Lignes directrices sur la qualité des données du Recensement de 2021

Recensement de la population, 2021

Date de diffusion : le 29 mars 2023

Statistique Canada
Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Une version HTML **est aussi disponible.**
This publication is also available in English.

Date de diffusion : le 29 mars 2023
Nº 98-26-0006 au catalogue, numéro 2021001
ISBN 978-0-660-44669-1

Table des matières

1. Introduction

Pour le Recensement de la population de 2021, la stratégie de diffusion des indicateurs de qualité a été entièrement repensée dans le but d'offrir des renseignements plus détaillés concernant la qualité des données. De nouveaux indicateurs de qualité accompagneront les produits de données, qui aideront les utilisateurs à mieux évaluer la qualité des données et à déterminer dans quelle mesure les renseignements disponibles répondent à leurs besoins. Les Lignes directrices concernant la qualité des données du Recensement de 2021 donnent un aperçu des indicateurs de qualité disponibles, de leur définition et de la façon de les interpréter.

Une série d'indicateurs de qualité accompagne les produits de données du Recensement de la population de 2021. Le taux de non-réponse totale (NRT) associé aux variables géographiques d'intérêt est présenté dans chaque tableau. Pour la première fois en 2021, les taux de non-réponse et d'imputation par question sont disponibles pour les niveaux détaillés de géographie normalisés dans des produits d'indicateurs de qualité des données. De plus, lorsque c'est techniquement possible, les intervalles de confiance sont fournis pour les estimations du questionnaire détaillé.

Le but de fournir des indicateurs de qualité des données est de brosser un portrait détaillé de la qualité des données, qui peut être influencée par des erreurs dues à la non-réponse, des erreurs de traitement et des erreurs d'échantillonnage. Les indicateurs fournis pour le Recensement de 2021 concernent l'exactitude des données, l'une des six dimensions de la qualité des données définies dans les Lignes directrices concernant la qualité de Statistique Canada. L'exactitude des renseignements statistiques est la mesure dans laquelle le renseignement décrit correctement ce qu'il devait évaluer. Les indicateurs de qualité des données font partie de la dimension « intelligibilité » de la qualité. Ils sont fournis pour que les utilisateurs de données soient informés de la qualité des renseignements statistiques diffusés et puissent juger de la pertinence et des limites des données en fonction de leurs besoins.

Le présent document fournit tous les renseignements requis pour comprendre et interpréter les indicateurs de qualité des données pour le Recensement de 2021. Les indicateurs de qualité des données sont présentés en détail, avec des lignes directrices qui permettent de s'en servir correctement. La section 2 fournit des considérations générales. La section 3 définit le taux de NRT et le compare avec le taux global de non-réponse (TGN) qui a été utilisé pour déclarer la non-réponse dans les cycles de recensement antérieurs. La section 4 décrit les indicateurs de qualité des données d'après les tableaux du lieu de résidence. La section 5 aborde les indicateurs de qualité par question, tandis que la section 6 aborde les indicateurs de qualité selon la variance. Enfin, la section 7 présente des pratiques exemplaires et des recommandations à l'égard de l'utilisation des indicateurs de qualité des données pour le Recensement de 2021.

2. Considérations générales

Cette section présente des considérations générales qui s'appliquent à bon nombre, voire à la totalité, des indicateurs de qualité décrits dans le présent document. Ces considérations sont importantes pour comprendre le contexte dans lequel les indicateurs de qualité des données sont produits, car cela peut influencer leur interprétation.

2.1 Au sujet du Recensement de la population

Le Recensement de la population comprend deux principales composantes qui ont chacune leur propre plan de sondage et leurs particularités. Les deux composantes sont étroitement liées aux questionnaires utilisés pour recueillir des renseignements sur les répondants : le questionnaire abrégé et le questionnaire détaillé. Pour le Recensement de 2021, le questionnaire abrégé 2A a été utilisé pour dénombrer tous les résidents habituels de 75 % des logements privés. Le questionnaire détaillé 2A-L, qui comprend aussi les questions du questionnaire abrégé 2A, a été utilisé pour dénombrer un échantillon de 25 % des ménages privés au Canada. Dans le cas des ménages privés dans les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées, le questionnaire 2A-R a été utilisé pour dénombrer 100 % de la population.

Les estimations produites à partir des réponses aux questions qui sont posées dans les deux questionnaires (c.-à-d. le contenu du questionnaire abrégé) sont obtenues à l'aide d'un **recensement** de la population. Ainsi, tous les ménages contribuent à un chiffre donné. Dans le présent document, de telles estimations sont appelées des « estimations du questionnaire abrégé » (données intégrales). De même, les indicateurs de qualité des données calculés avec tous les ménages sont appelés des « indicateurs du questionnaire abrégé ».

Les estimations produites à partir des réponses à au moins une question spécifique au questionnaire détaillé sont obtenues à partir d'une **enquête-échantillon**. Dans ce cas, seuls les ménages répondants de l'échantillon du questionnaire détaillé contribuent à l'estimation. Ces estimations sont appelées des « estimations du questionnaire détaillé » (25 % de données-échantillon). De même, les indicateurs de qualité des données calculés à partir des ménages échantillonnés sont appelés des « indicateurs du questionnaire détaillé ». Les estimations et les indicateurs du questionnaire détaillé sont pondérés pour qu'ils représentent l'ensemble de la population cible de l'enquête.

La **population cible** de la composante du recensement est la population totale du Canada, qui comprend toutes les personnes qui ont habituellement un lieu de résidence au Canada, et certains citoyens canadiens et immigrants reçus qui vivent à l'extérieur du pays. Cette population cible peut être divisée en trois catégories : les personnes qui vivent dans des logements privés, les personnes qui vivent dans des logements collectifs et les personnes qui vivent dans les logements à l'extérieur du Canada. La population cible de la composante de l'enquête ne comprend que les personnes qui vivent dans des logements privés.

Dans la plupart des régions, l'échantillon du questionnaire détaillé est sélectionné selon un plan d'échantillonnage systématique stratifié : le questionnaire détaillé est distribué au quart des ménages qui vivent dans des logements privés, et les ménages sélectionnés se voient attribuer un poids de sondage égal à quatre. Les poids de sondage sont ensuite corrigés pour compenser les non-réponses totales et sont calibrés sur les totaux choisis obtenus à l'aide du recensement. Les poids finaux sont restreints de manière à se situer entre 1 et 20. Dans les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées, le questionnaire détaillé est distribué à chaque ménage. Les non-réponses totales sont compensées par imputation, et les ménages ont un poids final égal à un.

Étant donné qu'elles proviennent d'une enquête-échantillon, les estimations du questionnaire détaillé sont sujettes aux erreurs d'échantillonnage. La variance échantillonnale reflète la variabilité des estimations en raison de l'utilisation d'un échantillon au lieu de la population totale. Cette variance est donc estimée selon une méthode statistiquement adéquate, c'est-à-dire qui tient compte du plan de sondage et de la stratégie d'estimation. Elle est estimée à l'aide d'une méthode de réplication.

