Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Reproduire fidèlement des microdonnées structurées : l'exemple de la synthèse de données hiérarchiques

par Héloïse Gauvin

Date de diffusion : le 29 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Reproduire fidèlement des microdonnées structurées: l'exemple de la synthèse de données hiérarchiques

Héloïse Gauvin[1]

Résumé

La Directive sur le gouvernement ouvert du gouvernement du Canada vise à garantir aux Canadiens un accès accru aux données et à de l'information gouvernementales. Une solution pour les données ouvertes repose sur les fichiers synthétiques dits intelligents, qui conservent autant de valeur analytique que possible tout en prenant en compte les enjeux de confidentialité qui découlent de la collecte de renseignements personnels.

Dans les dernières années, Statistique Canada a acquis une expertise reconnue dans la production de fichiers de données synthétiques à grande valeur analytique. Dans un projet en cours, nous relevons un nouveau défi avec la synthèse d'une base de données préservant les structures hiérarchiques sous forme de familles, où les enregistrements sont liés et partagent des traits communs qui doivent être maintenus. Ce sont des défis que nous rencontrons également lors de la synthèse de données structurées telles que les données d'entreprises.

Cet article présente les défis et les solutions mises en place pour construire des données synthétiques avec de telles structures hiérarchiques. L'application de cette stratégie sera illustrée avec le développement d'une base de données synthétiques soutenant le développement de politiques relatives aux revenus de retraite. Cette base de données comprend plus de 20 variables et 8 millions d'enregistrements structurés en environ 4 millions d'unités familiales. Nous présenterons comment les structures familiales ont été préservées, nous discuterons des défis pratiques et techniques inhérents au développement d'une base de données aussi grande et complexe, nous discuterons du risque et de l'utilité des données, et nous présenterons des pistes de recherche futures.

Mots Clés : Données synthétiques à grande valeur analytique; structures familiales; solution moderne d'accès aux données.

1. Introduction

1.1 Données synthétiques

Puisque le terme 'données synthétiques' est utilisé assez librement, il est important de définir ce que sont les données synthétiques dans le cadre de ce projet. Les données synthétiques sont des données générées sciemment dans le but de préserver la valeur analytique d'un jeu de données originales tout en s'assurant que la confidentialité soit protégée. Par valeur analytique, on sous-entend que les mêmes conclusions statistiques seront obtenues à partir des données synthétiques ou du fichier original lorsque des analyses seront menées alors que ces analyses ne sont pas connues à l'avance. Le but sous-jacent des données synthétiques est d'avoir plus de données utiles, plus facilement accessibles par le public que les autres options d'accès aux données actuelles, tout en évitant la divulgation d'informations confidentielles. Le défi du développement des données synthétiques repose sur la capacité à préserver le contenu du fichier en termes de corrélations et d'interactions entre les variables, que nous ayons identifié ces relations au préalable ou pas.

[1]Héloïse Gauvin, Statistique Canada, 100 promenade du pré Tunney, Ottawa (Ontario), Canada, K1A 0T6 (heloise.gauvin@statcan.gc.ca).
Non-responsabilité : Le contenu de cet article présente la position de l'auteure, mais pas nécessairement celle de Statistique Canada.

1.1.1 Processus de synthèse des données

L'approche que nous avons choisie pour la synthèse des données repose sur une prémisse importante : le jeu de données d'origine n'est qu'un des innombrables jeux de données possibles, qui sont tous considérés comme des occurrences indépendantes de la même distribution statistique. Notons que d'autres approches existent et que cet article se concentre sur l'approche mise en place à Statistique Canada jusqu'à présent (Sallier, 2020). C'est un processus en deux étapes : 1) identifier la distribution statistique régissant les données originales et 2) utiliser le modèle obtenu pour générer aléatoirement un nouvel ensemble de coordonnées de données, c.-à-d. les données synthétiques.

1.2 Présentation et objectifs du projet

Statistique Canada développe présentement un nouveau modèle canadien de microsimulation des revenus de retraite en partenariat avec Emploi et développement social Canada et HEC Montréal (École des hautes études commerciales de Montréal). Ce projet vise à développer un modèle de simulations ouvert, que les chercheurs pourront exécuter depuis leur ordinateur personnel en utilisant une base de données ouverte et non confidentielle. La base de données originales consiste en une partie du questionnaire long du recensement 2016. Le modèle de microsimulation pourra être utilisé pour analyser les changements apportés à la politique du Régime de pensions du Canada (RPC). Notre objectif est de créer aux fins de ce projet une version synthétique de la base de données. Ce jeu de données synthétiques présentera une haute valeur analytique et sera diffusé à l'extérieur d'un environnement sécurisé.

