped into three categories: fitting-based methods, simulation-based methods, and mixed methods.

6.2.1.1 Fitting-based synthesizers

Fitting-based synthesizers are the most popular in population synthesis practice. They generally belong to one of three categories: Synthetic Reconstruction (SR), Combinatorial Optimization (CO), and optimization-based methods.

SR population synthesizers use IPF (Deming & Stephan, 1940) and Monte Carlo simulations to weight the records of a microdata sample so that the frequencies of control variables’ categories published by national statistical agencies are met. Improvements of SR synthesizers were introduced thanks to various extensions of the IPF algorithm. Tweaking, categories aggregation, and frequencies interpolation techniques were used to warrant IPF convergence despite incorrect

zero cells that may occur due to existing types of agents not being observed in the microdata sample used. Multilevel synthesis algorithms such as Iterative Proportional Updating (IPU) (X. Ye et al., 2009), hierarchical IPF (hIPF) (Müller & Axhausen, 2011) and entropy optimization (Bar-Gera et al., 2009) made simultaneous control of constraints for multiple types of agents (e.g., households and individuals) possible. Konduri et al. (2016) and Moreno and Moeckel (2018) introduced multiresolution SR synthesizers allowing sociodemographic attributes observed at various geographic resolutions to be simultaneously controlled.

CO-based synthesis operates by initially drawing a synthetic population from the microdata sample, then, progressively updating each synthetic agent to improve the synthetic population's fit to the real population. CO-based synthesizers suggested in the literature are based on various algorithms such as simulated annealing (Voas and Williamson, 2000), hill climbing (Williamson et al., 1998; Abraham et al., 2012), genetic algorithm (Williamson et al., 1998; Birkin et al., 2006) and greedy heuristic (Srinivasan et al., 2008). Better synthetic population could be generated through CO compared to IPF (Ryan et al., 2009). However, the computational complexity of CO methods and their high risk to yield a solution that corresponds to a local optimum (Lee and Fu, 2011; Prédhumeau and Manley, 2023) limited their popularity in population synthesis practice. Abraham et al., (2012) and Ma and Srinivasan (2015) introduced multilevel CO-based synthesizers achieving fit at both individuals and households' levels simultaneously.

Recently, Fournier et al. (2021) introduced an optimization-based reweighting method using Cyclic Coordinate Descent to optimally weight the microdata sample households so that the real population's characteristics are reproduced. The proposed algorithm was found to be more efficient and accurate than its SR equivalent, the IPU algorithm (Fournier et al., 2021). An extension of this algorithm where any number of levels, i.e., agent types, and geographic resolutions can simultaneously and efficiently be controlled was then conceived (Khachman et al., 2022).

Despite some sample-free synthesizers introduced (Gargiulo et al., 2010; Barthelemy and Toint, 2013), the main weakness of SR techniques lies in that, in most cases, they strongly rely on the availability, but also on the quality of the microdata sample feeding the synthesizer. However, available disaggregate samples are usually of an extremely limited size (5% or less of the population) which raises questions regarding the representativeness of the corresponding populations (Prédhumeau and Manley, 2023).

6.2.1.2 Simulation-based synthesizers

Simulation-based or statistical learning methods operate in two steps. The joint distribution of chosen control variables is first estimated based on all sociodemographic datasets available. Synthetic agents are then sampled according to the estimated distribution. Various simulation-based synthesizers were introduced. Markov Chain Monte Carlo (MCMC) (Farooq et al., 2013) and Hidden Markov Model (HMM) (Saadi et al., 2016) were used for joint distribution estimation purposes. However, since a Markov-based synthesizer requires prior knowledge of control variables interdependencies (Zhou et al., 2022), Sun and Erath (2015) and Zhang et al. (2019) introduced alternative data-driven approaches based on Bayesian Networks. Recently, various deep learning algorithms such as Variational Auto-Encoder (Borysov et al., 2019), Wasserstein Generative Adversarial Network (Garrido et al., 2020), Conditional Variational Auto-Encoder (Borysov and Rich, 2021), Directed Acyclic Tabular Generative Adversarial Network (Lederrey et al., 2021), Composite Travel Generative Adversarial Network (Badu-Marfo et al., 2022), and Conditional Generative Adversarial Networks (Johnsen et al., 2022). Deep generative learning was found to be particularly performant when dealing with the sampling zeros problems (Lederrey et al., 2021).

The main advantage of simulation-based algorithms to generate sampling or incorrect zeros, i.e., types of agents that exist in the population but are not recorded in the microdata sample (Farooq et al., 2013). However, this feature leads also to generating structural zeros, i.e., types of agents that do not exist neither in the population nor in the sample (Kim and Bansal, 2023). Moreover, simulation-based synthesizers achieve a better fit to the multivariate joint distribution, but at the expense of diverging from the observed marginal distribution (Yaméogo et al., 2021). Finally, overfitting to sample's noise is another main concern when simulation-based approaches are used (Basu and Ferreira, 2020; Lederrey et al., 2021).

6.2.1.3 Mixed synthesizers

Mixed synthesizers were conceived with the aim of getting the best of fitting-based and simulation-based synthesizers' worlds, i.e., an optimal fit to both multivariate and marginal distributions (Sun and Erath, 2015; Sun et al., 2018). The idea is to first generate a synthetic population using a statistical learning algorithm so that the pool of synthetic agents is not limited to what is recorded in the microdata sample. Then, the synthetic population obtained is used to seed a fitting-based

population synthesizer that perfects its fit to the marginal distribution observed. Mixed synthesizers consisting of a hierarchical MCMC + Generalized Raking framework (Casati et al., 2015), an HMM + IPF framework (Saadi et al., 2018), and Bayesian Networks + Generalized Raking frameworks (Ilahi and Axhausen, 2019; Zhou et al., 2022) were introduced. Fournier et al. (2021) also tried both MCMC and Bayesian Networks to enrich a pool of synthetic households and individuals that were separately generated. They then jointly reweighted the synthetic agents using an optimization-based algorithm.

6.2.2 Explicit spatialization

Individuals' mobility behaviors are determined by both their sociodemographic attributes as well as those of the household they belong to, and their household's geographic location. Hence, precisely and accurately locating synthetic households is essential to warrant a realistic microsimulation of mobility behaviors. Most spatialized population synthesis methods suggested in the literature are limited to zonal spatialization where synthetic households are assigned to small zones such as disaggregate administrative areas (Yaméogo et al., 2023; Nejad et al., 2021; Long and Shen, 2015, p. 115–131; Anderson et al., 2014) or cells of a pre-defined grid (Ji and Wan, 2021; Su et al., 2010). While zonal spatialization is suitable for many applications, georeferenced synthetic populations remain necessary for various fine-grained and long-term urban planning applications including mobility planning, residential and workplace choices modelling, urban energy planning, disaster management, homeland security, facility placement, and public health (Chen et al., 2021; Liu et al., 2021; Pajares et al., 2021; Chapuis et al., 2018; Thomson et al., 2018; Zhu et al., 2018; Bast et al., 2015; Muñoz and Peters, 2014). Explicit spatialization techniques were thus introduced to assign a point location, i.e., (X;Y) coordinates, to synthetic households.

To generate a spatialized synthetic population, two approaches can be used: the Sequential Approach (SA) where synthetic population generation and spatialization are sequentially performed, or the Integrated Approach (IA) where the population is directly generated at the fine-grained resolution desired. Figure 6.1 shows the classification of existing explicitly spatialized population synthesis methods established in this section.

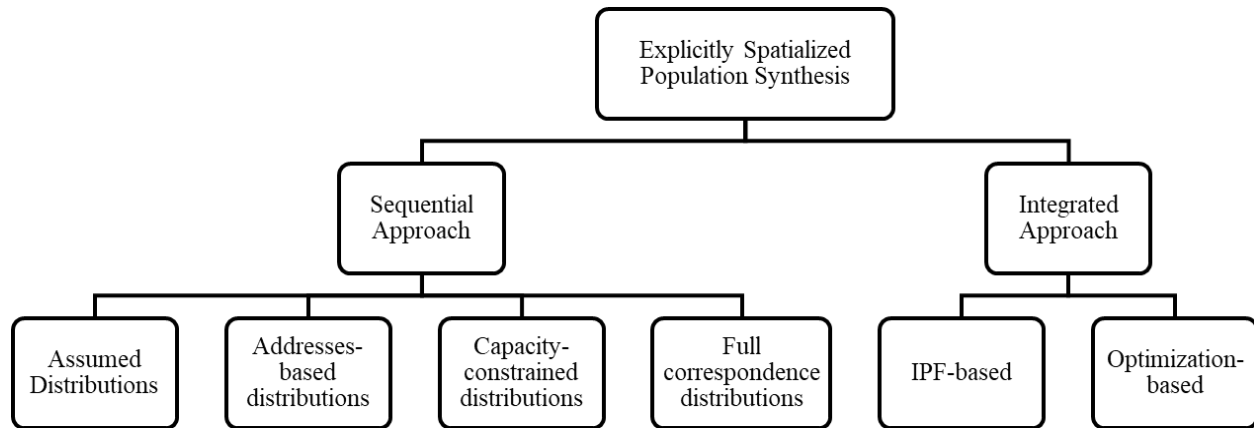


Figure 6.1 Classification of existing explicitly spatialized population synthesis methods

6.2.2.1 Sequential Approach

Explicitly spatialized population synthesis methods proposed in the literature are mostly based on the SA. A synthetic population is first generated at a coarse geographic resolution and synthetic households are then assigned to buildings or dwelling units within each synthesis geographic area.

The SA-based explicit spatialization techniques can be classed into four categories according to the extent to which the built environment is considered in the spatialization process:

- **Assumed distributions:** this category includes explicit spatialization methods that do not consider the built environment at all. Synthetic households are assigned to coordinates (X;Y) within their corresponding geographic areas using simple geographic distributions such as uniform or weighted distributions within a polygon (Zhang et al., 2019; Gallagher et al., 2018; Swarup and Marathe, 2016) or along the road network (Jiang et al., 2022; Gallagher et al., 2018). These methods yield obviously unrealistically spatialized synthetic populations as synthetic households are not assigned to buildings and can get located in non-residential areas or next to bridges or tunnels.
- **Addresses-based distributions:** this category includes explicit spatialization methods where only buildings' addresses are considered. Synthetic households are randomly distributed on buildings within their geographic area regardless of the buildings' characteristics, e.g., capacity, type, usage, etc. (Zhang et al., 2019; Sallard et al., 2021; Thomson et al., 2018). Besides not respecting buildings' capacities, these methods may assign households to non-

residential parts of the built environment such as non-residential buildings, detached garages, garden houses or floor commerce units.

- Capacity-constrained distributions: this category includes explicit spatialization methods where only building capacity related variables are used to constrain the synthetic households' distribution process (Pajares et al., 2021; Chapuis et al., 2018; Zhu and Ferreira, 2014). More realistic explicit spatialization can be performed using these methods.
- Full-correspondence distributions: this category includes methods where buildings and synthetic households' characteristics are considered to perform explicit spatialization through various processes such as combinatorial optimization (Muñoz and Peters, 2014), matching heuristics (Zhou et al., 2022), probabilistic modelling (Liu et al., 2021), etc. The most realistic explicit spatialization is obviously achieved using methods belonging to this category.

The SA-based explicitly spatialized population synthesis methods show two main limitations. First, tackling the synthetic population generation and spatialization as independent processes usually leads to inconsistencies between the synthetic population and the built environment marginals (Khachman et al., 2022). Some synthetic households may thus not be assigned a location. To ensure the explicit spatialization process completeness, artificial mitigation measures were proposed including relaxing rules and constraints (Zhou et al., 2022; Chapuis et al., 2018), correcting negatives values (Chapuis et al., 2018) and using capacity-constrained algorithms (Liu et al., 2021; De Palma et al., 2007). Second, these methods show transferability issues as they mainly consist of ad-hoc applications highly relying on rich spatial datasets availability, extensive knowledge of the local context (Liu et al., 2021), and locally calibrated models (Khachman et al., 2022).

6.2.2.2 Integrated approach

Explicitly spatialized population synthesis using the Integrated Approach (IA) consists of directly generating the synthetic population at the building or dwelling unit scale. When the control variables' marginals are not all available at the building or dwelling unit resolution which is generally the case, multiresolution control is required so that each control variable is controlled at the resolution where its marginals are available. To the best of authors' knowledge, only few methods can potentially be used to generate a synthetic population at the building scale directly

(Khachman et al., 2022; Moreno and Moeckel, 2018; Konduri et al., 2016; Zhu and Ferreira, 2014), and a single method was actually used to do so (Khachman et al., 2022). There are two types of multiresolution population synthesis methods that can be used to perform IA-based explicitly spatialized population synthesis:

- IPF-based methods such as the enhanced-IPU algorithm where two resolutions can be simultaneously controlled (Konduri et al., 2016) and the three-resolution control algorithm (Moreno and Moeckel, 2018). Zhu and Ferreira (2014) also mentioned that generation and explicit spatialization of a synthetic population could be integrated in a single IPF. These methods can theoretically be used to generate a synthetic population at the building scale directly. However, that remains practically difficult for computational reasons as an extremely high-dimensional matrix is obtained building or dwelling unit-resolution constraints are considered in the fitting process.
- Optimization-based methods such as the flexible framework proposed by Khachman et al. (2022). This method is more efficient than IPF-based methods when dealing with huge matrices (Khachman et al., 2022).

The IA-based explicitly spatialized population synthesis methods show two main limitations. The first one lies in the fact that population synthesis inputs, i.e., control variables' marginals and microdata sample, availability at the building or the dwelling unit resolution is generally extremely limited, mainly for privacy reasons. Marginals lacking at the building or dwelling unit resolution are considered at the resolution where they are available using multiresolution algorithms. However, the microdata sample unavailability remains problematic. In fact, microdata samples available from national statistical agencies are usually disseminated at a coarse scale and are thus not representative of smaller areas (Yaméogo et al., 2023; Nejad et al., 2021). For instance, the Canadian Public Use Microdata File (PUMF) is only available at the census metropolitan area scale. Using it for population synthesis for a smaller area does not warrant the generation of a synthetic population that is representative of the smaller area's real population. This representativeness problem gets worse when the smaller area considered is as fine as a building or a dwelling unit. Explicitly spatialized microdata samples are sometimes collected as part of National Travel Surveys such as the Origin-Destination Surveys (ODS) cyclically carried out in Quebec, Canada. Microdata samples for small areas can thus be extracted from these surveys'

samples. However, this does not make population synthesis at disaggregate spatial resolutions necessarily possible as 1) there is no guarantee of records' availability in all the small areas considered, e.g., many buildings will not include a surveyed household, and 2) the small areas' microdata samples, when available, will include an extremely limited number of sociodemographic attributes' combinations raising thus both representativeness and convergence issues (Guo and Bhat, 2007).

The second challenge encountered with IA-based explicitly spatialized population synthesis is the difficulty to store and handle huge matrices obtained when the numerous marginals at the building or the dwelling unit resolution are considered. This usually requires virtual machines or particular data structures that would compromise the synthesis process efficiency (Khachman et al., 2022). SA and IA limitations are summarized in Table 6.1. Choosing the best approach to perform explicitly spatialized population synthesis, depends on rich spatialized data availability in the study area. Based on their specifications, the SA should be preferred to the IA when rich spatialized datasets allowing to perform accurate full correspondence spatialization are available at the building/dwelling unit scale. That allows to optimize both the representativeness and the spatial precision of the synthetic population generated. The IA-based methods showing minimal data requirements remain useful when such datasets are partially or totally unavailable at the building/dwelling unit resolution.

Table 6.1 SA and IA limitations

Problem	SA	IA
Inconsistencies between the synthetic population and the built environment	X	
Transferability issues	X	
Representativeness damaged due to limited/inexistent synthesis inputs		X
Computational issues		X

6.2.3 Machine learning algorithms

Two machine learning algorithms are used in this paper: Random Forest (Breiman, 1996) and Partitioning Around Medoids (PAM) (Kaufman & Rousseeuw, 1990). This section includes an overview of these two algorithms.

6.2.3.1 Random Forest

Random Forest (Breiman, 1996) is a supervised learning algorithm that can be used for both regression and classification problems. It is an ensemble learning method that makes predictions based on multiple decision trees. Decision trees' outcomes are averaged in regression problems, and their most predicted class is selected in classification problems. The Random Forest models draw their strength from relying on a forest of relatively uncorrelated decision trees (Yiu, 2019).

In fact, a decision tree is highly sensitive to the data it is trained on, and the features used to split its nodes (Yiu, 2019). Hence, a Random Forest model uses bootstrap aggregation (Breiman, 1996) and feature randomness (Breiman, 2001) to minimize correlation among the decision trees it relies on. Bootstrap aggregation consists of each decision tree in the forest being generated using a random selection of N observations that are drawn with replacement from the original N -size training set. Feature randomness consists of generating each decision tree in the forest considering a random subset of the features initially considered. Varying both training sets and features makes the decision trees generated relatively uncorrelated hence the robustness of Random Forest models' predictions (Yiu, 2019).

6.2.3.2 Clustering

K-medoids clustering is an unsupervised learning algorithm that is used to divide a dataset into k clusters each represented by an object in the cluster called "medoid" (Kassambara, 2017). PAM (Kaufman & Rousseeuw, 1990) is the most common k -medoids algorithm. It operates in two steps: build and swap (Kassambara, 2017). In the build step, k initial medoids are randomly chosen. A dissimilarity matrix is then calculated, and each object is assigned to its closest medoid. In the swap step, non-medoid objects within each cluster are iteratively considered as medoids and a dissimilarity indicator is computed. The object yielding the minimal dissimilarity within the cluster is elected as its new medoid. If at least one cluster's medoid is updated at the end of the swap step, each object is reassigned to its nearest medoid, and the swap step is run again. Otherwise, convergence is reached.

A main advantage of k -medoids clustering over the renowned k -means clustering is that the former is less sensitive to outliers (Kassambara, 2017). That is because when k -means is used, a cluster is represented by a mean point (centroid) that is not necessarily an object in the cluster, and the mean is sensitive to outliers. However, PAM is inefficient on large datasets. To reduce computation time and memory storage required when clustering large datasets, an extension of PAM known as

Clustering Large Applications (CLARA) algorithm (Kaufman and Rousseeuw, 1990) can be used instead. CLARA runs PAM on multiple random samples of the same size that are extracted from the original large dataset. A dissimilarity indicator is computed for the whole dataset considering each set of medoids and the set yielding the minimal dissimilarity is selected. The objects of the original datasets are finally assigned to the closest medoid of the selected set (Kassambara, 2017).

6.2.4 Contributions

This paper is a continuation of the research work published in Khachman et al. (2022). The latter proposed an IA-based framework to generate explicitly spatialized synthetic populations that is, among others, particularly useful when only few marginals are available at the building/dwelling unit resolution. This paper is meant to introduce a new SA-based method that leverages rich spatialized datasets available to achieve a realistic explicit spatialization of synthetic households through a well-fitted full correspondence distribution without damaging the synthetic population's representativeness. A synthetic population of better quality is thus obtained.

The new method shows four main novelties compared to its predecessors. First, it leverages performant unsupervised (CLARA) and supervised (Random Forest) learning algorithms in a clustering + classification framework to accurately predict the location of synthetic households at the building scale directly. Second, it leverages the rich spatial information usually collected as part of surveys to perform accurate predictions of synthetic households' locations. Third, CLARA's ability to handle large datasets adds efficiency and larger transferability to the proposed framework accuracy. In fact, the more spatially precise a dataset is, the larger its size. Hence, when applied to sizeable study areas, CLARA allows to perform efficient clustering of buildings/dwelling units and/or households where other clustering algorithms such as k-means and PAM may fail. Finally, CLARA's low sensitivity to outliers guarantees a realistic clustering thus further improving the realism of the explicit spatialization performed.

6.3 Methodology

The six-step framework shown in Figure 6.2 is meant to generate explicitly spatialized synthetic populations. This framework takes advantage of powerful machine learning algorithms and exploits the spatial information generally collected as part of surveys to achieve a realistic spatialization of synthetic populations at the building resolution. In Figure 6.2, cylinders symbolize

datasets and rectangles refer to processing steps. The proposed framework's steps are detailed in this section.

6.3.1 Step 1: population synthesis

The first step is to generate a zonal synthetic population. The proposed framework requires the microdata sample to be explicitly spatialized. To guarantee optimal representativeness, only observations within a geographic area are used to generate its synthetic population. Moreover, the synthesis is required to be run using a Synthetic Reconstruction algorithm so that synthetic populations made of clones of surveyed agents are generated. Households and people marginals can be extracted at any geographic resolution(s). However, the more disaggregate the synthesis reference resolution, the more realistic and spatially precise the synthetic population (Khachman et al., 2021b).

6.3.2 Step 2: spatial data processing

In this step, spatial datasets (e.g., property assessment rolls, addresses' coordinates, etc.) are processed using appropriate data fusion techniques to generate a dataset of buildings and their characteristics (BLD).

6.3.3 Step 3: nearest neighbor join

Each household in the explicitly spatialized microdata sample has (X;Y) coordinates. These coordinates often slightly differ from corresponding buildings' coordinates whether for privacy reasons, or simply due to the independence of data collection processes. In this step, each household in the microdata sample is considered to occupy a dwelling unit in the closest building to it. A joint household-building (HH-BLD) dataset is thus obtained.

The HH-BLD dataset is used in Step 4 and Step 5 to compute $P_{\text{distance}}(\text{HH} \cap \text{BLD})$ and $P_{\text{cluster}}(\text{HH} \cap \text{BLD})$ which are considered to be the two components of the global probability of each surveyed household's synthetic clones to live in each building within their synthesis geographic area. The global probability is calculated at the end of Step 5 and the spatial distribution is then performed in Step 6 based on the set of global probabilities computed for each synthetic household.