Dans les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées, le choix de traiter les non-réponses totales par imputation plutôt que par repondération a aussi un effet sur la variabilité des estimations. Cette variabilité est également appelée la variance due à l'imputation. La méthode de réplication utilisée pour estimer la variance échantillonnale dans d'autres régions a été adaptée pour estimer la variance due à l'imputation dans les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées. Dans les deux cas, des poids de rééchantillonnage sont utilisés pour produire des estimations de la variance, lesquelles sont utilisées pour calculer des intervalles de confiance.

De plus amples renseignements sur la pondération et l'estimation seront disponibles dans le *Rapport technique sur l'échantillonnage et la pondération, Recensement de la population, 2021*, Statistique Canada, produit nº 98-306-X au catalogue.

2.2 Pondération des indicateurs de qualité

Les indicateurs de qualité concernant les données du questionnaire détaillé, notamment le taux de NRT du questionnaire détaillé et les indicateurs de qualité par question pour le contenu du questionnaire détaillé, sont pondérés de façon à ce qu'ils représentent la population cible de l'enquête, et non seulement les unités formant l'échantillon. Les indicateurs de qualité pondérés fournissent une mesure de la qualité à laquelle on s'attendrait si toute la population avait été dénombrée. Ils donnent plus d'informations sur la qualité des estimations connexes que les indicateurs de qualité non pondérés. Le poids de sondage ou le poids final peut être utilisé pour mettre au point des indicateurs de qualité pondérés. Le poids utilisé pour un indicateur précis est fourni dans la section correspondante.

2.3 Unités statistiques et population d'intérêt

Certains indicateurs de qualité des données sont directement associés à une estimation ou à un groupe d'estimations correspondant à la même variable d'intérêt, par exemple les taux par question. Dans ce cas, les mêmes unités et la population d'intérêt sont utilisées dans le calcul des estimations et des indicateurs de qualité des données connexes. Des renseignements détaillés sur les unités statistiques et la population d'intérêt par sujet sont fournis à l'annexe 1.3 du *Guide du Recensement de la population*, Statistique Canada, produit nº 98-304-X au catalogue.

2.4 Évaluation de la qualité à l'étape de l'estimation

Les taux de non-réponse peuvent être calculés à l'étape de la collecte ou à l'étape de l'estimation. Cette dernière est généralement plus utile pour les utilisateurs de données qui veulent déterminer si les données sont d'une qualité suffisante pour répondre à leurs besoins. Les taux de réponse de la collecte sont aussi disponibles à l'échelle du pays, des provinces et des territoires pour le Recensement de 2021, mais les taux de non-réponse décrits dans le présent document ont tous été obtenus à l'étape de l'estimation.

Dans le contexte du recensement, cela signifie que les taux de non-réponse sont calculés en tenant compte de la classification finale du statut d'occupation du logement. Autrement dit, les logements qui ont un statut final de logement privé occupé, qu'ils soient désignés comme étant occupés pendant la collecte ou imputés comme tel durant le traitement, sont pris en compte dans les taux de réponse à l'étape de l'estimation.

La classification du statut d'occupation du logement est basée sur l'analyse des données recueillies pendant les opérations sur le terrain et des données fournies par les répondants, et, dans la plupart des régions du pays, elle est également ajustée en fonction des résultats de l'Enquête sur la classification des logements (ECL). De plus amples renseignements sur l'ECL sont fournis au chapitre 9 du *Guide du Recensement de la population*. Lorsque le statut d'occupation des logements est finalisé, la procédure d'imputation des ménages au complet est appliquée afin d'imputer les données pour les logements occupés des non-répondants.

2.5 Sources d'erreur

Les sources d'erreur auxquelles répondent les indicateurs actuellement disponibles sont principalement attribuables à la non-réponse, à l'imputation et à l'échantillonnage. Il existe d'autres sources d'erreur, par exemple les erreurs de couverture et de mesure, qui ne sont pas évaluées par les indicateurs de qualité disponibles, mais qui peuvent aussi influer sur la qualité des chiffres et des estimations du Recensement de 2021. De plus amples renseignements sur les sources d'erreur possibles sont fournis au chapitre 9 du *Guide du Recensement de la population*. De plus amples renseignements sur les erreurs de couverture seront disponibles dans le *Rapport technique sur la couverture, Recensement de la population, 2021*, Statistique Canada, produit nº 98-303-X au catalogue.

3. Taux de non-réponse totale

La non-réponse totale survient lorsque toutes les questions sont sans réponse pour un logement ayant reçu un questionnaire ou lorsqu'un questionnaire retourné ne contient pas le contenu minimal (c.-à-d. qu'il contient des renseignements insuffisants pour continuer le traitement). Elle est mesurée par le taux de NRT, qui est l'indicateur principal de la qualité qui accompagne chaque produit diffusé du Recensement de la population de 2021. En ce sens, il remplace le TGN, qui a été utilisé lors du Recensement de la population de 2016 et des cycles précédents.

Le TGN combinait la non-réponse totale et la non-réponse partielle alors que le taux de NRT tient uniquement compte de la non-réponse totale. La non-réponse partielle survient lorsque les réponses à certaines questions ne sont pas disponibles pour un ménage répondant. La non-réponse partielle est maintenant comptabilisée séparément (voir la section 5). Cette nouvelle méthode permet de comparer la qualité des données entre les variables, ce que le TGN ne permettait pas.

3.1 Définitions

Pour chaque région géographique, deux taux de NRT distincts sont calculés. Il existe donc deux définitions de ce taux : le taux de NRT non pondéré et le taux de NRT pondéré par le plan de sondage. Supposons que Q_i est une variable à l'échelle des ménages qui prend une valeur négative si l'unité i n'appartient pas à la population cible, prend la valeur 0 si l'unité i appartient à la population cible et n'a pas répondu, et prend la valeur 1 si l'unité i appartient à la population cible et a répondu.

Le taux de NRT non pondéré (UTNR) est utilisé pour calculer le taux de NRT au questionnaire abrégé. Le taux de NRT non pondéré est calculé à l'aide de la formule suivante :

$$UTNR = 100 \times \left(1 - \frac{\sum_{(i:Q_i>0)} 1}{\sum_{(i:Q_i\geq 0)} 1} \right).$$

Le taux de NRT pondéré (WTNR) est utilisé pour calculer le taux de NRT au questionnaire détaillé. Supposons que d_i désigne le poids de sondage de l'unité i. Le taux de NRT pondéré est calculé à l'aide de la formule suivante :

$$WTNR = 100 \times \left(1 - \frac{\sum_{(i:Q_i>0)} d_i}{\sum_{(i:Q_i\geq 0)} d_i} \right).$$

3.2 Comparaison avec le taux global de non-réponse utilisé dans les cycles de recensement précédents

Le taux de NRT de 2021 et le TGN des cycles de recensement précédents remplissent le même objectif : mesurer la portée de la non-réponse dans une région donnée. Conceptuellement, la différence observée entre le TGN d'un recensement précédent et le taux de NRT de 2021 pour une région donnée peut être décomposée en deux parties : la différence attribuable au changement de définition et la différence effective du taux de non-réponse entre les deux cycles. Toutefois, il est impossible de décrire exactement la relation entre les deux indicateurs. D'une part, la taille du ménage influence le TGN, mais n'influence pas le taux de NRT. L'effet de cette différence dans la définition devrait diminuer à mesure que la taille de la population augmente. D'autre part, le TGN comprend la non-réponse partielle et est donc généralement plus élevé que le taux de NRT (il peut toutefois être plus faible).