2. Créer des données synthétiques avec une structure familiale

2.1 Méthode générale de synthèse des données

L'approche de synthèse des données adoptée pour ce projet est la spécification conditionnelle complète (SCC) (Drechsler, 2011). Cette méthode est celle qui a été utilisée par le passé dans de tels projets à Statistique Canada (Sallier, 2020). La SCC suppose que la valeur analytique du fichier original découle de la distribution conjointe des données. Pour préserver l'information, la SCC décompose la distribution conjointe multidimensionnelle en une série de distributions conditionnelles à une dimension. Si on considère un jeu de données avec *p* variables, cette approche se traduit par la formule suivante :

$$(1)\quad f_{X_1,\, X_2,\ldots,\, X_p} = f_{X_1} \times f_{X_2|X_1} \times \ldots \times f_{X_p|X_1,\, X_2,\ldots,\, X_{p-1}}$$

Essentiellement, au lieu d'essayer d'expliquer à la fois toutes les relations qui existent dans l'ensemble de données, la synthèse procède à leur capture séquentielle, une variable à la fois. En pratique, la méthode suit les étapes suivantes :

1. Prendre un échantillon aléatoire simple de $x_{1,obs}$ et le définir tel que $x_{1,syn}$
2. Ajuster un modèle $f(x_{2,obs}|x_{1,obs})$ et tirer $x_{2,syn}$ de $f(x_{2,syn}|x_{1,syn})$
3. Ajuster un modèle $f(x_{3,obs}|x_{1,obs}, x_{2,obs})$ et tirer $x_{3,syn}$ de $f(x_{3,syn}|x_{1,syn}, x_{2,syn})$
4. Ainsi de suite, jusqu'à $f(x_{p,syn}|x_{1,syn}, x_{2,syn}, \ldots, x_{p-1,syn})$,
   où $x_{i,obs}$ désigne la i^{e} variable observée et $x_{i,syn}$ désigne la i^{e} variable synthétique.

Cette approche est disponible dans la librairie R *synthpop* (Nowok et coll. 2016) qui est celle utilisée pour le projet. Les modèles choisis sont un mélange de modèles paramétriques et d'arbres de décision, selon la variable à modéliser. L'ordre dans lequel les variables sont synthétisées est important puisqu'il influence la qualité de la synthèse. Notons qu'il est préférable d'effectuer la synthèse des variables d'intérêt à la fin. Notons aussi que pour minimiser les risques de réidentification et faciliter la création de leur version synthétique, les données sont prétraitées (réduction de l'information présente à différents niveaux, voir section 4.2).

2.2 Processus de synthèse de données hiérarchiques

Jusqu'à présent, la synthèse de données était un processus plutôt simple, où les observations (ici des individus) étaient d'abord analysées pour générer ensuite des observations équivalentes mais synthétiques. Toutefois, dans le cadre de ce projet, la structure des observations (le fait que les observations ne soient pas toutes indépendantes les unes des autres) joue un rôle important pour les simulations. En effet, les revenus de retraite sont influencés par le fait d'avoir eu des enfants (clause pour élever des enfants) ou d'avoir un conjoint (séparation des revenus en cas de divorce, par exemple). La question est donc : comment conserver les corrélations entre les variables sachant que les individus eux-mêmes sont liés (présence de couples et de familles)?

Pour y arriver, nous séparons le problème en quatre types de structures familiales (1. les personnes hors famille de recensement (personnes vivant seules), 2. les couples sans enfants, 3. les couples avec enfants et 4. les familles monoparentales. Le processus de synthèse se fait en deux phases : nous générons d'abord les structures familiales (première synthèse) et ensuite nous appliquons le processus de synthèse pour chaque type de structure (deuxième synthèse). Autrement dit, nous divisons les données, générons les structures (création des familles) et ensuite nous remplissons ces structures (création des personnes).

Concrètement, pour la première phase, la synthèse se concentre sur cinq variables, soit 1) le type de structure familiale, 2) la taille de la famille, 3) l'âge du premier membre du couple, 4) la différence d'âge et 5) le type de couple (de sexe opposé ou de même sexe). Notons que les deux dernières variables sont présentes seulement s'il y a présence d'un couple. De plus, la taille de la famille est toujours de 1 pour les personnes hors famille de recensement et de 2 pour les couples sans enfants.

