Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Techniques de couplage d'enregistrements pour identifier les logements du Recensement canadien de 2021 dans le nouveau Registre statistique des immeubles

par Martin Lachance, Tanvir Quadir et Darryl Janes

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Techniques de couplage d'enregistrements pour identifier les logements du Recensement canadien de 2021 dans le nouveau Registre statistique des immeubles

Martin Lachance, Tanvir Quadir et Darryl Janes[1]

Résumé

La réconciliation des logements du recensement de 2021 avec le nouveau Registre statistique des immeubles (RSIm) a présenté des défis de couplage. Le Recensement de la population a recueilli des renseignements sur divers types de logements. Pour une grande partie de la population, les adresses postales, utilisées pour communiquer avec les gens et recueillies comme coordonnées, jouaient un rôle central. Parallèlement, l'environnement des registres a évolué. L'agence passe du Registre des adresses (RA) au Registre statistique des immeubles (RSIm), contenant les adresses postales et les adresses municipales, tout en couvrant les immeubles non résidentiels. La réconciliation a été effectuée à l'aide d'une combinaison de systèmes, notamment le nouveau Moteur d'appariement aux registres (MAR) pour les cas difficiles. Le MAR contient différents comparateurs de chaînes sophistiqués pertinents. Une méthode de couplage déterministe, tout en incorporant certaines connaissances sur les données comme l'entropie, fut employée. Grâce aux métadonnées, les experts en appariement ont aussi pu réduire le nombre de faux positifs et le nombre de faux négatifs.

Mots et phrases clés : métadonnées; comparateur de chaînes; faux positif; faux négatif.

1. Introduction

Tous les cinq ans, Statistique Canada recueille des renseignements au moyen d'un recensement. De plus, l'agence assure la mise à jour régulière d'un système de registres pour compléter les données du recensement. Le Registre des adresses (RA) (Statistique Canada, 2019) est l'un de ces registres servant à appuyer le recensement et les opérations d'enquête au sein de l'agence. Jusqu'à présent, il a servi à la collecte et l'entreposage des adresses et coordonnées des logements résidentiels privés (Statistique Canada, 2022) et collectifs (Statistique Canada, 2022). Après le Recensement de la population de 2021, le RA a été abandonné en faveur d'un nouveau registre : le Registre statistique des immeubles (RSIm) (Gagné et coll., 2018). Ces registres seront abordés à la section 3.

Au fur et à mesure de l'élaboration du RSIm, des renseignements sont attribués à chaque immeuble concernant son emplacement. L'existence de plusieurs types d'adresses au Canada complique l'entreposage et le couplage des adresses. La plupart des adresses, surtout dans les régions urbaines, sont de style adresse de voirie avec un numéro sur la rue et un suffixe, un nom de rue, le type de rue, la direction de la rue et le numéro d'appartement, s'il y a lieu (p. ex. « 203 - 12, rue Principale Nord »). Cependant, certaines adresses rurales utilisent d'autres méthodes, comme une description (p. ex. « maison blanche avec boîte aux lettres verte »), une intersection (p. ex. « Autoroute 4 et rue Cherry ») ou une borne kilométrique (p. ex. « km 1606, Autoroute de l'Alaska »). De nombreux propriétaires en région rurale fournissent comme coordonnées d'envoi un numéro de case postale (p. ex. « CP 205 ») ou un numéro de route rurale. Les adresses de bureau de poste posent un problème, car elles n'indiquent pas l'emplacement de la propriété et elles sont susceptibles de changer. Dans les provinces des Prairies du Manitoba, de la Saskatchewan, de l'Alberta et dans la région de la rivière de la Paix en Colombie-Britannique, certaines adresses rurales sont identifiées au moyen d'un système « quart de section, section, canton, rang et méridien » (QSTRM) (p. ex. « SW-21-24-8-W3 »).

[1]Martin Lachance, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, martin-p.lachance@statcan.gc.ca; Tanvir Quadir, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, tanvir.quadir@statcan.gc.ca; Darryl Janes, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, darryl.janes@statcan.gc.ca

Dans certaines collectivités du Grand Nord, les adresses peuvent être identifiées simplement par des numéros d'immeuble (p. ex. « édifice B42 ») sans nom de rue.

Le Recensement de la population de 2021 est une excellente source de mises à jour des adresses résidentielles. En effet, il couvre les régions rurales et urbaines de toutes les provinces et de tous les territoires. Après le lancement du RSIm, les logements du recensement furent considérés comme la première source de mises à jour de ce registre. La nouveauté du registre et celle du Moteur d'appariement aux registres (MAR) ont constitué des éléments clés du processus de réconciliation. Cet article traite brièvement du recensement canadien, du registre des adresses et du registre statistique des immeubles. Il présentera ensuite les outils et les méthodes de couplage utilisés pour effectuer la réconciliation. L'article s'intéressera davantage à la méthodologie qu'aux résultats.