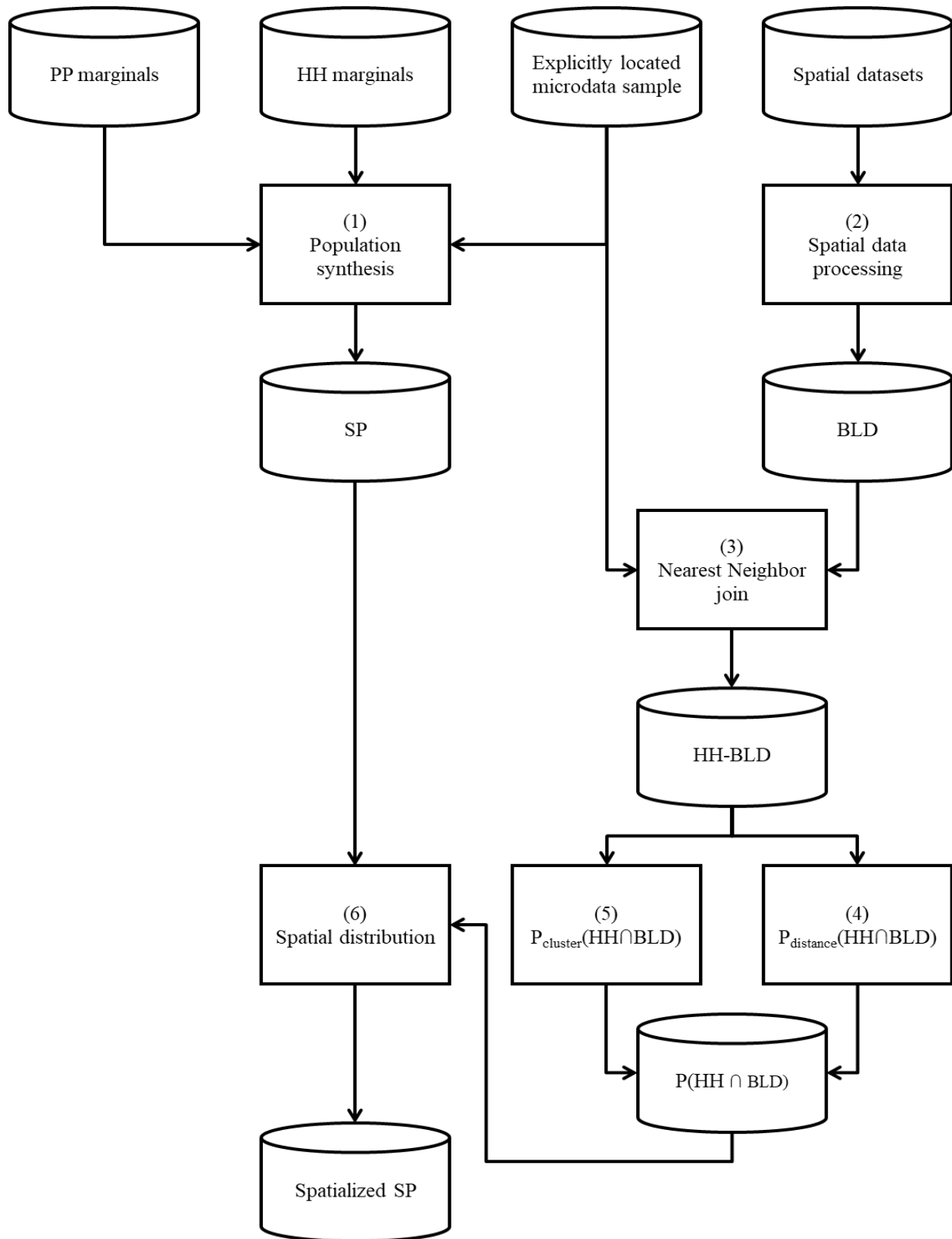


Figure 6.2 Proposed explicitly spatialized population synthesis framework

6.3.4 Step 4: $P_{\text{distance}}(HH \cap BLD)$

This step is dedicated to calculate the probability $P_{\text{distance}}(HH \cap BLD)$ of each surveyed household's synthetic clones to live in each building within their synthesis geographic area, considering the household-building distance only. The microdata sample spatial information, i.e., surveyed households' locations, is used in this step. Various probability distributions including different types of kernels can be used. This step is optional and can be skipped if the observed distance between a surveyed household and a building within its synthesis geographic area is assumed not to influence the probability of its synthetic clones to live in that building.

6.3.5 Step 5: $P_{\text{cluster}}(HH \cap BLD)$

The objective of this step is to calculate the probability $P_{\text{cluster}}(HH \cap BLD)$ of each surveyed household's synthetic clones to live in each building within their synthesis geographic area, considering the cluster to which the building belongs. The microdata sample spatial information, i.e., surveyed households' locations, is also used in this step. The buildings located within the study area are first clustered. Each surveyed household is thus assigned the cluster of the building where it lives. A classification algorithm is then calibrated to calculate the probabilities for each surveyed household's synthetic clones to be assigned to each building cluster previously defined according to the household's characteristics.

The global probability of each surveyed household's synthetic clones to live in each building is finally calculated as per Eq. 1:

$$P(HH \cap BLD) = P_{\text{distance}}(HH \cap BLD) \times P_{\text{cluster}}(HH \cap BLD) \quad (1)$$

6.3.6 Step 6: spatial distribution

The spatial distribution is performed within each synthesis geographic area independently. Synthetic households are distributed on buildings within their synthesis areas in descending order of $P(HH \cap BLD)$.

6.3.7 Assessment

The proposed framework is assessed through the evaluation of the quality of each of its components. Various indicators are used to assess the quality of 1) the synthetic population

generated in Step 1, 2) the clustering performed in Step 5, 3) the classification performed in Step 5, and 4) the spatial distribution performed in Step 6. The synthetic population generated is assessed using the Average Relative Error (ARE) calculated at the synthesis reference resolution as per Eq. 2:

$$ARE(\%) = \frac{\frac{\sum_{i=1}^n |\hat{b}_i - b_i|}{b_i} \times 100}{n} \quad (2)$$

where $\hat{b}_i$ is the i^{th} estimated value, b_i is the i^{th} observed value, and n is the number of values compared. Estimated and observed values compared in this case are synthetic and observed totals for each control variable's category. Synthetic totals and marginal totals are also graphically compared. The quality of the building clustering performed in Step 5 was assessed using the Average Silhouette Width (ASW) metric (Rousseeuw, 1987) computed as per Eq. 3:

$$ASW = \frac{\sum_{i=1}^n s(i)}{n} \quad (3)$$

where n is the total number of data points in the clustered database and $s(i)$ is the silhouette of each data point i . ASW takes values between -1 and 1 where -1 reflects the worst clustering performance and 1 reflects the best clustering performance. The multi-class classification performed in Step 5 is evaluated using the accuracy, i.e., the proportion of correct predictions performed, and the macro-average F1 score (MacroF1) computed as per Eq. 4:

$$MacroF1 = \frac{\sum_{c=1}^C F1_c}{C} \quad (4)$$

where $F1_c$ is the F1-score for the class c and C is the total number of classes. MacroF1 is a practical indicator of multi-class classification models as it summarizes a model's performance in a single number, and it is not sensitive to classes' sizes since each $F1_c$ is given the same weight in the average calculation. The spatial distribution performed in Step 6 cannot be assessed by comparison to a ground truth as the real explicitly spatialized synthetic population is not disseminated for privacy reasons. In fact, if the explicitly spatialized real population was available, there would have been no need for explicitly spatialized population synthesis. Hence, to assess the spatial distribution, we first verify if any artificial post-processing such as assigning a synthetic household

to a building out of its synthesis area was required. We also assess the extent to which the final spatial distribution diverged from the constraint-free most probable one. The percentage of synthetic households that were finally assigned to their most probable building cluster is thus calculated.

6.4 Case study

The population synthesis framework proposed in this paper is applied to generate an explicitly spatialized synthetic population for Montreal Island in Canada, that includes the city of Montreal and other municipalities. This section details the application of each step of the framework described hereabove to the study area.

6.4.1 Step 1: population synthesis

The microdata sample for Montreal Island was extracted from the ODS 2018 sample. The ODS sample contains the explicit location of each surveyed household as well as many variables related to surveyed households, including the socio-demographic of each member and trips made by each individual. The area covered by the ODS (Figure 6.3 - white) is larger than Montreal Island (Figure 6.3 - yellow). Montreal Island is classed by Statistics Canada as a Census Division (CD). It is comprised of 41 Municipal Boroughs (MBs) (Figure 6.4a) and 3202 Dissemination Areas (DAs) (Figure 6.4b).

The MB is the weighting area of the ODS. Marginal totals are thus extracted at the MB resolution by summing households' weights in each MB for each attribute category. Since ODS households are explicitly spatialized, some variables disseminated by Statistics Canada (Statistics Canada, 2019b) at the DA resolution were added to each household using spatial joins. This helped further characterizing each household making it even more unique through linking it to its neighbourhood. Household and people variables extracted from the ODS 2018 sample and the 2016 Canadian Census of Population, as well as their corresponding categories, and control status, are listed in Table 6.2. The reweighting optimization-based method introduced by Fournier et al. (2021) and extended by Khachman et al. (2022) to multilevel and multiresolution population synthesis is used to generate a synthetic population for each MB of Montreal Island. The synthesis problem takes the $AX=B$ format where A is seeded using the microdata sample, B contains the marginal totals and X contains the households' optimized weights. The optimization is carried out using the

efficient and performant Cyclic Coordinate Descent algorithm (Fournier et al., 2021) and the decimal weights obtained are converted to integers using the Truncate, Replicate, Sample (TRS) method (Lovelace & Ballas, 2013). The TRS being a probabilistic algorithm, a thousand runs are performed and the one with minimal errors is retained (Khachman et al., 2022).

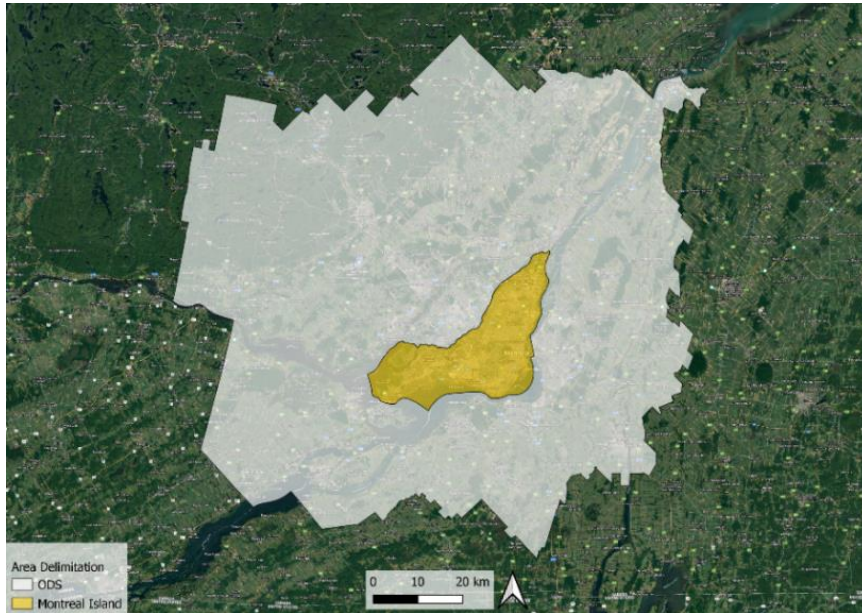


Figure 6.3 ODS area and Montreal Island area delimitations

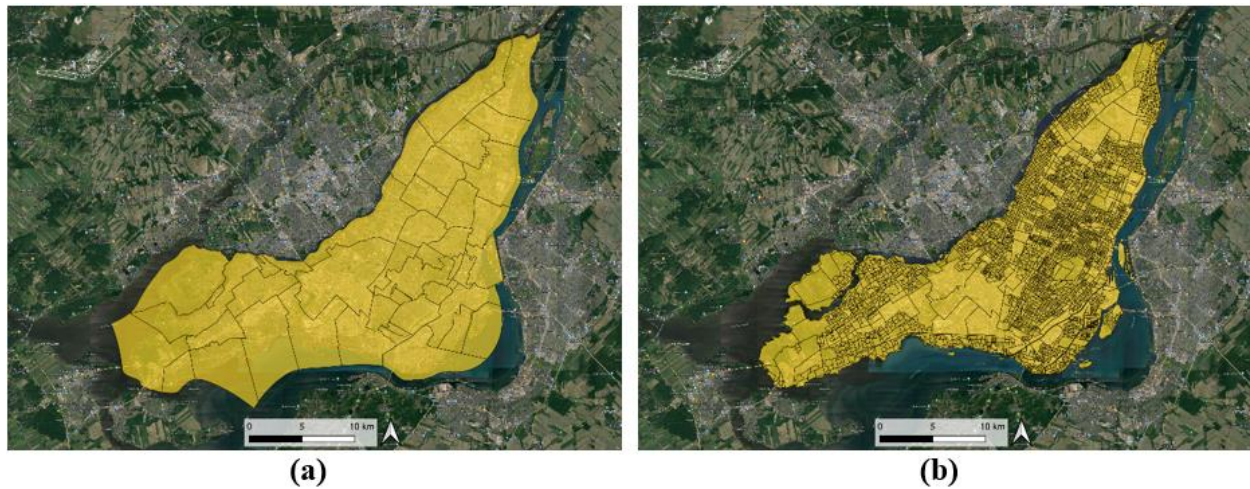


Figure 6.4 Montreal Island (a) MBs and (b) DAs

Table 6.2 Household and people variables

Variable	Definition	Categories	Description	Control
hh_id	Household unique identifier	-	-	-

Table 6.2 Household and people variables (suite)

Variable	Definition	Categories	Description	Control
hh_homelat	Household home latitude	-	-	-
hh_homelon	Household home longitude	-	-	-
hh_weight	Household weight	-	-	-
hh_count	Household count	1	1 household	✓
hh_size	Household size	1	1 person	✓
		2	2 people	
		3	3 people	
		4	4 people	
		5	5+ people	
hh_income	Household income	1	Less than 30k\$	✓
		2	[30k\$; 60k\$[	
		3	[60k\$; 90k\$[	
		4	[90k\$; 120k\$[	
		5	[120k\$; 150k\$[	
		6	[150k\$; 180k\$[	
		7	[180k\$; 210k\$[	
		8	210k\$ and more	
hh_motorization	Household motorization	1	1 car	✓
		2	2 cars	
		3	3 cars	
		4	4 cars	
		5	5+ cars	
hh_da_id	Household's DA unique identifier	-	-	-
hh_da_medinc	Median income in household's DA	-	-	-
hh_da_ppdens	Population density in household's DA	-	-	-
pp_id	People unique identifier	-	-	-
pp_count	People count	1	1 person	✓
pp_age	People age group	1	[0;5[	✓
		2	[5;10[	
		3	[10;15[	
		4	[15;20[	
		5	[20;25[	
		6	[25;30[	
		7	[30;35[	
		8	[35;40[	
		9	[40;45[	
		10	[45;50[	
		11	[50;55[	
		12	[55;60[	

Table 6.2 Household and people variables (suite)

Variable	Definition	Categories	Description	Control
		13	[60;65[	
		14	[65;70[	
		15	[70;75[	
		16	75+	
pp_sex	People sex	1	Men	✓
		2	Women	
pp_occupation	People main occupation	1	Full-time worker	✓
		2	Part-time worker	
		3	Student	
		4	Retired	
		5	Other	
pp_drlicense	People driving license	0	Without driving license	✓
		1	With driving license	
pp_transpass	People transit pass	0	Without transit pass	✓
		1	With transit pass	
pp_carshsub	People carsharing service subscription	0	Without carsharing subscription	✓
		1	With carsharing subscription	
pp_bikeshsub	People bikesharing service subscription	0	Without bikesharing subscription	✓
		1	With carsharing subscription	

6.4.2 Step 2: spatial data processing

We used the method described in Khachman et al. (2022) to generate the Montreal Island's built environment based on the addresses (Ville de Montréal, 2023a) and property assessment units (Ville de Montréal, 2023c; Ministry of Municipal Affairs and Housing of Quebec, 2023) publicly accessible. Some variables disseminated by Statistics Canada (Statistics Canada, 2019b) at the DA resolution were also added to each building using spatial joins. This helps further characterizing spatial locations making each building even more unique through linking it to its neighbourhood. The building attributes retained are detailed in Table 6.3.

Table 6.3 Building attributes

Variable	Definition	Categories	Description
bld_id	Building unique identifier	-	-
bld_lat	Building latitude	-	-
bld_lon	Building longitude	-	-

Table 6.3 Building attributes (suite)

Variable	Definition	Categories	Description
bld_size	Number of dwelling units in the building	-	-
bld_area	Building area	-	-
bld_value	Building value	-	-
bld_landvalue	Building land value	-	-
bld_da_id	Building's DA unique identifier	-	-
bld_da_dumedvalue	Median dwelling unit value in buildings' DA	-	-

6.4.3 Step 3: nearest neighbor join

The ODS 2018 sample coordinates slightly differ from the buildings' coordinates. Hence, a household-building dataset is generated through joining each ODS 2018 surveyed household to its nearest building. The dataset obtained includes ODS and building variables listed in Table 6.2 and Table 6.3. It is worth mentioning that at the beginning, many other ODS and building variables were retained. However, only variables that were found to be useful following an extensive sensitivity analysis were mentioned in the paper for the sake of conciseness.

6.4.4 Step 4: $P_{\text{distance}}(\text{HH} \cap \text{BLD})$

Based on knowledge of the local context, it was assumed that within each MB, the further a building is from a surveyed household, the lower the probability a synthetic clone of that surveyed household is living in it. A distance matrix summarizing the distances between each surveyed household and all its synthesis geographic area's buildings is first computed. Various distance measures (e.g., direct distance, distance on the road network, etc.) can be used. Then, a function determining the probability of each surveyed household to live in each building according to the distance separating them is defined. This function can be based on various formulations, e.g., binary, inverse, exponential, logarithmic, etc. and can include kernels of both types: fixed, e.g., $P_{\text{distance}}(\text{HH} \cap \text{BLD})=1$ within a pre-determined radius R , and $P_{\text{distance}}(\text{HH} \cap \text{BLD})=0$ beyond R , and adaptive, e.g., $P_{\text{distance}}(\text{HH} \cap \text{BLD})=1$ for a pre-determined number of closest neighbors N , and $P_{\text{distance}}(\text{HH} \cap \text{BLD})=0$ beyond N . We used an exponential formulation with a fixed kernel based on direct distance to estimate, in each MB, the probability $P_{\text{distance}}(\text{HH}_i \cap \text{BLD}_j)$ of the i^{th} surveyed household's clones in the synthetic population to live in the j^{th} building as per Eq. 5:

$$P_{distance}(HH_i \cap BLD_j) = \frac{e^{d_{max}}}{e^{d_{max}} - 1} \times \frac{1}{e^d} + \frac{1}{1 - e^{d_{max}}} \quad (5)$$

where d_{max} is the distance between HH_i and the furthest building to it within its MB, and d_{ij} is the distance between HH_i and BLD_j . The probabilities calculated for each synthetic household are finally normalized so that the sum of the probabilities of each synthetic household to live in each building of its MB is equal to 1. The normalization is performed as per Eq. 6:

$$P_{distance}(HH_i \cap BLD_j) = \frac{P_{distance}(HH_i \cap BLD_j)}{\sum_{j=1}^{j=B} P_{distance}(HH_i \cap BLD_j)} \quad (6)$$

6.4.5 Step 5: $P_{cluster}(HH \cap BLD)$

This step comprises two components: clustering and classification. Buildings are first assigned to different clusters and the probability of each surveyed household's synthetic clones to live in each building cluster is then calculated using a classification model.

6.4.5.1 Clustering

The CLARA algorithm is run to separate the buildings into a pre-determined number of clusters considering relevant building attributes, i.e., building attributes deemed to influence a household's residential choice. An extensive sensitivity analysis was carried out to determine the optimal: 1) combination of building attributes, 2) number of clusters (nbclust), 3) distance metric (dmetric), and 4) sample size (sampsize) to be used. Based on ASW comparisons, four variables were found to be decisive in the clustering process: *bld_size*, *bld_area.pu*, *bld_totvalue.pa*, and *bld_da_dumedvalue*. The variables *bld_size* and *bld_da_dumedvalue* are defined in Table 6.3. *bld_area.pu* refers to the average area per dwelling unit in the building and is calculated as per Eq. 7:

$$bld_area.pu = \frac{bld_area}{bld_size} \quad (7)$$

bld_totvalue.pa refers to the value of the area unit in the building and is calculated as per Eq. 8:

$$bld_totvalue.pa = \frac{bld_value + bld_landvalue}{bld_area} \quad (8)$$

To mitigate the eventual impact of random sampling, each nbclust*dmetric*sampsize combination was run 10 times (i.e., using 10 different seeds in R programming language). The values tried for each parameter are summarized in Table 6.4. In total 1800 parameter combinations were tested which was possible thanks to CLARA efficient implementation in R package “cluster” (Maechler et al., 2022) and parallel computing.

Table 6.4 CLARA tuning parameters

Parameter	Number of values	Values
seed	10	100; 200; 300; 400; 500; 600; 700; 800; 900; 1000
nbclust	9	2; 3; 4; 5; 6; 7; 8; 9; 10
dmetric	2	Euclidian; Manhattan
sampsize	10	100; 200; 300; 400; 500; 600; 700; 800; 900; 1000

The average ASW over the 10 runs/seeds of each nbclust*dmetric*sampsize combination was used to assess each combination. Usually, the nbclust showing the maximum ASW is used for clustering (Rousseeuw, 1987). However, this rule is not blindly applicable to explicitly spatialized population synthesis practice. In fact, the more clusters there are the more precise the explicit spatialization would become. Extreme cases may help grasp the previous statement. In fact, if all buildings belong to a single cluster, the distribution of households over buildings is random with regard to the building’s cluster, while if the number of clusters is equal to the number of buildings, each household is specifically assigned to a unique building. However, increasing the number of clusters should be done cautiously as it might damage the ASW (i.e., the clustering quality) as well as the predictive quality of the discrete choice model later applied to calculate $P_{\text{cluster}}(\text{HH} \cap \text{BLD})$.

Thus, when the nbclust maximizing the ASW is judged to be low, we suggest using the nbclust minimizing the loss of ASW per additional cluster (ASWL) instead of the nbclust maximizing ASW. If the highest ASW is achieved with nbclust=j, the ASWL for nbclust=i $\forall i > j$ is calculated using Eq. 9:

$$ASWL_i = \frac{ASW_j - ASW_i}{i - j} \quad \forall i > j \quad (9)$$

where j is the number of clusters achieving the maximum ASW, i is a number of clusters that is greater than j, ASW_j is the ASW associated to j (i.e., the maximum ASW) and ASW_i is the ASW associated to i.