La deuxième phase dépend largement du type de structure devant être remplie. Pour les personnes hors famille de recensement, la première phase nous renseigne sur l'âge de l'individu. Toutes les autres caractéristiques de l'individu doivent être générées dans la deuxième phase. Les variables à synthétiser sont regroupées sous différentes catégories et les variables des différentes catégories sont synthétisées dans l'ordre suivant: les variables démographiques, d'immigration et de mobilité, de scolarité, de travail, de revenus et de prestations du régime de pension.

2.2.1 Les couples sans enfants

L'ordre de synthèse des données est le même pour les autres types de structures à quelques variations près. Pour les couples, la première phase fournit l'information sur l'âge du premier membre du couple, la différence d'âge entre les deux membres du couple et le type de couple. Par conséquent, l'âge qu'aurait le deuxième membre du couple est déduit de l'âge du premier membre et la différence d'âge entre eux. Ensuite, le type de couple nous informe sur le sexe de chacun des membres. Une fois ces variables déduites, nous poursuivons avec les autres variables, en alternant entre le premier et le deuxième membre du couple, pour toujours considérer l'information générée préalablement. Nous procédons ainsi puisque nous avons observé dans les données originales que les couples ne sont pas formés aléatoirement. Par exemple, les personnes immigrantes forment plus souvent un couple avec une autre personne immigrante; de la même façon, les données montrent qu'il est plus fréquent de voir une personne ayant obtenu un diplôme universitaire former un couple avec une autre personne ayant le même genre de profil. Autrement dit, qui se ressemble s'assemble et pour que les données synthétiques reflètent cet aspect, l'information du couple au complet est utilisée au fil de la synthèse. Concrètement, prenons l'exemple de l'année d'immigration du deuxième membre du couple. Nous ajusterons un modèle conditionnel sur les informations démographiques de cette personne, sur celle du couple, sur les variables de migration de la première personne déjà modélisées, sur celles de la deuxième personne et enfin sur la même variable (l'année d'immigration, s'il y a lieu) de la première personne.

2.2.2 Les couples avec enfants et les familles monoparentales

Pour les couples avec enfants et les familles monoparentales, la procédure est similaire. Dans les deux cas, nous générons d'abord les parents. Pour les couples avec enfants, nous procédons comme pour les couples sans enfants. Pour les familles monoparentales, le parent est traité comme une personne hors famille de recensement. Ensuite, nous

passons aux enfants pour lesquels toute l'information ou une partie est générée selon leur âge. Nous commençons par créer les enfants selon la taille de la famille, puis nous générons leurs âges et sexes. Ensuite, nous générons les autres variables en conditionnant sur les parents et en respectant les contraintes lorsqu'il y en a. Par exemple, certaines variables sont absentes ('NA') lorsque les enfants sont âgés de moins de 15 ans. Enfin, lorsque toutes les structures familiales sont complétées, elles sont combinées en un seul jeu de données.

3. Résultats

3.1 Vue d'ensemble des données

Rappelons qu'ici les données originales représentent un échantillon de la population canadienne, que certains enregistrements ont été retirés et que les données ont été prétraitées. Sachant que nous avons décidé de modéliser d'abord des unités familiales, les deux jeux de données (originales et synthétiques) présentent le même nombre de *familles* (4 236 024), mais le nombre d'*individus* est légèrement différent (8 651 677 pour les données originales et 8 655 913 pour les données synthétiques). Lorsque nous regardons la distribution des unités familiales selon les types de structures familiales, nous constatons que la distribution est similaire entre les données synthétiques et les données originales. Il en va de même si l'on observe la distribution du nombre d'individus dans chaque type de structure familiale (Tableau 3.1-1).

Tableau 3.1-1
Nombre d'individus dans chaque structure familiale pour les données originales et les données synthétiques