2. Recensement de la population de 2021

Le Registre statistique des immeubles (RSIm) comprend les immeubles résidentiels et non résidentiels du Canada. Il nécessite des renseignements à jour sur les adresses aux fins de planification, de conception et d'exécution de nombreuses opérations statistiques menées par Statistique Canada. Il contribue à l'établissement des bases de sondage et sert à la collecte et à la validation des données. Parmi les sources servant à mettre à jour le RSIm, le recensement est considéré comme un contributeur important en raison de son incidence importante sur la couverture, en particulier pour les régions insuffisamment couvertes par les fichiers administratifs. Certains logements résidentiels qui doivent être représentés dans le RSIm ne peuvent pas être identifiés à partir de sources administratives. Ils ne sont trouvés que lors des opérations du recensement, surtout dans les régions rurales ou dans les petits immeubles à logements multiples. Dans le cadre du processus de mise à jour et d'actualisation du RSIm, le Recensement de la population de 2021 a été couplé au RSIm afin d'améliorer sa couverture et de mettre à jour les attributs des immeubles et des unités d'immeuble.

Aux fins du couplage entre le Recensement de 2021 et le RSIm, la liste des adresses des logements du recensement a été établie. Deux sources principales ont été combinées : le Système principal de commande (SPC) et la base de données des réponses (BDR). Le SPC est un ensemble d'adresses établi avant la collecte, initialisé à partir du Registre des adresses (RA) et qui comptait environ 98,53 % de toutes les adresses. La BDR est un ensemble d'adresses post-collecte contenant les réponses entrées par les répondants, par voie électronique ou sur papier, et couvrant 1,46 % des adresses supplémentaires. Notons que l'univers de contrôle et d'imputation (CI) du recensement de 2021 représente l'ensemble définitif valide de logements. Il confirme l'existence et la validité des adresses du SPC et de la BDR. Environ 0,01 % des cas étaient des cas marginaux qui n'appartenaient pas à l'une ou l'autre des sources (SPC ou BDR) : ils n'ont pas été pris en compte dans le couplage. L'image retenue du recensement contenait environ 16,7 millions de logements. Il existe trois catégories de logements du recensement : les logements privés (97,42 %), les logements collectifs primaires (0,23 %) et les logements associés aux logements collectifs (2,35 %).

Le mode de collecte du recensement est important, car il pourrait avoir une incidence sur l'intégralité et l'exactitude des adresses et, par conséquent, sur les stratégies de couplage et les taux de couplage. Trois principales méthodes de collecte ont été utilisées aux fins du Recensement de la population de 2021 : l'envoi par la poste (86,10 %), l'envoi par la poste et livraison à la porte (EPLP) (6,07 %) et des méthodes autres que l'envoi par la poste (listage/livraison, recensement par interview, réserves, sites saisonniers) (7,83 %) (Ha et coll., 2022). La définition de ces catégories est associée à la distribution du questionnaire dans une unité de collecte (Statistique Canada, 2022).

Il y avait des différences régionales au niveau des adresses. Au Canada, environ 95 % des adresses sont de type municipal. Les variations sont régionales ou se trouvent dans les régions rurales, notamment : QSTRM dans l'Ouest canadien, adresses descriptives souvent observées dans certaines provinces de l'Est et système de numérotation des bâtiments propre aux régions des territoires du Nord.

3. Les registres

Pour ce qui est des renseignements sur l'emplacement et l'adresse, Statistique Canada passe d'un registre des adresses (RA) à un registre statistique des immeubles (RSIm). Le RA a été utilisé par Statistique Canada pendant environ trois décennies. Au départ, il s'agissait d'utiliser des fichiers téléphoniques pour appuyer les enquêtes. Durant cette période, il a stocké divers renseignements sur les logements résidentiels et collectifs du Canada. Cela comprenait les adresses

postales et municipales, certains renseignements géographiques et les numéros de téléphone associés au logement. Il contenait les codes postaux et les noms des municipalités déterminés par la Société canadienne des postes. Le RA ne conservait aucun renseignement personnel sur les occupants ou le propriétaire d'un immeuble à l'adresse enregistrée.

Un identifiant d'adresse unique, appelé AR_UID, était attribué à chaque logement du RA. Le RA était un registre d'adresses. L'identifiant pouvait représenter un logement individuel, mais il aurait aussi pu désigner les unités individuelles d'un bâtiment à logements multiples, comme un immeuble d'appartements ou de condominiums. Un immeuble de 100 unités résidentielles, appartements ou bureaux aurait 100 identifiants sur le RA.

Jusqu'en 2021, le RA était régulièrement mis à jour avec de nouveaux renseignements et de nouvelles adresses dans les secteurs en croissance. Par sa nature évolutive, il jouait un rôle vital dans les opérations d'enquête de Statistique Canada. Le recensement canadien l'utilisait pour déterminer les méthodes de collecte et les adresses des questionnaires à distribuer. Après chaque recensement, le RA était réinitialisé avec le contenu du recensement.

Avec l'évolution de la technologie, le RSIm a été conçu pour remplacer le RA en raison du besoin croissant d'information et d'avoir de meilleurs moyens de couplage avec d'autres registres de Statistique Canada, comme le Registre des entreprises (RE). Le développement du RSIm a commencé avant le Recensement de 2021. Il sera le registre central de renseignements sur les immeubles et les adresses pour les enquêtes à venir après le Recensement de 2021, ainsi que pour la préparation du Recensement de la population de 2026.