6.4.5.2 Classification

The classification was performed using a Random Forest model. To assess the quality of the Random Forest model used to calculate $P_{\text{cluster}}(\text{HH} \cap \text{BLD})$, the joint HH-BLD database obtained in Step 3 is separated into three subsets: training dataset (60%), cross-validation dataset (20%), and test dataset (20%). The model is first trained on the training dataset. Its performance on the cross-validation dataset is then assessed for parameters tuning purposes. Training and cross-validation are repeated using various parameters until an optimal model is found. The final model obtained is then applied on the test dataset to evaluate how it would perform on new datasets and detect any potential overfitting problems.

An extensive sensitivity analysis was carried out using training and cross-validation datasets to determine the optimal: 1) combination of predictors, 2) number of randomly sampled candidate variables at each split (mtry), 3) minimum size of terminal nodes (nodesize), and 4) number of trees (ntree) to be used. At each iteration, accuracy and macroF1 were calculated for both training (acc_{train}, macroF1_{train}) and cross-validation (acc_{cv}, macroF1_{cv}) datasets to evaluate bias (i.e., underfitting) and variance (i.e., overfitting) problems. Based on accuracy and macroF1 comparisons, the following set of predictors was retained: hh_size, hh_incmed, hh_meanage, hh_da_hhmedinc, and hh_da_ppdens. hh_size, hh_da_medinc and hh_da_ppdens are defined in Table 6.2. hh_incmed is derived from hh_income defined in Table 6.2 through replacement of each hh_income value by the center of its corresponding range. hh_meanage refers to the mean age of people in each household.

To mitigate an eventual impact of random sampling, each mtry*nodesize*ntree combination was run 10 times (i.e., using 10 different seeds in R programming language). The values tried for each parameter are summarized in Table 6.5. In total 10,000 parameter combinations were tested which was possible thanks to Random Forest efficient implementation in R package “randomForest” (Liaw & Wiener, 2002) version 4.7-1.1 (2022), and parallel computing.

Table 6.5 Random Forest tuning parameters

Parameter	Number of values	Values
seed	10	100; 200; 300; 400; 500; 600; 700; 800; 900; 1000
mtry	10	1; 2; 3; 4; 5; 6; 7; 8; 9; 10
nodesize	10	5; 10; 15; 20; 25; 30; 35; 40; 45; 50
ntree	10	100; 200; 300; 400; 500; 600; 700; 800; 900; 1000

The average acc_{cv} , and $macroF1_{cv}$ over the 10 runs/seeds of each $mtry \times nodesize \times ntree$ combination were used to assess each combination. The combinations showing the maximum $macroF1_{cv}$ are first selected. Out of them, only the combinations with the maximum acc_{cv} are retained. Finally, the remaining combination with minimum $mtry$, $nodesize$, and $ntree$ was selected. The selected model was used to calculate the probability $P_{cluster}(HH \cap BLD)$ of each surveyed household's synthetic clones to live in each building cluster of their MB.

Finally, the $P(HH \cap BLD)$ global probability of each synthetic household to live in each building of its MB was calculated by multiplying $P_{cluster}(HH \cap BLD)$ and the normalized $P_{distance}(HH \cap BLD)$ obtained at Step 4.

6.4.6 Step 6: spatial distribution

The spatial distribution is performed within each synthesis geographic area independently. Each synthesis geographic area comprises H unique synthetic households and B buildings. We refer to unique synthetic households with the index i where $i=1$ to H , and to each building with the index j where $j=1$ to B . A database listing all the unique $H \times B$ synthetic household-building pairs is first established. Each pair constituted of the i^{th} unique synthetic household (HH_i) and the j^{th} unique building (BLD_j) is assigned $P(HH_i \cap BLD_j)$, the probability of the unique synthetic household HH_i to live into the building BLD_j calculated at the end of step 5. Moreover, the HH_i 's weight (W_{HH_i}), i.e., the number of its clones in the synthetic population generated in step 1, and the number of flats in BLD_j (F_{BLD_j}) are assigned to each HH_i - BLD_j pair in the initial database.

A four-step iterative process is then run. First, the HH_i - BLD_j with the highest $P(HH_i \cap BLD_j)$ is considered. If multiple HH_i - BLD_j pairs have the highest $P(HH_i \cap BLD_j)$, a single pair among them is randomly selected. Second, the minimum of W_{HH_i} and F_{BLD_j} (N) is calculated and HH_i - BLD_j is repeated N times in the spatialized synthetic population. Third, the initial database is updated by subtracting N from both W_{HH_i} and F_{BLD_j} for all pairs including HH_i , BLD_j or both. Fourth, HH_i - BLD_j pairs with $W_{HH_i} = 0$ and/or $F_{BLD_j} = 0$ are deleted. The same process is then repeated until no HH_i - BLD_j remains in the database. Finally, the characteristics of each HH_i and BLD_j are transferred from the microdata sample and the buildings' database to their corresponding pairs in the spatialized synthetic population established.

6.5 Results

The results of population synthesis zonal generation, buildings' clustering, households' classification, and households' explicit spatialization processed are detailed in this section. The performance of the explicitly spatialized population synthesis method introduced in this paper is also assessed through appropriate indicators measuring the framework components' quality.

6.5.1 Population synthesis

The synthetic population generated in Step 1 consists of a database listing synthetic households and people with their corresponding sociodemographic attributes. The only spatial information available for each synthetic household at this point is the MB to which it belongs. The synthetic population generated shows a low ARE of 2.85%, and the synthetic totals seem to correspond to the observed totals nearly perfectly as shown in Figure 6.5. As mentioned previously, any other fitting-based synthesizer can be used at this step.

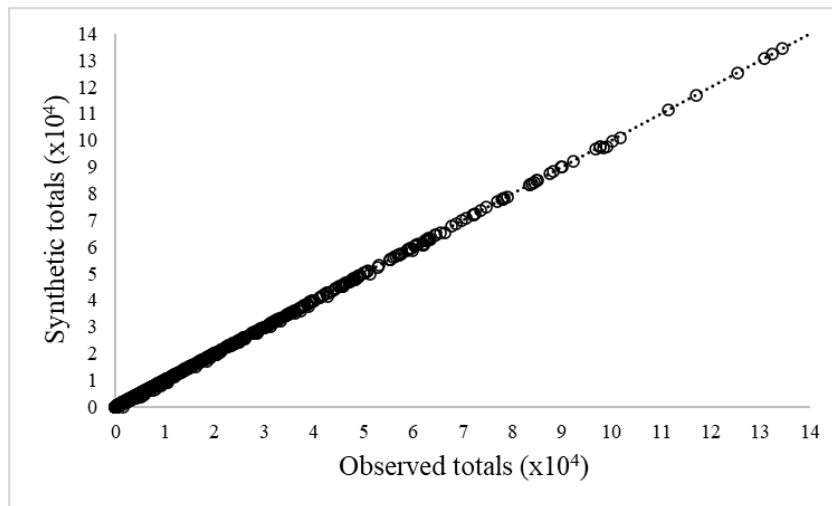


Figure 6.5 Comparison of synthetic and observed marginal totals

6.5.2 Clustering

The maximum ASW is obtained with nbclust=2 as shown in Figure 6.6. However, the precision of the explicit spatialization is not judged to be sufficient when buildings are separated into two clusters only. Hence, the ASWL indicator is calculated for 3 to 10 clusters. The ASWL obtained for each nbclust are shown in Figure 6.7. The optimal number of clusters thus found is 8 as it shows

the lowest ASWL. The dmetric*sampsize combination maximizing ASW for nbclust=8 is finally used to achieve optimal clustering of the buildings' dataset. The clustering parameters and quality measure are listed in Table 6.6. Clustering results are summarized in Table 6.7.

Table 6.6 Retained clustering model characteristics

seed	dmetric	samples	ASW
100	euclidean	300	0.48

Table 6.7 Buildings' clustering results

Building type	%	Mean bld_size (dwellings)	Mean bld_area.pu (sq.m.)	Mean bld_totvalue.pa (k\$/sq.m.)	Mean bld_da_dumedvalue (k\$)
T1	22.56	2.54	106.67	2.75	270.86
T2	28.40	2.02	122.07	2.97	378.80
T3	10.02	28.58	76.14	2.46	338.81
T4	3.10	1.64	223.82	8.17	860.58
T5	9.19	1.94	213.46	3.34	596.41
T6	7.85	156.80	88.89	7.06	355.82
T7	17.51	2.13	125.26	3.21	475.16
T8	1.36	1.38	287.25	6.90	1,444.62

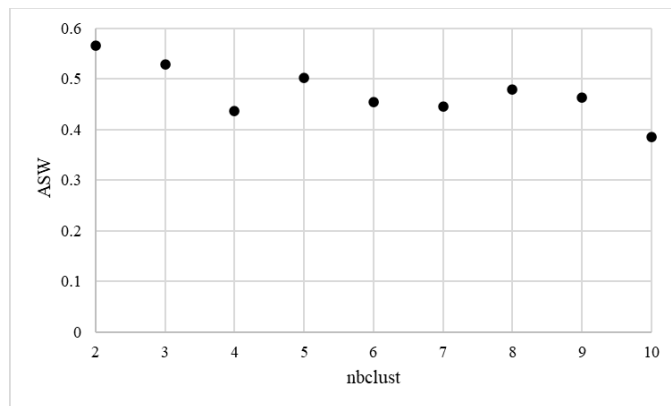


Figure 6.6 ASW per nbclust

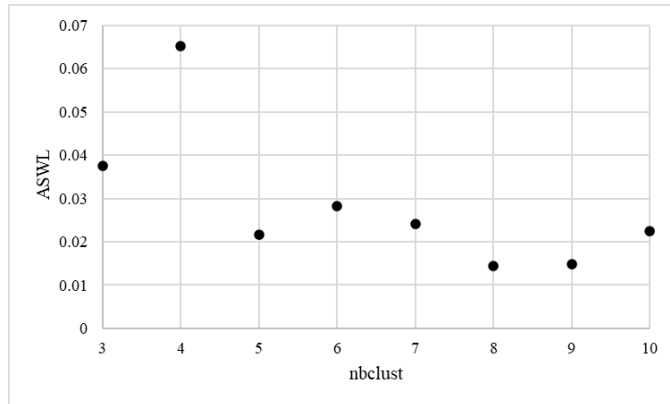


Figure 6.7 ASWL per nbclust

6.5.3 Classification

The parameters of the retained classification model and its quality measures are listed in Table 6.8. High accuracies and macroF1 scores are achieved on all datasets. The selected model is thus robust with regard to potential bias and variance issues and reliable $P_{\text{cluster}}(\text{HH} \cap \text{BLD})$ are obtained. The confusion matrix of the test dataset is shown in Figure 6.8. It confirms the reliability of the retained model's predictions.

Table 6.8 Retained Random Forest model characteristics

seed	mtry	nodesize	ntree	acc _{train}	acc _{cv}	acc _{test}	MacroF1 _{train}	MacroF1 _{cv}	MacroF1 _{test}
300	5	5	900	0.95	0.82	0.82	0.94	0.79	0.79

	T1	T2	T3	T4	T5	T6	T7	T8
T1	88.6%	1.9%	4.8%	0.0%	0.6%	2.8%	1.1%	0.1%
T2	2.4%	91.6%	3.2%	0.0%	0.2%	1.6%	1.0%	0.0%
T3	22.4%	20.6%	33.8%	0.8%	2.9%	11.1%	8.1%	0.3%
T4	2.9%	2.4%	1.5%	89.3%	1.0%	1.9%	0.5%	0.5%
T5	2.3%	3.9%	2.6%	0.2%	83.4%	2.6%	4.9%	0.0%
T6	10.9%	10.0%	9.8%	1.5%	2.7%	60.7%	3.3%	1.2%
T7	2.0%	3.1%	2.1%	0.0%	0.5%	1.4%	91.0%	0.0%
T8	3.3%	1.1%	1.1%	0.0%	0.0%	0.0%	0.0%	94.4%

Figure 6.8 Confusion matrix of the retained model applied to the test dataset

6.5.4 Spatial Distribution

A database listing synthetic households and people with their coordinates and sociodemographic attributes is obtained. Table 6.9 shows example records from the explicitly spatialized synthetic population produced. The mapping of the Montreal Island explicitly spatialized synthetic population is shown in Figure 6.9.

Table 6.9 Example records from the spatialized synthetic population

hh_id	hh_homelat	hh_homelon	hh_bld	hh_mb	hh_size	hh_income	hh_motorization
1	45.50436	-73.56346	128795	101	1	3	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1
10000	45.51601	-73.565	147949	101	5	4	1

pp_id	pp_age	pp_sex	pp_occupation	pp_drlicense	pp_transpass	pp_carshsub	pp_bikeshsub
11	5	1	3	1	1	0	0
100001	8	1	5	1	1	0	1
100002	6	2	1	1	0	0	0
100003	9	1	2	1	1	0	0
100004	3	1	3	0	1	0	0
100005	1	1	5	0	1	0	0
100006	1	1	5	0	1	0	0

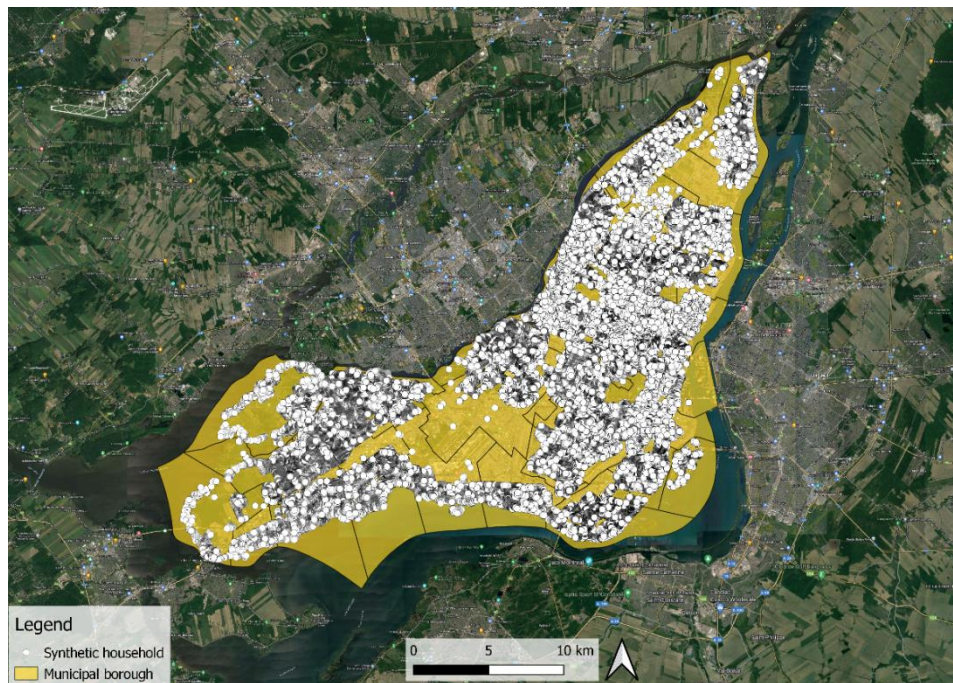


Figure 6.9 Montreal Island explicitly spatialized synthetic population

The algorithm managed to locate all synthetic households within their corresponding MB and did not require any artificial post-processing or constraint relaxation. Moreover, 78.9% of the synthetic households were assigned to their most probable building cluster and 11.3% of the synthetic households were assigned to their second most probable building cluster as shown in Figure 6.10. The spatial distribution was hence found not to significantly diverge from the most probable constraint-free distribution.

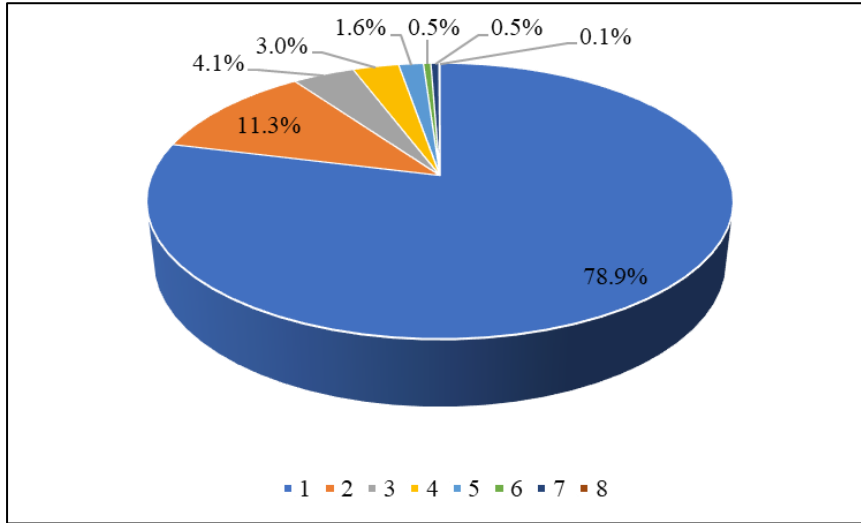


Figure 6.10 Distribution of synthetic households per building cluster probability rank

6.6 Conclusion and outlook

To run reliable simulations of the population's mobility behaviors, explicitly spatialized synthetic populations are required. That is because both, the sociodemographic attributes and the spatial location of a household have an important impact on the mobility behaviors of the individuals who compose it. However, the effort devoted to improving synthetic populations' spatial distributions remains limited compared to the research works that focused on improving their sociodemographic representativeness.

This paper contributes to improving synthetic populations' spatial distributions through a new explicitly spatialized population synthesis framework allowing to assign accurate (X;Y) coordinates to each synthetic household. The proposed method uses performant and efficient machine learning algorithms to exploit the spatial information usually collected as part of surveys achieving thus accurate predictions of synthetic households' explicit locations. It is based on the

SA where the population is first generated at an aggregate resolution then distributed on buildings/dwelling units instead of an IA where the synthetic population is directly generated at the building resolution. The new method clusters the study area's buildings and predicts the probability of each synthetic household to live in each building according to both the building's cluster and the distance between its observed clone in the microdata sample and the building. Synthetic households are then distributed on buildings in descending order of association probability within buildings' capacities limits.

The proposed framework shows various advantages compared to existing explicitly spatialized population synthesis frameworks. First, being based on the SA, the sociodemographic representativeness of the synthetic population and the generation efficiency are not compromised as they are when the IA is used due to the lack of synthesis inputs and the multiplicity of constraints at disaggregate resolutions. Second, it is, to the best of authors' knowledge, the first clustering + classification framework that leverages performant unsupervised (CLARA) and supervised (Random Forest) learning algorithms to accurately predict the location of synthetic households at the building scale directly. Third, it allows taking advantage of the rich spatial information usually collected as part of surveys in making accurate predictions of synthetic households' locations. Fourth, the proposed method shows larger transferability than other SA-based methods, including to sizeable study areas, thanks to CLARA's ability to cluster large datasets of buildings/dwelling units and/or households where other clustering algorithm may fail. Its transferability remains however limited compared to IA-based methods not relying neither on the availability of an explicitly spatialized microdata sample nor on locally calibrated models. Finally, the proposed framework is based on a reliable clustering of buildings as CLARA's sensitivity to outliers is low.

The new framework was applied to generate an explicitly spatialized synthetic population for Montreal Island. All the components of the framework were found to perform well with a 2.85% ARE for the zonal synthetic population generated, a 0.48 ASW for the CLARA model retained, a 0.79 MacroF1_{test} for the Random Forest model retained, and 78.9% of the synthetic households being distributed on a building of their most probable cluster. Moreover, the proposed framework managed to successfully locate all synthetic households in a building within their respective synthesis areas.

Despite the various advantages of the introduced method, it also shows some limitations. On the one hand, the zonal synthetic population should be generated using synthetic reconstruction techniques so that the probabilities can be computed for surveyed households' synthetic clones based on their observed characteristics and locations in the microdata sample. On the other hand, the results assessment performed remains incomplete as a general problem faced when performing explicitly spatialized population synthesis is the lack of an observed population at the building/dwelling unit resolution that can be used for validation purposes.

A potential extension of the research works presented in this paper would be adapting the proposed framework to simulation-based synthesizers such as those leveraging deep generative learning algorithms or mixed synthesizers. This would allow to generate a synthetic population showing optimal and diverse sociodemographic and spatial representativeness simultaneously.

Acknowledgments

The authors wish to acknowledge the contribution and financial support of the Mobilité research chair partners: Ministère des transports du Québec (MTQ), Société de Transport de Montréal (STM), Autorité Régionale de Transport Métropolitain, exo, and Ville de Montréal.

Author contribution statement

The authors confirm contribution to the paper as follows:

Mohamed Khachman: conceptualization, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, writing – original draft, writing – review & editing.

Catherine Morency: conceptualization, funding acquisition, validation, writing – review & editing.

Francesco Ciari: conceptualization, validation, writing – review & editing.

All authors reviewed the results and approved the final version of the manuscript.

CHAPITRE 7 DISCUSSION GÉNÉRALE

L'objectif de ce chapitre est de discuter les contributions des travaux de recherche menés dans le cadre de cette thèse. Ainsi, les objectifs et contributions du projet de recherche sont d'abord rappelés avant d'en mener une discussion approfondie. Les limitations des travaux de recherche et les perspectives de recherche qui permettraient d'en compléter les contributions sont aussi évoquées dans ce chapitre.

7.1 Synthèse des contributions de recherche

Comme l'indique son titre « Synthèse spatialisée de la population pour fins de modélisation des comportements de mobilité », l'objectif structurant de cette thèse est de contribuer à l'amélioration de la modélisation des comportements de mobilité d'une population en proposant un cadre méthodologique permettant d'améliorer la qualité des populations synthétiques qui les alimentent. Les caractéristiques sociodémographiques d'une population et leurs distributions spatiales sont toutes les deux déterminantes des comportements de mobilité qui y sont observés. Ainsi, la qualité d'une population synthétique destinée à alimenter un modèle de la demande en transport est définie par son ajustement aux caractéristiques sociodémographiques de la population réelle et la précision avec laquelle elle reproduit les distributions spatiales observées de ces caractéristiques. Cependant, tel que le montre la revue de littérature détaillée dans le Chapitre 2, la majorité des contributions visant à améliorer la qualité des populations synthétiques résument cette qualité à l'ajustement sociodémographique de celles-ci faisant abstraction de la précision spatiale. Pour combler ce manque, les travaux de cette thèse ont principalement été axés sur l'amélioration de la précision spatiale des populations synthétiques tout en veillant à ne pas en compromettre l'ajustement sociodémographique. L'objectif est qu'une amélioration apportée à la précision spatiale de la synthèse contribue à l'amélioration de la qualité globale de la population synthétique produite. Les contributions des travaux de recherche permettant d'atteindre l'objectif structurant de la thèse sont résumées dans cette section.