Les taux de NRT du cycle du Recensement de 2016 ont été calculés pour une étude comparative du TGN et du taux de NRT. Cette étude a révélé qu'il existe une forte corrélation positive entre les deux indicateurs et que leur

différence est généralement inférieure à 5 %. De plus grandes différences ont été observées plus souvent pour les taux relatifs au questionnaire détaillé que pour les taux relatifs au questionnaire abrégé.

Recommandation : Quand le TGN d'un cycle précédent et le taux de NRT de 2021 sont comparés, les différences de moins de 5 % peuvent être considérées comme étant uniquement attribuables au changement de définition.

Dans les cycles de recensement précédents, les régions présentant un TGN supérieur à un certain seuil ont été supprimées des produits diffusés (le seuil utilisé en 2016 était de 50 %). Ce type de suppression de données fondée sur la qualité a été abandonné en 2021.

Recommandation : Les données provenant des régions où le taux de NRT est supérieur à 50 % doivent être utilisées avec prudence.

4. Indicateurs de qualité des données aux fins de mise en tableaux selon les régions géographiques du lieu de résidence

Afin de donner un aperçu de la qualité des données associées à une région géographique, un code numérique à cinq chiffres représentant cinq indicateurs de qualité des données est attribué à chaque région géographique normalisée dans la base de données du recensement. Par exemple, le code à l'échelle nationale est 20000. Ces indicateurs de qualité des données figurent dans les tableaux en fonction des régions géographiques du lieu de résidence. Ils peuvent être utilisés afin de déterminer les régions pour lesquelles des données ont été supprimées pour des raisons précises et d'obtenir des renseignements au sujet du niveau de NRT dans la région. Le code numérique à cinq chiffres et ses composantes sont décrits plus en détail ci-dessous.

4.1 Indicateurs des réserves et établissements partiellement dénombrés

Le premier chiffre du code numérique à cinq chiffres des indicateurs de qualité des données indique si la région géographique dans le tableau comprend une région partiellement dénombrée. Dans le cadre du Recensement de la population de 2021, de même que dans le cadre de recensements antérieurs, le dénombrement n'a pas pu être pleinement effectué pour certaines réserves et certains établissements. Ces réserves et ces établissements partiellement dénombrés, ainsi que les régions géographiques supérieures qui les englobent, sont indiqués dans les produits. Bien que les données du recensement ne soient pas disponibles pour les réserves et les établissements partiellement dénombrés, les régions elles-mêmes sont comprises dans les hiérarchies des régions géographiques normalisées dans la base de données du recensement. Pour obtenir la liste des réserves et des établissements partiellement dénombrés de 2021, voir l'annexe 1.5 du *Guide du Recensement de la population*.

4.2 Indicateurs de non-réponse totale

L'ampleur du taux de NRT dans la région géographique associée à un tableau a une incidence sur la qualité des données. Afin de guider les utilisateurs, sa portée a été divisée en catégories, comme il est montré à la section 4.4. Le deuxième chiffre du code numérique à cinq chiffres des indicateurs de qualité des données contient la catégorie du taux de NRT relatif au questionnaire abrégé, et le quatrième chiffre contient la catégorie du taux de NRT relatif au questionnaire détaillé. Comme il a été mentionné dans la section 3.2, **il faut utiliser les données associées aux régions dont le taux de NRT est supérieur à 50 % avec prudence**. Une remarque à ce sujet accompagne les produits de données.

Les deuxième et quatrième chiffres sont aussi utilisés lorsque des données ont été supprimées afin de respecter les exigences en matière de confidentialité de la *Loi sur la statistique*.

4.3 Indicateurs relatifs à la suppression liée à la confidentialité des données sur le revenu

Dans certaines régions géographiques, les données sur le revenu du questionnaire abrégé ou du questionnaire détaillé sont supprimées afin de respecter les exigences en matière de confidentialité de la *Loi sur la statistique*. Les indicateurs relatifs à la suppression liée à la confidentialité des données sur le revenu tirées du questionnaire abrégé et du questionnaire détaillé indiquent si des données sur le revenu ont été supprimées pour une région donnée. Ces indicateurs sont fournis, respectivement, par le troisième chiffre et le cinquième chiffre du code numérique à cinq chiffres des indicateurs de qualité des données. Les règles de contrôle de la divulgation statistique relatives aux variables du revenu sont différentes des règles concernant les autres types de variables et exigent donc des indicateurs de suppression distincts.

4.4 Résumé des indicateurs de qualité des données pour les tableaux selon les régions géographiques du lieu de résidence

Le tableau 1 ci-dessous décrit le code numérique à cinq chiffres des indicateurs de qualité des données et son contenu pour les tableaux tirés du questionnaire abrégé, et le tableau 2 décrit les indicateurs tirés du questionnaire détaillé. Pour n'importe lequel des cinq chiffres, un zéro est la valeur par défaut de l'indicateur correspondant.

Tableau 1
Indicateurs de la qualité des données pour le questionnaire abrégé du Recensement de 2021

Caractère numérique	Description	Valeur	Description de la valeur
Premier (0XXXX)	Indicateur de dénombrement partiel	0	Valeur implicite. Sans objet.
		1	Réserve ou établissement partiellement dénombré (supprimé).
		2	Ne comprend pas les données du recensement pour une ou plusieurs réserves ou pour un ou plusieurs établissements partiellement dénombrés.

Tableau 1
Indicateurs de la qualité des données pour le questionnaire abrégé du Recensement de 2021

Caractère numérique	Description	Valeur	Description de la valeur
Deuxième (X0XXX)	Indicateur relatif à la qualité des données pour le questionnaire abrégé	0	Valeur implicite. Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale inférieur à 10 %.
		1	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 10 %, mais inférieur à 20 %.
		2	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 20 %, mais inférieur à 30 %.
		3	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 30 %, mais inférieur à 40 %.
		4	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 40 %, mais inférieur à 50 %.
		5	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 50 % (**à utiliser avec prudence**).
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au questionnaire abrégé.
Troisième (XX0XX)	Indicateur relatif à la suppression liée à la confidentialité des données sur le revenu pour le questionnaire abrégé	0	Valeur implicite. Aucune suppression appliquée aux données sur le revenu pour le questionnaire abrégé.
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au revenu du questionnaire abrégé.
Quatrième (XXX0X)	Sans objet	0	Valeur implicite. Sans objet.
Cinquième (XXXX0)	Sans objet	0	Valeur implicite. Sans objet.

