Normes et lignes directrices relatives à la confidentialité et à la qualité des données (version publique)

Enquête nationale auprès des ménages de 2011

Table des matières

Page

1. Introduction

1.1 Contexte

L'information qui a été précédemment recueillie au moyen du questionnaire du recensement complet (à participation obligatoire) a été recueillie dans le cadre de la nouvelle Enquête nationale auprès des ménages (ENM) à participation volontaire. Environ 4,5 millions de ménages ont reçu le questionnaire de l'ENM.

Les données de l'ENM diffusées font l'objet de divers processus automatisés et manuels visant à déterminer s'il y a lieu de les supprimer. Ces processus sont mis en œuvre principalement pour deux raisons : 1) pour garantir la non-divulgation de l'identité et des caractéristiques des répondants (notion représentée par le terme **confidentialité** dans la suite du présent document) et 2) pour restreindre la diffusion de données de qualité inacceptable (notion représentée par l'expression **qualité des données** dans la suite du présent document).

En outre, des règles de suppression de données peuvent être appliquées pour des motifs propres à des produits en particulier, habituellement liés au formatage. Le terme **produit** renvoie en général à des tableaux. Les données peuvent être soit modifiées dans le produit, soit supprimées complètement conformément aux règles de suppression appliquées.

2. Règles relatives à la confidentialité (non-divulgation)

La présente section décrit les diverses règles utilisées pour garantir la confidentialité (ou non-divulgation) de l'identité et des caractéristiques des répondants. Toutes les données de l'ENM sont assujetties à des règles (de non-divulgation).

2.1 Suppression des régions pour les régions géographiques normalisées[1] et non normalisées

La suppression des régions a pour objet d'éliminer toutes les données sur les caractéristiques pour les régions géographiques dont la population est inférieure à un seuil donné.

Le seuil de population défini pour toutes les régions normalisées[1] ou les regroupements de régions normalisées, sauf les îlots, les côtés d'îlot et les régions définies selon le code postal, est de 40. Par conséquent, aucune caractéristique ou donnée totalisée ne doit être diffusée pour les régions dont la population est inférieure à 40 habitants.

1. Pour plus de renseignements sur les régions normalisées, reportez-vous au *Dictionnaire du recensement* de 2011.

Dans le cas des régions définies selon le code postal à six caractères (région de tri d'acheminement – unité de distribution locale [RTA-UDL]), des régions géocodées et des régions personnalisées constituées d'îlots, de côtés d'îlot ou d'UDL, le seuil de population est établi à 100 personnes. Par conséquent, aucune caractéristique ou donnée totalisée ne doit être diffusée si la population totale de la région est inférieure à 100 habitants. En général, la population des îlots et des côtés d'îlot urbains (un côté d'une rue, situé entre deux intersections consécutives) sera trop faible pour respecter les seuils de population établis. Les données relatives aux regroupements d'îlots ou de côtés d'îlot dont la population est supérieure au seuil fixé peuvent être extraites à l'aide d'un tableau personnalisé.

Veuillez consulter la section 2.2 Règles d'agrégation minimales des codes postaux pour les règles supplémentaires applicables aux données relatives aux codes postaux.

2.1.1 Univers de la population utilisés pour les routines de suppression

La population à l'étude pour toutes les totalisations des données est l'estimation de la population dans les ménages privés de l'ENM.

Dans le cas des données sur le lieu de travail, la population à l'étude est la population active occupée ayant un lieu de travail habituel ou ayant travaillé à domicile.

Univers de la population utilisés pour les routines de suppression

ENM	Régions géographiques des LDT
Estimation de la population dans les ménages privés	Population active occupée ayant un lieu habituel de travail ou ayant travaillé à domicile

Pour les totalisations de l'ENM qui sont fondées sur les géographies ou les régions du lieu de travail, tous les critères doivent être fondés sur les estimations de la population active occupée ayant un lieu habituel de travail ou ayant travaillé à domicile. En l'occurrence, les seuils de population de 40, 100 et 250 sont les estimations de la population active occupée ayant un lieu habituel de travail ou ayant travaillé à domicile, au lieu de la population des régions. Dans le cas des totalisations qui portent à la fois sur des régions géographiques du lieu de résidence et sur des régions géographiques du lieu de travail, les seuils de population (40, 100 et 250) s'appliquent à la fois aux données sur le lieu de résidence (population) et sur le lieu de travail (population active occupée ayant un lieu de travail habituel ou ayant travaillé à domicile).

2.2 Règles d'agrégation minimales des codes postaux

En plus des règles de confidentialité concernant la diffusion des données de l'Enquête nationale auprès des ménages assorties des codes postaux, les règles suivantes s'appliquent aux codes postaux. Ces règles s'inscrivent dans le cadre de la clause 03.01 (n) de l'Entente de licence commerciale concernant l'utilisation des codes postaux à des fins autres que la correspondance conclue entre Statistique Canada et la Société canadienne des postes.