7.1.1 Indicateurs de précision spatiale et de qualité globale

L'objectif principal du projet de recherche étant d'améliorer la précision spatiale et par conséquent la qualité globale de la synthèse de population, des indicateurs permettant de les mesurer sont

requis. De tels indicateurs n'existent toutefois pas dans la littérature. Ainsi, la première contribution du projet de recherche consiste en la définition d'indicateurs de qualité permettant de mesurer la précision spatiale et la qualité globale de la synthèse. L'indicateur de qualité globale permet d'intégrer l'ajustement sociodémographique et la précision spatiale d'une population synthétique dans un chiffre unique. Cela est important notamment dans un contexte où l'amélioration de la précision spatiale compromet l'ajustement sociodémographique de la population synthétique générée et vice-versa.

Ces indicateurs sont définis dans le cadre de l'article constituant le Chapitre 4 de cette thèse. Outre son originalité, cette contribution s'avère primordiale pour les travaux subséquents et les contributions qui en découlent. En effet, elle permet d'évaluer correctement la qualité du produit de différents paramétrages et/ou méthodologies de synthèse afin de tirer des conclusions généralisables et d'établir les meilleures pratiques.

7.1.2 Impact de la résolution de référence sur la qualité de synthèse

Plusieurs choix effectués avant l'exécution d'un algorithme de synthèse de population sont susceptibles d'impacter la qualité de la population synthétique obtenue. Le choix de la résolution géographique de référence en fait partie. Ainsi, une analyse de sensibilité est menée pour mesurer l'impact de différents choix de résolution de référence sur l'ajustement sociodémographique, la précision spatiale et la qualité globale de la synthèse en utilisant les indicateurs constituant la première contribution du projet de recherche (section 7.1.1). Les expérimentations menées ont permis de constater que la désagrégation de la résolution de référence entraîne une amélioration de la précision spatiale et une détérioration de l'ajustement sociodémographique de la population synthétique générée. Toutefois, pour la même augmentation du degré de désagrégation de la résolution de référence, l'amplitude de la détérioration de l'ajustement sociodémographique mesurée reste largement inférieure à celle du gain en précision spatiale. L'utilisation de la résolution géographique la plus désagrégée permet ainsi d'optimiser la qualité globale de la population synthétique produite.

Les analyses de sensibilité et les résultats obtenus menant à cet avancement des connaissances quant à l'impact du choix de la résolution de référence sur les différents indicateurs de qualité de la population synthétique générée sont aussi présentés dans l'article du Chapitre 4.

7.1.3 Méthode d'harmonisation des distributions marginales

La détérioration de la qualité de l'ajustement sociodémographique accompagnant une désagrégation de la résolution géographique de référence serait causée par deux facteurs. Il s'agit d'une part de l'augmentation exponentielle des contraintes marginales à respecter émanant de l'augmentation du nombre d'unités géographiques lors de la désagrégation de la résolution de référence, et d'autre part des incohérences intra et inter-résolutions des distributions marginales (section 4.3.5) dont l'amplitude augmente aux échelles les plus désagrégées. Aucune mitigation ne peut être appliquée au premier facteur puisqu'il est intrinsèquement lié à l'adoption d'une résolution géographique de référence désagrégée dans le cadre d'un exercice de synthèse. Cependant, le deuxième facteur, étant essentiellement le résultat de traitements de données effectués par les instituts nationaux de la statistique pour préserver la vie privée des individus, peut être sujet à des mesures permettant d'en limiter les impacts négatifs sur la qualité de la synthèse. Ainsi, une méthode d'harmonisation intra et inter-résolution des distributions de probabilités marginales alimentant l'algorithme de synthèse est proposée et l'impact de l'utilisation de données harmonisées sur la qualité de la synthèse est mesuré pour des résolutions de références et des zones d'étude variées. Les résultats montrent que l'harmonisation des distributions marginales permet effectivement d'améliorer la qualité de la population synthétique produite et que son effet est plus important aux échelles les plus désagrégées où les incohérences intra et inter-résolution deviennent plus importantes.

Les analyses comparatives menées entre la synthèse basée sur des distributions marginales brutes et celle basée sur des distributions marginales harmonisées à des échelles et dans des zones d'étude variées sont détaillées dans le Chapitre 4.

7.1.4 Cadre méthodologique intégré

Les contributions précédentes ont permis de définir une meilleure pratique permettant d'optimiser la qualité d'une population synthétique générée consistant à synthétiser à la résolution géographique la plus désagrégée possible en utilisant des distributions de probabilités marginales harmonisées. Pour maximiser la qualité d'une population synthétique, il faudrait alors idéalement la générer à l'échelle du bâtiment voire de l'unité de logement avec des données harmonisées.

Les méthodes de synthèse spatialisée existantes sont basées sur l'approche séquentielle où une population synthétique est d'abord générée à une résolution zonale et chaque ménage synthétique est ensuite distribué sur un bâtiment ou une unité de logement de sa zone de synthèse. Les méthodes de synthèse spatialisées les plus précises utilisent des algorithmes variés tels qu'une optimisation combinatoire, une heuristique de correspondance ou une modélisation probabiliste pour associer chaque ménage synthétique à son bâtiment ou son unité de logement d'accueil. Ces méthodes ne sont toutefois pas utilisables pour générer une population synthétique pour l'île de Montréal ou pour tout autre zone d'étude où elles n'ont pas été développées à la base parce qu'elles dépendent d'un format et d'une abondance locaux de données spatialisées, de connaissances du contexte local d'application, et parfois de la performance de modèles localement calibrés. Il s'agit d'un problème de non-transférabilité intrinsèquement lié à la nature séquentielle des méthodes de synthèse spatialisée existantes. Un autre problème principal de l'approche séquentielle réside dans l'indépendance des processus de génération et de spatialisation dont peuvent émerger des problèmes d'incohérence entre la population synthétique et l'environnement bâti. En effet, les données relatives à l'environnement bâti et les données sociodémographiques sont généralement publiées par des entités distinctes et ne présentent par conséquent aucune garantie de cohérence. Ces incohérences impliquent généralement l'utilisation de mesures artificielles telles que des relaxations de contraintes ou du post-traitement de données afin de réussir à spatialiser l'ensemble des ménages synthétiques générés.

Pour remédier à ces deux problèmes une nouvelle contribution est apportée. Il s'agit de l'introduction d'une approche dite intégrée de synthèse spatialisée de la population où la population synthétique est directement générée à l'échelle du bâtiment. Cette approche présente plusieurs caractéristiques intéressantes constituant chacune une contribution originale.

7.1.4.1 Algorithme multi-résolution flexible et efficace

Générer une population synthétique à l'échelle du bâtiment directement implique l'utilisation d'un algorithme multi-résolution puisque les variables dont les distributions de probabilités marginales peuvent être observées ou dérivées à l'échelle du bâtiment sont limitées. Ainsi, pour ne pas compromettre la qualité de la population synthétique obtenue en négligeant des variables de contrôle importantes non observées à l'échelle du bâtiment, un algorithme multi-résolution permettant de contrôler chaque variable d'intérêt à l'échelle la plus désagrégée où sa distribution

marginale existe est utilisé. Peu efficaces, les algorithmes multi-résolution existants ne peuvent pas être appliqués sur une matrice intégrant des contraintes à l'échelle du bâtiment à cause de la taille colossale que cette dernière pourrait atteindre. De plus, limitées à un nombre maximal de résolutions préfixé par le développeur qui ne dépasse pas trois résolutions, les algorithmes multi-résolution existants ne répondraient pas à tous les besoins puisqu'un nombre supérieur de résolutions à contrôler pourrait être requis selon la disponibilité et le format des données disponibles dans la zone d'étude considérée.

Une caractéristique intéressante du cadre méthodologique proposé est qu'il est basé sur un nouvel algorithme consistant en une extension multi-résolution d'un algorithme efficace de synthèse par optimisation. Cet algorithme est conçu de manière flexible permettant à l'utilisateur de contrôler n'importe quel nombre de niveaux (i.e., types d'agents) et de résolutions géographiques requis.

7.1.4.2 Intégration de la génération et de la spatialisation

Une autre caractéristique intéressante du cadre méthodologique proposé réside dans l'intégration de la génération et de la spatialisation dans un processus unique incluant une harmonisation des distributions marginales observées à l'échelle du bâtiment avec les distributions marginales du recensement. Cela permet d'éviter les problèmes d'incohérences caractéristiques des méthodes de synthèse spatialisée basées sur l'approche séquentielle.

7.1.4.3 Transférabilité optimale

De plus, contrairement à l'approche séquentielle requérant une abondance de données spatialisées notamment pour la calibration de modèles, l'approche proposée présente des besoins minimes en termes de données spatialisées ce qui permet d'en optimiser la transférabilité. En effet, pour des zones d'étude où les données spatialisées accessibles ne sont pas abondantes, la distribution marginale d'une seule variable à l'échelle du bâtiment est suffisante pour pouvoir rouler l'algorithme. L'efficacité de l'algorithme multi-résolution introduit contribue aussi à la transférabilité du cadre méthodologique proposé permettant son application dans des zones d'étude variées sans problèmes de calculs.

7.1.4.4 Meilleure qualité de synthèse

Le cadre méthodologique intégré génère plus d'erreurs d'ajustement sociodémographique qu'une méthode basée sur l'approche séquentielle pour des raisons relevant notamment de la dimension de la matrice (i.e., de la multiplicité des contraintes) engendrée par l'utilisation de l'approche intégrée. Toutefois, la réduction en termes d'erreurs de spatialisation découlant de l'utilisation de l'approche intégrée est largement plus importante que l'augmentation d'erreurs d'ajustement sociodémographique notamment dans un contexte où les incohérences entre la population synthétique générée et l'environnement bâti sont fortement présentes. Ainsi, en plus de sa transférabilité, son efficacité et sa flexibilité, le cadre méthodologique proposé permet d'améliorer la qualité globale de la population synthétique générée.

Les différentes composantes du nouveau cadre méthodologique de synthèse spatialisée de la population sont décrites dans le Chapitre 5. Les analyses comparatives menées pour tester la performance de l'algorithme y sont aussi détaillées.

7.1.5 Cadre méthodologique séquentiel

La qualité d'une population synthétique dépend fortement du degré auquel l'échantillon désagrégé utilisé est représentatif de la population réelle. Or, générer une population synthétique par l'approche intégrée à l'échelle du bâtiment directement implique généralement l'utilisation d'un échantillon zonal pour chacun des bâtiments. Cet échantillon n'est par conséquent généralement pas représentatif de la population de chaque unité de synthèse, en l'occurrence la population de chaque bâtiment. La représentativité sociodémographique de la population synthétique est alors susceptible d'être dégradée même si un ajustement satisfaisant aux distributions marginales peut mathématiquement être atteint. De plus, la dimension de l'exercice d'optimisation matricielle qu'implique l'approche intégrée pourrait empêcher son utilisation dans des contextes où la matrice devient intraitable.

Les données spatialisées deviennent de plus en plus accessibles aux planificateurs et aux chercheurs voire au public. Ainsi, lorsque la zone d'étude considérée en dispose suffisamment pour qu'une synthèse spatialisée de qualité par l'approche séquentielle puisse y être effectuée, utiliser cette approche permettrait de préserver la représentativité sociodémographique de la population synthétique générée et d'éviter tout problème potentiel d'évolutivité. Les méthodes existantes de

synthèse spatialisée basées sur l'approche séquentielle n'étant pas transférables, un nouveau cadre méthodologique séquentiel de synthèse spatialisée de la population est développé pour la zone d'étude. Celui-ci permet de tirer profit de la richesse des données spatialisées accessibles pour optimiser la qualité de la population synthétique spatialisée produite sans en compromettre la représentativité sociodémographique. Il est opéré comme suit :

- Une population synthétique zonale est générée en utilisant pour chacune des zones de synthèse un échantillon local représentatif de sa population ;
- Les bâtiments de la zone d'étude sont partitionnés selon les attributs déterminants des choix résidentiels des ménages ;
- La probabilité de chaque ménage synthétique d'appartenir à chaque bâtiment de sa zone de synthèse est estimée. Elle est calculée selon le groupe du bâtiment déterminé à l'étape précédente et la distance entre le clone du ménage synthétique dans la population réelle et le bâtiment considéré ;
- Les paires ménages-bâtiments dans chacune des zones de synthèse sont sélectionnées par ordre décroissant de leur probabilité d'occurrence en veillant à respecter les fréquences des ménages synthétiques et les capacités d'accueil des bâtiments.

Se démarquant sur plusieurs aspects, le cadre méthodologique de synthèse spatialisée proposé constitue une autre contribution originale des travaux de recherche. Ses caractéristiques sont détaillées ci-dessous.

7.1.5.1 Approche séquentielle

Basée sur l'approche séquentielle, la méthode de synthèse spatialisée de la population introduite permet d'éviter les problèmes potentiellement rencontrés lors d'une synthèse basée sur l'approche intégrée. Il s'agit notamment d'éviter que la représentativité sociodémographique de la population synthétique produite ne soit dégradée et que l'efficacité du processus de synthèse ne soit compromise compte tenu du manque d'intrants et de la multiplicité des contraintes à l'échelle du bâtiment.

7.1.5.2 Structure basée sur l'apprentissage automatique

Le cadre méthodologique proposé est le premier de son type (partitionnement + classification) tirant profit de la performance d'algorithmes d'apprentissage automatique non supervisé (CLARA) et supervisé (Random Forest) pour prédire avec précision l'emplacement des ménages synthétiques à l'échelle du bâtiment directement.

7.1.5.3 Transférabilité acceptable

L'efficacité de l'algorithme « CLARA » et sa robustesse face aux valeurs aberrantes ainsi que la performance de l'algorithme « Random Forest » rendent possible l'application du cadre méthodologique proposé dans d'autres zones d'études où la taille et la nature de la base de données relative à l'environnement bâti et de l'échantillon désagrégé peuvent être différentes. De plus, au lieu d'être effectué en arrière-plan, le calibrage approfondi des paramètres des modèles de partitionnement et de classification utilisés permettant d'en optimiser les performances est intégré dans le cadre méthodologique proposé. Cela confère à la méthode de synthèse spatialisée proposée un degré de transférabilité meilleur que celui des autres méthodes basées sur l'approche séquentielle mais qui n'atteint pas celui de l'approche intégrée.

7.1.5.4 Distribution sans relaxation de contraintes

Contrairement à de nombreuses méthodes de synthèse spatialisée existantes basées sur l'approche séquentielle, la méthode proposée ne requiert aucun post-traitement artificiel des données pour compléter la distribution des ménages synthétiques sur les bâtiments de leurs zones de synthèse respectives. Tous les ménages synthétiques ont été localisés avec succès dans un bâtiment de leurs zones de synthèse respectives. De plus, la majorité des ménages synthétiques a été localisée dans un bâtiment de son groupe de bâtiments d'accueil le plus probable.

7.1.5.5 Évaluation adaptée du partitionnement

Le nombre de groupes optimal à considérer dans un exercice de partitionnement peut être déterminé en utilisant différents indicateurs dont la largeur de silhouette moyenne ou « Average Silhouette Width » (ASW). Lorsque l'ASW est utilisée, le nombre optimal de groupes est celui qui la maximise. Or, dans le cadre d'un exercice de synthèse spatialisée, un nombre réduit de groupes de bâtiments est synonyme de perte de précision spatiale. Ainsi, lorsque le nombre de groupes

maximisant l'ASW est jugé insuffisant pour obtenir une précision spatiale satisfaisante, l'utilisation d'un nouvel indicateur dérivé de l'ASW est recommandée. Il s'agit de la perte en ASW par groupe additionnel (ASWL) qui reflète la dégradation moyenne en termes de qualité de partitionnement par groupe ajouté au nombre de groupes maximisant l'ASW. En d'autres termes, c'est le ratio de la perte en qualité de partitionnement par le gain en précision spatiale. Le nombre optimal de groupes est celui qui minimise l'ASWL.

7.1.5.6 Exploitation optimale de l'information spatiale

Le cadre méthodologique proposé permet une exploitation optimale de l'information spatiale contenue dans la base de données décrivant l'environnement bâti et dans l'échantillon désagrégé explicitement spatialisée de l'enquête O-D dans la performance d'une synthèse spatialisée fiable. À l'étape de génération de la population synthétique zonale, le détail spatial disponible dans l'enquête O-D est utilisé pour filtrer les données à l'échantillon local de la zone de synthèse préservant ainsi la représentativité sociodémographique de la population synthétique générée. La précision spatiale de l'enquête O-D est à nouveau exploitée dans l'association des ménages enquêtés à leurs bâtiments respectifs et par conséquent dans l'établissement d'une estimation fiable des probabilités d'association de chacun des ménages synthétiques à chaque type de bâtiment défini lors de leur partitionnement. Enfin, les distances séparant le clone de chaque ménage synthétique dans l'échantillon O-D des bâtiments de sa zone de synthèse sont dérivées de l'information spatiale précise disponible. Elles sont pour la première fois incluses dans le calcul des probabilités d'association de chaque ménage synthétique aux bâtiments de sa zone d'étude.

Les différentes composantes du nouveau cadre méthodologique séquentiel sont décrites dans le Chapitre 6. Les indicateurs de qualité utilisés pour évaluer la performance de chacune de ses composantes y sont aussi décrits. La méthode proposée est utilisée pour générer une population synthétique explicitement spatialisée pour l'île de Montréal et les indicateurs de performance calculés sont tous satisfaisants.

7.2 Discussion des contributions de recherche

Les contributions du projet de recherche sont en partie exploratoires et en partie méthodologiques. Les contributions exploratoires ont pour thématique principale la détermination et la caractérisation de l'impact de la résolution de référence sur la qualité de la synthèse de population afin d'évaluer

l'intérêt de générer une population synthétique explicitement spatialisée, i.e., à l'échelle du bâtiment ou de l'unité de logement. Une fois l'intérêt de la synthèse spatialisée confirmé, les contributions méthodologiques sont entamées. Il s'agit de proposer des solutions de synthèse explicitement spatialisée adaptées à différents contextes. Une discussion des contributions exploratoires et méthodologiques apportées est menée dans cette section.

7.2.1 Intérêt de la synthèse explicitement spatialisée

L'argumentation liée à l'intérêt de la synthèse explicitement spatialisée est basée sur une démonstration des retombées de la désagrégation de la résolution de référence sur la qualité de la population synthétique générée. Les retombées similaires observées dans plusieurs régions métropolitaines de recensement analysées ont permis de tirer des conclusions générales quant à l'impact de la résolution de référence sur l'ajustement sociodémographique, la précision spatiale et par conséquent la qualité globale de la population synthétique. Bien que cela soit suffisant pour établir une meilleure pratique en ce qui a trait au choix de la résolution de référence de synthèse, les observations recueillies ont été interprétées dans le but d'amorcer la compréhension des variations et interactions des variables suivantes : 1) le degré d'agrégation de la résolution géographique de référence, 2) les incohérences inter-résolutions, 3) les erreurs d'ajustement sociodémographique, et 4) les erreurs de spatialisation. L'idée est de préparer le terrain à l'établissement de modèles explicatifs des variations et interactions de ces paramètres qui pourraient être ultimement utilisés pour prédire les tendances qui seraient observées dans une zone d'étude quelconque et effectuer par conséquent un choix de résolution de référence adapté.

7.2.1.1 Observations

Des populations synthétiques ont été générées pour les trois plus grandes régions métropolitaines de recensement du Canada (Toronto, Montréal et Vancouver) en utilisant des résolutions de référence présentant différents degrés d'agrégation. La qualité de chacune des populations synthétiques générées en termes d'ajustement sociodémographique, de précision spatiale et de qualité globale a été mesurée. Les résultats ont montré que la désagrégation de la résolution de référence mène à une augmentation des erreurs d'ajustement (sociodémographique) et à une réduction des erreurs de spatialisation. L'amplitude de la réduction en termes d'erreurs de spatialisation entraînée demeure toutefois largement plus importante que l'augmentation en termes

d'erreurs d'ajustement. Ainsi, l'utilisation de la résolution de référence la plus désagrégée est prouvée optimiser la qualité globale de la population synthétique et est par conséquent définie comme meilleure pratique pour la synthèse de population. L'intérêt de la synthèse explicitement spatialisée est alors démontré.

7.2.1.2 Interprétations

La similarité des tendances observées dans les différentes régions métropolitaines étudiées était la base sur laquelle la proposition de meilleure pratique généralisable a été émise. Bien que cela était suffisant pour les fins des travaux de recherche menés dans le cadre de cette thèse, une investigation de ces tendances a été amorcée dans le but d'en fournir des interprétations plausibles. L'idée est de tenter de transposer les relations de corrélation observées en des relations de causalité permettant de mieux comprendre les facteurs régissant l'impact de la résolution de référence sur la qualité de la synthèse. Un choix éclairé de résolution de référence optimale selon les caractéristiques de la zone d'étude considérée pourra par conséquent être effectué.

Plusieurs questions liées à la caractérisation de l'impact de la résolution de référence sur la qualité de la population synthétique sont investiguées. La première observation questionnée est l'augmentation des incohérences inter-résolutions lorsque la résolution de référence est désagrégée. En effet, plus l'échelle à laquelle les distributions marginales du recensement sont distribuées est fine, plus le risque d'identification des ménages et/ou individus est important, et plus les altérations des distributions marginales par les instituts nationaux de la statistique sont fréquentes pour des raisons relevant de la protection de la vie privée. La deuxième observation discutée est celle de la détérioration de l'ajustement sociodémographique lorsque la résolution de référence est désagrégée. L'augmentation des incohérences inter-résolutions et l'augmentation du nombre de contraintes à respecter à chaque fois que la résolution de référence devient plus fine sont des explications plausibles de cette tendance. Le troisième constat abordé est la diminution des erreurs de spatialisation lorsque la résolution de référence est désagrégée. En effet, la synthèse de la population à une échelle géographique implique implicitement l'homogénéité spatiale aux échelles plus désagrégées de la population générée dans chacune des unités de synthèse. Or, les distributions spatiales des attributs de la population synthétique aux résolutions plus désagrégées que celle utilisée comme résolution de référence ne sont généralement pas homogènes. Ainsi, plus la

résolution de référence utilisée est agrégée, plus l'hypothèse d'homogénéité spatiale est forte, plus les erreurs de spatialisation sont par conséquent susceptibles d'être importantes, et vice versa.