Source : Statistique Canada, Recensement de la population, 2021.

Tableau 2
Indicateurs de la qualité des données pour le questionnaire détaillé du Recensement de 2021

Caractère numérique	Description	Valeur	Description de la valeur
Premier (0XXXX)	Indicateur de dénombrement partiel	0	Valeur implicite. Sans objet.
		1	Réserve ou établissement partiellement dénombré (supprimé).
		2	Ne comprend pas les données du recensement pour une ou plusieurs réserves ou pour un ou plusieurs établissements partiellement dénombrés.
Deuxième (X0XXX)	Indicateur relatif à la qualité des données pour le questionnaire abrégé	0	Valeur implicite. Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale inférieur à 10 %.
		1	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 10 %, mais inférieur à 20 %.
		2	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 20 %, mais inférieur à 30 %.
		3	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 30 %, mais inférieur à 40 %.
		4	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 40 %, mais inférieur à 50 %.
		5	Indice de la qualité des données indiquant, pour le questionnaire abrégé, un taux de non-réponse totale supérieur ou égal à 50 % (**à utiliser avec prudence**).
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au questionnaire abrégé.
Troisième (XX0XX)	Indicateur relatif à la suppression liée à la confidentialité des données sur le revenu pour le questionnaire abrégé	0	Valeur implicite. Aucune suppression appliquée aux données sur le revenu pour le questionnaire abrégé.
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au revenu du questionnaire abrégé.

Tableau 2
Indicateurs de la qualité des données pour le questionnaire détaillé du Recensement de 2021

Caractère numérique	Description	Valeur	Description de la valeur
Quatrième (XXX0X)	Indicateur relatif à la qualité des données pour le questionnaire détaillé	0	Valeur implicite. Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale inférieur à 10 %.
		1	Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale supérieur ou égal à 10 %, mais inférieur à 20 %.
		2	Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale supérieur ou égal à 20 %, mais inférieur à 30 %.
		3	Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale supérieur ou égal à 30 %, mais inférieur à 40 %.
		4	Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale supérieur ou égal à 40 %, mais inférieur à 50 %.
		5	Indice de la qualité des données indiquant, pour le questionnaire détaillé, un taux de non-réponse totale supérieur ou égal à 50 % (**à utiliser avec prudence**).
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au questionnaire détaillé.
Cinquième (XXXX0)	Indicateur relatif à la suppression liée à la confidentialité des données sur le revenu pour le questionnaire détaillé	0	Valeur implicite. Aucune suppression appliquée aux données sur le revenu pour le questionnaire détaillé.
		9	Confidentiel en vertu des dispositions de la *Loi sur la statistique* pour les données relatives au revenu du questionnaire détaillé.

Source : Statistique Canada, Recensement de la population, 2021.

5. Indicateurs de qualité par question

Les indicateurs de qualité par question sont des mesures de la qualité des données spécifiques à chaque question. Dans ce contexte, une « question » désigne les questions du questionnaire abrégé et du questionnaire détaillé et les variables relatives au revenu, à l'immigration ou à la mobilité pour lesquelles les données proviennent principalement des dossiers administratifs et des dossiers du Recensement de 2016, et non des réponses aux questionnaires du Recensement de 2021. Dans la présente section, nous faisons la distinction entre les taux relatifs au questionnaire abrégé calculé à partir des membres de la population ayant rempli le questionnaire abrégé, et les taux relatifs au questionnaire détaillé calculés à partir de l'échantillon.

Les indicateurs de qualité fournis par question permettent de quantifier deux sources d'erreur connexes dans les données : les erreurs dues à la non-réponse et les erreurs d'imputation. Plus précisément, les indicateurs de qualité par question disponibles pour le Recensement de 2021 sont le taux de non-réponse par question, le taux d'imputation par question et, pour les variables liées au revenu, l'incidence de l'imputation par question. La présente section fournit des détails au sujet de leur disponibilité, de leur définition et de leurs particularités, ainsi que des lignes directrices sur leur interprétation. De plus amples renseignements sur la façon d'interpréter ensemble ces indicateurs sont présentés à la section 7.

5.1 Disponibilité

Les indicateurs de qualité par question sont calculés à certains niveaux géographiques précis. Pour le Recensement de 2021, ils sont publiquement disponibles dans les tableaux de données, plus précisément pour la qualité des données des hiérarchies des régions géographiques normalisées suivantes :

- Canada, provinces et territoires, régions métropolitaines de recensement (RMR), agglomérations de recensement (AR) et subdivisions de recensement dans les RMR et les AR;
- Canada, provinces et territoires, divisions de recensement et subdivisions de recensement.

Les indicateurs de qualité par question peuvent également être obtenus pour d'autres niveaux géographiques au moyen de demandes personnalisées.

Pour le Recensement de 2016, le taux d'imputation par question et l'incidence de l'imputation par question étaient disponibles à l'échelle du pays, des provinces et des territoires dans leurs guides de référence respectifs. Les taux de non-réponse par question sont une nouveauté du Recensement de 2021.

Les indicateurs de qualité par question disponibles sont présentés à l'annexe 1. Dans le cas des variables qui sont calculées à partir de réponses à plus d'une question, les indicateurs de qualité des données ne sont pas directement disponibles. Dans ces cas, les utilisateurs devraient se référer aux indicateurs liés aux questions portant sur le sujet d'intérêt. Pour en savoir plus, voir les guides de référence pertinents.

5.2 Définitions générales

Les définitions utilisées pour produire les indicateurs de qualité des données par question sont présentées dans cette section. Les lignes directrices sur leur interprétation sont présentées dans la section suivante. Le calcul des indicateurs de qualité des données par question exige des variables indicatrices à l'échelle des unités associées à chaque question. Les unités désignent les personnes ou les ménages, selon le sujet de la question, que ce soit une caractéristique concernant la personne ou le ménage. Supposons que Y est la variable d'intérêt associée à une question et que y_i est la valeur qu'elle prend pour l'unité i. Supposons également que Z est une variable indicatrice associée à Y et que z_i est la valeur qu'elle prend pour l'unité i.

Pour calculer le taux de non-réponse, la variable indicatrice Z indique si l'unité était un répondant ou un non-répondant à la question : $z_i < 0$ si l'unité i est hors du champ d'enquête pour la variable Y, $z_i = 0$ si l'unité i a répondu et $z_i = 1$ si l'unité i n'a pas répondu. De plus amples renseignements au sujet de la non-réponse sont fournis à la section 5.3.

Pour calculer le taux d'imputation, la variable indicatrice Z indique si la réponse à la question a été imputée ou non pour l'unité : $z_i < 0$ si l'unité i est hors du champ d'enquête pour la variable Y, $z_i = 0$ si la réponse de l'unité i n'a pas été imputée et $z_i = 1$ si elle a été imputée. De plus amples renseignements au sujet de l'imputation sont fournis à la section 5.4.