Contrairement au RA qui emmagasinait l'information en fonction d'adresses, le RSIm a été conçu pour stocker l'information concernant une structure physique de bâtiment, tout en contenant l'information du RA. En fait, près d'un an avant le jour du recensement de 2021, le RSIm a été initialisé au moyen d'un choix sélectif de fichiers administratifs, dont la plupart étaient déjà utilisés dans les mises à jour du RA. La version initiale du RSIm a été couplée au RA pour évaluer sa couverture. Le RSIm contient le nom, le type et la taille d'un immeuble. Le registre contient également des attributs géographiques liés à l'emplacement des immeubles, comme les géographies du recensement (Statistique Canada, 2022), les coordonnées et la géographie QSTRM, s'il y a lieu. L'autre avantage du RSIm est qu'il contiendra les immeubles résidentiels, les immeubles à vocation commerciale et ceux à usage mixte, tandis que le RA se limitait aux adresses résidentielles.

Chaque immeuble comportera une ou plusieurs unités. Par exemple, un immeuble d'appartements contenant un bureau et 20 appartements comptera 21 unités d'immeuble. Le RSIm emmagasine des renseignements sur l'unité d'immeuble, comme l'adresse, les coordonnées contact connues et son statut actif.

La transition du RA au RSIm a été facilitée par les données du Recensement de 2021. Premièrement, une réconciliation classique entre les adresses des logements du recensement et le RA a été effectuée. Une base formée de l'ensemble de données contenant des renseignements du RA et du recensement a ensuite été créée comme d'habitude, même si le RA ne devait pas être réinitialisé. Par la suite, on a effectué la réconciliation des logements du Recensement de 2021 avec le RSIm. Les logements du recensement ont été couplés directement au RSIm, tout en utilisant leur connexion au RA pour tirer parti de la réconciliation RA-RSIm. Le processus de réconciliation entre les logements du Recensement de la population de 2021 et le RSIm est présenté dans les sections suivantes.

4. Réconciliation

Il y a deux raisons importantes de réconcilier les logements du Recensement de la population de 2021 au Registre statistique des immeubles (RSIm). La première est de veiller à ce que le RSIm offre une excellente couverture des immeubles résidentiels et des unités d'immeuble partout au Canada. La deuxième est que les utilisateurs du recensement et des enquêtes ont besoin du meilleur lien possible entre les logements du recensement et le RSIm.

Le couplage de réconciliation a été effectué par province ou territoire, en tenant compte des divers systèmes d'adressage utilisés à l'échelle du Canada mentionnés précédemment. Une fois le couplage entre les logements du recensement et le RSIm terminé, les logements du recensement avec suffisamment de renseignements sur leur emplacement et qui n'ont pu être trouvés sur le RSIm ont été « nouvellement créés » ou ajoutés temporairement comme nouveaux enregistrements sur le RSIm en attendant leur confirmation. Les logements du recensement qui n'avaient pas assez d'information et qui n'ont pu être trouvés ont dû suivre un processus visant à déterminer le meilleur

emplacement géographique possible (province, ville ou niveau inférieur). Cette étape est effectuée parce que les utilisateurs du recensement ont toujours besoin de savoir, au minimum, où se trouvent ces logements. Cet article ne décrit pas ce processus d'identification des emplacements géographiques.

4.1 Les systèmes

Statistique Canada utilise deux systèmes principaux pour coupler les adresses au RSIm, soit le Moteur de traitement des adresses (MTA) et le Moteur d'appariement aux registres (MAR).

Depuis de nombreuses années, le MTA sert à effectuer des couplages d'adresses tirées de données administratives avec le Registre des adresses (RA). Le système met l'accent sur le prétraitement intensif des données (c.-à-d. l'analyse syntaxique, la normalisation) et le couplage vise principalement à parfaitement apparier les attributs nettoyés. Les couplages sont effectués par identificateurs de rue, comme le nom de rue, son type et sa direction, les numéros de voirie et les suffixes, les numéros d'appartement ou d'unité, les noms de municipalité et les codes postaux aux fins d'envoi par la poste, les géographies du recensement, comme les unités de collecte et les subdivisions de recensement, et les numéros de téléphone. Les numéros de téléphone n'ont pas été utilisés dans le processus de réconciliation du RSIm quand le MTA était utilisé aux fins de couplage.

Le MTA est un système de procédures SAS qui identifie d'abord les composantes d'une adresse et les sépare au besoin. Par exemple, « Unité 201, 129A Saddle Road Est, Red Fish, Alberta, T8G7K9 » serait analysé selon ses composantes, à savoir Appartement = « 201 », NuméroDeVoirie = « 129 », SuffixeDeVoirie = « A », NomDeRue = « Saddle », TypeDeRue = « Road », DirectionDeRue = « Est », Municipalité = « Red Fish », Province = « Alberta » et CodePostal = « T8G7K9 ». Il normalise ensuite les composantes (p. ex. « Road » devient « RD » et « Alberta » devient « AB »). De plus, il applique une transformation du nom de rue et crée une variable dérivée pour faciliter l'appariement. Ce nom de rue transformé et codé est appelé RASK (pour *Road Attribute Search Key*) ou clé de recherche d'attribut de route.