Bien qu'elles fournissent une certaine compréhension des tendances observées, les interprétations proposées ne permettent pas de quantifier les interactions entre le degré d'agrégation de la résolution de référence, les incohérences inter-résolutions, les erreurs d'ajustement et les erreurs de spatialisation. En effet, le degré d'agrégation demeure non défini de manière quantitative et la variation des erreurs d'ajustement, par exemple, en fonction du degré d'agrégation de la résolution de référence ne peut être modélisée. Il est par conséquent impossible à ce stade de prédire précisément l'impact d'une résolution de référence sur la qualité de la synthèse dans d'autres zones d'étude.

7.2.1.3 Modélisations

Une approche purement quantitative est alors tentée. Il s'agit d'examiner l'évolution des erreurs de spatialisation en fonction des erreurs d'ajustement dans les trois régions métropolitaines étudiées. Les erreurs d'ajustement sont considérées refléter le degré d'agrégation spatiale de la résolution géographique de référence. Un nombre plus élevé d'erreurs d'ajustement est synonyme d'un degré d'agrégation moins important de la résolution de référence. Malgré les similarités entre les tendances linéaires décroissantes observées pour les trois régions métropolitaines, une modélisation généralisable de la variation des erreurs de spatialisation en fonction des erreurs d'ajustement demeure compliquée à établir compte tenu du nombre limité de zones étudiées (i.e., de la taille limitée des données d'entraînement). De plus, pour une modélisation numérique de l'impact du degré d'agrégation de la résolution de référence sur la qualité de la population synthétique, plusieurs extensions des travaux de recherche menés dans le cadre de cette thèse sont requises.

Il s'agit par exemple de définir un indicateur numérique caractérisant directement l'agrégation d'une résolution géographique. Une modélisation généralisable des variations des incohérences inter-résolutions, des erreurs d'ajustement et des erreurs de spatialisation en fonction de l'indicateur d'agrégation défini pourrait alors être établie. Un autre indicateur caractérisant cette fois-ci l'hétérogénéité spatiale d'une zone d'étude serait aussi utile pour expliquer les variations des fréquences des erreurs observées. En effet, plus une zone d'étude est hétérogène aux échelles les plus désagrégées, plus les erreurs de spatialisation lors d'une synthèse aux résolutions agrégées

sont importantes, et plus l'on gagne en termes de précision spatiale de la population synthétique en désagrégeant la résolution de référence utilisée. Dans le cas théorique d'une zone d'étude présentant une homogénéité spatiale parfaite, une désagrégation de la résolution de référence pourrait engendrer plus de pertes en ajustement sociodémographique que de gains en précision spatiale et détériorer par conséquent la qualité de la population synthétique. Les retombées de la variation du degré d'agrégation de la résolution de référence sur la qualité de la synthèse dépendent donc de l'hétérogénéité spatiale de la zone d'étude. Il est ainsi important d'inclure un indicateur mesurant l'hétérogénéité spatiale d'une zone d'étude dans la modélisation de l'impact de la résolution de référence sur la qualité de la synthèse de population afin de déterminer pour chaque zone d'étude le degré de désagrégation fournissant le meilleur compromis entre l'ajustement sociodémographique et la précision spatiale de la population synthétique. La résolution géographique la plus désagrégée est toutefois estimée demeurer celle qui offre ce meilleur compromis compte tenu de l'hétérogénéité naturelle des zones d'étude. L'utilisation d'un autre indicateur reflétant la complexité de l'exercice d'optimisation (i.e., multiplicité des contraintes) permettrait aussi de raffiner la modélisation quantitative de l'impact de la résolution de référence sur la qualité de la population synthétique.

Tel que mentionné précédemment, les contributions exploratoires apportées dans le cadre de la thèse sont suffisantes pour la démonstration des bénéfices de la synthèse explicitement spatialisée introduisant les contributions méthodologiques subséquentes. Une exploration quantitative plus approfondie de l'interaction entre les différentes caractéristiques de la résolution de référence et les différents indicateurs de qualité de la synthèse permettrait toutefois de prédire et caractériser plus précisément l'impact de la résolution géographique de référence considérée sur la qualité de la population synthétique générée pour une zone d'étude quelconque. Le degré de désagrégation spatiale et par conséquent la résolution géographique de référence offrant la meilleure qualité globale de synthèse pourraient ainsi être prédits pour n'importe quelle zone d'étude grâce à des modèles robustes entraînés sur un nombre suffisant d'observations. Compte tenu de l'hétérogénéité spatiale des distributions sociodémographiques naturellement observée dans une zone d'étude, la synthèse explicitement spatialisée est supposée toujours optimiser la qualité de la population synthétique. Une approche quantitative permettrait toutefois d'identifier un cas extrême tel qu'une zone d'étude présentant une homogénéité spatiale où la synthèse à l'échelle du bâtiment ou de l'unité de logement n'offrirait pas la meilleure qualité globale de synthèse.

7.2.2 Choix méthodologique

Les cadres méthodologiques intégré et séquentiel constituent les contributions méthodologiques principales de ce projet de recherche. Ces cadres méthodologiques de synthèse spatialisée de la population présentent des caractéristiques différentes et sont par conséquent appropriés à des contextes de synthèse différents. Cela rend la contribution apportée plus générale puisque le regroupement des méthodologies proposées permet de fournir un choix méthodologique de synthèse spatialisée incluant un choix approprié pour tous les cas de figure que l'on pourrait rencontrer. Ainsi, pour une exploitation optimale du choix méthodologique fourni, il s'avère essentiel de définir des règles permettant de sélectionner, parmi les cadres méthodologiques proposés et leurs extensions possibles, la solution la plus appropriée à la synthèse d'une population synthétique spatialisée selon les caractéristiques de la zone d'étude considérée. Dans cette section, les caractéristiques des deux cadres méthodologiques proposés sont d'abord brièvement rappelées et une méthodologie de sélection du cadre le plus approprié à une zone d'étude est ensuite proposée.

7.2.2.1 Cadres méthodologiques proposés

Tel que détaillé précédemment, les approches intégrée et séquentielle présentent chacune des limitations potentielles qui doivent être considérées lors de la sélection de l'approche de synthèse spatialisée la plus appropriée à une zone d'étude. En effet, un cadre méthodologique basé sur l'approche séquentielle serait plus vulnérable face aux incohérences entre l'environnement bâti et la population synthétique zonale générée et présenterait des problèmes de transférabilité compte tenu de sa dépendance vis-à-vis de la richesse en données spatialisées disponibles et de la performance de modèles localement calibrés. Alors qu'un cadre méthodologique basé sur l'approche intégrée permettrait d'éviter ces problèmes, il s'avère susceptible d'en causer d'autres. En effet, une génération directe à l'échelle du bâtiment serait sujette à un problème d'évolutivité compte tenu de la multiplicité des contraintes notamment entraînée par le nombre des bâtiments dans la zone d'étude. De plus, l'utilisation de l'approche intégrée pourrait aussi nuire à la représentativité sociodémographique de la population synthétique compte tenu du manque d'intrants notamment désagrégés à une échelle aussi fine que celle du bâtiment. Ces problématiques ne seraient généralement pas rencontrées lorsque l'approche séquentielle est adoptée pour la synthèse puisque la population synthétique est générée à une échelle zonale impliquant beaucoup moins de contraintes en utilisant des intrants locaux aux zones de synthèse. Les limitations

potentielles de la synthèse spatialisée par les approches séquentielle et intégrée sont résumées dans la Figure 7.1.

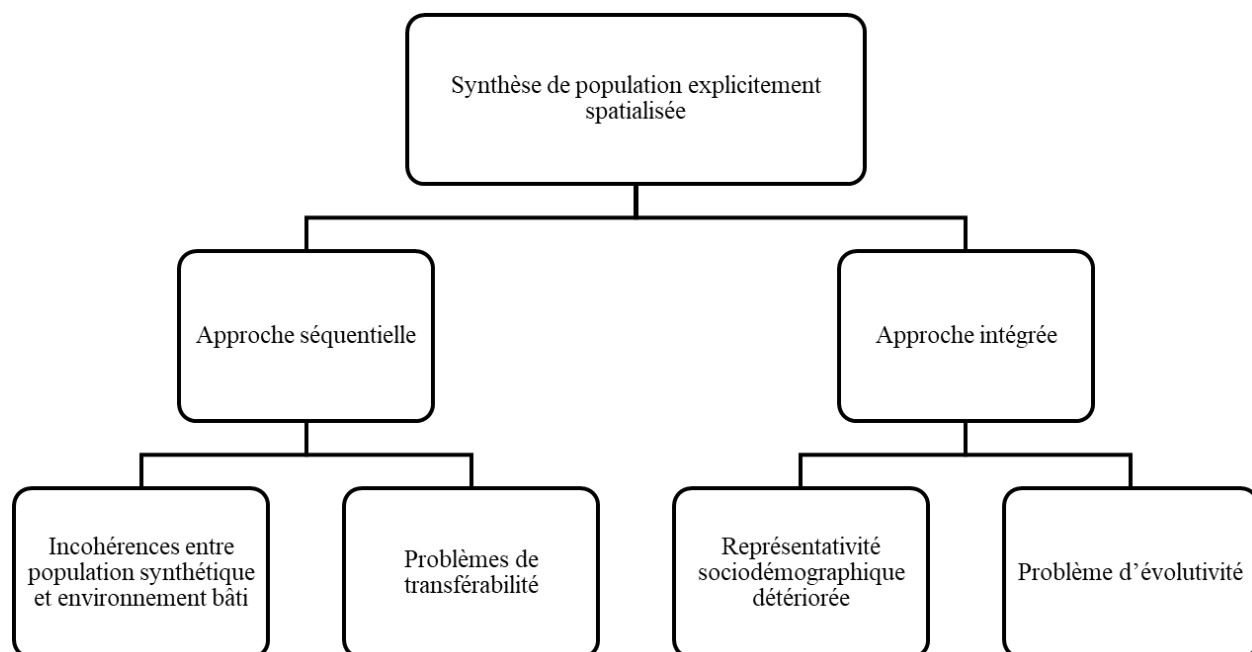


Figure 7.1 Limitations potentielles des approches de synthèse explicitement spatialisée de la population

7.2.2.2 Arbre de décision

Cette section constitue une sorte de mode d'emploi pour le choix de méthodes de synthèse spatialisée de la population proposé dans le cadre de cette thèse. En effet, pour optimiser le profit tiré des contributions méthodologiques apportées, il est important de savoir identifier la méthode la plus appropriée selon les caractéristiques de la zone d'étude pour laquelle une population synthétique explicitement spatialisée est à générer.

La motivation première de la conception de méthodes de synthèse spatialisée basées sur les approches séquentielle et intégrée fut la disponibilité non uniforme des données spatialisées d'une zone d'étude à une autre. En effet, contrairement aux données sociodémographiques généralement accessibles auprès des instituts nationaux de la statistique, la disponibilité et la richesse des données explicitement spatialisées relatives à la population (e.g., échantillon explicitement spatialisé) et à l'environnement bâti varient fortement d'une zone d'étude à une autre. Ainsi, dans l'optique de fournir un choix méthodologique de synthèse spatialisée adapté aux différents contextes, un cadre

méthodologique intégré avec des besoins minimaux en termes de données, et un autre séquentiel permettant une exploitation optimale de la richesse en données spatialisées d'une zone d'étude ont été proposés. Toutefois, l'avancement des travaux de recherche a permis d'identifier d'autres facteurs importants à considérer lors de la sélection d'une méthode de synthèse spatialisée parmi celles proposées. Il s'agit du degré de représentativité sociodémographique convoité au niveau de la population synthétique, de la dimension de l'exercice de génération de la population synthétique à l'échelle du bâtiment directement, de la qualité des prédictions des modèles localement calibrés utilisés dans la spatialisation explicite d'une population synthétique zonale, et de la qualité de l'environnement bâti reflétée par sa cohérence avec les données sociodémographiques utilisées dans la génération de la population synthétique.

Les principes de sélection recommandés sont résumés par l'arbre de décision de la Figure 7.2 permettant de choisir l'approche optimale de synthèse spatialisée de la population selon les caractéristiques de la zone d'étude considérée.

D'abord, pour préserver la représentativité sociodémographique de la population synthétique, la méthode basée sur l'approche séquentielle est préférée à celle basée sur l'approche intégrée si les autres conditions permettant son application sont réunies. La première condition est la disponibilité de données spatialisées dans la zone d'étude considérée. Il s'agit notamment de données décrivant l'environnement bâti permettant l'établissement de distributions marginales à l'échelle du bâtiment (ou de l'unité de logement) et d'au moins un échantillon désagrégé explicitement spatialisé où figurent des caractéristiques des ménages et des individus déterminantes de leurs choix résidentiels et de leurs comportements de mobilité – puisque la population synthétique est produite pour des fins de modélisation de comportements de mobilité. Si de telles données ne sont pas disponibles, une approche séquentielle par correspondance totale ne peut être appliquée et l'approche intégrée doit être utilisée.

La deuxième condition d'utilisation de l'approche séquentielle est la performance locale des modèles utilisés. Cela englobe la qualité du partitionnement des bâtiments de la zone d'étude et la qualité de prédiction du groupe de bâtiments auquel appartient chacun des ménages synthétiques. Même un modèle performant tel que Random Forest demeure dépendant de la qualité des données sur lesquelles il est entraîné et voit par conséquent sa performance varier d'une zone d'étude à une

autre. Ainsi, si la performance locale du modèle s'avère pauvre, il serait préférable d'utiliser la méthodologie basée sur l'approche intégrée.

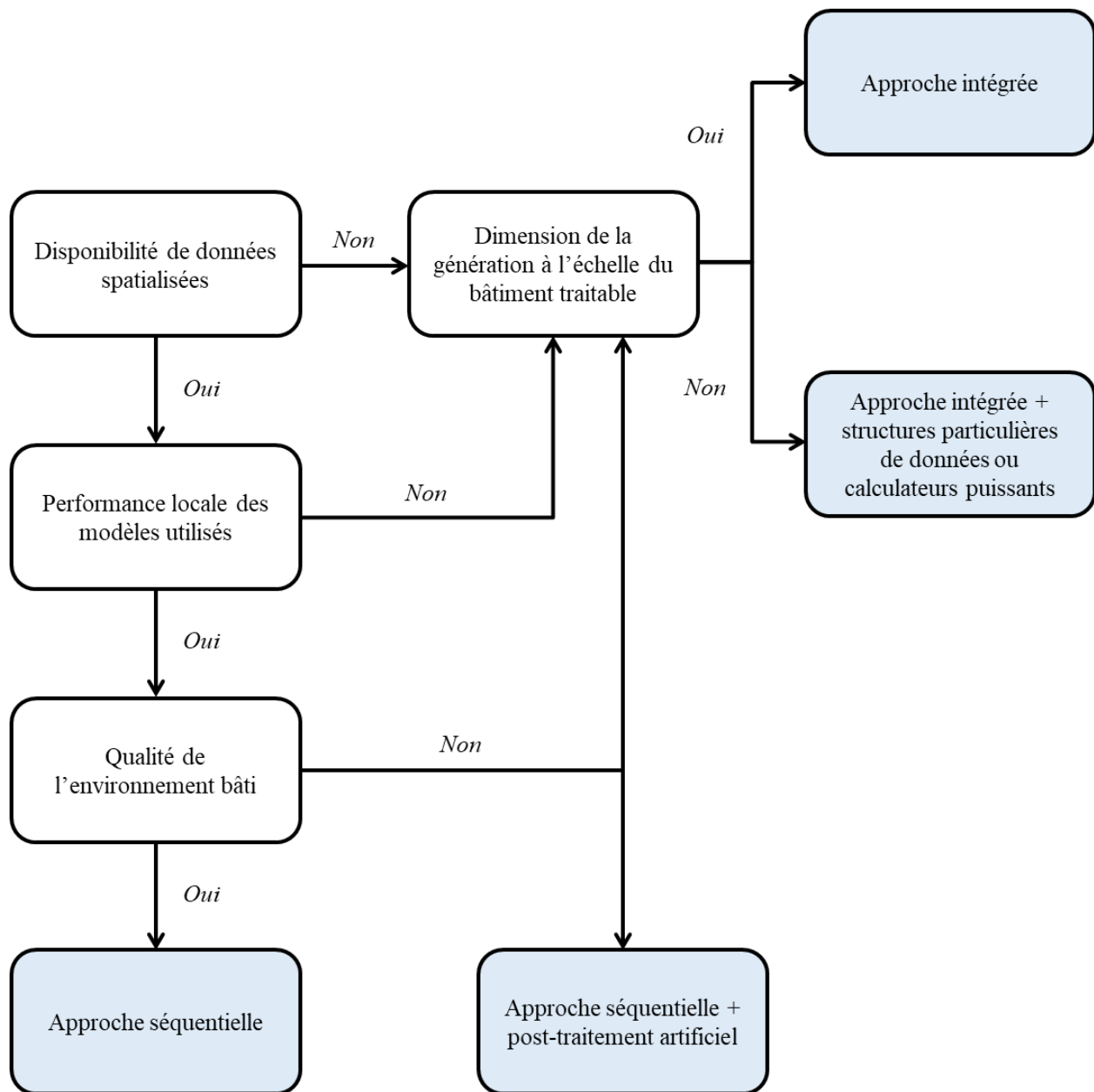


Figure 7.2 Arbre de décision permettant d'identifier l'approche de synthèse spatialisée de la population la plus appropriée

La troisième condition d'utilisation de l'approche séquentielle est la qualité de l'environnement bâti utilisé. Celle-ci dépend de la qualité et de la richesse des distributions marginales disponibles à l'échelle du bâtiment et/ou de la précision et de la qualité des données dont ces distributions

peuvent être dérivées. Si des incohérences entre l’environnement bâti et la population synthétique zonale générée sont jugées limitées, l’approche séquentielle peut quand-même être utilisée mais un post-traitement artificiel des données pour compléter la spatialisation de l’ensemble des ménages synthétiques serait probablement requis. Si les incohérences sont majeures, l’approche intégrée est privilégiée.

Lorsque les trois conditions précédemment décrites sont réunies, l’approche séquentielle préservant la représentativité sociodémographique est considérée la plus appropriée pour une synthèse spatialisée de la population de la zone d’étude. Dans les cas où l’approche intégrée est recommandée, si un problème d’évolutivité est décelé, l’utilisation de structures particulières de données requérant moins de ressources en termes de stockage et/ou de calcul (e.g., données clairsemées) et/ou de calculateurs puissants permettrait de compléter la synthèse spatialisée de la population de la zone d’étude.

7.3 Limitations et perspectives

Les travaux de recherche menés dans le cadre de cette thèse ont permis d’avancer les connaissances relatives à l’impact de la résolution de référence sur la qualité de la synthèse de population et d’accomplir des progrès méthodologiques dans la pratique de la synthèse explicitement spatialisée. Cependant, comme toute contribution scientifique, les contributions de cette thèse sont sujettes à des limitations. Plusieurs extensions des travaux de recherche permettraient ainsi de compléter les avancées apportées à l’état de l’art et les progrès méthodologiques réalisés. Les limitations et les perspectives des travaux de recherche sont détaillées dans cette section.

7.3.1 Limitations

Les limitations des contributions apportées dans le cadre de cette thèse sont multiples. Certaines sont générales, i.e., émanent de la démarche générale de recherche et concernent l’ensemble des contributions scientifiques, et d’autres sont spécifiques à une ou plusieurs des contributions du projet de recherche.

7.3.1.1 Limitations générales

Deux limitations générales des travaux de recherche menés dans le cadre de cette thèse sont identifiées : la qualité non optimale des intrants utilisés et l'unicité du type d'algorithmes de synthèse testés.

7.3.1.1.1 *Qualité non optimale des intrants*

Une limitation générale réside dans les défis rencontrés au niveau des données utilisées. Sur le plan des intrants sociodémographiques, les échantillons désagrégés utilisés incluent de nombreux ménages et/ou individus ne disposant pas de valeurs pour certaines variables de contrôle. Pour résoudre ce problème il est possible soit d'éliminer ces enregistrements de l'échantillon utilisé soit de garder ces enregistrements et d'imputer les valeurs manquantes. La première solution est susceptible de réduire drastiquement la taille de l'échantillon et la deuxième pourrait altérer la structure de corrélation entre les variables de contrôle caractérisant la population réelle. La qualité de la synthèse pourrait ainsi être fortement compromise et un équilibre entre la préservation de la taille de l'échantillon et la préservation de la structure de corrélation des variables de contrôle dans la population réelle doit être trouvé afin de limiter la détérioration de la qualité de synthèse. Un autre problème rencontré au niveau des intrants sociodémographiques est celui des incohérences entre la catégorisation de l'échantillon désagrégé utilisé et celle des totaux du recensement. Lorsqu'une catégorisation harmonisée ne peut être définie pour une variable de contrôle, on pourrait soit éliminer les enregistrements qui ne peuvent être assignés à aucune des catégories utilisées pour cette même variable dans les résumés du recensement soit estimer la catégorie de la variable dans les résumés du recensement à laquelle l'enregistrement appartiendrait. Ces solutions présentent toutefois des enjeux sur la qualité de la synthèse tel que détaillé précédemment. Enfin, lorsqu'une population synthétique est générée en utilisant l'échantillon de l'enquête O-D, les défis liés aux incohérences ne concernent pas les catégories des variables de contrôle uniquement, mais les variables de contrôle en elles-mêmes. En effet, pour être contrôlée, une variable doit être présente à la fois dans les totaux du recensement et dans l'échantillon désagrégé. Or, le nombre de variables d'intérêt communes entre le recensement et l'enquête O-D est extrêmement limité. Ainsi, pour générer une population synthétique en utilisant l'échantillon de l'enquête O-D il faudrait soit se limiter aux variables communes entre l'enquête O-D et le recensement, soit se baser sur la pondération de l'enquête pour dériver les distributions marginales des variables d'intérêt. Celles-

ci ne seraient toutefois disponibles qu'à l'échelle du secteur municipal constituant l'unité de pondération de l'enquête O-D.

Sur le plan de l'environnement bâti, la zone d'étude utilisée dans le projet de recherche (Montréal) ne dispose pas d'une base de données listant et caractérisant ses bâtiments et/ou unités de logement. Ainsi, une fusion de plusieurs bases de données moyennant une variété d'heuristiques basées sur des hypothèses plus ou moins fortes est nécessaire pour établir une base de données des bâtiments de l'île de Montréal. La qualité de la base de données obtenue pourrait alors être compromise ce qui dégraderait la qualité de la synthèse spatialisée.