Les indicateurs de qualité par question sont calculés de manière à ce qu'ils soient cohérents avec l'ensemble d'unités qui est considéré comme étant dans le champ d'enquête pour chaque question. Une unité est considérée comme faisant partie du champ d'enquête pour une question donnée si la question s'applique à cette unité et si l'unité appartient à la population d'intérêt liée à la question (voir la section 2.3). Les unités qui font partie du champ d'enquête sont les unités qui contribuent à l'estimation.

5.2.1 Taux relatifs au questionnaire abrégé

Les taux relatifs au questionnaire abrégé sont calculés pour l'ensemble de la population. Ces taux correspondent au taux de non-réponse et au taux d'imputation. En utilisant la notation ci-dessus, les taux relatifs au questionnaire abrégé sont obtenus en suivant la formule générale suivante :

$$NR \mid IMP = \frac{\sum_{(i:z_i=1)} 1}{\sum_{(i:z_i \geq 0)} 1}.$$

Le taux de non-réponse par question du questionnaire abrégé est défini comme le nombre d'unités dans le champ d'enquête dans la population d'intérêt qui n'ont pas répondu à la question, divisé par le nombre d'unités dans le champ d'enquête dans la population d'intérêt.

Le taux d'imputation par question du questionnaire abrégé est défini comme le nombre d'unités dans le champ d'enquête dans la population d'intérêt pour lesquelles la réponse à la question a été imputée, divisé par le nombre d'unités dans le champ d'enquête dans la population d'intérêt.

5.2.2 Taux relatifs au questionnaire détaillé

Les taux relatifs au questionnaire détaillé sont calculés à partir de l'échantillon de ce questionnaire. Ces taux correspondent au taux de non-réponse, au taux d'imputation et à l'incidence de l'imputation. En utilisant la notation ci-dessus, les taux de non-réponse et d'imputation relatifs au questionnaire détaillé sont calculés à l'aide de la formule générale suivante :

$$NR \mid IMP = \frac{\sum_{(i:z_i=1)} w_i}{\sum_{(i:z_i \geq 0)} w_i},$$

où w_i est le facteur de pondération final de l'unité i.

Le taux de non-réponse par question du questionnaire détaillé est défini comme la somme des poids finaux des unités dans le champ d'enquête dans la population d'intérêt qui n'ont pas répondu à la question, divisée par la somme des poids finaux des unités dans le champ d'enquête dans la population d'intérêt.

Le taux d'imputation par question du questionnaire détaillé est défini comme la somme des poids finaux des unités dans le champ d'enquête dans la population d'intérêt pour lesquelles la réponse à la question a été imputée, divisée par la somme des poids finaux des unités dans le champ d'enquête dans la population d'intérêt.

L'incidence de l'imputation fait intervenir une variable d'intérêt continue Y. Dans certains cas, la valeur de cette variable est obtenue en prenant la somme de diverses composantes. Supposons que Z^* est la variable associée à Y et que z_i^* est la valeur que prend l'unité i, de sorte que $z_i^* < 0$ lorsque l'unité i est hors du champ d'enquête pour la variable Y et z_i^* se situe autrement entre 0 et 1.

La variable Z^* indique dans quelle mesure les composantes de la variable Y ont été imputées. Plus précisément, $z_i^* = 0$ si aucune de ses composantes n'a été imputée pour l'unité i, $z_i^* = 1$ si toutes ses composantes ont été imputées et z_i^* prend une valeur entre 0 et 1 si une partie (mais pas la totalité) de ses composantes ont été imputées. Pour les variables d'intérêt qui ne sont pas extraites de composantes séparées, z_i^* prend une valeur de 0 ou de 1.

Pour une variable d'intérêt continue Y, l'incidence de l'imputation est obtenue à l'aide de la formule suivante :

$$INCIDENCE = \frac{\sum_{\left(i:z_i^*>0\right)} w_i z_i^* y_i}{\sum_{\left(i:z_i^*\geq 0\right)} w_i y_i}.$$

5.3 Taux de non-réponse par question

Le taux de non-réponse par question, tel qu'il est défini ci-dessus, est une mesure des renseignements qui sont manquants à cause de la non-réponse à une question en particulier. Si une réponse à une question n'est pas fournie pour une personne ou un ménage donné, cela peut être attribuable à une non-réponse totale ou à une non-réponse partielle. La non-réponse est considérée comme étant totale lorsqu'aucun questionnaire n'est retourné pour un ménage ou lorsqu'un questionnaire retourné ne répond pas aux critères de contenu minimal. La non-réponse partielle a lieu lorsque les réponses à certaines questions ne sont pas disponibles pour un ménage répondant.

Les types de cas de non-réponse pris en compte par le taux de non-réponse par question ne sont pas les mêmes pour les deux types de questionnaires. Pour les taux relatifs au questionnaire abrégé, les cas de non-réponse partielle et de non-réponse totale contribuent au taux de non-réponse. En ce qui concerne les taux relatifs au questionnaire détaillé, seuls les cas de non-réponse partielle sont compris, sauf pour les communautés des Premières Nations, les établissements métis, les régions inuites et d'autres régions éloignées, pour lesquels les cas de non-réponse totale et de non-réponse partielle sont pris en compte.

Il en est ainsi parce que la non-réponse totale est traitée différemment selon le type de questionnaire. La non-réponse totale au questionnaire abrégé est imputée, alors que le traitement de la non-réponse totale au questionnaire détaillé dépend de la région géographique du logement. Dans les régions où la fraction de sondage est égale à un quart, la non-réponse totale est compensée par la repondération des ménages répondants de manière à ce qu'ils représentent les non-répondants. Dans les régions où tous les ménages font partie de l'échantillon du questionnaire détaillé (communautés des Premières Nations, établissements métis, régions inuites et autres régions éloignées), la non-réponse totale au questionnaire détaillé est plutôt traitée par imputation.

Interprétation : En général, le taux de non-réponse par question peut être interprété comme étant la proportion des unités dans le champ d'enquête dans la population d'intérêt pour lesquelles il manque des renseignements à cause de la non-réponse. Les taux relatifs au questionnaire détaillé sont pondérés de façon à tenir compte du fait que ce questionnaire n'est distribué qu'à un échantillon de la population. Donc, dans ce cas, il s'agit d'une proportion estimée.

5.4 Taux d'imputation par question

Le taux d'imputation par question indique dans quelle mesure les réponses à une question donnée ont été imputées. L'imputation est utilisée pour remplacer les données manquantes en cas de non-réponse ou lorsqu'une réponse est jugée non valide. Elle est également utilisée pour les variables dont les données ont été obtenues à partir de dossiers administratifs. Dans les cas où un dossier administratif ne peut être couplé à un répondant pour fournir les renseignements nécessaires, l'imputation sert à produire les valeurs manquantes. Lorsqu'elle est utilisée correctement, l'imputation devrait réduire le biais attribuable à la non-réponse.