- Toutes les demandes doivent comprendre des lots d'au moins deux codes postaux. Seuls les codes postaux dont le second caractère est un zéro (codes postaux ruraux) font exception.
- On doit attribuer aux groupes de codes postaux un numéro ou une classification unique (p. ex., K1A 0T6, 0T7, 0T8 = Région personnalisée 1); selon les modalités du contrat mentionné ci-dessus, les clients ne peuvent pas recevoir de listes de codes postaux, seul le nom mentionné dans la requête du client peut être utilisé.
- Toutes les autres règles de confidentialité pour les extractions personnalisées s'appliquent, conformément à la section 2.1.

En outre, l'avertissement qui suit doit être inclus dans toutes les demandes de codes postaux personnalisées :

> Avis de non-responsabilité concernant la validation du code postal : Statistique Canada ne fait aucune déclaration et n'offre aucune garantie concernant l'exactitude des données relatives aux codes postaux^{MO} soumis à Statistique Canada.

Ces règles s'appliquent aussi aux demandes de codes postaux historiques.

2.3 Arrondissement aléatoire

Toutes les estimations des totalisations de l'ENM sont assujetties à un processus appelé l'arrondissement aléatoire. L'arrondissement aléatoire transforme toutes les estimations brutes en estimations arrondies aléatoirement. Ceci réduit la possibilité de révéler l'identité de personnes dans les totalisations.

Toutes les estimations supérieures à 10 sont arrondies à un multiple de 5, les estimations inférieures à 10 sont arrondies à un multiple de 10. Cela signifie que toutes les estimations inférieures à 10 seront toujours remplacées par 0 ou 10. Le tableau ci-après montre l'effet de l'arrondissement sur les estimations comportant une valeur inférieure à 10.

Tableau 2.1 Fréquence de l'arrondissement aléatoire

Estimation de	Sera arrondie à 0	Sera arrondie à 10
1	9 fois sur 10	1 fois sur 10
2	8 fois sur 10	2 fois sur 10
3	7 fois sur 10	3 fois sur 10
4	6 fois sur 10	4 fois sur 10
5	5 fois sur 10	5 fois sur 10
6	4 fois sur 10	6 fois sur 10
7	3 fois sur 10	7 fois sur 10
8	2 fois sur 10	8 fois sur 10
9	1 fois sur 10	9 fois sur 10
0	Toujours	Jamais

L'algorithme d'arrondissement aléatoire utilise une valeur de départ aléatoire pour déclencher le processus d'arrondissement pour les tableaux. Lorsque ces routines sont appliquées, il est possible que la même estimation dans le même tableau soit arrondie vers le haut dans une exécution et vers le bas dans une autre en raison de la méthode utilisée pour amorcer le processus.

2.4 Prévention de la divulgation pour les statistiques

Les statistiques (moyenne, somme, médiane, centile, ratio ou pourcentage) ne sont pas arrondies aléatoirement. Toutefois, l'inclusion des statistiques dans les tableaux où figurent également les estimations à partir desquelles elles ont été calculées peut se solder par la divulgation de l'identité des répondants. Pour éviter ce problème, on applique des méthodes de suppression pour les statistiques ou des calculs statistiques spéciaux.

2.4.1 Suppression de statistiques

On procède à la suppression de statistiques chaque fois que l'une ou l'autre des trois conditions suivantes est satisfaite :

1) Les statistiques relatives à une cellule seront supprimées si l'étendue des données (c.-à-d. le montant maximal en dollars de la cellule moins le montant minimal en dollars) divisée par le maximum des valeurs absolues est sous la valeur de seuil d'un paramètre. Cette méthode de suppression est appliquée uniquement à des statistiques calculées à partir des valeurs quantitatives mesurées en dollars ($), notamment le revenu ou la valeur des logements.

2) Dans le cas de toutes les variables quantitatives, une statistique est supprimée si le nombre d'enregistrements réel (non arrondi ni pondéré) à partir desquels elles ont été calculées est inférieur à 4. Pour les statistiques quantiles, un autre nombre minimum d'enregistrements s'applique : pour les quartiles, les quintiles ou les déciles, il faut 20 enregistrements et pour le centile, 400 enregistrements sont nécessaires.

3) Les statistiques pour une cellule seront supprimées si elle contient une valeur aberrante. Une cellule sera considérée comme ayant une valeur aberrante si la valeur absolue la plus grande divisée par la somme des valeurs absolues est au-dessus de la valeur de seuil d'un paramètre.

Note : Le nombre d'enregistrements utilisés aux fins du calcul n'est pas nécessairement égal au nombre d'enregistrements que comporte la cellule; il s'agit plutôt du nombre d'enregistrements applicables ou disponibles aux fins du calcul de la statistique figurant dans la cellule.

Exemple :

Supposons une cellule contenant les enregistrements suivants :

Les huit enregistrements dans la cellule représentent 47,6 personnes (la somme des facteurs de pondération). Puisque pour la variable « Salaires » seules des valeurs non nulles sont utilisées dans le calcul, la moyenne 22 727,27 $ sera supprimée parce que seulement trois enregistrements sont utilisés dans le calcul.

