Annexe B. Qualité des données, échantillonnage et pondération, confidentialité et arrondissement aléatoire

Modifiée le 4 décembre 2003

Qualité des données

Généralités

Le recensement de 2001 a été une entreprise complexe et de grande envergure. Bien que l'on ait déployé des efforts considérables pour assurer le respect de normes élevées au cours des opérations de la collecte et du traitement, il est inévitable que les estimations résultantes soient entachées d'erreurs. Les utilisateurs des données du recensement doivent savoir que ces erreurs existent et doivent avoir une idée générale de leurs principales composantes afin d'être en mesure de déterminer l'utilité des données produites et d'évaluer les risques qu'ils courent en tirant des conclusions ou en prenant des décisions à partir de ces données.

Des erreurs peuvent se produire pratiquement à toutes les étapes du recensement, depuis la préparation des documents jusqu'au traitement des données, en passant par l'établissement des listes de logements et la collecte des données. Certaines erreurs, qui surviennent par hasard, ont tendance à s'annuler lorsque les réponses fournies par les divers répondants sont agrégées pour un groupe assez important. Dans le cas d'erreurs de cette nature, l'estimation correspondante sera d'autant plus précise que le groupe visé sera grand. C'est pourquoi on conseille aux utilisateurs de faire preuve de prudence lorsqu'ils utilisent des estimations relatives à de petits groupes. Toutefois, certaines erreurs peuvent survenir de façon plus systématique et introduire un « biais » dans les estimations. Comme ce biais persiste quelle que soit la taille du groupe pour lequel les réponses sont agrégées et comme il est particulièrement difficile d'en mesurer l'importance, les erreurs systématiques posent pour la plupart des utilisateurs de données des problèmes plus graves que les erreurs aléatoires mentionnées plus haut.

En ce qui concerne les données du recensement en général, les principaux types d'erreurs sont les suivants :

– les **erreurs de couverture** qui se produisent lorsqu'on oublie des logements ou des personnes, qu'on les dénombre à tort ou qu'on les compte plus d'une fois;

– les **erreurs dues à la non-réponse** qui surviennent lorsqu'on n'a pu obtenir de réponses d'un certain nombre de ménages ou de personnes en raison d'une absence prolongée ou pour toute autre raison;

– les **erreurs de réponse** qui surviennent lorsque le répondant, ou parfois le recenseur, a mal interprété une question du recensement et a inscrit une mauvaise réponse ou s'est tout simplement trompé de case de réponse;

– les **erreurs de traitement** qui peuvent se produire à diverses étapes, notamment lors du **codage**, lorsque les réponses en lettres sont converties en codes numériques; lors de la **saisie des données**, lorsque les préposés à l'entrée des données transfèrent dans un format électronique les réponses figurant au questionnaire du recensement; lors de **l'imputation**, lorsqu'une réponse « valide », mais pas nécessairement exacte, est insérée dans un enregistrement par l'ordinateur pour remplacer une réponse manquante ou « invalide » (« valide » et « invalide » renvoient à la cohérence de la réponse, compte tenu des autres renseignements compris dans l'enregistrement);

– les **erreurs d'échantillonnage** qui s'appliquent uniquement aux questions supplémentaires figurant dans le questionnaire complet distribué à un échantillon de un cinquième des ménages. Ces erreurs résultent du fait que les réponses à ces questions supplémentaires, une fois pondérées pour représenter l'ensemble de la population, diffèrent inévitablement des réponses qu'on aurait obtenues si l'on avait posé ces questions à tous les ménages.

Les types d'erreur mentionnés plus haut ont tous une composante aléatoire et une composante systématique. Toutefois, la composante systématique de l'erreur d'échantillonnage est d'ordinaire très petite comparativement à sa composante aléatoire. Dans le cas des autres erreurs non dues à l'échantillonnage, tant la composante aléatoire que la composante systématique peuvent être importantes.