Diverses méthodes d'imputation ont été utilisées dans le traitement des données du Recensement de 2021. Ces méthodes, pour la plupart aléatoires, font appel aux valeurs déclarées des répondants pour produire les renseignements manquants. Les vérifications déterministes ne sont pas considérées comme de l'imputation et ne sont pas prises en compte dans le taux d'imputation par question ou dans l'incidence de l'imputation par question. Lorsque le problème est clair et non ambigu (c.-à-d. qu'il n'y a qu'une valeur raisonnable), les vérifications déterministes attribuent une valeur précise pour le résoudre. Cette méthode de vérification des données est employée dans certaines situations pour traiter la non-réponse partielle et les réponses incohérentes aux questions.

Interprétation : En général, le taux d'imputation par question peut être interprété comme étant la proportion des unités dans le champ d'enquête dans la population d'intérêt pour lesquelles les renseignements ont été imputés plutôt que déclarés. Ce taux ne tient pas compte des vérifications déterministes. Les taux relatifs au questionnaire détaillé sont pondérés de façon à tenir compte du fait que ce questionnaire n'est distribué qu'à un échantillon de la population. Donc, dans ce cas, il s'agit d'une proportion estimée.

5.5 Incidence de l'imputation par question

L'incidence de l'imputation est une mesure disponible pour les concepts liés au revenu.

Interprétation : L'incidence de l'imputation par question peut être interprétée comme étant la proportion du total de la variable pour laquelle les valeurs ont été imputées. Comme le taux d'imputation par question, l'incidence de l'imputation ne tient pas compte des vérifications déterministes. Pour les variables calculées à partir de diverses composantes, l'incidence de l'imputation tient aussi compte de la proportion des composantes qui ont été imputées.

L'incidence de l'imputation intègre les valeurs d'une variable, au lieu de mesurer uniquement la fraction des unités dans le champ d'enquête pour lesquelles la variable a été imputée, comme c'est le cas pour le taux d'imputation. Les plus grandes valeurs imputées de la variable imputée contribuent davantage à l'incidence de l'imputation de cette variable que ses plus petites valeurs imputées. Par exemple, une valeur de revenu d'emploi imputée de 200 000 $ par année contribue plus fortement à l'incidence de l'imputation pour le revenu d'emploi qu'une valeur imputée de 30 000 $ par année.

Certaines variables liées au revenu peuvent prendre des valeurs négatives. Les valeurs imputées peuvent elles aussi prendre des valeurs négatives dans de tels cas. Les valeurs imputées négatives tendent à diminuer l'incidence de l'imputation d'une valeur totale positive et pourraient même mener à une incidence de l'imputation négative. Il est plus difficile dans de tels cas d'interpréter cet indicateur. D'autres indicateurs de qualités, comme le taux d'imputation, ou l'incidence de l'imputation du revenu absolu pourraient aider à l'interprétation. Ces indicateurs ne sont pas diffusés, mais sont disponibles sur demande.

6. Indicateurs de la qualité fondés sur la variance

La variance est une mesure de l'incertitude d'une estimation produite à partir d'un échantillon. Comme elle est difficile à interpréter, les enquêtes fournissent habituellement aux utilisateurs de données d'autres indicateurs de la qualité dérivés de l'estimateur de variance, à savoir les erreurs-types, les coefficients de variation ou les intervalles de confiance. La section 2.1 fournit plus de détails sur l'estimation de la variance.

L'intervalle de confiance a été choisi comme indicateur de la qualité fondé sur la variance pour appuyer les estimations du questionnaire détaillé du Recensement de la population de 2021 parce qu'il permet aux utilisateurs de réaliser facilement une inférence statistique. Par conséquent, les intervalles de confiance accompagnent généralement les estimations du questionnaire détaillé dans les produits de données du Recensement de 2021.

6.1 Erreur-type

L'erreur-type associée à une estimation est la racine carrée de sa variance estimée. L'échelle est la même que celle de l'estimation elle-même.

6.2 Coefficient de variation

Le coefficient de variation associé à une estimation correspond au ratio de l'erreur-type à l'estimation. Il s'agit d'une mesure normalisée qui est exprimée en pourcentage de l'estimation.

6.3 Intervalle de confiance

Un intervalle de confiance est associé à un niveau de confiance. Un niveau de confiance par défaut est généralement établi pour une enquête ou dans un domaine d'études en fonction des besoins des utilisateurs. Pour le système de diffusion des données du recensement, le niveau de confiance par défaut a été établi à 95 %. Un intervalle de confiance de 95 % est un intervalle construit autour de l'estimation de telle façon que, si le processus ayant généré l'échantillon était répété de nombreuses fois, la valeur du paramètre d'intérêt dans la population serait contenue dans 95 % de ces intervalles.

L'intervalle de confiance habituel, aussi appelé intervalle de Wald, suppose que la distribution d'échantillonnage de l'estimateur est une distribution normale. L'intervalle de confiance de Wald de 95 % est déterminé en soustrayant ou en additionnant environ deux fois l'erreur-type à l'estimation. Lorsque la taille de l'échantillon est petite, ainsi que pour certaines statistiques comme les proportions et les comptes, l'hypothèse selon laquelle la distribution des estimateurs est normale est souvent enfreinte. Par conséquent, un intervalle de confiance déterminé de cette façon n'est pas approprié; son niveau de confiance réel est inférieur au niveau de confiance nominal de 95 % indiqué.

Par conséquent, les intervalles de confiance présentés avec les estimations du questionnaire détaillé du Recensement de la population de 2021 sont produits à l'aide de méthodes plus élaborées qui offrent un véritable niveau de confiance plus près du niveau nominal. Les méthodes utilisées pour produire les intervalles de confiance sont décrites ci-dessous. Bien que les intervalles de confiance fondés sur ces méthodes aient généralement de bonnes propriétés, tous les intervalles de confiance sont fondés sur des hypothèses qui ne peuvent pas être vérifiées. De plus amples détails au sujet des différentes méthodes utilisées pour construire les intervalles de confiance et de leurs hypothèses sont présentés dans le *Rapport technique sur l'échantillonnage et la pondération, Recensement de la population, 2021*, produit nº 98-306-X au catalogue de Statistique Canada.

6.3.1 Intervalle de confiance de Student

L'intervalle de confiance de Student est utilisé pour toutes les statistiques, sauf les proportions et les comptes. Il est fondé sur la loi de Student, une distribution de probabilité ayant un paramètre : le nombre de degrés de liberté. Lorsque le nombre de degrés de liberté est très élevé, les intervalles de Wald et de Student sont à peu près identiques. Cependant, ce n'est souvent pas le cas des estimations du questionnaire détaillé du recensement. Le nombre de degrés de liberté de la distribution t de Student est influencé par le plan d'échantillonnage, le nombre d'unités échantillonnées et la méthode d'estimation de la variance. Le nombre de degrés de liberté a une incidence sur la largeur de l'intervalle de confiance. Dans le cadre du Recensement de 2021, les degrés de liberté ont été estimés par le nombre de répétitions utilisées pour l'estimation de la variance et désigné par R.