7.3.1.1.2 Unicité du type d'algorithmes de synthèse

Compte tenu de leur performance, leur efficacité et leur transférabilité, les méthodes basées sur l'ajustement demeurent les plus populaires dans la pratique de la synthèse de population. Ainsi, une variété de méthodes basées sur l'ajustement a été utilisée pour toutes les étapes de la démarche de recherche de cette thèse. Malgré les avantages que cela présente notamment en termes de répliquabilité des méthodologies développées et de leur transférabilité à d'autres zones d'étude, les contributions exploratoires et méthodologiques du projet de recherche se trouvent limitées aux méthodes basées sur l'ajustement. Ainsi, des expérimentations et vérifications sont nécessaires si l'on veut étendre les connaissances introduites et les cadres méthodologiques proposés à la synthèse par simulation puisque celle-ci est opérée selon des principes différents de ceux de la synthèse par ajustement tel que détaillé dans le Chapitre 2.

7.3.1.2 Limitations spécifiques

Outre les limitations générales détaillées ci-dessus, des limitations spécifiques à chacune des contributions du projet de recherche sont identifiées. Les limitations spécifiques sont relatives à l'une des trois étapes de la démarche de recherche soit la détermination de l'impact de la résolution de référence sur la qualité de synthèse, le cadre méthodologique intégré introduit, et le cadre méthodologique séquentiel proposé.

7.3.1.2.1 Détermination de l'impact de la résolution de référence sur la qualité de synthèse

Les travaux entrepris pour déterminer l'impact de la résolution de référence sur la qualité de la synthèse de population sont présentés dans le Chapitre 4. La limitation principale des conclusions

tirées réside dans le nombre de régions métropolitaines de recensement investiguées. En effet, malgré la possibilité de tirer des conclusions générales en se basant sur des tendances observées dans trois grandes régions métropolitaines de recensement telles que Toronto, Montréal et Vancouver, une investigation d'un nombre plus important de régions métropolitaines permettrait d'une part de renforcer ces conclusions et d'autre part de les approfondir afin d'aboutir ultimement à une modélisation des interactions régissant les impacts de la résolution de référence sur la qualité de la synthèse observés.

Une autre limitation des conclusions tirées quant à l'impact de la résolution de référence sur la qualité de la synthèse découle du fait qu'elles sont basées sur des populations synthétiques produites par un seul algorithme de synthèse. Ainsi, outre augmenter le nombre de régions métropolitaines investiguées, varier les algorithmes de synthèse investis dans les expérimentations permettrait de renforcer et/ou de raffiner les conclusions tirées.

7.3.1.2.2 Cadre méthodologique intégré

Le cadre méthodologique intégré de synthèse spatialisée de la population est introduit dans le Chapitre 5. Malgré ses nombreux avantages, le cadre proposé présente deux limitations principales que sont la représentativité sociodémographique détériorée de la population synthétique produite et le problème d'évolutivité.

7.3.1.2.2.1 Représentativité sociodémographique détériorée

Lorsque le cadre méthodologique intégré est utilisé, la population synthétique est directement générée à l'échelle du bâtiment ou de l'unité de logement. Or, il est impossible de disposer d'un échantillon désagrégé pour chacune des unités géographiques de synthèse lorsqu'une résolution de référence aussi fine est utilisée. En effet, les échantillons de microdonnées publiés par les instituts nationaux de la statistique sont généralement relatifs à des découpages géographiques grossiers (e.g., la région métropolitaine de recensement pour le PUMF canadien) et les échantillons explicitement spatialisés d'enquêtes, lorsque disponibles, ne couvrent pas l'ensemble des bâtiments d'une zone d'étude. Une pratique courante consiste à emprunter la totalité d'un échantillon relatif à une zone plus ou moins agrégée pour la génération de la population de chacun des bâtiments appartenant à celle-ci. L'échantillon emprunté n'est ainsi pas forcément représentatif de la

population du bâtiment. Son utilisation pour la synthèse est par conséquent susceptible de dégrader la représentativité sociodémographique de la population synthétique produite.

7.3.1.2.2 Problème d'évolutivité

Un autre défi rencontré lors de l'utilisation du cadre méthodologique intégré réside dans la grande dimension des matrices engendrées par celui-ci. La raison principale expliquant la taille grandissante de la matrice à optimiser est l'augmentation exponentielle du nombre d'unités géographiques de synthèse lorsque la génération se fait directement à une échelle fortement désagrégée telle que celle du bâtiment. En effet, la matrice à optimiser lors de la génération doit inclure les distributions marginales pour l'ensemble des catégories des variables contrôlées dans chacune des unités géographiques de référence. De plus, ce problème de dimensionnalité est accentué lorsque l'information spatiale incluse dans l'échantillon de microdonnées utilisé est très peu précise. En effet, moins l'information spatiale dans l'échantillon désagrégé est précise, plus la taille de l'échantillon emprunté pour générer la population synthétique de chacune des unités géographiques de synthèse est grande. Par exemple, la seule information spatiale disponible dans le hPUMF étant la région métropolitaine d'appartenance des ménages, l'échantillon de l'ensemble de la région métropolitaine de Montréal est emprunté pour générer une population synthétique pour chacun des bâtiments qui y sont situés. Il est important de noter que toute augmentation de la taille de l'échantillon emprunté est amplifiée au niveau de la matrice de génération puisque cet échantillon est dupliqué pour chacune des unités géographiques de synthèse situées dans sa zone d'échantillonnage.

Lorsqu'un échantillon de microdonnées spatialisé tel que celui de l'enquête O-D est disponible, il est possible d'utiliser par exemple l'échantillon d'un secteur municipal pour générer les populations synthétiques des bâtiments qui y sont localisés afin de réduire la dimension de l'échantillon dupliqué. Toutefois, d'autres défis pourraient émerger d'une telle pratique tels que l'indisponibilité dans l'échantillon d'enquête des variables dont l'on souhaite contrôler les distributions marginales notamment aux échelles les plus désagrégées, et les problèmes de cases nulles et de variance découlant du nombre et de la variabilité des observations incluses dans l'échantillon qui pourraient entraver la convergence de l'algorithme de synthèse. De plus, la dépendance du cadre méthodologique intégré à la disponibilité d'un échantillon spatialisé réduirait

sa transférabilité d'où l'importance qu'il puisse être exécuté en utilisant l'échantillon spatialement peu précis du recensement.

Malgré sa grande dimension, la matrice impliquée par la synthèse directe à l'échelle du bâtiment demeure assez rapidement optimisable grâce à l'efficacité de l'algorithme de synthèse utilisé dans le cadre méthodologique intégré proposé. La problématique rencontrée est essentiellement au niveau du stockage de cette matrice. Dans l'application du cadre méthodologique intégré à l'île de Montréal détaillée dans le Chapitre 5, la synthèse de la population de chacune des aires de diffusion de l'île est traitée séparément pour alléger la taille de la matrice à optimiser. Le choix de l'aire de diffusion est basé sur le fait qu'il s'agit du découpage géographique le plus fin où l'ensemble des distributions marginales du recensement sont disséminées. Malgré cela, la dimension de la plus grande matrice à optimiser pour une aire de diffusion était d'environ 10^{10} . Face à ce problème, deux solutions sont possibles. La première consiste à utiliser des structures de données particulières telles que les matrices clairsemées pour réduire la taille de la matrice à optimiser. Cela devrait toutefois être appliqué avec précaution puisque ces structures pourraient ralentir la manipulation de la matrice. Cela ne devrait toutefois pas impacter l'efficacité de la synthèse. Si cette solution ne fonctionne pas, une machine virtuelle capable de stocker la matrice à optimiser devrait alors être utilisée.

Enfin, une limitation de la méthodologie de recherche notamment au niveau de l'évaluation de la performance du cadre méthodologique intégré proposé réside dans l'absence d'une observation complète de la population à des échelles aussi fines que celles du bâtiment ou de l'unité de logement qui peut être utilisée pour évaluer de manière exhaustive la précision spatiale de la population synthétique spatialisée produite.

7.3.1.2.3 Cadre méthodologique séquentiel

Un nouveau cadre méthodologique séquentiel de synthèse spatialisée de la population est introduit dans le Chapitre 6. Malgré ses nombreux avantages, ce cadre méthodologique présente quelques limitations. D'une part, comme toute autre méthode basée sur l'approche séquentielle, la transférabilité de la méthode proposée demeure limitée puisque son applicabilité dépend à la fois de la disponibilité d'un échantillon de microdonnées explicitement spatialisé et de la performance de modèles localement calibrés. D'autre part, la méthode séquentielle proposée ne peut être appliquée que lorsque la population synthétique est générée moyennant un algorithme de synthèse

basé sur l'ajustement pour que chacun des ménages synthétiques puisse hériter des probabilités calculées pour son clone dans l'échantillon de microdonnées spatialisé utilisé. Enfin, tel que mentionné pour le cadre méthodologique intégré, un problème généralement rencontré lorsqu'il s'agit de synthèse spatialisée de la population est l'absence d'une observation complète de la population à des échelles aussi fines que celles du bâtiment ou de l'unité de logement permettant de valider la précision spatiale de la population synthétique produite.

Compte tenu des avantages et limitations des cadres méthodologiques intégré et séquentiel proposé dans le cadre de cette thèse, une discussion du choix méthodologique à adopter selon les caractéristiques de la zone d'étude considérée est menée dans la section 7.2.2.

7.3.2 Perspectives

Comme toute contribution scientifique, les contributions apportées dans le cadre de cette thèse s'inscrivent dans la continuité de travaux précédents (voir Chapitre 2) et frayent le chemin à d'autres contributions permettant d'en mitiger les limitations d'une part et d'y apporter des extensions intéressantes d'autre part. Quelques perspectives des travaux de recherche identifiées sont détaillées dans cette section.

7.3.2.1 Détermination de l'impact de la résolution de référence sur la qualité de la synthèse

Tel que détaillé dans la section 7.3.1.2.1 portant sur les limitations des travaux entrepris pour déterminer l'impact de la résolution de référence sur la qualité de la synthèse, plusieurs améliorations de ces travaux permettant de renforcer ou de raffiner les conclusions tirées sont possibles. Il s'agit par exemple d'investiguer l'impact de la résolution de référence sur la qualité de populations synthétiques générées par une plus grande variété d'algorithmes de synthèse, dont des algorithmes de synthèse par simulation, dans un plus grand nombre de régions métropolitaines. Une autre amélioration potentielle consisterait à approfondir les analyses des interactions des paramètres régissant l'impact de la résolution de référence sur la qualité de la synthèse afin d'en établir une modélisation précise et fiable.

7.3.2.2 Cadre méthodologique intégré avec algorithme de génération mixte

Le cadre méthodologique intégré proposé est basé sur l'utilisation d'un algorithme multi-résolution de synthèse par ajustement. Or, malgré sa capacité à reproduire assez fidèlement les distributions

marginales observées dans le recensement, une limitation principale de cet algorithme réside dans son incapacité à générer des agents non observés dans le hPUMF. La variabilité des agents dans la population synthétique obtenue est alors limitée par celle observée dans le hPUMF et sa représentativité sociodémographique est par conséquent dépendante de la taille et de la qualité du hPUMF. La taille du hPUMF est toutefois extrêmement limitée. Une solution à explorer consisterait à intégrer les algorithmes de synthèse par simulation dans l'étape de génération de la population synthétique du cadre méthodologique intégré. L'idée serait par exemple d'exécuter un cadre méthodologique mixte où le hPUMF est d'abord enrichi par des agents générés par un algorithme hiérarchique de synthèse par simulation, puis investi dans la génération par l'algorithme multi-résolution de synthèse par ajustement inclus dans le cadre méthodologique intégré de synthèse spatialisée proposé.

7.3.2.3 Cadre méthodologique séquentiel avec échantillon O-D enrichi par appariement

L'exécution du cadre méthodologique séquentiel proposé requiert un échantillon de microdonnées explicitement spatialisé. L'information spatiale incluse dans le hPUMF publié par Statistique Canada étant extrêmement limitée, l'échantillon O-D explicitement spatialisé est utilisé. Or, un problème potentiellement rencontré lors de la génération d'une population synthétique en utilisant un échantillon provenant d'une entité autre que l'institut national de la statistique publiant les distributions marginales des variables d'intérêt réside dans le fait que certaines de ces variables pourraient ne pas être enquêtées et par conséquent ne pas figurer dans l'échantillon. Une piste à explorer pour faire face à cela serait d'exécuter une technique d'appariement ou « matching » entre l'échantillon O-D et le hPUMF dans le but d'estimer pour chacun des ménages et des individus inclus dans l'échantillon O-D les valeurs des attributs d'intérêt qui lui manquent en fonction des observations du hPUMF.

7.3.2.4 Optimisation de la représentativité sociodémographique

Les contributions en ce qui a trait à l'amélioration de la représentativité sociodémographique des populations synthétiques étant abondantes dans la littérature, la thèse a été essentiellement axée sur l'introduction et l'optimisation de la précision spatiale de la synthèse tout en veillant à ne pas compromettre la représentativité sociodémographique et par conséquent la qualité globale des populations synthétiques produites. Il serait toutefois intéressant d'inclure dans le processus d'optimisation de la qualité de la synthèse des analyses de sensibilité portant sur des paramètres

affectant la représentativité sociodémographique de la population synthétique. Il s'agit par exemple de tenter différentes combinaisons de variables de contrôle, différentes catégorisations de ces variables, différents croisements des catégories de variables contrôlées (dans le but d'optimiser l'ajustement à la distribution conjointe), différents algorithmes de synthèse dont des algorithmes de synthèse par simulation, etc. L'idée est d'identifier les choix en termes de génération et de spatialisation permettant de simultanément optimiser la représentativité sociodémographique et la précision spatiale de la population synthétique.

7.3.2.5 Évaluation des impacts sur les comportements de mobilité simulés

Les critères d'évaluation utilisés dans le cadre de cette thèse sont tous relatifs à la population synthétique en elle-même. Toutefois, tel qu'expliqué précédemment, la population synthétique produite dans le cadre de ce projet de recherche est destinée à alimenter des microsimulations des comportements de mobilité. Ainsi, il aurait été utile de mesurer l'impact des variétés de scénarios et de méthodologies testées sur la représentativité des comportements de mobilité simulés de ceux observés dans la population réelle. En effet, un choix qui paraît fortement affecter un indicateur de qualité de la population synthétique pourrait avoir un effet moins important sur la simulation de comportements de mobilité alimentée par cette population synthétique et vice versa.

7.3.2.6 Synthèse de population dans le futur

La modélisation des comportements de mobilité est couramment investie dans la prévision de la demande en transport. En effet, une prévision adéquate de la demande en transport constitue un prérequis à l'établissement de politiques stratégiques permettant de fournir une offre en mobilité suffisante et durable. Ainsi, une perspective naturelle des travaux menés dans le cadre de cette thèse serait d'étendre les contributions méthodologiques apportées à la synthèse explicitement spatialisée de la population à un horizon futur.

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

Les déplacements sont à la base de notre activité quotidienne. Les problèmes entravant la mobilité des personnes impactent ainsi l'ensemble de la population qui par conséquent aspire en continu à disposer d'une offre de mobilité améliorée voire optimale. Or, pour proposer des solutions efficaces aux problèmes de mobilité il faut commencer par les identifier et les comprendre. Il est ainsi nécessaire d'établir une modélisation fiable des comportements de mobilité de la population afin de diagnostiquer correctement les différents enjeux et d'identifier les facteurs qui les régissent. Une fois établi, un modèle de qualité permet de tester une variété de scénarios de réponse et d'en mesurer les impacts pour effectuer un choix éclairé.

Les approches modernes de modélisation de la demande en transport sont des approches désagrégées basées sur des processus de prise de décision individuels relevant des caractéristiques des agents modélisés, des contraintes de leur environnement, et des règles régissant les interactions des agents entre eux d'une part, et avec leur environnement d'autre part. Ces modèles sont alimentés, entre autres, par des bases de données listant et caractérisant l'ensemble des individus de la population. Une telle base de données pourrait théoriquement être obtenue par la compilation des formulaires de recensement complétés par la population. Cette compilation demeure toutefois non accessible aux chercheurs et aux planificateurs pour des raisons relevant de la protection de la vie privée, d'où le recours aux populations synthétiques.

La synthèse de population est le processus par lequel les résumés du recensement et les fichiers de microdonnées généralement disséminés par les instituts nationaux de la statistique sont utilisés pour produire une liste complètement énumérée et caractérisée d'agents (ménages et/ou individus) non identifiables présentant un comportement de mobilité statistiquement cohérent avec celui de la population réelle. Deux facteurs exercent une influence importante sur les choix de mobilité d'un individu : la localisation géographique de son ménage, et ses caractéristiques sociodémographiques ainsi que celles du ménage auquel il appartient. Ainsi, pour une simulation réaliste des comportements de mobilité, la population synthétique utilisée en intrant devrait présenter une cohérence statistique avec la population réelle sur les aspects sociodémographique et spatial.

Les travaux du projet de recherche s'inscrivent dans le processus d'amélioration de la modélisation de la demande en transport. L'objectif général de la thèse consiste à optimiser la qualité des populations synthétiques destinées à alimenter des simulations réalistes des comportements de

mobilité. La majorité des travaux recensés dans la littérature partageant ce même objectif général résumant la qualité de la population synthétique à sa représentativité sociodémographique et négligent par conséquent la notion de précision spatiale. Ainsi, pour combler le manque existant, les travaux menés dans le cadre de cette thèse sont axés sur l'optimisation de la précision spatiale des populations synthétiques tout en veillant à ne pas compromettre leur représentativité sociodémographique afin d'aboutir à la meilleure qualité globale possible.

Pour atteindre l'objectif de la thèse, une revue de la littérature portant sur la pratique de la modélisation des comportements de mobilité, les méthodes de génération et les méthodes de spatialisation de populations synthétiques est d'abord effectuée (Chapitre 2). La méthodologie générale en trois étapes de la thèse est décrite dans le Chapitre 3. La première étape consiste à déterminer l'impact de la résolution de référence sur la qualité de la synthèse en termes d'ajustement sociodémographique, de précision spatiale et de qualité globale. Le but est de définir une meilleure pratique en ce qui a trait au choix de la résolution de référence. Les travaux effectués dans le cadre de cette étape sont détaillés dans le Chapitre 4. La deuxième étape consiste en la proposition d'un cadre méthodologique universel de synthèse spatialisée de la population permettant de combler les lacunes en termes de transférabilité des méthodes de synthèse spatialisée existantes. En effet, le cadre méthodologique proposé tire son universalité ou sa transférabilité de ses exigences minimales en termes de données spatialisées. Ce cadre méthodologique est détaillé dans le Chapitre 5. Dans la troisième étape de la démarche de recherche, un autre cadre méthodologique de synthèse spatialisée permettant cette fois-ci l'exploitation de la richesse en données spatialisées de certaines zones d'étude est proposé. L'idée est de tirer profit des données spatialisées lorsque disponibles dans une zone d'étude, afin d'optimiser la qualité globale de la synthèse spatialisée de la population. Ce cadre méthodologique est décrit dans le Chapitre 6. Une discussion des travaux incluant un arbre de décision permettant de choisir le cadre méthodologique à appliquer selon les caractéristiques de la zone d'étude est menée dans le Chapitre 7. Ce chapitre inclut aussi les limitations de la méthodologie utilisée et des conclusions tirées ainsi que quelques perspectives du projet de recherche. Enfin, le présent chapitre (Chapitre 8) termine cet ouvrage en récapitulant le contenu. Il s'agit notamment des contributions de recherche apportées, des limitations et contraintes, et des perspectives de recherche.

8.1 Contributions

Plusieurs contributions scientifiques originales sont apportées par les travaux menés dans le cadre de ce projet de recherche. La préoccupation principale de cette thèse étant l'optimisation de la précision spatiale et par conséquent de la qualité globale de la synthèse de population, la première contribution apportée consiste naturellement en la définition d'indicateurs de précision spatiale et de qualité globale.

L'impact du choix de la résolution de référence sur la qualité de la synthèse est ensuite mesuré en utilisant les indicateurs introduits. Les conclusions tirées des investigations menées constituent la deuxième contribution de cette thèse. Il s'agit notamment de l'amélioration de la précision spatiale et de la détérioration de l'ajustement sociodémographique accompagnant la désagrégation de la résolution de référence. L'amplitude de l'amélioration en termes de précision spatiale est toutefois largement plus importante que celle de la détérioration en termes d'ajustement sociodémographique. La désagrégation de la résolution de référence est alors prouvée améliorer la qualité globale de la population synthétique et l'utilisation de la résolution géographique la plus désagrégée est définie comme meilleure pratique en ce qui a trait au choix de la résolution de référence pour la génération d'une population synthétique. L'intérêt de la synthèse explicitement spatialisée est ainsi démontré.

L'un des facteurs expliquant l'augmentation des erreurs d'ajustement sociodémographique accompagnant la désagrégation de la résolution de référence est l'augmentation des incohérences intra et inter-résolution au niveau des distributions marginales disséminées par les instituts nationaux de la statistique aux échelles les plus désagrégées. Ainsi, dans le but de limiter la détérioration de l'ajustement sociodémographique et d'améliorer par conséquent la qualité globale de la synthèse, un algorithme d'harmonisation des incohérences intra et inter-résolution est proposé. Celui-ci constitue la troisième contribution de la thèse.

L'intérêt de la synthèse explicitement spatialisée étant prouvé et les méthodes de synthèse spatialisée existantes faisant montre d'une transférabilité extrêmement limitée, un nouveau cadre méthodologique de synthèse spatialisée est proposé. En effet, les méthodes de synthèse spatialisée existantes suivent un cadre méthodologique séquentiel consistant à effectuer une génération zonale suivie d'une spatialisation explicite des agents synthétiques. Elles requièrent une abondance de données spatialisées qui servent à la calibration de modèles locaux investis dans la prédiction de la

localisation ponctuelle de chaque ménage synthétique entre autres. Les données spatialisées n'étant pas systématiquement disponibles et uniformes dans toutes les zones d'étude et la qualité de la spatialisation étant dépendante de modèles localement calibrés, la transférabilité des méthodes séquentielles existantes demeure extrêmement limitée. Ainsi, la préoccupation principale lors de la conception du cadre méthodologique proposé était de remédier au problème de transférabilité des méthodes de synthèse spatialisée existantes. Les étapes de génération et de spatialisation sont alors intégrées de manière que la population synthétique soit générée à l'échelle du bâtiment directement. La méthode proposée présente des besoins minimes en termes d'intrants consistant en la distribution marginale d'une seule variable d'intérêt à l'échelle du bâtiment, et n'utilise pas de modèles localement calibrés. Cela lui confère une transférabilité optimale.