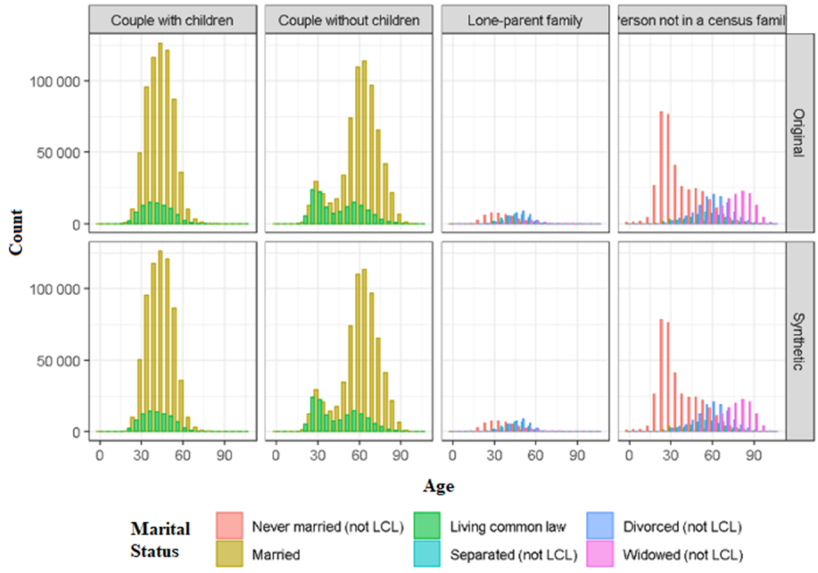
Individus dans chaque structure familiale	Données	
	originales	synthétiques
Personne hors famille de recensement	1 854 855	1 852 585
Couple sans enfants	2 198 240	2 201 314
Couple avec enfants	3 734 342	3 734 804
Famille monoparentale	864 240	867 210
Total	**8 651 677**	**8 655 913**

3.2 Variables synthétiques

Pour que l'on puisse considérer que le fichier synthétique soit de haute valeur analytique, il faut vérifier que les similitudes avec les données originales aillent plus loin que la structure générale du fichier. Plus spécifiquement, il faut s'assurer que les analyses sur les données originales et synthétiques mènent à des conclusions statistiques quasi identiques. Dans le cas présent, nous savons que les données seront utilisées afin de mener des simulations, mais il est difficile de savoir exactement quelles analyses seront conduites ainsi que leurs conclusions et donc quelles corrélations doivent être préservées. Nous serons capables de comparer les résultats lorsque le modèle de microsimulation sera prêt. En attendant, nous avons procédé à une première évaluation de la valeur analytique en s'assurant que les distributions univariées, bivariées et multivariées concordaient entre les données originales et les données synthétiques.

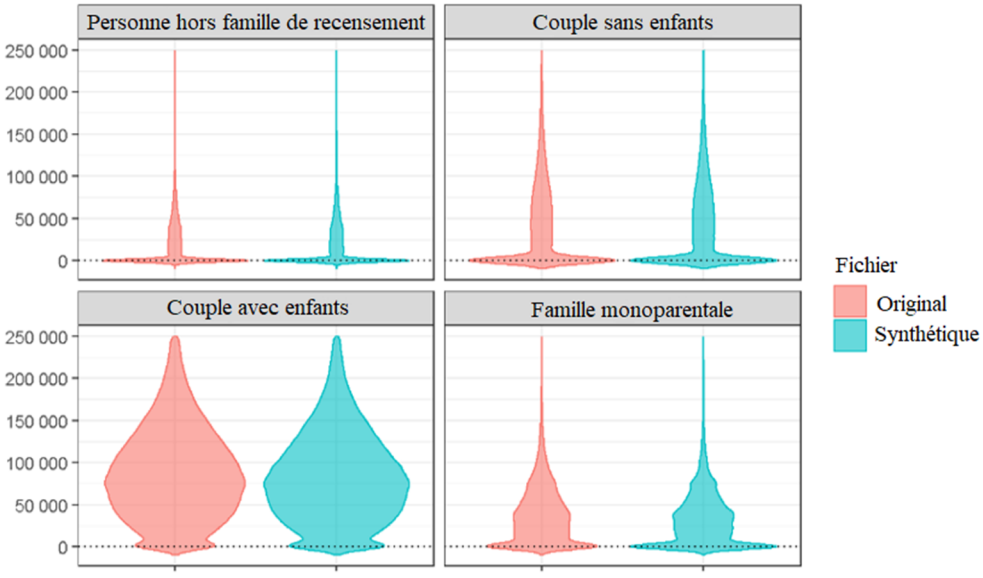
Pour illustrer cette démarche, nous présentons deux figures. D'abord, la figure 3.2-1 présente des histogrammes des âges pour chaque type de structures familiales, selon l'état matrimonial. Nous constatons que les distributions issues des données synthétiques semblent identiques à celles des données originales. Notons que les deux figures présentent des résultats pour une région uniquement.

Figure 3.2-1
Distribution des âges par type de structure familiale et selon le statut marital, données originales et données synthétiques



Puisque l'âge et l'état matrimonial sont parmi les premières variables à être modélisées, nous avons choisi de présenter également des résultats pour une variable survenant plus tard dans le processus de synthèse, soit le revenu (somme des revenus d'emploi des parents de la famille). La figure 3.2-2 présente le revenu familial par type de famille pour les données originales et synthétiques. Encore une fois, les distributions originales et synthétiques concordent. Notons que l'échelle de l'axe des revenus est coupée et présente seulement les valeurs de 0 à 250 000$.