La borne inférieure (LB) et la borne supérieure (UB) de l'intervalle de confiance de Student de 95 % pour un paramètre de population d'intérêt θ sont données par :

$$LB = \hat{\theta} - t \times \widehat{SE}(\hat{\theta}),$$
$$UB = \hat{\theta} + t \times \widehat{SE}(\hat{\theta}),$$

où

- $\hat{\theta}$ est l'estimation de θ;
- t est le $97,5^{e}$ centile de la distribution t de Student à R degrés de liberté;
- $\widehat{SE}(\hat{\theta})$ est l'erreur-type de $\hat{\theta}$.

6.3.2 Intervalle de confiance de Wilson modifié pour les proportions

La méthode d'intervalle de confiance de Wilson modifié est utilisée pour les statistiques de type proportion. Les bornes LB et UB d'un intervalle de confiance de Wilson modifié de 95 % pour une statistique de type proportion p sont données par :

$$LB = \frac{\hat{p} + t^2/2n_e}{1 + t^2/n_e} - \frac{t\sqrt{\hat{p}(1-\hat{p}) + t^2/4n_e}}{\sqrt{n_e}\,(1 + t^2/n_e)},$$

$$UB = \frac{\hat{p} + t^2/2n_e}{1 + t^2/n_e} + \frac{t\sqrt{\hat{p}(1-\hat{p}) + t^2/4n_e}}{\sqrt{n_e}\,(1 + t^2/n_e)},$$

où

- $\hat{p}$ est l'estimation de p ;
- t est le $97,5^e$ centile de la distribution t de Student à R degrés de liberté;
- $n_e = \min(\mathrm{n}/\mathrm{deff}(\hat{p}), n)$ est la taille effective de l'échantillon;
- $\mathrm{deff}(\hat{p}) = \frac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/\mathrm{n}}$ est l'effet de plan estimé;
- n est la taille de l'échantillon dans le champ de l'enquête;
- $\hat{V}(\hat{p})$ est la variance estimée de $\hat{p}$.

Les avancées théoriques et de nombreuses études par simulation[1,2] ont montré que cette méthode a de bonnes propriétés dans la plupart des situations et offre un meilleur résultat que les intervalles de confiance de Wald et de Student lorsque les hypothèses ne tiennent pas.

6.3.3 Intervalle de confiance de Wilson modifié pour les comptes

La méthode d'intervalle de confiance de Wilson modifié est utilisée pour les comptes estimés. Les bornes LB et UB d'un intervalle de confiance de Wilson modifié de 95 % pour un compte Y sont données par :

$$LB = \hat{Y} + t^2\frac{1}{2}\frac{\hat{V}(\hat{Y})}{\hat{Y}} - \sqrt{t^2\hat{V}(\hat{Y}) + \left(t^2\frac{1}{2}\frac{\hat{V}(\hat{Y})}{\hat{Y}}\right)^2},$$

$$UB = \hat{Y} + t^2\frac{1}{2}\frac{\hat{V}(\hat{Y})}{\hat{Y}} + \sqrt{t^2\hat{V}(\hat{Y}) + \left(t^2\frac{1}{2}\frac{\hat{V}(\hat{Y})}{\hat{Y}}\right)^2},$$

où

- $\hat{Y}$ est une estimation de Y
- t est le $97,5^e$ centile de la distribution t de Student à R degrés de liberté;
- $\hat{V}(\hat{Y})$ est la variance estimée de $\hat{Y}$.

1. Phillip S. Kott et D. Andrew Carr, 1997, « Developing an Estimation Strategy for a Pesticide Data Program », *Journal of Official Statistics*, vol. 13, nº 4, p. 367 à 383 (disponible seulement en anglais).
2. Elisabeth Neusy et Harold Mantel, 2016, « Confidence Intervals for Proportions Estimated from Complex Survey Data », *Proceedings of the Survey Methods Section*, Assemblée annuelle de la Société statistique du Canada, juin 2016 (disponible seulement en anglais).

Les avancées théoriques et de nombreuses études par simulation[3] ont montré que cette méthode a de bonnes propriétés dans la plupart des situations et offre un meilleur résultat que les intervalles de confiance de Wald et de Student lorsque les hypothèses ne tiennent pas.

6.4 Interprétation à l'aide d'intervalle de confiance

Lorsqu'un utilisateur dispose d'un intervalle de confiance de 95 % pour un paramètre d'intérêt, il peut dire qu'il est confiant à 95 % que le paramètre réel de population se trouve à l'intérieur de l'intervalle. Par exemple, si l'estimation du revenu moyen d'emploi est 40 000 $ et que son intervalle de confiance à 95 % varie de 35 000 $ à 45 000 $, l'utilisateur peut dire qu'il est confiant à 95 % que le revenu moyen d'emploi de la population se situe entre 35 000 $ et 45 000 $.

7. Pratiques exemplaires et recommandations

La présente section fournit plus de détails sur la façon d'utiliser ensemble les indicateurs et sur leurs interrelations, ainsi que des recommandations générales à l'égard de l'interprétation.

7.1 Stratégie globale

Pris ensemble, les indicateurs de qualité des données disponibles fournissent des renseignements sur la qualité globale des chiffres et des estimations du Recensement de 2021. Afin d'avoir le meilleur portrait possible de la qualité des données, les utilisateurs de données doivent consulter toute la série d'indicateurs pertinents. À titre de rappel, le taux de NRT et les indicateurs de qualité des données à cinq chiffres accompagnent chaque tableau. Pour les tableaux produits à partir du questionnaire détaillé, les intervalles de confiance accompagnent habituellement les estimations. Les indicateurs de qualité des données par question, y compris les taux de non-réponse et d'imputation par question, sont disponibles dans les tableaux de données expressément pour la qualité des données.

7.2 Interprétation des taux de non-réponse

La non-réponse est une source potentielle de biais dans les chiffres du recensement et les estimations du questionnaire détaillé. Le biais survient lorsque les caractéristiques des répondants diffèrent de celles des non-répondants. Malheureusement, le biais ne peut pas être mesuré directement, parce que les caractéristiques des non-répondants sont généralement inconnues.

Le taux de NRT et les taux de non-réponse par question indiquent le risque qu'un biais important soit causé par la non-réponse, et l'ampleur de ce risque. Pour un profil donné de non-répondants, un taux de non-réponse plus faible indique un risque moins élevé de biais dû à la non-réponse et, par conséquent, des chiffres et des estimations plus fiables.