Exemple de huit enregistrements montrant le poids appliqué et les salaires de chaque répondant

Numéro d'enregistrement	Facteur de pondération	Salaires ($)
1	5,5	16 500
2	2,9	345 600
3	8,1	12 900
4	6,2	0
5	6,6	0
6	5,9	0
7	5,4	0
8	6,9	0

4) Pour toutes les variables quantitatives, toutes les statistiques sont supprimées si la somme des facteurs de pondération est inférieure à 10.

2.4.2 Calculs statistiques spéciaux

1) La valeur prise par la statistique n'est jamais arrondie, sauf dans le cas des fréquences.

2) Toutes les statistiques fondées sur un classement par ordre des valeurs (médianes, centiles) sont calculées de la façon habituelle, c'est-à-dire jamais arrondies.

3) Dans le cas des sommes, si le programme calcule une valeur exprimée en dollars, un nombre de semaines, un nombre d'heures ou un âge, il multiplie alors la moyenne non arrondie du groupe en question par la fréquence pondérée arrondie. Autrement, le programme arrondit la somme pondérée.

Lorsqu'il faut faire une division (moyennes, pourcentages, ratios, etc.), le programme doit appliquer la règle énoncée au point 3) tant au numérateur qu'au dénominateur avant d'effectuer la division.

Note : Les statistiques fondées sur un classement par ordre des valeurs, telles que la médiane et les centiles, sont toujours calculées au moyen d'interpolations linéaires. Ces statistiques ne sont donc pas fiables dans le cas de cellules comportant de faibles estimations. C'est pourquoi aucune autre règle relative à la confidentialité ne leur est appliquée.

Note : La moyenne d'une valeur en dollars, d'un nombre de semaines, d'un nombre d'heures ou d'un âge n'est pas modifiée par l'arrondissement, parce que le numérateur correspond au produit de la moyenne réelle multipliée par la fréquence arrondie, et que le dénominateur correspond à la fréquence arrondie. Les deux fréquences s'annulent l'une l'autre de sorte que la moyenne réelle n'est pas modifiée.

2.5 Suppression des estimations de l'ENM en raison de la protection de la confidentialité des renseignements

À la section 2.3, il est question d'arrondissement aléatoire des estimations des totalisations de l'ENM. L'arrondissement aléatoire est utilisé afin de protéger la confidentialité dans les estimations. L'analyse des données de l'ENM a révélé que même si on applique l'arrondissement aléatoire, nous risquons, dans certains cas, de diffuser des données comportant des risques élevés de divulgation.

Ces risques élevés se manifestent car l'ajustement de la non-réponse de l'ENM a exigé un large éventail de poids. Les poids élevés pourraient permettre aux individus ayant de rares caractéristiques d'être facilement identifiables dans un tableau, spécialement si leurs caractéristiques sont connues du grand public.

Afin de minimiser ces risques, une règle relative aux estimations semblables à la règle pour les variables quantitatives décrite à la section 2.3 a été mise en œuvre. L'estimation de la cellule sera supprimée si le nombre d'enregistrements comprenant l'attribut ou la combinaison d'attributs représenté par la cellule (non arrondie et non pondérée) est inférieur à 4. Dans ces cas, la cellule affichera le nombre 0 plutôt que la valeur supprimée, et par conséquent, ne se distinguera pas d'une cellule vide véritable.

Exemple :

Supposons que nous avons les enregistrements suivant pour une géographie donnée :

Exemple de 15 enregistrements montrant le poids appliqué et l'âge de chaque répondant

Numéro d'enregistrement	Facteur de pondération	Âge
1	6,5	20
2	4,9	22
3	8	25
4	6,8	26
5	5,4	27
6	6,1	27
7	4,7	27
8	5,7	29
9	2,8	32
10	6,8	36
11	41,1	39
12	5	39
13	81,4	40
14	5,1	50
15	3,2	54

En appliquant l'arrondissement aléatoire seulement, les estimations de l'ENM seraient publiées (arrondies aléatoirement) comme illustré dans le tableau suivant. Le nombre d'enregistrements ne serait jamais publié, mais sert seulement à démontrer l'effet de la règle.

Exemple d'estimations qui seraient publiées sans l'application de la suppression fondée sur la cellule

	Étendue d'âge				
Valeurs	20 à 29	30 à 39	40 à 49	50 à 59	Total
Estimation de l'ENM	50	55	80	10	195
Nombre d'enregistrements	8	4	1	2	15

La suppression fondée sur les estimations dans les cellules supprime les étendues d'âge 40 à 49 et 50 à 59, car aussi peu que quatre (4) enregistrements possèdent l'attribut en question et le résultat est le tableau suivant. Le total demeure inchangé puisque la cellule entière représente au moins quatre personnes.