Bien que la distribution d'une seule variable à l'échelle du bâtiment soit suffisante pour l'application du cadre méthodologique proposé, plusieurs variables d'intérêt sont généralement contrôlées dans le cadre d'un exercice de synthèse de population. Or, les distributions marginales des variables d'intérêt ne sont généralement pas toutes disponibles à une échelle aussi fine que celle du bâtiment. Ainsi, un algorithme multi-résolution efficace est requis pour contrôler les variables d'intérêt aux différentes résolutions où leurs distributions marginales sont disséminées. Les algorithmes multi-résolution existants sont toutefois limités à trois résolutions géographiques et présentent une efficacité restreinte. Une extension d'un algorithme efficace de synthèse par optimisation à la synthèse multi-résolution est alors introduite. En plus de son efficacité, l'algorithme proposé est flexible, i.e., il permet à l'utilisateur de contrôler n'importe quel nombre de niveaux (i.e., types d'agents) et de résolutions désiré. Cela contribue aussi à la transférabilité du cadre méthodologique proposé à des zones d'études où les distributions marginales des variables d'intérêt sont disséminées à plus de trois résolutions géographiques. Enfin, l'intégration de la génération et de la spatialisation dans un processus unique permet d'éviter les problèmes d'incohérences entre la population synthétique et l'environnement bâti optimisant par conséquent la précision spatiale de la population synthétique produite. Ce cadre méthodologique multi-résolution intégré, efficace, flexible, transférable et performant constitue la quatrième contribution de cette thèse.

La génération d'une population synthétique par l'approche intégrée proposée à une résolution de référence aussi fine que celle du bâtiment fait toutefois face à deux enjeux principaux. D'une part la représentativité sociodémographique de la population synthétique pourrait potentiellement être

détériorée compte tenu du manque d'échantillons spécifiques aux unités géographiques de synthèse. En effet, l'information spatiale incluse dans les échantillons de microdonnées diffusés par les instituts nationaux de la statistique est souvent peu précise (e.g., seulement la région métropolitaine d'appartenance d'un ménage est spécifiée dans l'échantillon canadien). De plus, lorsque disponibles, les échantillons explicitement spatialisés ne couvrent pas l'ensemble des unités géographiques de synthèse, i.e., l'ensemble des bâtiments. D'autre part, la génération d'une population synthétique à l'échelle du bâtiment directement pourrait entraîner un problème d'évolutivité. En effet, le nombre de contraintes à considérer simultanément augmente drastiquement lorsque la génération est effectuée à l'échelle du bâtiment et la matrice de génération pourrait par conséquent atteindre une dimension intraitable. Pour éviter les enjeux de la synthèse intégrée, un nouveau cadre méthodologique séquentiel est proposé. Il constitue la cinquième contribution de cette thèse. Une population synthétique zonale est d'abord générée en utilisant un échantillon d'enquête explicitement spatialisé. La population de chaque unité géographique de synthèse est alors générée en utilisant un échantillon local. Les bâtiments de la zone d'étude sont ensuite partitionnés selon leurs caractéristiques influençant le choix résidentiel des ménages par le biais de l'algorithme robuste et efficace « CLARA ». La probabilité de chacun des ménages synthétiques d'habiter dans chacun des bâtiments de la même unité géographique de synthèse est alors estimée en fonction des caractéristiques du ménage influençant son choix résidentiel, du groupe dans lequel est classé le bâtiment à l'issue de l'étape précédente, et de la distance séparant le clone du ménage synthétique dans l'échantillon désagrégé du bâtiment. L'algorithme performant « Random Forest » est utilisé pour l'estimation de ces probabilités. Les ménages synthétiques sont enfin distribués sur les bâtiments de leur unité géographique de synthèse par ordre décroissant de probabilité d'appartenance tout en veillant à respecter les contraintes de capacité.

Outre éviter les limitations potentielles du cadre méthodologique intégré en termes de représentativité sociodémographique et d'efficacité, le cadre méthodologique séquentiel proposé se démarque de ses prédécesseurs sur plusieurs aspects. Il s'agit d'abord du premier cadre méthodologique partitionnement-classification tirant profit d'algorithmes d'apprentissage non supervisé (CLARA) et supervisé (Random Forest) performants pour prédire avec précision la localisation d'un ménage synthétique à l'échelle du bâtiment directement. Il permet ensuite d'exploiter la richesse en information spatiale des données d'enquête afin d'améliorer la qualité des prédictions émises. De plus, le cadre méthodologique proposé démontre une transférabilité

supérieure à celle des cadres méthodologiques séquentiels de synthèse spatialisée existants. En effet, l'algorithme « CLARA » permet de partitionner des bases de données de bâtiments et/ou de ménages à grandes dimensions qui demeurent intraitables pour d'autres algorithmes de partitionnement. Cela confère à la méthodologie proposée une transférabilité à différentes zones d'études peu importe le nombre de bâtiments et/ou de ménages qui y sont localisés. Cette transférabilité demeure toutefois inférieure à celle du cadre méthodologique intégré proposé précédemment puisque ce dernier ne requiert ni une richesse en données spatialisées dans la zone d'étude ni des modèles localement calibrés performants. Le cadre méthodologique proposé est aussi basé sur un partitionnement fiable grâce à la robustesse de l'algorithme « CLARA » face aux valeurs aberrantes. Un nouvel indicateur de qualité du partitionnement adapté à l'exercice de synthèse de population est introduit pour garantir la sélection d'un partitionnement permettant d'optimiser la précision spatiale de la population synthétique. Enfin, contrairement à d'autres méthodes de synthèse spatialisée existantes, la méthode proposée s'est avérée capable d'assigner chaque ménage synthétique à un bâtiment de son unité géographique de synthèse sans ni relaxations de contraintes ni post-traitements artificiels. La majorité des ménages synthétiques a été localisée dans un bâtiment de son groupe de bâtiments d'accueil le plus probable.

La disponibilité et le format non uniformes des données spatialisées d'une zone d'étude à une autre sont à l'origine de la proposition des deux cadres méthodologiques intégré et séquentiel. En effet, ces deux cadres permettent de répondre à des besoins différents en termes de synthèse spatialisée de la population. Le cadre méthodologique intégré est conçu pour la synthèse spatialisée dans une zone d'étude disposant de peu de données spatialisées relatives aux ménages et/ou aux bâtiments, tandis que le cadre méthodologique séquentiel permet une exploitation optimale de la richesse en données spatialisées d'une zone d'étude. Plusieurs facteurs sont à considérer lors du choix du cadre méthodologique le plus adapté à un exercice de synthèse spatialisée. Les contributions de cette thèse sont alors conclues par la définition d'un arbre de décision permettant d'effectuer un choix méthodologique optimal selon les caractéristiques de la zone d'étude considérée. Afin d'éviter de compromettre la représentativité sociodémographique de la population synthétique et l'efficacité du processus de synthèse, le cadre méthodologique séquentiel devrait être priorisé lorsque les conditions suivantes sont réunies : des données spatialisées sont disponibles dans la zone d'étude, les modèles de partitionnement et de classification localement calibrés sont performants, et les incohérences entre l'environnement bâti et la population synthétique sont limitées voire

inexistantes. Lorsqu'au moins l'une des conditions précédentes n'est pas respectée, l'approche intégrée devrait être utilisée. Si un problème d'évolutivité est décelé lors de l'utilisation de l'approche intégrée, la synthèse spatialisée de la population de la zone d'étude peut être complétée en utilisant des structures particulières de données et/ou des calculateurs puissants.

8.2 Limitations et contraintes

Les travaux de recherche menés dans le cadre de cette thèse présentent plusieurs limitations. Certaines sont générales à l'ensemble de la démarche de recherche et d'autres sont spécifiques à une ou plusieurs contributions. Les limitations générales résident notamment dans deux aspects : la qualité non optimale des intrants et l'unicité du type d'algorithmes de synthèse utilisés dans le cadre du projet de recherche.

Sur le plan de la qualité des intrants, plusieurs problèmes sont décelés au niveau des données sociodémographiques utilisées dont les valeurs d'attributs manquantes pour certains ménages et/ou individus dans les échantillons désagrégés utilisés pour la synthèse, les incohérences entre la catégorisation de l'échantillon désagrégé utilisé et celle des totaux du recensement, et le nombre extrêmement limité de variables communes entre le recensement et l'échantillon d'enquête O-D. Ces problèmes imposent soit une purification des données par élimination des enregistrements problématiques et/ou des variables qui ne sont pas trouvables au niveau des données agrégées et désagrégées simultanément, soit une correction des données par imputation ou recatégorisation des valeurs d'attributs problématiques. Ces deux mesures ont toutefois des effets indésirables. La purification est susceptible de réduire drastiquement la taille de l'échantillon et/ou le nombre de variables contrôlées et la correction pourrait altérer la structure de corrélation entre les variables de contrôle caractérisant la population réelle. Les deux mesures sont alors susceptibles d'altérer la qualité de la synthèse. Un équilibre entre la préservation de la taille de l'échantillon et la préservation de la structure de corrélation des variables de contrôle dans la population réelle doit par conséquent être trouvé afin de limiter la détérioration de la qualité de synthèse. Les données relatives à l'environnement bâti utilisées présentent aussi quelques problèmes. En effet, ces données sont établies moyennant une variété d'heuristiques basées sur des hypothèses plus ou moins fortes puisque la zone d'étude considérée dans le projet de recherche (Montréal) ne dispose pas d'une base de données listant et caractérisant ses bâtiments et/ou unités de logement. La qualité de la base de données établie pourrait alors être compromise ce qui dégraderait la qualité de la

synthèse spatialisée. Sur le plan du type d'algorithmes de synthèse utilisés, l'ensemble de la démarche de recherche a été complétée en utilisant des méthodes de synthèse basées sur l'ajustement compte tenu de leur performance, leur efficacité et leur transférabilité. Les contributions exploratoires et méthodologiques du projet de recherche se trouvent ainsi limitées aux méthodes basées sur l'ajustement et ne peuvent être généralisées aux méthodes basées sur la simulation sans expérimentations et vérifications additionnelles.

En plus de ces limitations générales, les différentes contributions apportées par les travaux menés dans le cadre de cette thèse présentent chacune des limitations spécifiques. D'abord, les conclusions tirées quant à l'impact de la résolution de référence sur la qualité de la synthèse sont basées sur l'application d'un seul algorithme de synthèse à un nombre relativement limité de régions métropolitaines. Ensuite, le cadre méthodologique intégré présente deux limitations principales : une représentativité sociodémographique potentiellement détériorée de la population synthétique produite et un éventuel problème d'évolutivité. Le problème de représentativité sociodémographique détériorée découle du fait qu'un échantillon de microdonnées représentatif de chacune des unités géographiques de synthèse (les bâtiments) n'est généralement pas disponible. Un échantillon relatif à une zone plus ou moins agrégée est alors emprunté pour la génération de la population synthétique de chacun des bâtiments qui y sont localisés. L'échantillon emprunté n'étant pas forcément représentatif de la population du bâtiment, son utilisation est susceptible de dégrader la représentativité sociodémographique de la population synthétique produite. Le problème d'évolutivité découle de la grande dimension des matrices engendrées par l'utilisation du cadre méthodologique intégré. Le nombre élevé d'unités géographiques de synthèse à une échelle de référence aussi fine que celle du bâtiment et la duplication d'un échantillon zonal emprunté pour chacun des bâtiments situés de sa zone sont les facteurs principaux expliquant la croissance substantielle de la matrice de génération lorsqu'une résolution géographique fine telle que celle des bâtiments est adoptée comme résolution de référence. Enfin, par sa nature séquentielle, le cadre méthodologique séquentiel proposé présente des enjeux de transférabilité comparé au cadre méthodologique intégré. En effet, son applicabilité dépend à la fois de la disponibilité d'un échantillon de microdonnées explicitement spatialisé et de la performance de modèles localement calibrés. De plus, le concept méthodologique introduit est dépendant de l'utilisation d'un algorithme de synthèse par ajustement à l'étape de génération pour que chacun des ménages synthétiques puisse ensuite hériter des probabilités calculées pour son clone dans

l'échantillon de microdonnées spatialisé utilisé. Il reste à mentionner qu'un problème généralement rencontré lors d'un exercice de synthèse spatialisée de la population réside dans l'absence d'une observation précise de la population entière permettant de valider la précision spatiale de la population synthétique produite.

8.3 Perspectives de recherche

Plusieurs extensions possibles des travaux menés dans le cadre de cette thèse permettraient de mitiger les limitations des contributions apportées et/ou de compléter l'avancement des connaissances réalisé.

Pour ce qui est de la détermination de l'impact de la résolution de référence sur la qualité de la synthèse, une investigation de différents algorithmes de synthèse, dont des algorithmes de synthèse par simulation, dans un plus grand nombre de régions métropolitaines permettrait de renforcer ou de raffiner les conclusions tirées. Les analyses des interactions des paramètres régissant l'impact de la résolution de référence sur la qualité de la synthèse pourraient aussi être approfondies afin d'en établir une modélisation précise et fiable.

La performance du cadre méthodologique intégré pourrait être améliorée, notamment sur le plan de la représentativité sociodémographique, par l'utilisation d'un algorithme de génération mixte. L'idée est d'enrichir le hPUMF par des agents générés par un algorithme hiérarchique de synthèse par simulation puis d'investir le hPUMF enrichi obtenu dans la génération par l'algorithme multi-résolution de synthèse par ajustement inclus dans le cadre méthodologique intégré de synthèse spatialisée proposé. Cela permettrait d'augmenter la variabilité des agents synthétiques qui est actuellement limitée aux observations peu nombreuses du hPUMF.

La performance du cadre séquentiel proposé pourrait être améliorée en exécutant une technique d'appariement entre l'échantillon O-D et le hPUMF. L'idée est d'estimer pour chacun des ménages et des individus inclus dans l'échantillon O-D les valeurs des attributs d'intérêt qui lui manquent en fonction des observations du hPUMF. Un échantillon O-D avec plus de variables de contrôle figurant dans les totaux du recensement est ainsi obtenu ce qui permettrait d'améliorer la représentativité sociodémographique de la population synthétique générée.

L'amélioration de la représentativité sociodémographique des populations synthétiques pourrait être plus amplement considérée par l'inclusion d'analyses de sensibilité portant sur des paramètres

affectant la représentativité sociodémographique de la population synthétique dans le processus d'optimisation de la qualité de la synthèse. Ainsi, les variables contrôlées, leurs catégorisations, les croisements éventuels des catégories de variables contrôlées, les algorithmes de synthèse, etc. pourraient être variés et les impacts des différentes combinaisons mesurés et analysés. L'idée est d'identifier les choix en termes de génération et de spatialisation permettant de simultanément optimiser la représentativité sociodémographique et la précision spatiale de la population synthétique.

L'évaluation des choix de synthèse pourrait être étendue à leurs impacts sur les microsimulations des comportements de mobilité. En effet, les populations synthétiques spatialisées étant produites pour alimenter des microsimulations des comportements de mobilité, il serait utile de mesurer l'impact des variétés de scénarios et de méthodologies testées sur la représentativité des comportements de mobilité simulés de ceux observés dans la population réelle. Les choix de synthèse pourraient ainsi être raffinés selon les impacts mesurés.

Enfin, la modélisation des comportements de mobilité étant souvent investie dans la prévision de la demande en transport, il serait intéressant de tester une application des contributions de la thèse à la synthèse explicitement spatialisée de la population à un horizon futur. Un exercice de synthèse dans le futur présenterait des enjeux différents de ceux rencontrés lors d'une synthèse dans le présent notamment au niveau de la disponibilité et de la nature des intrants alimentant le cadre méthodologique de synthèse spatialisée, et de la validation des résultats.

BIBLIOGRAPHIE

- Abraham, J.E., Stefan, K.J., & Hunt, J.D. (22-26 janvier 2012). *Population synthesis using combinatorial optimization at multiple levels* [Communication de conférence]. Transportation Research Board 91st Annual Meeting, Washington D.C.
- Adnan, M., Pereira, F.C., Azevedo, C.M.L., Basak, K., Lovric, M., Feliu, R.S., Zhu, Y., Ferreira, J., Zegras, C., & Ben-Akiva, M.E. (10-14 janvier 2016). *SimMobility: a multi-scale integrated agent-based simulation platform* [Communication de conférence]. Transportation Research Board 95th Annual Meeting, Washington, D.C. http://web.mit.edu/cami/Public/CLA_SimMobility_Framework_TRB_2016.pdf
- Anderson, B. (2012). Estimating small-area income deprivation: an iterative proportional fitting approach. Dans Tanton, R., & Edwards, K. (édit.), *Spatial microsimulation: a reference guide for users* (p. 49-67). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4623-7_4
- Anderson, W., Guikema, S., Zaitchik, B., & 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. <https://doi.org/10.1371/journal.pone.0100037>
- Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2014(1), 85–91. <https://doi.org/10.3141/2014-11>
- Autorité Régionale de Transport Métropolitain (ARTM) (2018). *Enquête Origine-Destination 2018* [Données brutes non publiées]. ARTM.
- Autorité Régionale de Transport Métropolitain (ARTM) (2023). *Enquête Origine-Destination 2013*. <https://www.artm.quebec/eod/2013/>
- Atkins, S. (1977). Transportation planning: is there a road ahead? *Traffic Engineering and Control*, 18(2).
- Auld, J.A., Mohammadian, A.K., & Wies, K. (2009). Population synthesis with subregion-level control variable aggregation. *Journal of Transportation Engineering*, 135(9), 632–639. [https://doi.org/10.1061/\(ASCE\)TE.1943-5436.0000040](https://doi.org/10.1061/(ASCE)TE.1943-5436.0000040)

- Badu-Marfo, G., Farooq, B., & Patterson, Z. (2022). Composite travel generative adversarial networks for tabular and sequential population synthesis. *IEEE Transactions on Intelligent Transportation Systems*, 23(10). <https://doi.org/10.1109/TITS.2022.3168232>
- Balakrishna, R., Sundaram, S., Lam, J. (12-16 janvier 2020). *An enhanced and efficient population synthesis approach to support advanced travel demand models* [Communication de conférence]. Transportation Research Board 99th Annual Meeting, Washington, D.C.
- Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place*, 11(1), 13–34. <https://doi.org/10.1002/psp.351>
- Ballas, D., Clarke, G., Dorling, D., & Rossiter, D. (2007). Using SimBritain to model the geographical impact of national government policies. *Geographical Analysis*, 39(1), 44–77. <https://doi.org/10.1111/j.1538-4632.2006.00695.x>
- Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., & Nagel, K. (2009). MATSim-T: architecture and simulation times. Dans Bazzan, A., & Klügl, F. (édit.), *Multi-Agent Systems for Traffic and Transportation Engineering*, (p. 57–78). IGI Global, Pennsylvania. <https://doi.org/10.4018/978-1-60566-226-8.ch003>
- Bar-Gera, H., Konduri, K., Sana, B., Ye, X., & Pendyala, R.M. (11-15 janvier 2009). *Estimating survey weights with multiple constraints using entropy optimization methods* [Communication de conférence]. Transportation Research Board 88th Annual Meeting, Washington, D.C. http://rampendyala.weebly.com/uploads/5/0/5/4/5054275/popsynentropyapproach_bargerakondurisanayependyala_transportationrev.pdf
- Barthelemy, J., & Toint, P.L. (2013). Synthetic population generation without a sample. *Transportation Science*, 47(2), 266–279. <http://dx.doi.org/10.1287/trsc.1120.0408>
- Bast, H., Storandt, S., & Weidner, S. (3-6 novembre 2015). *Fine-grained population estimation* [Communication de conférence]. 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. <https://doi.org/10.1145/2820783.2820828>
- Basu, R., & Ferreira, J. (2020). Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research. Procedia*, 48, 1674–1693. <https://doi.org/10.1016/j.trpro.2020.08.207>

- Beckman, R.J., Baggerly, K.A., & McKay, M.D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6), 415–429. [https://doi.org/10.1016/0965-8564\(96\)00004-3](https://doi.org/10.1016/0965-8564(96)00004-3)
- Birkin, M., Turner, A., & Wu, B. (28-30 juin 2006). *A synthetic demographic model of the UK population: methods, progress and problems* [Communication de conférence]. Second International Conference on e-Social Science, Manchester, U.K. <https://www.researchgate.net/publication/252883471>
- Bonnel, P. (2002). *Prévision de la demande de transport* [Thèse de doctorat, Economies et finances - Université Lumière - Lyon II]. TEL archives ouvertes. <https://theses.hal.science/tel-00268919>
- Borysov, S.S., & Rich, J. (2021). Introducing synthetic pseudo panels: application to transport behaviour dynamics. *Transportation*, 48, 2493–2520. <https://doi.org/10.1007/s11116-020-10137-5>
- Borysov, S.S., Rich, J., & 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. <https://doi.org/10.1016/j.trc.2019.07.006>
- Bowman, J.L., Bradley, M., Castiglione, J., & Yoder, S.L. (12-16 janvier 2014). *Making advanced travel forecasting models affordable through model transferability* [Communication de conférence]. Transportation Research Board 93rd Annual Meeting, Washington, D.C.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Calka, B., Bielecka, E., & Zdunkiewicz, K. (2016). Redistribution population data across a regular spatial grid according to buildings characteristics. *Geodesy and Cartography*, 65(2), 149–162. <https://doi.org/10.1515/GEOCART-2016-0011>
- Casati, D., Müller, K., Fourie, P.J., Erath, A., Axhausen, K.W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record: Journal of the Transportation Research Board*, 2493(1), 107–116. <https://doi.org/10.3141/2493-12>