Après la transformation de l'adresse, le MTA tente itérativement d'apparier les composantes de l'adresse à la base de données PCODE qui contient les renseignements de la Société canadienne des postes, ainsi qu'aux renseignements de la Base nationale de données géographiques (BNDG) de Statistique Canada. Si l'emplacement de l'adresse est trouvé, les attributs géographiques du recensement deviennent connus pour cet enregistrement. Enfin, l'adresse et ses attributs géographiques sont appariés au RA. Le résultat final contient l'adresse segmentée, divers renseignements concernant la géographie et les rues, les indicateurs de qualité des appariements et l'identifiant AR_UID.

Les adresses de voirie peuvent prendre diverses formes. Les abréviations, la ponctuation, les autres noms de municipalité ou de rue, et les façons d'indiquer l'unité ou l'appartement peuvent causer des difficultés à l'analyse syntaxique. Malgré cela, le MTA réussit très bien à traiter la plupart de ces types d'adresses. Il traite également très bien les adresses avec code postal.

Le MTA est un outil précieux pour relier les adresses de voirie au RSIm, en particulier les adresses qui se trouvent dans la base de données PCODE. Il a été utilisé pour effectuer très rapidement le couplage de la plupart des adresses du Recensement de 2021 avec le RSIm. Malheureusement, le MTA ne réussit pas aussi bien à traiter d'autres types d'adresses, comme les adresses descriptives, la géographie QSTRM ou les adresses de collectivités nordiques. Il est également en difficulté quand les éléments d'adresse comportent des fautes d'orthographe, des troncations ou des mots supplémentaires (p. ex. « 129, chemin Saddle, rez-de-chaussée »). Un deuxième outil, appelé *Moteur d'appariement aux registres*, a ensuite été utilisé pour effectuer la réconciliation des cas restants qui ne pouvaient pas être couplés au moyen du MTA.

Le MAR est un système résidant dans l'infonuagique de Statistique Canada. Il a été spécialement conçu aux fins de couplage de données administratives avec le système de registres, y compris le RSIm. Il est doté d'une boîte à outils de comparateurs de chaînes sophistiqués avec efficiences intégrées pour trouver des paires d'appariement. Il comporte également des outils axés sur les métadonnées pour contrôler le nombre de paires faussement positives et faussement négatives.

L'un des avantages du MAR par rapport au MTA est qu'il ne dépend pas des bases de données PCODE ou de la BNDG pour effectuer le couplage. En effet, il peut coupler les adresses directement au RSIm sans se préoccuper de l'état postal, ce qui en fait un outil très utile pour les couplages d'adresses municipales. Le MAR peut traiter les adresses postales comme les adresses municipales. Il a servi à traiter les cas d'adresses de voirie que le MTA ne pouvait pas coupler, en particulier les adresses absentes des bases de données PCODE ou de la BNDG, ou qui pouvaient contenir des éléments susceptibles d'être mieux traités par une stratégie d'appariement plus robuste et la boîte à outils des comparateurs. La section 4.2 traite du MAR plus en détail.

4.2 Le moteur d'appariement aux registres

Le Moteur d'appariement aux registres (MAR) utilise comme point de départ le progiciel de couplage déterministe hiérarchique SAS MixMatch (Lachance, 2014). MixMatch a été développé à l'interne à Statistique Canada. Au départ, il s'agissait d'un projet de recherche conçu pour traiter divers types de couplages, notamment pour les personnes, les adresses ou les entreprises. Il a été utilisé dans des couplages pour les bases de sondage de l'agriculture et des listes de propriétaires, entre autres. La force du logiciel réside dans ses comparateurs de chaînes sophistiqués et une approche axée sur les métadonnées pour maximiser le couplage, tout en prévenant la formation de faux liens causés par la présence de noms légitimes qui se ressemblent. Cependant, MixMatch n'a pas été conçu pour traiter de grands volumes de données. Le choix de la technologie a donc été étudié. Pour coupler les adresses, il a fallu trouver une méthode de couplage déterministe qui utilise certaines des caractéristiques avantageuses du MTA, tout en palliant ses carences. Le MAR est conçu en Java et s'exécute sur l'infonuagique. La plupart des idées appliquées dans MixMatch ont été retenues pour le MAR et la méthodologie du logiciel continue d'évoluer en fonction des besoins des utilisateurs.

Les principales étapes de tout couplage sont le prétraitement des données, l'identification des liens potentiels (c.-à-d. les paires candidates) entre deux sources de données et le post-traitement des liens trouvés. Le post-traitement consiste à résoudre les liens trouvés en conservant les liens corrects selon une série de critères, qu'une approche de couplage déterministe ou probabiliste ait été employée ou non. Le MAR a été conçu ainsi.