Exemple d'estimations qui seraient publiées avec l'application de la suppression fondée sur la cellule

	Étendue d'âge				
Valeurs	20 à 29	30 à 39	40 à 49	50 à 59	Total
Estimation de l'ENM	50	55	0	0	195
Nombre d'enregistrements	8	4	1	2	15

La principale raison d'être de cette règle est de prévenir la divulgation des renseignements personnels liés à certaines personnes. Dans l'exemple ci-dessus, s'il n'y a qu'une seule personne dans une région donnée qui se situe dans l'étendue d'âge 40 à 49 ans, le fait que cette personne soit un répondant de l'ENM ne sera pas divulgué, et par le fait même, minimise le risque que plus de renseignements provenant de l'ENM au sujet de cette personne ne soient divulgués.

3. Pratiques relatives à la confidentialité

3.1 Suppression de régions pour les données sur les caractéristiques du revenu

La règle de suppression des régions consiste à remplacer les données sur les caractéristiques du revenu par un « x » dans le cas des régions géographiques dont la population et/ou le nombre de ménages est inférieur à un seuil déterminé.

Si une totalisation de l'ENM contient des données quantitatives sur le revenu (p. ex., le revenu total, les salaires), des données qualitatives fondées sur les notions de revenu (p. ex., état du faible revenu avant impôt) ou des données dérivées fondées sur les variables quantitatives du revenu (p. ex., les indices) pour des personnes, des familles ou des ménages, la règle suivante s'applique : les données sur les caractéristiques du revenu sont remplacées par un « x » pour les régions dans lesquelles la population est inférieure à 250 ou dans lesquelles le nombre de ménages privés est inférieur à 40. Le seuil de ménage privé ne s'applique pas aux totalisations fondées sur les géographies du lieu de travail.

3.2 Ajustement des estimations sur le lieu de travail pour des raisons de confidentialité

Il est possible d'obtenir des estimations personnalisées du lieu de travail pour les îlots de diffusion. Ces estimations seront rajustées pour renforcer le caractère confidentiel des données. De fait, toutes les estimations de la population active occupée ayant un lieu de travail habituel ou ayant travaillé à domicile seront arrondies à un multiple de 5. Cet ajustement sera cependant contrôlé. Cela signifie que les agrégats (totaux) des estimations rajustées de population ajustés pour les aires de diffusion seront toujours inférieurs ou supérieurs de moins de 5 unités des valeurs réelles.

3.3 Ajustement des estimations de la population de jour pour des raisons de confidentialité

Les estimations de la population de jour seront déterminées en ajoutant à la population résidant dans un secteur donné les travailleurs qui résident ailleurs, mais qui travaillent dans le secteur en question, puis en soustrayant du résultat obtenu les travailleurs qui résident dans le secteur, mais qui travaillent ailleurs. Le nombre de travailleurs est fondé sur la population active occupée ayant un lieu de travail habituel ou ayant travaillé à domicile. Les estimations de la population de jour seront rajustées afin de mieux protéger le caractère confidentiel des données par une méthode d'arrondissement contrôlé des estimations à un multiple de 5.

3.4 Prévention de la divulgation

Il faut également tenir compte des risques de divulgation directe ou par recoupements au moment de déterminer le contenu des produits. L'évaluation du risque nécessite la prise en considération d'un certain nombre de facteurs. Les données détaillées relatives aux diverses variables, le recoupement des variables et l'échelon géographique selon lequel les données sont réparties auront tous une incidence sur le niveau de risque. Par exemple, il se peut que la production de totalisations indiquant le nombre de pièces que compte le logement et la tranche de valeur détaillée du logement selon diverses

caractéristiques des membres du ménage ne pose pas de risque de divulgation dans le cas de grandes régions géographiques, mais que ce risque augmente pour les unités géographiques de niveau inférieur.

La méthode la plus largement utilisée pour prévenir la divulgation consiste à définir le contenu qui convient à un niveau géographique donné. Il est aussi possible de définir des seuils de population croissants ou de procéder à des suppressions manuelles, au besoin. Étant donné que ces méthodes sont liées aux exigences propres des divers produits, elles ne sont pas appliquées par les systèmes de suppression automatisés.

3.5 Totalisation du Recensement de l'agriculture

Les données du Recensement de l'agriculture et de l'Enquête nationale auprès des ménages (ENM) sont appariées en utilisant l'information géographique et les caractéristiques des exploitants agricoles (c.-à-d. l'âge et le sexe). Le taux d'appariement s'établit à 95 % environ et on effectue une pondération afin de tenir compte des cas de non-appariement. Il est possible d'obtenir des données sur tous les membres des ménages dont un des membres est un exploitant agricole.

Les données du Recensement de l'agriculture comprennent le type de ferme, les ventes agricoles, les superficies des terres de culture et les nombres de têtes de bétail, tandis que l'ENM fournit des données socioéconomiques, y compris la scolarité, le revenu et la profession des familles et des membres du ménage. Les produits normalisés dont l'élaboration est prévue d'avance sont offerts à l'échelle provinciale seulement.

Les produits personnalisés sont disponibles aux niveaux provinciaux. Les données sont soumises à un arrondissement aléatoire et à une suppression des valeurs inférieures à un seuil précisé pour assurer la protection de la confidentialité. Les suppressions sont effectuées manuellement dans le cas des cellules comportant une valeur inférieure à un seuil déterminé.