- Chapuis, K., Taillandier, P., Misslin, R., & Drogoul, A. (2018). Gen*: a generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science*, 32(6), 1194–1210. <https://doi.org/10.1080/13658816.2018.1440563>
- Chen, H., Wu, B., Yu, B., Chen, Z., Wu, Q., Lian, T., Wang, C., Li, Q., & Wu, J. (2021). A new method for building-level population estimation by integrating LiDAR, nighttime light, and POI data. *Journal of Remote Sensing*, 2021, 9803796. <https://doi.org/10.34133/2021/9803796>
- Choupani, A.A., & Mamdoohi, A.R. (2016). Population synthesis using iterative proportional fitting (IPF): a review and future research. *Transportation Research Procedia*, 223–233. <https://doi.org/10.1016/j.trpro.2016.11.078>
- Deming, W.E., & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4), 427–444. <https://doi.org/10.1214/aoms/1177731829>
- De Palma, A., Picard, N., & Waddell, P. (2007). Discrete choice models with capacity constraints: an empirical analysis of the housing market of the greater Paris region. *Journal of Urban Economics*, 62(2), 204–230. <https://doi.org/10.1016/j.jue.2007.02.007>
- Deville, J.C., Särndal, C.E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423), 1013–1020.
- Douglass, B., & Lee, J. (1973). Requiem for large-scale models. *Journal of the American Institute of Planners*, 39(3), 163–178. <https://doi.org/10.1080/01944367308977851>
- Fabre, L., & Morency, C. (2019). Enriching travel demand forecasting models with a household typology. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(4), 975–987. <https://doi.org/10.1177/0361198119838978>
- Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, 243–263. <https://doi.org/10.1016/j.trb.2013.09.012>
- Farrell, N., Morrissey, K., & O'Donoghue, C. (2012). Creating a spatial microsimulation model of the Irish local economy. Dans Tanton, R., & Edwards, K. (édit.) *Spatial microsimulation: a reference guide for users* (p. 105–125). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4623-7_7

- Fournier, N., Christofa, E., Akkinapally, A.P., & Azevedo, C.L. (2021). Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, 48, 1061–1087. <https://doi.org/10.1007/s11116-020-10090-3>
- Frick, M., & Axhausen, K.W. (25-26 mars 2004). *Generating synthetic populations using IPF and Monte Carlo techniques: some new results* [Communication de conférence]. 4th Swiss Transport Research Conference, Ascona, Switzerland. <https://doi.org/10.3929/ethz-a-004753115>
- Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>
- Gallagher, S., Richardson, L.F., Ventura, S.L., & Eddy, W.F. (2018). SPEW: synthetic populations and ecosystems of the world. *Journal of Computational and Graphical Statistics* 27(4), 773–784. <https://doi.org/10.1080/10618600.2018.1442342>
- Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PLoS ONE*, 5(1), e8828. <https://doi.org/10.1371/journal.pone.0008828>
- Garrido, S., Borysov, S.S., Pereira, F.C., & Rich, J. (2020). Prediction of rare feature combinations in population synthesis: application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120, 102787. <https://doi.org/10.1016/j.trc.2020.102787>
- Guo, J., & Bhat, C. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014(1), 92–101. <https://doi.org/10.3141/2014-12>
- Harding, A., Vidyattama, Y., & Tanton, R. (2011). Demographic change and the needs-based planning of government services: projecting small area populations using spatial microsimulation. *Journal of Population Research*, 28, 203–224. <https://doi.org/10.1007/s12546-011-9061-6>
- Harland, K., Heppenstall, A., Smith, D., 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), 1–25. <http://dx.doi.org/10.18564/jasss.1909>

- Huang Z, & Williamson P. (2001). *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata* [Rapport technique]. Department of Geography: University of Liverpool. http://pcwww.liv.ac.uk/~william/microdata/workingpapers/hw_wp_2001_2.pdf
- Hunsinger, E. (2008). *Iterative proportional fitting for a two-dimensional table* [Présentation PowerPoint]. Alaska Department of Labor and Workforce Development. <https://edyhsgr.github.io/IPFDescription/AKDOLWDIPFTWOD.pdf>
- Hynes, S., Morrissey, K., O'donoghue, C., & Clarke, G. (2009). A spatial micro-simulation analysis of methane emissions from Irish agriculture. *Ecological Complexity*, 6(2), 135–146. <https://doi.org/10.1016/j.ecocom.2008.10.014>
- Ilahi, A., & Axhausen, K.W. (2019). Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. *Regional Studies, Regional Science*, 6(1), 623–636. <https://doi.org/10.1080/21681376.2019.1687011>
- Ireland, C.T., Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55(1), 179–188. <https://doi.org/10.2307/2334462>
- Ji, Z., & Wan, Y. (2021). A novel method for socioeconomic data spatialization. *Spatial Statistics*, 43, 100501. <https://doi.org/10.1016/j.spasta.2021.100501>
- Johnsen, M., Brandt, O., Garrido, S., & Pereira, F.C. (2022). Population synthesis for urban resident modeling using deep generative models. *Neural Computing and Applications*, 34, 4677–4692. <https://doi.org/10.1007/s00521-021-06622-2>
- Jiang, N., Crooks, A.T., Kavak, H., Burger, A., & Kennedy, W.G. (2022). A method to create a synthetic population with social networks for geographically-explicit agent-based models. *Computational Urban Science*, 2(7). <https://doi.org/10.1007/s43762-022-00034-1>
- Kagho, G.O., Balac, M., & Axhausen, K.W. (2020). Agent-based models in transportation planning: current state, issues, and expectations. *Procedia Computer Science*, 170, 726–732. <https://doi.org/10.1016/j.procs.2020.03.164>
- Kalter, M.J.O., & 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. <https://doi.org/10.18757/ejtir.2016.16.4.3166>

Kane, L., & Behrens, R. (15-19 juillet 2002). *Transport planning models – an historical and critical review* [Communication de conférence]. 21st South African Transport Conference, Pretoria, South Africa. <https://repository.up.ac.za/bitstream/handle/2263/7834/037.pdf>

Kassambara, A. (2017). *Practical guide to cluster analysis in R: unsupervised machine learning* (1^{ère} éd.). STHDA.

Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470316801>

Khachman, M., Morency, C., & Ciari, F. (25 mai 2021a). *Enjeux de la synthèse multi-résolution de la population* [Communication de conférence]. Colloque Annuel de la Chaire Mobilité, Montréal, Québec, Canada.

Khachman, M., Morency, C., & Ciari, F. (2021b). Impact of the geographic resolution on population synthesis quality. *ISPRS International Journal of Geo-Information*, 10, 790. <https://doi.org/10.3390/ijgi10110790>

Khachman, M., Morency, C., & Ciari, F. (2022). Integrated multiresolution framework for spatialized population synthesis. *Transportation*. <https://doi.org/10.1007/s11116-022-10358-w>

Khachman, M., Morency, C., & Ciari, F. (2023). A novel machine learning-based spatialized population synthesis framework. Submitted to *Transportation*.

Kim, E.-J., & Bansal, P. (2023). A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148, 104053. <https://doi.org/10.1016/j.trc.2023.104053>

Kitamura, R., Chen, C., Pendyala, R.M., & Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27, 25–51. <https://doi.org/10.1023/A:1005259324588>

Konduri, K.C., You, D., Garikapati, V.M., & Pendyala, R.M. (2016). Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions.

Transportation Research Record: Journal of the Transportation Research Board, 2563(1), 40–50.
<https://doi.org/10.3141/2563-08>

Larousse. (s.d.). Modèle. Dans *Larousse.fr : encyclopédie et dictionnaires gratuits en ligne*. Consulté le 1 août 2023, tiré de <https://www.larousse.fr/dictionnaires/francais/mod%C3%A8le/51916#181897>

Lederrey, G., Hillel, T., & Bierlaire, M. (12-14 septembre 2021). *DATGAN: Integrating expert knowledge into deep learning for population synthesis* [Communication de conférence]. 21st Swiss Transport Research Conference, Monte Verità / Ascona. https://www.strc.ch/2021/Lederrey_EtAl.pdf

Lee, D.H., & Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record: Journal of the Transportation Research Board* 2255(1), 20–27. <https://doi.org/10.3141/2255-03>

Liaw, A., & Wiener, M. (2002) Classification and Regression by randomForest. *R News* 2(3), 18-22.

Little, R.J.A., & Wu, M.M. (1991). Models for contingency tables with known marginals when target and sampled populations differ. *Journal of the American Statistical Association*, 86(413), 87-95. <https://doi.org/10.2307/2289718>

Liu, J., Ma, X., Zhu, Y., Li, J., He, Z., & Ye, S. (2021). Generating and visualizing spatially disaggregated synthetic population using a web-based geospatial service. *Sustainability*, 13(3), 1587. <https://doi.org/10.3390/su13031587>

Long, Y., & Shen, Z. (2015). *Geospatial Analysis to Support Urban Planning in Beijing*. Springer, Cham. https://doi.org/10.1007/978-3-319-19342-7_6

Loo, B.P.Y., & Lam, W.W.Y. (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. <http://dx.doi.org/10.1080/18128602.2011.643509>

Lovelace, R., & Ballas, D. (2013). Truncate, replicate, sample: a method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41, 1–11. <https://doi.org/10.1016/j.compenvurbsys.2013.03.004>

- Ma, L., & 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.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2022). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4.
- Miller, E.J., & Roorda, M.J. (2003). Prototype model of household activity-travel scheduling. *Transportation Research Record: Journal of the Transportation Research Board*, 1831(1), 114–121. <https://doi.org/10.3141/1831-13>
- Ministère des Affaires Municipales et de l'Habitation (MAMH) (2010). *À quoi sert l'évaluation foncière municipale ?* <https://www.mamh.gouv.qc.ca/evaluation-fonciere/evaluation-fonciere-municipale-au-quebec/a-quoi-ca-sert/>
- Ministère des Affaires Municipales et de l'Habitation (MAMH) (2020). *Rôles d'évaluation foncière du Québec* [Données brutes non publiées]. MAMH.
- Ministère des Affaires Municipales et de l'Habitation (MAMH) (2022). *Répertoire des données du rôle d'évaluation, format ouvert* (24 mars 2022) [Guide]. Données Québec 2022. <https://www.donneesquebec.ca/recherche/dataset/roles-d-evaluation-fonciere-du-quebec>
- Ministère des Affaires Municipales et de l'Habitation (MAMH) (2023). *Rôles d'évaluation foncière du Québec* (Version du 13 juillet 2023) [Ensemble de données]. Données Québec 2022. <https://www.donneesquebec.ca/recherche/dataset/roles-d-evaluation-fonciere-du-quebec>
- Mobility Analytics Research Group (MARG) (2016). *PopGen: Synthetic Population Generator* (Version 2.0) [Logiciel]. <http://www.mobilityanalytics.org/popgen.html>
- Mohammadian, A.K., Javanmardi, M., & Zhang, Y. (2010). Synthetic household travel survey data simulation. *Transportation Research Part C: Emerging Technologies*, 18, 869–878. <https://doi.org/10.1016/j.trc.2010.02.007>
- Moreno, A.T., & Moeckel, R. (2018). Population synthesis handling three geographical resolutions. *ISPRS International Journal of Geo-information*, 7(5), 174. <https://doi.org/10.3390/ijgi7050174>

- Mosteller, F. (1968). Association and Estimation in Contingency Tables. *Journal of American Statistical Association*, 63(321), 1-28. <https://doi.org/10.2307/2283825>
- Müller, K., & Axhausen, K.W. (1-3 septembre 2010). *Population Synthesis for Microsimulation: State of the Art* [Communication de conférence]. 10th Swiss Transport Research Conference, Ascona, Switzerland. <https://doi.org/10.3929/ethz-a-006132973>
- Müller, K., & Axhausen, K.W. (30 août – 2 septembre 2011). *Hierarchical IPF: generating a synthetic population for Switzerland* [Communication de conférence]. 51st Congress of the European Regional Science Association, Barcelona, Spain. <https://www-sre.wu.ac.at/ersa/ersaconfs/ersa11/e110830aFinal00305.pdf>
- Müller, K., & Axhausen, K.W. (2012) Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs- und Raumplanung*, 821, 1–21. <https://doi.org/10.3929/ethz-a-010192943>
- Muñoz, E., & Peters, I. (2014). Constructing an urban microsimulation model to assess the influence of demographics on heat consumption. *International Journal of Microsimulation*, 7(1), 127–157. <https://doi.org/10.34196/ijm.00096>
- Nejad, M.M., Erdogan, S., & Cirillo, C. (2021). A statistical approach to small area synthetic population generation as a basis for carless evacuation planning. *Journal of Transport Geography*, 90. <https://doi.org/10.1016/j.jtrangeo.2020.102902>
- Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A: Economy and Space*, 27(3), 425–446. <https://doi.org/10.1068/a270425>
- Otani, N., Sugiki, N., Vichiensan, V., & Miyamoto, K. (2012). Modifiable attribute cell problem and solution method for population synthesis in land use microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2302(1), 157-163. <https://doi.org/10.3141/2302-17>
- Pajares, E., Muñoz Nieto, R., Meng, L., & Wulfhorst, G. (2021). Population disaggregation on the building level based on outdated census data. *ISPRS International Journal of Geo-Information*, 10, 662. <https://doi.org/10.3390/ijgi10100662>

- Panori, A., Ballas, D., Psycharis, Y. (2017). SimAthens: a spatial microsimulation approach to the estimation and analysis of small area income distributions and poverty rates in the city of Athens, Greece. *Computers, Environment and Urban Systems*, 63, 15–25. <http://dx.doi.org/10.1016/j.compenvurbsys.2016.08.001>
- Prédhumeau, M., & Manley, E. (2023). A synthetic population for agent-based modelling in Canada. *Scientific Data*, 10, 148 (2023). <https://doi.org/10.1038/s41597-023-02030-4>
- Pritchard, D.R., & Miller, E.J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39, 685–704. <https://doi.org/10.1007/s11116-011-9367-4>
- Raux, C., & Chapleau R. (1997). Conclusion. Dans Bonnel P, Chapleau R, Lee-Gosselin M, & Raux C (édit.), *Les enquêtes de déplacements urbains : mesurer le présent, simuler le futur* (p. 459-468), Programme Rhône-Alpes Recherches en Sciences Humaines, Lyon. <https://shs.hal.science/halshs-00271106>
- Rich, J., & Mulalic, I. (2012). Generating synthetic baseline populations from register data. *Transportation Research Part A: Policy and Practice*, 46(3), 467-479. <https://doi.org/10.1016/j.tra.2011.11.002>
- Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population synthesis: comparing the major techniques using a small, complete population of firms. *Geographical Analysis* 41(2), 181–203. <https://doi.org/10.1111/j.1538-4632.2009.00750.x>
- Saadi, I., Farooq, B., Mustafa, A., Teller, J., & Cools, M. (2018). An efficient hierarchical model for multi-source information fusion. *Expert Systems With Applications*, 110, 352–362. <https://doi.org/10.1016/j.eswa.2018.06.018>
- Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov model-based population synthesis. *Transportation Research Part B: Methodological*, 90, 1–21. <https://doi.org/10.1016/j.trb.2016.04.007>

- Sallard, A., Balać, M., & Hörl, S. (2021). An open data-driven approach for travel demand synthesis: an application to São Paulo. *Regional Studies, Regional Science*, 8(1), 371-386. <https://doi.org/10.1080/21681376.2021.1968941>
- Smith, L., Beckman, R., Anson, D., Nagel, K., & Williams, M.E. (17-21 avril 1995). *TRANSIMS: transportation analysis and simulation system* [Communication de conférence]. Fifth National Conference on Transportation Planning Methods Applications, Seattle, Washington. <https://doi.org/10.2172/88648>
- Srinivasan, S., Ma, L., & Yathindra, K. (2008). *Procedure for forecasting household characteristics for input to travel-demand models* (Rapport TRC-FDOT-64011-2008) [Rapport technique]. University of Florida, Florida Department of Transportation, Gainesville, FL.
- Statistique Canada (2018). *Dictionnaire, Recensement de la population, 2016* (98-301-X2016001). <https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/98-301-x2016001fra.pdf>
- Statistique Canada (2019a). *Fichier de microdonnées à grande diffusion (FMGD) du Recensement de 2016, Fichier hiérarchique* (2016001) [Ensemble de données]. <https://www150.statcan.gc.ca/n1/fr/catalogue/98M0002X2016001>. La responsabilité liée aux calculs, à l'utilisation et à l'interprétation de ces données est entièrement celle de l'auteur ou des auteurs.
- Statistique Canada (2019b). *Profil du recensement, Recensement de 2016* (98-316-X2016001) [Ensemble de données]. <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=F>
- Su, M.D., Lin, M.C., Hsieh, H.I., Tsai, B.W., & Lin, C.H. (2010). Multi-layer multi-class dasymetric mapping to estimate population distribution. *Science of the Total Environment*, 408(20), 4087–4816. <https://doi.org/10.1016/j.scitotenv.2010.06.032>
- Sun, L., & Axhausen, K.W. (2016). Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation Research Part B: Methodological*, 91, 511–524. <https://doi.org/10.1016/j.trb.2016.06.011>
- Sun, L., & Erath, A. (2015). A Bayesian Network Approach for Population Synthesis. *Transportation Research Part C: Emerging Technologies*, 61, 49–62. <https://doi.org/10.1016/j.trc.2015.10.010>

- Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, 114, 199–212. <https://doi.org/10.1016/j.trb.2018.06.002>
- Swarup, S., & Marathe, M.V. (9-13 mai 2016). *Generating synthetic populations for social modeling* [Tutoriel]. 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Singapore. https://biocomplexity.virginia.edu/sites/default/files/staff/AAMAS_2016_generating_synthetic_populations_for_social_modeling_full_tutorial.pdf
- Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation*, 7(1), 4–25. <https://doi.org/10.34196/ijm.00092>
- Tanton, R., Vidyattama, Y., Nepal, B., & McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(4), 931–951. <https://doi.org/10.1111/j.1467-985X.2011.00690.x>
- Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of synthetic complex data: The R package simPop. *Journal of statistical software*, 79(10), 1–38. <https://doi.org/10.18637/jss.v079.i10>
- Thomson, D. R., Kools, L., & Jochem, W. C. (2018). Linking synthetic populations to household geolocations: a demonstration in Namibia. *Data*, 3(3), 30. <https://doi.org/10.3390/data3030030>
- Trépanier, M. (1999). *Modélisation totalement désagrégée et orientée-objet appliquée aux transports urbains* [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. <https://publications.polymtl.ca/8672/>
- Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33 (1), 213–239. <https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x>
- Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17(1), 33–51. <https://doi.org/10.1177/0962280207081238>
- Vidyattama, Y., Tanton, R., & Biddle, N. (2013). *Small area social indicators for the indigenous population: synthetic data methodology for creating small area estimates of indigenous disadvantage* (NATSEM Working Paper Series 13/24) [Rapport technique]. University of Canberra, National Centre for Social and Economic Modelling. <http://www.natsem.canberra.edu.au/files/download?id=1180>

- Ville de Montréal (2023a). *Adresse ponctuelle*. <https://donnees.montreal.ca/dataset/adresses-ponctuelles>
- Ville de Montréal (2023b). *Bâtiments*. <https://donnees.montreal.ca/dataset/batiment-2d>
- Ville de Montréal (2023c). *Unités d'évaluation foncière*. <https://donnees.montreal.ca/dataset/unites-evaluation-fonciere>
- Voas, D., & 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. [https://doi.org/10.1002/1099-1220\(200009/10\)6:5<349::AID-IJPG196>3.0.CO;2-5](https://doi.org/10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5)
- Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2), 177–200. <https://doi.org/10.1080/13615930120086078>
- Waddell, P. (2002). UrbanSim: modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3), 297–314. <https://doi.org/10.1080/01944360208976274>
- Ward, K. (2020). *ipfr: list balancing for reweighting and population synthesis* (Version 1.0.2.) [R package]. <https://CRAN.R-project.org/package=ipfr>
- Williamson, P., Birkin, M., & Rees, P.H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A: Economy and Space*, 30(5), 785–816. <https://doi.org/10.1068/a300785>
- Xie, Y., Weng, A., & Weng, Q. (2015). Population estimation of urban residential communities using remotely sensed morphologic data. *IEEE Geoscience and Remote Sensing Letters*, 12(5), 1111–1115. <https://doi.org/10.1109/LGRS.2014.2385597>
- Yaméogo, B.F., Gastineau, P., Hankach, P., & Vandanjon, P.O. (2021). Comparing methods for generating a two-layered synthetic population. *Transportation Research Record: Journal of the Transportation Research Board*, 2675(1), 136–147. <https://doi.org/10.1177/0361198120964734>
- Yaméogo, B.F., Hankach, P., Vandanjon, P.O., & Gastineau, P. (2023). Allocating synthetic population to a finer spatial scale: An integer quadratic programming formulation. *Environment*

and Planning B: Urban Analytics and City Science, 50(2), 515–540. <https://doi.org/10.1177/23998083221120019>

Ye, P., Tian, B., Lv, Y., Li, Q., & Wang, F.Y. (2022). On iterative proportional updating: limitations and improvements for general population synthesis. *IEEE Transactions on Cybernetics*, 52(3), 1726–1735. <https://doi.org/10.1109/tcyb.2020.2991427>

Ye, X., Konduri, K.C., Pendyala, R.M., Sana, B., & Waddell, P. (11-15 janvier 2009). *A methodology to match distributions of both household and person attributes in the generation of synthetic populations* [Communication de conférence]. Transportation Research Board 88th Annual Meeting, Washington, D.C. https://www.researchgate.net/publication/228963837_Methodology_to_match_distributions_of_both_household_and_person_attributes_in_generation_of_synthetic_populations

Yiu, T. (12 juin 2019). *Understanding random forest*. Towards Data Science. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

Zhang, D., Cao, J., Feygin, S., Tang, D., Shen, Z.J., & Pozdnoukhov, A. (2019). Connected population synthesis for transportation simulation. *Transportation Research Part C: Emerging Technologies*, 103, 1–16. <https://doi.org/10.1016/j.trc.2018.12.014>

Zhou, M., Li, J., Basu, R., & Ferreira, J. (2022). Creating spatially-detailed heterogeneous synthetic populations for agent-based microsimulation. *Computers, Environment and Urban Systems*, 91, 101717. <https://doi.org/10.1016/j.compenvurbsys.2021.101717>

Zhu, Y., Diao, M., Ferreira, J., & Zegras, P.C. (2018). An integrated microsimulation approach to land-use and mobility modeling. *Journal of Transport and Land Use*, 11(1), 633–659. <https://doi.org/10.5198/jtlu.2018.1186>

Zhu, Y., & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2429(1), 168–177. <https://doi.org/10.3141/2429-18>

Zhu, Y., & Ferreira, J. (2015). Data Integration to create large-scale spatially detailed synthetic populations. Dans Geertman, S., Ferreira, Jr.J., Goodspeed, R., & Stillwell, J. (édit.), *Planning Support Systems and Smart Cities. Lecture Notes in Geoinformation and Cartography*, (p. 121-141). Springer, Cham. https://doi.org/10.1007/978-3-319-18368-8_7

Zhuge, C., Li, X., Ku, C.A., Gao, J., & Zhang, H. (2017). A heuristic-based population synthesis method for micro-simulation in transportation. *KSCE Journal of Civil Engineering*, 21(6), 2373–2383. <https://doi.org/10.1007/s12205-016-0704-1>