La boîte à outils de comparateurs de chaînes du MAR contient des comparateurs très connus comme la distance de Levenshtein (Navarro, 2001), la distance de Jaro-Winkler (Jaro, 1989; Winkler, 2006) et *la plus longue sous-séquence commune*. Cette dernière sert de base à des comparateurs plus sophistiqués. Quelques comparateurs de variables numériques ont été progressivement ajoutés, notamment la distance euclidienne nécessaire pour calculer les distances géographiques rapprochées lorsqu'on traite des coordonnées dans la projection de Lambert. Il est aussi possible de convertir des coordonnées de latitude et de longitude et employer ces valeurs converties dans le comparateur. Les comparateurs multi-mots sont une des forces du MAR. Le MAR contient actuellement des familles de comparateurs stricts et de comparateurs souples. Les comparateurs stricts fournissent probablement des liens de meilleure qualité en réduisant le nombre de faux positifs au détriment de l'augmentation du nombre de liens manqués (faux négatifs). Les comparateurs souples fourniront plus de liens, au risque que certains liens soient des faux positifs. Les comparateurs multi-mots se sont déjà révélés très utiles dans le traitement d'adresses, en particulier quand elles ne sont pas segmentées, incorrectement segmentées ou qu'elles contiennent des renseignements superflus.

Chaque comparateur donne un score quantitatif entre 0 et 1, ainsi qu'un score descriptif appelé P-R-I-C-E. La partie P-R-I indique le niveau de concordance dans l'ordre (Parfait appariement, Ressemblance autrement, Inclusion autrement), tandis que les lettres C-E (Conflits et mots en Excès) indique le niveau de non-concordance une fois l'algorithme terminé. Le score P-R-I-C-E est très utile pour accepter ou rejeter des liens et les départager dans les couplages d'adresses.

De plus, la mise en œuvre et l'utilisation du MAR reposent sur une méthode axée sur les métadonnées. Pour le prétraitement, l'utilisateur du MAR peut fournir des tables utilitaires pour exprimer une série de transformations. Par exemple, une normalisation de type de rue peut se produire si « STREET » devient « ST ». De plus, à l'étape où des liens potentiels sont détectés, l'utilisateur peut se servir de son expertise pour établir la stratégie de couplage du MAR. Cela peut jouer un rôle crucial lors de l'appariement d'une variable de nom. Le concept de « table d'exclusion » de MixMatch est utilisé dans le MAR pour « exclure » au préalable les faux liens. L'utilisateur indique une liste de paires de valeurs qui se ressemblent, mais pour lesquelles il ne veut pas de lien. Donnons (Martin, Marvin) comme exemple de ce type de paires quand on essaie d'apparier « Marvin Street » à « Martin Street ». Nous sommes en présence de deux noms de rue légitimes et non de fautes d'orthographe. À l'inverse, l'utilisateur peut vouloir lier deux valeurs

essentiellement équivalentes mais qui ne se ressemblent pas. L'utilisateur informe alors le MAR qu'il doit « inclure » la paire de valeurs dans le couplage au moyen d'une « table d'inclusion ». Par exemple, une région rurale pourrait être définie par deux noms très différents. La fusion des municipalités en est un autre exemple, quand l'ancien nom continue d'être en usage. Pour obtenir un appariement, l'utilisateur doit indiquer la paire de noms équivalents dans une table d'inclusion. Cette approche fait diminuer le nombre de faux négatifs, soit le nombre de liens potentiellement corrects qui sont rejetés. Il est aussi possible d'élargir les tables d'inclusion à des variables d'appariement de nature différente, comme un nom de ville avec un code postal rural. Dans ce cas, un appariement de valeurs indiquerait que les adresses se trouvent dans la même zone. Dans ce type d'utilisation, l'utilisateur doit interpréter le lien avec prudence.

Tableau 4.2-1
Comparaison d'informations complètes sur la rue au moyen d'un comparateur multi-mots

		Information complète sur la rue				
Paire d'enregistrements	Source	Nom de la rue	Type de rue	Dir. de rue	Score MAR	MAR P-R-I-C-E
1	Recensement	52328, RANGE RD 234 BOX 156	ROAD		0,60	P=3 (RANGE, ROAD, 234), R=0 (aucun appariement de coquille), I=0, C=0, E=4 (52328, RD, BOX, 156)
1	RSIm	Range Road 234				
2	Recensement	Blue Ribon Lane by the Lake			0,65	P=2 (Blue, Lane), R=1 (Ribon), I=0, C=0, E=3 (by, the, Lake)
2	RSIm	Blue Ribbon	LANE			

L'identification des liens potentiels repose sur une série de règles. Par exemple, nous pourrions chercher un appariement parfait sur le numéro de voirie, un appariement partiel sur le nom de rue, le type et la direction, ou un appariement parfait sur la ville. Pour chaque condition d'une règle comportant une variable administrative et une variable du RSIm (p. ex. nom de rue, type et direction combinés), le MAR applique la même séquence de vérifications. Premièrement, s'il manque une valeur d'un côté, une décision est prise sur la condition (réussite/échec). S'il ne manque aucune valeur, le MAR vérifie si une table d'exclusion est utilisée. Dans ce cas, si nous avons une occurrence sur une paire de valeurs dans la table, alors la condition échoue. Si aucune table de ce type n'est utilisée ou si nous n'avons pas d'occurrence sur une paire de valeurs, le MAR applique le comparateur sélectionné par l'utilisateur. Il se peut que le comparateur traite les coquilles et utilise une table d'inclusion. Par exemple, l'utilisation d'un comparateur multi-mots pourrait coupler des rues pour chaque paire d'enregistrements indiquée dans le tableau 4.2-1. Par ailleurs, le MAR est efficace dans l'utilisation d'index sur le RSIm pour traiter les appariements parfaits. Le Recensement de la population de 2021 comportait des adresses présentant divers types de problèmes : renseignements mal placés, erreurs de numérisation pour les questionnaires papier, information partielle, etc. Le MAR a été précieux dans le traitement des cas plus difficiles.