Les totalisations sont toutes vérifiées à l'interne par le personnel du Recensement de l'agriculture.

3.6 Fichiers de microdonnées à grande diffusion (FMGD)

Les produits FMGD de l'ENM de 2011 consisteront en deux fichiers de microdonnées : le fichier des particuliers et le fichier hiérarchique. Le fichier des particuliers contiendra les enregistrements d'environ 3 % de la population canadienne et le fichier hiérarchique contiendra les enregistrements de 1 % de la population dans les ménages privés.

Les fichiers de microdonnées sont uniques parmi les produits de l'ENM car ils sont les seuls qui permettent aux utilisateurs d'avoir accès aux données non agrégées. Les FMGD sont donc de puissants

outils de recherche. Chaque fichier comprend un grand nombre de variables. Les utilisateurs peuvent regrouper et manipuler ces variables en fonction de leurs besoins en données et de leurs exigences en matière de recherche. Des totalisations qui ne sont pas incluses dans d'autres produits de l'ENM peuvent être créées, ou des relations entre les variables peuvent être analysées en utilisant divers outils d'analyse.

Les fichiers de microdonnées à grande diffusion (FMGD) de l'ENM permettent d'accéder rapidement à une vaste base de données sociales et économiques concernant le Canada et sa population. Ils consistent en échantillons de réponses anonymes au questionnaire de l'ENM (formules N1, N2). Les FMGD contiennent des informations statistiques sur les Canadiens, les familles et les ménages auxquels ils appartiennent et les logements dans lesquels ils vivent et permettent aux chercheurs d'étudier les relations entre ces univers.

Statistique Canada est tenu de protéger la confidentialité des données qu'il recueille. Étant donné la nature même d'un fichier de microdonnées, diverses mesures sont mises en place afin de respecter cet engagement. Le Comité de la diffusion des microdonnées examine toutes les demandes concernant la diffusion de microdonnées.

Les données pour de petites régions géographiques ne seront pas disponibles dans ces fichiers. L'utilisateur trouvera l'information uniquement pour certaines régions métropolitaines de recensement, les provinces et les territoires. Les données variables sont agrégées afin de préserver la confidentialité, tout en fournissant le plus de détails possibles pour maintenir la valeur analytique du fichier. Certaines des valeurs de variables sensibles sont supprimées parce que si on les combinait, on pourrait les utiliser pour révéler l'identité d'une personne, d'une famille ou d'un ménage. En outre, toutes les variables quantitatives de valeurs en dollars sont assujetties à un arrondissement aléatoire et on leur attribue une valeur plafond et une valeur plancher.

4. Pratiques relatives à la qualité des données

La section ci-après décrit les méthodes utilisées pour restreindre la diffusion des données de l'ENM de qualité inacceptable.

4.1 Mesures de la qualité des données

4.1.1 Indicateurs de la qualité des données relatifs aux totalisations selon les géographies du lieu de résidence

4.1.1.1 Indicateurs de la qualité des données

Des indicateurs de la qualité des données sont liés à toutes les régions géographiques normalisées du lieu de résidence sur lesquelles des données sont diffusées. Dans les environnements de base de données de l'ENM, les indicateurs de la qualité des données consistent en un champ numérique à cinq chiffres. Dans la base de données et dans les produits électroniques parcourus à l'aide de Beyond 20/20, ces indicateurs sont affichés en utilisant un code numérique à cinq chiffres (exemple : 1 0 0 1 0). Sur le site Web de l'ENM, les régions partiellement dénombrées sont indiquées aux utilisateurs au moyen de signes conventionnels. Les signes conventionnels utilisés pour l'ENM de 2011 sont décrits à la section 6.4.

4.1.1.1.1 Régions partiellement dénombrées

En 2011, il y avait au total 36 réserves indiennes et établissements indiens qui étaient partiellement dénombrés dans l'ENM. Pour ces réserves ou établissements, le dénombrement de l'ENM n'était soit pas permis, soit a été interrompu avant qu'il puisse être terminé, ou il n'était pas possible en raison d'événements naturels (plus particulièrement des feux de forêt dans le Nord de l'Ontario).

Il n'y a aucunes données concernant les réserves indiennes et les établissements indiens partiellement dénombrés dans la base de données de l'ENM. Les régions géographiques de niveaux supérieurs comprenant ces régions sont identifiées dans les produits de l'ENM.

Bien que les données de l'ENM ne soient pas disponibles pour les réserves indiennes et les établissements indiens partiellement dénombrés, les régions elles-mêmes sont comprises dans les hiérarchies géographiques normalisées des bases de données de l'ENM. Les logiciels d'extraction et de totalisation peuvent extraire ces régions, mais sans les données. Pour les géographies du lieu de travail, ces régions seront supprimées.

4.1.1.1.2 Régions incomplètement dénombrées

Toutes les régions géographiques qui renferment un secteur partiellement dénombré sont considérées comme des régions incomplètement dénombrées. On signale aux utilisateurs à l'aide d'un indicateur que ces régions contiennent des secteurs partiellement dénombrés.