Le taux de NRT et les taux de non-réponse par question qui s'appliquent devraient être consultés, parce qu'ils peuvent apporter différentes perspectives sur la qualité des données. Pensons à une région où le taux de NRT est élevé et où le taux de non-réponse est faible pour une question précise. Cela peut se produire lorsqu'une question est hors du champ d'enquête pour un sous-groupe important de la population qui présentait un faible taux de réponse. Dans une telle situation, il y a un risque de biais à l'égard des caractéristiques connexes. Par exemple, si le taux de NRT est élevé pour les personnes au chômage et qu'il y a beaucoup de chômeurs dans la région, le

3. Elisabeth Neusy, Sarah-Anne Savard, Mike Hidiroglou et Vincent Martin, 2021, « Modified Wilson Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation », présentation au Comité consultatif des méthodes statistiques, mai 2021, document interne, Statistique Canada (disponible seulement en anglais).

taux de NRT sera élevé. Si, en même temps, les personnes occupant un emploi ont bien répondu, le taux de non-réponse par question pour les questions relatives au travail pourrait être faible. Étant donné qu'il existe un risque de biais, les utilisateurs doivent interpréter avec prudence les données sur le travail dans cette région.

En revanche, pensons à une région où le taux de NRT est faible et où le taux de non-réponse par question est élevé pour une question précise. Si le taux d'imputation pour la question est également élevé, cela peut indiquer qu'il existe un risque de biais en ce qui concerne la caractéristique d'intérêt. Dans ce cas, les utilisateurs doivent interpréter les données avec prudence, même si le taux de NRT est faible.

Lorsqu'ils comparent le taux de NRT avec le taux de non-réponse par question, les utilisateurs doivent être conscients des différences dans leur définition. Premièrement, ces deux taux ne sont pas fondés sur les mêmes unités statistiques : les unités utilisées dans le calcul du taux de NRT sont les ménages, alors que les unités employées dans le calcul du taux de non-réponse par question peuvent être des personnes ou des ménages, selon la question. De plus, le dénominateur du taux de NRT est toute la population cible, alors que le dénominateur du taux de non-réponse par question est le sous-ensemble d'unités qui sont dans le champ d'enquête pour la question et qui font partie de la population d'intérêt.

7.3 Interprétation des taux d'imputation

Le taux d'imputation indique si la quantité des valeurs imputées est importante par rapport à la quantité des valeurs déclarées. L'incidence de l'imputation peut être interprétée similairement en remplaçant la quantité par la somme. Plus le taux est élevé, plus il y a lieu de s'interroger sur la qualité des estimations et le risque de biais.

Cependant, les taux eux-mêmes n'indiquent pas le niveau de qualité des données imputées. Une évaluation de la stratégie d'imputation, si elle est possible, fournit des renseignements supplémentaires au sujet de la qualité des estimations. Lorsque les modèles d'imputation sont basés sur l'utilisation de renseignements auxiliaires dont la corrélation avec la caractéristique d'intérêt (l'approche privilégiée pour le Recensement de la population de 2021) a été bien établie, on peut conclure que les valeurs imputées sont assez précises. Dans ce cas, un taux d'imputation élevé ne sous-entend pas nécessairement que la qualité est discutable.

7.4 Relation entre le taux de non-réponse et le taux d'imputation

Comme l'imputation est utilisée pour traiter la non-réponse, le taux d'imputation par question est fortement lié au taux de non-réponse par question. Quoi qu'il en soit, dans certaines circonstances, ces deux taux ne sont pas équivalents.

Par exemple, une unité pourrait être considérée comme répondante et imputée si la réponse déclarée a été jugée non valide pendant le traitement. Si c'est souvent le cas pour une question, cela peut faire en sorte que le taux d'imputation soit plus élevé que le taux de non-réponse.

Inversement, il se peut qu'une unité ne soit considérée ni comme répondante ni comme imputée. Cela se produit quand des vérifications déterministes sont appliquées pour traiter les cas de non-réponse qui peuvent être résolus d'une manière unique. Si c'est souvent le cas pour une question, cela peut faire en sorte que le taux de non-réponse soit plus élevé que le taux d'imputation.

8. Conclusion

En bref, les produits de données du Recensement de la population de 2021 comportent de nombreux indicateurs de qualité des données. Le taux de NRT est indiqué dans chaque tableau, et de plus amples renseignements

sur la qualité au niveau géographique de la région sont disponibles pour les tableaux en fonction des régions géographiques du lieu de résidence. Les taux de non-réponse et d'imputation par question sont disponibles pour les niveaux de géographie inférieurs normalisés dans des tableaux séparés. Enfin, pour la plupart des estimations provenant du questionnaire détaillé, les intervalles de confiance sont aussi fournis.

Les indicateurs de qualité des données sont indiqués pour que les utilisateurs puissent évaluer la pertinence des données par rapport à leurs besoins. Leur définition et les lignes directrices concernant leur interprétation ont été présentées dans ce document. En général, les données du Recensement de la population de 2021 sont d'une très bonne qualité, mais dans certains cas, les données doivent être utilisées avec prudence. Il est fortement recommandé que les utilisateurs consultent tous les indicateurs de qualité des données disponibles pour mieux juger de la qualité des produits de données qui les intéressent.

Annexe 1

Tableau A1.1
Indicateurs de qualité des données par question disponibles en ce qui a trait au contenu du questionnaire abrégé (à l'exclusion des questions sur le revenu)

Sujet	Question	Taux de non-réponse	Taux d'imputation	Incidence de l'imputation
Âge, sexe à la naissance et genre	Âge	Oui	Oui	Non
	Sexe à la naissance	Oui	Oui	Non
	Genre	Oui	Oui	Non
État matrimonial	État matrimonial (légal)	Oui	Oui	Non
	Union libre	Oui	Oui	Non
Familles et ménages	Relation avec la personne 1	Oui	Oui	Non
Langue	Connaissance des langues officielles	Oui	Oui	Non
	Toutes les langues parlées à la maison	Oui	Oui	Non
	Langue parlée le plus souvent à la maison	Oui	Oui	Non
	Langue maternelle	Oui	Oui	Non
Expérience militaire canadienne	Expérience militaire canadienne	Oui	Oui	Non
Type de logement (collectif ou privé)	Type de construction résidentielle	Oui	Oui	Non

Source : Statistique Canada, Recensement de la population, 2021.

Tableau A1.2
Indicateurs de qualité des données par question disponibles en ce qui a trait au sujet du revenu

Question	Taux de non-réponse	Taux d'imputation	Incidence de l'imputation
Contenu du questionnaire abrégé			
Revenu total pour 2020	Oui	Non	Non
Revenu du marché pour 2020	Oui	Non	Non
Revenu d'emploi pour 2020	Oui	Non	Non
Transferts gouvernementaux pour 2020	Oui	Non	Non
Revenu après impôt pour 2020	Oui	Non	Non
Revenu total pour 2019	Oui	Non	Non
Revenu d'emploi pour 2019	Oui	Non	Non
Contenu du questionnaire détaillé			
Revenu total pour 2020	Oui	Non	Oui
Revenu du marché pour 2020	Oui	Non	Oui
Revenu d'emploi pour 2020	Oui	Non	Oui
Transferts gouvernementaux pour 2020	Oui	Non	Oui
Revenu après impôt pour 2020	Oui	Non	Oui
Revenu total pour 2019	Oui	Non	Oui
Revenu d'emploi pour 2019	Oui	Non	Oui

Source : Statistique Canada, Recensement de la population, 2021.