4.3 Le processus de couplage

La réconciliation des logements du Recensement de 2021 avec le RSIm a été effectuée au moyen d'une approche agile. Dans la première vague, une grande partie des logements ayant une adresse de voirie ont été liés au RSIm. Un premier produit donnant une image du contenu du RSIm a ensuite été mis à la disposition des utilisateurs. Puis, dans une deuxième vague ayant mené à un second produit, des techniques plus compliquées ont été employées, avec l'aide du MAR, pour essayer de trouver les logements avec adresse de voirie qui n'avaient pas encore été appariés au RSIm. La deuxième vague s'est également occupée des autres types d'adresses, autres que de voirie (p. ex. QSTRM, Nord, adresse descriptive), et des logements collectifs (p. ex. résidences pour personnes âgées, prisons, etc.). L'article présente ici seulement la résolution des adresses de voirie.

La première vague consistait en une combinaison de méthodes, qui utilisait à la fois le MTA et le MAR. Le MTA a d'abord été utilisé pour essayer de lier rapidement le plus grand nombre possible de logements au RSIm. À la vague 1, les liens directs entre les logements du recensement et le RSIm ont été trouvés, mais certains liens indirects ont également été trouvés. Les liens directs consistaient à coupler les adresses directement aux adresses du RSIm, tandis que les liens indirects consistaient à établir des liens au RSIm par l'intermédiaire du RA. Des liens indirects étaient possibles en raison du lien entre le recensement et le RA obtenu à partir de la réconciliation classique entre les adresses

des logements du recensement et le RA, ainsi qu'en raison du lien entre le RA et le RSIm tiré de l'initialisation du RSIm. Dans le processus de résolution, les liens directs étaient jugés de meilleure qualité que les liens indirects.

Parmi les méthodes de couplage direct, le MTA a servi à lier les adresses normalisées des logements du recensement au RSIm afin d'essayer d'identifier les unités d'immeuble. Pour certains enregistrements n'ayant pu être liés en raison de problèmes de segmentation de l'adresse, les adresses brutes furent utilisées lors d'une deuxième exécution du MTA. Les liens directs comprenaient également l'utilisation du numéro de téléphone sur le RSIm au niveau de l'unité d'immeuble. Quand ce numéro appariait à un logement du recensement, la géographie était comparée entre le recensement et le RSIm. Une stratégie de couplage utilisant le MAR complétait le portrait des liens directs au RSIm. Cette stratégie fut appliquée aux logements n'ayant pas été couplés lors de la deuxième exécution du MTA, afin de tirer parti des outils sophistiqués du MAR pour traiter les cas plus difficiles. Les autres méthodes de couplage employées ont fourni des liens indirects. Certaines méthodes utilisaient les numéros de téléphone et d'autres comprenaient des couplages supplémentaires réalisés à l'aide du MTA. Ces couplages furent considérés de deuxième, troisième et quatrième niveau de qualité. Les adresses de Postes Canada provenant des fichiers de Point de remise (PDR) nourrissaient le RA et le RSIm. Ces adresses ont leur propre identifiant, stable dans le temps, renforçant la connexion AR-RSIm obtenue par couplage. Au total, six méthodes de couplage indirect furent employées.

Dans le processus de résolution de la vague 1, chaque lien individuel a été classé d'abord en fonction du nombre de méthodes de couplage fournissant le lien, puis en fonction de la qualité du lien. Après le classement, une décision a été prise sur le lien du haut de la liste. Un lien de qualité 1 trouvé par au minimum deux méthodes était considéré comme un lien solide. Les liens plus faibles nécessitaient des vérifications supplémentaires. Pour ce faire, l'identification du bon immeuble était la première étape. Ensuite, les liens dans lesquels l'immeuble pouvait être confirmé faisaient l'objet d'un processus de validation semblable qui cherchait à confirmer l'unité d'immeuble. Dans certains cas, seuls les liens à l'immeuble pouvaient être validés. Tout au long de ce processus de validation en deux étapes, les doublons possibles sur le RSIm ont été détectés au moyen de liens multiples entre les logements du recensement et les immeubles du RSIm. Ils furent mis de côté pour être examinés. Des paires de doublons possibles du côté du recensement furent également repérées, par exemple, quand plusieurs logements du recensement étaient fortement liés à une même unité d'immeuble du RSIm. Cependant, contrairement à celles du RSIm, les données du recensement sont figées dans le temps et il était donc seulement possible de signaler les cas potentiels de logements du recensement en double. La résolution de liens de « plusieurs à un » et de « un à plusieurs » nécessitait des règles complexes, qui tenaient compte du fait que certaines rues sont connues sous plusieurs noms ou sous des pseudonymes. Plus de 15 millions d'adresses de voirie de logements du recensement ont été couplées à une unité d'immeuble au cours de la vague 1. Un peu plus d'un demi-million de logements du recensement en attente ont ensuite été traités par la vague 2.