4.1.1.1.3 Taux globaux de non-réponse

Le taux global de non-réponse (TGN) est un indicateur de la qualité des données qui combine les non-réponses complètes et les non-réponses partielles à l'enquête. Un TGN plus faible indique que le risque d'un biais de non-réponse est plus faible, c.-à-d. un moindre risque de manque d'exactitude. Les taux globaux de non-réponse sont déterminés pour chacune des régions géographiques de l'ENM. Ces

régions sont indiquées dans la base de données selon le taux de non-réponse. Les régions qui présentent un taux global de non-réponse supérieur ou égal à 50 % sont supprimées des produits de données normalisés, mais seront disponibles sur demande personnalisée. Les régions géographiques comportant un taux global de non-réponse inférieur à 50 % sont indiquées dans les totalisations, mais non supprimées. Dans les produits électroniques, un indicateur numérique ainsi que le taux global de non-réponse réel sont fournis.

Tableau 4.1 Indicateurs de la qualité des données pour le lieu de résidence – ENM de 2011

Caractère numérique	Description	Indicateur	Description de l'indicateur
1er (0XXXX)	**Indicateur de dénombrement partiel**	0	Valeur implicite
		1	Réserve indienne ou établissement indien partiellement dénombré (supprimées)
		2	Ne comprend pas les données de l'Enquête nationale auprès des ménages pour une ou plusieurs réserves indiennes ou établissements indiens partiellement dénombrés
2e (X0XXX)	**Sans objet**	0	Valeur implicite
3e (XX0XX)	**Sans objet**	0	Valeur implicite
4e (XXX0X)	**Indicateur relatif à la qualité des données**	0	Indice de la qualité des données signalant un taux global de non-réponse inférieur à 50 %
		1	Indice de la qualité des données signalant un taux global de non-réponse supérieur ou égal à 50 % (supprimé)
5e (XXXX0)	**Sans objet**	0	Valeur implicite

4.1.2 Indicateurs de la qualité des données relatifs aux totalisations selon les géographies du lieu de travail

Comme indiqué à la section 4.1.1.1.3, les taux globaux de non-réponse (TGN) sont déterminés pour chacune des régions géographiques de l'ENM. Par conséquent, les régions géographiques du lieu de travail (LDT) ont leurs propres taux globaux de non-réponse. Les TGN du LDT sont fondés sur la population âgée de 15 ans et plus ayant travaillé à un moment quelconque entre janvier 2010 et mai 2011, soit à un lieu habituel de travail ou à la maison, situé dans la région géographique donnée du lieu de travail, tandis que les TGN des régions géographiques du lieu de résidence (LDR) sont fondés sur la population habitant dans la région. Ainsi, les régions géographiques du lieu de travail pourraient avoir des valeurs de taux globaux de non-réponse différentes comparativement à la région géographique du lieu de résidence équivalente. Par exemple, le taux global de non-réponse pour le lieu de travail de la

subdivision de recensement de Toronto pourrait ne pas être le même que le taux global de non-réponse pour le lieu de résidence de la subdivision de recensement de Toronto.

Les TGN des LDT comme les TGN des LDR sont une estimation et non une valeur absolue. Les TGN sont donc assujettis à une certaine variance. Cependant, les TGN du LDT ont plus de variabilité que les TGN du LDR parce que l'estimation de la population du lieu de travail est assujettie à une grande variance étant donné que la population du lieu de travail ne peut être calibrée à une population du lieu de travail dénombrée dans le cadre du recensement, contrairement à l'estimation de la population du lieu de résidence qui elle peut être calibrée à une population du lieu de résidence dénombrée dans le cadre du recensement.

Tout comme pour les régions géographiques du lieu de résidence, les données pour les régions géographiques du lieu de travail qui ont un taux global de non-réponse de 50 % ou plus seront supprimées dans les produits normalisés, mais seront offertes à titre de demande personnalisée. Toutefois, il est important de noter que pour les produits normalisés, des données pourraient être offertes pour certaines régions géographiques du lieu de résidence (si leur taux global de non-réponse est inférieur à 50 %), alors qu'elles ne le seraient pas pour la région géographique du lieu de travail équivalente (si cette région géographique du lieu de travail équivalente a un taux global de non-réponse de 50 % ou plus), et vice-versa.

L'indicateur de qualité des données pour le lieu de travail utilise le quatrième chiffre du code numérique à cinq chiffres.

Tableau 4.2 Indicateur de qualité des données pour le lieu de travail, ENM de 2011

Caractère numérique	Description	Indicateur	Description de l'indicateur
4e (XXX0X)	**Indicateur relatif à la qualité des données**	0	Indice de la qualité des données indiquant un taux global de non-réponse inférieur à 50 %
		1	Indice de la qualité des données indiquant un taux global de non-réponse supérieur ou égal à 50 % (supprimé)

4.2 Autres méthodes de suppression relatives à la qualité des données

Les méthodes de suppression mentionnées jusqu'à maintenant sont suffisantes pour supprimer les régions pour lesquelles la qualité des données est inacceptable et pour signaler les données de qualité inférieure dans la plupart des produits de données de l'ENM. Toutefois, le secteur qui établit les spécifications ou le secteur chargé de la production peut demander que des règles de suppression additionnelles en raison de la qualité des données soient appliquées pour certains produits : par exemple,

en augmentant les seuils de population ou en supprimant des répartitions ou des cellules. Il s'agit de règles de suppression qui s'appliquent à des produits en particulier; elles ne font donc pas partie des systèmes de suppression automatisés. Dans tous les cas, il est nécessaire d'utiliser un processus manuel.