Figure 3.2-2
Distribution des revenus familiaux par type de structures familiales, données originales et données synthétiques



Enfin, notons que nous avons aussi pris le soin de vérifier que les distributions des variables au sein des couples concordent. Tel que mentionné plus haut, les couples ne semblent pas formés au hasard dans la vie réelle et donc les données synthétiques se doivent de refléter ce fait.

4. Discussion

4.1 Utilité

Les résultats (comparaison des données originales et synthétiques à l'aide d'analyses univariées, bivariées, de statistiques descriptives telles que minimum, maximum, moyenne, médiane, écart-type, etc) démontrent que la valeur analytique semble bien conservée. L'utilité spécifique à ce projet sera confirmée lorsque nous pourrons comparer les résultats des microsimulations effectuées à partir des données originales et synthétiques. L'utilité des données synthétiques pourra même aller au-delà de ce projet, selon la qualité de celles-ci. Autrement dit, si la synthèse a bien été réalisée, les analyses possibles sur les données synthétiques seront illimitées.

4.2 Gestion du risque pour la confidentialité

Avant de parler du risque que pourraient poser les données synthétiques, notons que nous avons déjà mis de l'avant certaines approches classiques pour préserver la confidentialité des données. D'abord, nous avons agrégé plusieurs variables (par exemple la fréquentation scolaire est passée de 9 catégories à 3), nous avons réduit l'information géographique disponible et les variables monétaires ont été arrondies et bornées aux extrêmes. De plus, puisque les données proviennent du recensement long, elles sont issues d'un échantillon systématique d'un logement canadien privé sur quatre au Canada. Par conséquent, un enregistrement unique sur la base de données originales n'est pas nécessairement unique dans la population. Cela signifie que la reproduction d'enregistrements originaux n'est pas nécessairement problématique. Il faut prendre en considération leur unicité au sein de la population d'origine de même que les informations présentées pour mesurer le risque allégué.

4.3 Évaluation du risque pour la confidentialité

Pour tenter d'évaluer le risque posé par les données synthétiques, nous avons effectué un appariement exact sur toutes les variables. Cet appariement nous permet de constater que près de 70% des données synthétiques sont nouvelles (c'est-à-dire des enregistrements absents des données originales). L'autre partie (le 30% restant) consiste en des enregistrements identiques aux données originales. De ceux-ci, il y en a 10% qui sont des enregistrements uniques sur la base de données originales, ce qui nous mène à 3% du total des enregistrements. Enfin, de ces enregistrements, le tiers provient de personnes hors famille de recensement (≈1%), le reste fait partie d'une famille ou d'un couple. Il est important de le souligner, car en se penchant sur l'analyse des familles ou des couples, nous sommes arrivés au constat qu'aucun de ceux-ci n'est reproduit dans son entièreté. Autrement dit, un enregistrement peut sembler le même au niveau individuel mais lorsque nous considérons le reste de sa famille ou l'autre membre du couple, il n'est pas identique à l'original. Il reste à évaluer si cette fraction d'enregistrements (≈1%) reproduits à l'identique et qui sont rares sur les données originales peut causer problème en matière de confidentialité.

4.4 Prochaines étapes

L'analyse du risque pour la confidentialité n'est pas terminée. Nous cherchons maintenant à déterminer quel est le risque de ré-identification et comment quantifier l'unicité des enregistrements. Concrètement, nous cherchons à évaluer le risque lié à la capacité de déduire des informations si nous en connaissons déjà certaines et si certaines combinaisons de variables amènent un risque plus important. Notre attention est portée aussi sur le risque perçu. Nous souhaitons le minimiser et nous évaluons différentes stratégies pour y parvenir. Enfin, lorsque l'évaluation sera terminée, nous procéderons à la diffusion de la première version de la base de données synthétiques et nous aborderons des défis supplémentaires comme la synthèse de données historiques.

Bibliographie

Drechsler, J. 2011. *Synthetic Data Sets for Statistical Disclosure Control*. New York: Springer.

Nowok, B., Raab, G. M., and Dibben, C. 2016. *synthpop : Bespoke creation of synthetic data in R.* Journal of statistical software, 74: 1–26. [https://www.jstatsoft.org/article/view/v074i11]

Sallier, K. 2020. *Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis*, Statistical Journal of the IAOS, vol 36, no. 4, pp. 1059-1066.