La deuxième vague de réconciliation a été effectuée uniquement au moyen du MAR. Cette vague ciblait les adresses de voirie les plus difficiles. C'est pourquoi il fallait des stratégies de couplage plus élaborées que celles de la vague 1. Cette fois-ci, une méthode en deux étapes a été employée. Au lieu d'établir un lien direct avec les unités d'immeuble immédiatement, on a d'abord essayé de trouver le bon immeuble, puis on a utilisé une autre stratégie de couplage pour trouver l'unité appropriée à l'intérieur de l'immeuble.

Après avoir trouvé les liens entre logements du recensement et immeubles du RSIm, il a fallu résoudre ces liens, sans aucun doute la partie la plus difficile du processus. Des règles furent appliquées, en s'appuyant sur plusieurs variables, chacune ayant différents degrés de précision d'appariement.

Premièrement, pour chaque lien obtenu, un score fut attribué à chaque paire de variables comparées (p. ex. les noms de rue du recensement et les noms de rue du RSIm) au moyen d'un comparateur et un score P-R-I-C-E était obtenu. Par exemple, le lien n^{o} 1 pouvait obtenir un score de 0,67 pour la comparaison des noms de rue, tandis que le lien n^{o} 2 pouvait également obtenir un score de 0,67, mais avec un score P-R-I-C-E différent. À l'aide de ce descripteur, nous pouvions rejeter les liens comprenant des Conflits, qui sont principalement de mauvais liens (p. ex. la « **rue** Blue Jays » comparativement à la « **voie** Blue Jays »). Lorsque nous étions beaucoup plus stricts dans l'acceptation des paires, nous pouvions également rejeter les liens ayant le plus de bruit, qui sont représentés par les comptes dans le R-I du P-R-I-C-E. Cela était particulièrement vrai si la table d'exclusion était significativement incomplète et entraînait un couplage de fausses coquilles (p. ex. « rue **Annie** » ou « rue **Angie** »). Enfin, les mots en Excès peuvent représenter des précisions supplémentaires sur les composantes de l'adresse et avoir ainsi un effet négatif minimal. Ils ont été

acceptés dans les appariements de noms de rue complets. Le score quantitatif (ici, 0,67) était calculé au moyen d'un ratio de la portion de concordance pondérée (P-R-I) sur les composantes P-R-I-C-E pondérées.

Bien qu'une approche déterministe ait servi à coupler les logements du recensement au RSIm, une certaine connaissance des données a été introduite. Cela a joué un rôle dans la deuxième vague. Chaque variable entrant dans une règle de résolution a été pondérée au moyen d'une mesure d'*entropie*, laquelle est un attribut de variable. Une variable ayant une grande diversité de valeurs a une grande puissance discriminante, ce qui se traduit par une entropie élevée. Les variables à forte entropie ont été utilisées dans les règles d'appariement. Notons toutefois que les erreurs d'orthographe peuvent augmenter le nombre de valeurs distinctes pour une variable donnée, ce qui gonfle l'entropie. Afin de minimiser ce biais, l'entropie des variables est calculée à partir du RSIm, qui est considéré comme la référence la « plus propre ». De plus, la fréquence d'une valeur en particulier prise par une variable doit être utilisée en combinaison avec l'entropie. Par exemple, la variable de nom de rue a une entropie élevée. Toutefois, certaines valeurs comme « Main », dans « Main Street », ont des fréquences élevées parce qu'il y a une rue principale dans la plupart des municipalités. Pour la valeur « Main », la puissance discriminante du nom de rue apportée par l'entropie est perdue. Malheureusement, les fréquences n'ont pas pu être mises en œuvre à temps pour la deuxième vague de réconciliation. Seule l'entropie a pu être utilisée.

Après avoir tenu compte de tout ce qui précède, la force d'un lien pour les adresses de voirie reposait principalement sur trois éléments clés : (1) un appariement parfait sur le numéro de voirie, (2) un appariement partiel sur l'information de la rue et (3) un appariement sur au moins une géographie. Une série de règles furent ensuite appliquées, classant les liens des plus forts aux plus faibles. Enfin, une résolution des liens multiples fut menée, en tenant compte de la force du lien.

Après confirmation d'un lien à un immeuble du RSIm, nous avons tenté de trouver l'unité appropriée à l'intérieur de l'immeuble. La première tentative consistait à trouver un appariement parfait sur les numéros d'appartement. Malheureusement, comme on pouvait s'y attendre, cette opération n'était pas très réussie, car il s'agissait des cas les plus difficiles. Les petits immeubles à logements multiples sont particulièrement difficiles à résoudre. Il s'agit habituellement de bâtiments qui ont été subdivisés en unités plus petites. Par exemple, le numéro de voirie « 20 » est devenu 20 et 20A, ou 20A et 20B, et ainsi de suite. Souvent, ces unités ont tendance à ne pas figurer dans les fichiers administratifs. Il semblerait que la plupart soient trouvées uniquement via les opérations de recensement sur le terrain. Elles sont également très difficiles à résoudre par un processus automatisé. Idéalement, pour des cas aussi complexes, si nous avions eu des renseignements sur les personnes de l'unité d'immeuble du RSIm, nous aurions peut-être pu coupler ces logements du recensement aux unités d'immeuble du RSIm. Nous n'avons cependant pas pu le faire à l'étape de réconciliation. Un autre problème est attribuable au mauvais ordonnancement des suffixes de voirie ou des numéros d'unité ou d'appartement. Par exemple, si le « A » dans « 20A » est un suffixe de voirie, alors l'adresse représente un immeuble sur le RSIm. S'il s'agit d'une unité, l'adresse représente une unité dans un immeuble. Nous avons donc élaboré une stratégie spéciale en trois étapes pour traiter ces cas. Premièrement, nous avons combiné le numéro de voirie, le suffixe de voirie et le numéro d'unité en une seule variable. Nous l'avons fait des deux côtés : pour le recensement et pour le RSIm. Ensuite, nous avons distribué l'information sur deux nouvelles variables, pour les deux : les chiffres seulement dans une première variable, le reste dans l'autre. Enfin, nous avons exigé un appariement sur les variables de type caractère et numérique, au moyen de comparateurs multi-mots pour traiter les éléments d'information supplémentaires. Cette méthode a permis d'établir seulement 10 000 liens au niveau des unités d'immeuble sur les 66 700 nouveaux liens obtenus au niveau de l'immeuble, mais notre confiance dans la qualité des liens était élevée.