4.2.1 Suppression de répartitions

La suppression de répartitions constitue l'exemple le plus fréquent d'autres méthodes de suppression visant à garantir la qualité des données. Cette méthode de suppression est utilisée dans certains **produits normalisés sur le revenu**, lorsque les répartitions du revenu sont supprimées parce que le **nombre total d'unités** (personnes, familles, ménages) dans la répartition du revenu est inférieur à **250**. Une variante de cette méthode est appliquée aux produits normalisés qui renferment des statistiques uniquement sur le nombre, la médiane et la moyenne du revenu d'emploi ou du revenu total.

4.3 Calcul des statistiques d'ordre

Les médianes et, de façon plus générale, les quantiles sont calculés au moyen d'interpolations linéaires. L'intervalle de quantile (c'est-à-dire l'intervalle dans lequel figure la valeur du quantile) est déterminé au moyen de deux méthodes fondées sur le genre des valeurs attribuées aux variables statistiques :

1) Des variables dont les valeurs peuvent comporter des décimales et des variables dont les valeurs sont exprimées en dollars

   L'intervalle de quantile est construit de façon à ce que les erreurs relatives découlant de l'utilisation de l'interpolation linéaire soit inférieures à 0,78 %. Par exemple, si le quantile réel est 30 000,00 $, l'erreur imputable à l'utilisation de l'algorithme intégré est inférieure à 234,00 $.

2) Variables dont les valeurs sont des nombres entiers non exprimés sous forme de dollars

   Pour ces variables, l'intervalle de quantile correspond toujours à une unité (1). Par exemple, si le quantile réel est 23,46, l'interpolation est appliquée à l'intervalle [23, 24].

4.4 Règle sur la qualité des données pour la diffusion des données relatives à la population âgée de 100 ans et plus

Les données sur la population âgée de 100 ans et plus ne peuvent être diffusées en année d'âge. Pour les demandes personnalisées qui exigent une ventilation plus détaillée que celle des produits de données normalisés, dans lesquels on groupe la population âgée de 100 ans et plus, la seule ventilation détaillée possible est comme suit, et ne peut être fournie que pour « Canada » :

Population totale âgée de 100 ans et plus

100 ans à 104 ans

105 ans à 109 ans

110 ans et plus

4.5 Règle sur la qualité des données pour la diffusion des données relatives aux couples de même sexe et de sexe opposé

Les questionnaires du Recensement de 2011 et de l'Enquête nationale auprès des ménages de 2011 ont utilisé, pour la première fois, une réponse précise sur les liens entre les membres du ménage afin de déterminer le nombre de couples mariés de même sexe. L'analyse des données sur les couples mariés de même sexe a montré qu'une surestimation de ce type de familles et d'état matrimonial a pu survenir. L'Enquête nationale auprès des ménages de 2011 montre un total de 63 920 couples de même sexe au Canada, dont 20 280 sont des couples mariés. À l'échelle nationale, l'écart de la surestimation de ces deux estimations varie entre 0 et 3 800.

Les niveaux géographiques tels que le Canada, les provinces, les territoires et les régions métropolitaines de recensement (RMR) affichent des estimations généralement plus élevées, donc on s'attend à ce que la surestimation potentielle soit relativement petite; toutefois, il faut quand même interpréter les données avec prudence.

À un niveau géographique moins élevé, la même surestimation potentielle peut être relativement importante, alors non seulement faut-il interpréter les données avec prudence, mais, certaines règles de suppression limitent leur publication. Ces règles s'appliquent aux données du Recensement de 2011 et aux données de l'Enquête nationale auprès des ménages de 2011.

Premièrement, la ventilation des données sur les couples de même sexe ou les couples de sexe opposé selon la situation conjugale, c.-à -d. s'ils sont mariés ou vivant en union libre, ne doit pas être diffusée pour des régions géographiques autres que le Canada, les provinces, les territoires et les RMR.

Deuxièmement, les données qui identifient les couples de même sexe ou les couples de sexe opposé (au total, couples mariés ou vivant en union libre) pour toutes les régions où la population compte moins de 5 000 habitants (tel qu'établi lors de l'ENM de 2011) ne doivent pas être diffusées.

En résumé,

- Toutes les données pour les couples de même sexe ou les couples de sexe opposé pour le Canada, les provinces, les territoires et les régions métropolitaines de recensement (RMR) peuvent être diffusées, toutefois elles doivent être interprétées avec prudence.

- Les données sur les couples de même sexe ou les couples de sexe opposé peuvent être diffusées pour d'autres régions géographiques si leur population compte 5 000 habitants ou plus, à condition que la ventilation selon la situation conjugale (couples mariés ou vivant en union libre) ne soit pas incluse.
- Aucunes données identifiant les couples de même sexe ou les couples de sexe opposé ne peuvent être diffusées pour les régions comptant une population de moins de 5 000 habitants.

5. Suppression – Réserves indiennes

5.1 Suppression de N2 des réserves indiennes

On supprime aussi des données lorsque certaines questions ne sont pas posées à tous les répondants. Les questions sur la citoyenneté (question 10), le statut d'immigrant reçu (question 11) et l'année d'immigration (question 12) n'ont pas été posées aux personnes qui vivent dans des réserves indiennes et dans des établissements indiens et qui ont été dénombrées à l'aide du questionnaire N2 de l'ENM de 2011. Cependant, il est possible qu'une subdivision de recensement (SDR) ou une région géographique de niveau inférieur ait été dénombrée à l'aide du questionnaire N2 (population dans la réserve) et du questionnaire N1 (population hors réserve). Dans ce cas, il a fallu appliquer les règles suivantes afin de déterminer si toutes les données portant sur la citoyenneté et l'immigration devaient être supprimées pour cette SDR (ou région géographique de niveau inférieur) :

1) Si les estimations de la population dénombrée à l'aide du questionnaire N1 sont supérieures aux estimations de la population dénombrée à l'aide du questionnaire N2 (d'après les résultats pondérés), les estimations portant sur la citoyenneté et l'immigration doivent être incluses.

2) Si les estimations de la population dénombrée à l'aide du questionnaire N2 sont supérieures ou égales aux estimations de la population dénombrée à l'aide du questionnaire N1 (d'après les résultats pondérés), les estimations portant sur la citoyenneté et l'immigration doivent être exclues.

Par conséquent, les données portant sur la citoyenneté, le statut d'immigrant reçu et la période d'immigration sont supprimées pour les réserves indiennes et les établissements indiens au niveau de la subdivision de recensement et aux niveaux géographiques inférieurs où la majorité de la population a été dénombrée à l'aide du questionnaire N2. Ces données sont toutefois comprises dans les totaux pour les régions géographiques plus grandes, telles que les divisions de recensement et les provinces.

Pour obtenir une liste complète des réserves indiennes et des établissements indiens pour lesquels les données portant sur la citoyenneté, le statut d'immigrant reçu et la période d'immigration ont été supprimées, veuillez vous reporter à : http://www12.statcan.gc.ca/nhs-enm/2011/ref/sup_N2-fra.cfm.

6. Suppression des données – Autre

Comme il est mentionné à la section 5, on supprime des données lorsque certaines questions ne sont pas posées à tous les répondants. En outre, des règles de suppression de données peuvent être appliquées pour des raisons propres à des produits en particulier, par exemple, la taille des produits ou les contraintes liées au support de diffusion utilisé.

6.1 Fréquence de déclaration

On se sert de la fréquence de déclaration pour ordonner ou classer les données sur les caractéristiques (variables) selon la taille dans les produits. On peut ainsi choisir d'inclure dans un produit uniquement les « n » catégories les plus importantes d'une caractéristique.

6.2 Suppression des zéros

La suppression des zéros consiste à retirer les enregistrements dans lesquels toutes les estimations sont des zéros. Cette mesure sert à réduire la taille d'un produit en supprimant toutes les lignes de la matrice de sortie renfermant des zéros.

6.3 Suppression des données sur le lieu de travail dans les tableaux de flux

Ce qui suit est une ligne directrice uniquement et il n'est pas obligatoire de la mettre en œuvre dans tous les tableaux de flux. L'objet principal consiste à ce que l'application aux produits de données normalisés soit affichée sur Internet.

Le processus de suppression de données sur le lieu de travail permet de retirer des tableaux de flux (lieu de travail) les enregistrements renfermant des estimations très faibles. On peut l'utiliser pour choisir, en vue de les inclure dans un produit, uniquement les enregistrements indiquant une estimation de flux de navettage supérieure à une valeur seuil. La valeur seuil implicite est 20. Il s'agit d'exigences qui s'appliquent typiquement à des produits particuliers; elles ne font donc pas partie des systèmes de suppression automatisés.

6.4 Signes conventionnels relatifs à la qualité et à la confidentialité dans les tableaux de données

Les produits normalisés de l'ENM contiendront deux signes conventionnels dans les tableaux propres à Statistique Canada. Le tableau ci-après montre chaque signe conventionnel et donne sa description.

Signes conventionnels relatifs à la qualité et à la confidentialité dans les tableaux de données

Signes conventionnels	Description
...	n'ayant pas lieu de figurer
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*

Note : Lorsque la suppression de statistiques est appliquée (p. ex., MMM, nombre d'enregistrements, estimation, valeur aberrante), un zéro s'affichera au lieu du symbole « x ». Cela est effectué afin d'éliminer le risque de divulgation par recoupements.