5. Conclusion

La transition simultanée au nouveau Registre statistique des immeubles (RSIm) et à un nouveau système de couplage, le Moteur d'appariement aux registres (MAR), a présenté des défis et des possibilités d'apprentissage, mais a surtout permis l'élaboration de techniques qui améliorent la qualité de la base du RSIm. Premièrement, dans toute élaboration de stratégie de couplage, il est essentiel d'explorer et de comprendre les données. Les logements du Recensement canadien de 2021 contenaient une diversité d'adresses qui ne figuraient pas dans des fichiers administratifs. Il était également d'importance capitale de comprendre la structure des données du RSIm. Deuxièmement, les couplages comportant des adresses peuvent être beaucoup plus complexes que les couplages concernant des entreprises, lesquels sont eux-mêmes considérés comme plus difficiles à traiter que les couplages au niveau de la personne. C'est la plus

grande leçon que nous avons tirée du processus de réconciliation. Pour les couplages au niveau de la personne, on dispose généralement de plusieurs attributs de personne et d'autres variables, y compris les composantes de l'adresse ou un résumé de l'information de localisation. Concernant les couplages d'entreprises, le manque de variables est fréquent : souvent, on dispose seulement du nom de l'entreprise, sans compter les répétitions dues à la structure organisationnelle multiniveau des entreprises. Au Canada, une adresse s'exprime de multiples façons et elles détiennent le plus grand nombre de « noms ». De fait, pour tous les types de couplages, les « noms » sont les variables représentant les plus grands défis. Au minimum, il existe deux noms lorsqu'on traite une adresse au Canada : le nom de rue et le nom de municipalité. Des obligations juridiques à l'échelle provinciale et municipale empêchent heureusement les répétitions. Pour conclure, le MAR nous a fourni des outils souples et puissants pour coupler les adresses, en particulier les noms, notamment grâce aux comparateurs multi-mots et à une approche axée sur les métadonnées qui contrôlaient le nombre de faux positifs et de faux négatifs.

Bibliographie

Gagné, P., Pignal, J., Quadir, T. et C. Wolfe C. (2018), « Towards a Register-Centric Statistical System: Recent Developments at Statistics Canada », document présenté au Symposium international sur les questions de méthodologie de 2018 de Statistique Canada, Ottawa, Canada.

Ha, B., Lee, M., Mayda, M. et J. Wong. (2022), « 2021 AR Snapshot and Reconciliation Project », rapport non publié, Ottawa, Canada : Division de l'infrastructure et d'intégration des données, Statistique Canada.

Jaro, M. A. (1989), « Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida », *Journal of the American Statistical Association*, 84(406), p. 414-420.

Lachance, M. (2014), « Useful Functionalities for Record Linkage », document présenté au Symposium international sur les questions de méthodologie de 2014 de Statistique Canada, Ottawa, Canada.

Lachance, M. (2014), MixMatch 1.2 – User Guide, rapport non publié, Ottawa, Canada : Statistique Canada.

Navarro, G. (2001), « A Guided Tour to Approximate String Matching », *ACM Computing Surveys (CSUR)*, 33(1), p. 31-88.

Statistique Canada (2019), « Enquêtes et programmes statistiques – Registre des adresses (RA) », article publié sur le site Web de Statistique Canada. https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&Id=1260700

Statistique Canada (2022), « Dictionnaire, Recensement de la population, 2021 – Logement privé », article publié sur le site Web de Statistique Canada. https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-fra.cfm?ID=dwelling-logements005

Statistique Canada (2022), « Dictionnaire, Recensement de la population, 2021 – Logement collectif », article publié sur le site Web de Statistique Canada. https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/Definition-fra.cfm?ID=dwelling-logements002

Statistique Canada (2022), « Recensement de la population – Géographie du recensement », article publié sur le site Web de Statistique Canada. https://www12.statcan.gc.ca/census-recensement/2021/geo/index-fra.cfm

Winkler, W. E. (2006), « Overview of Record Linkage and Current Research Directions », *Tech. Report. Statistics #2006-2*. Statistical Research Division, U.S. Census Bureau.