Erreurs de couverture

Les erreurs de couverture ont une incidence directe sur la précision des chiffres du recensement, c'est-à-dire sur la taille des divers univers du recensement : la population, les familles, les ménages et les logements. Bien que des mesures aient été prises pour corriger certaines erreurs identifiables, les chiffres définitifs sont toujours entachés d'une certaine erreur parce que des personnes ou des logements ont été oubliés, dénombrés à tort ou comptés plus d'une fois.

L'oubli de logements ou de personnes se traduit par un **sous-dénombrement**. Des logements peuvent être oubliés en raison soit d'une mauvaise interprétation des limites du secteur de dénombrement (SD), soit qu'ils n'ont pas l'apparence de logements ou soit qu'ils semblent inhabitables. Des personnes peuvent être oubliées parce que leur logement est oublié ou classé comme inoccupé, ou parce que le répondant a mal interprété les instructions concernant les personnes à inclure sur le questionnaire. Enfin, certaines personnes peuvent être oubliées parce qu'elles n'ont pas de domicile habituel et qu'elles n'ont pas passé la nuit du recensement dans un logement.

Le dénombrement à tort ou le double compte de logements ou de personnes se traduit par un **surdénombrement**. Il peut y avoir surdénombrement de logements lorsque des constructions impropres à l'habitation sont classées comme logements (dénombrement à tort), lorsqu'il existe une certaine ambiguïté au sujet des limites des SD ou lorsque des unités d'habitation (par exemple, des chambres) sont comptées séparément plutôt que d'être considérées comme faisant partie d'un seul logement (double compte). Les personnes peuvent être comptées plus d'une fois parce que leur logement a été compté deux fois ou parce que les lignes directrices concernant les personnes à inscrire dans le questionnaire ont été mal interprétées. À l'occasion, il arrive qu'une personne ne faisant pas partie de l'univers de la population du recensement, comme un résident étranger ou une personne fictive, soit dénombrée à tort. En moyenne, le surdénombrement est moins susceptible de se produire que le sous-dénombrement; les chiffres des logements et des personnes sont donc probablement légèrement sous-estimés.

Pour le recensement de 2001, trois études permettent de mesurer l'erreur de couverture. Dans le contexte de l'Étude sur la classification des logements, on a de nouveau visité des logements classés comme inoccupés afin de vérifier s'ils étaient réellement inoccupés le jour du recensement, et des logements dont le ménage a été classé comme non répondant afin de déterminer le nombre de résidents habituels et leurs caractéristiques. Les chiffres définitifs du recensement ont ensuite été corrigés pour tenir compte des personnes ou des ménages oubliés parce que leur logement avait été classé par erreur comme inoccupé. Il est aussi possible que les chiffres du recensement aient été corrigés pour tenir compte des logements dont le ménage a été classé comme non répondant. En dépit de ces ajustements, les chiffres définitifs peuvent tout de même être entachés d'un certain sous-dénombrement. Le sous-dénombrement tend à être plus élevé pour certains segments de la population comme les jeunes adultes (plus particulièrement ceux de sexe masculin) et les personnes récemment immigrées. L'Étude de la contre-vérification des dossiers permet de mesurer le sous-dénombrement résiduel pour le Canada, de même que pour chaque province et chaque territoire. L'Étude sur le surdénombrement a pour objet d'étudier les erreurs de surdénombrement. Ensemble, les résultats de l'Étude de la contre-vérification des dossiers et de l'Étude sur le surdénombrement fournissent une estimation du sous-dénombrement net.

Autres erreurs non dues à l'échantillonnage

Alors que les erreurs de couverture ont une incidence sur le nombre d'unités comprises dans les divers univers du recensement, d'autres erreurs influent sur les caractéristiques de ces unités.

Il est parfois impossible d'obtenir une réponse complète d'un ménage, même si le logement a été classé comme étant occupé et un questionnaire y a été livré. Il se peut que les membres du ménage aient été absents pendant toute la période du recensement ou, en de rares occasions, que le membre responsable du ménage ait refusé de remplir le questionnaire. Il arrive plus souvent que le questionnaire soit retourné, mais qu'il y ait des questions laissées sans réponse. Des efforts sont déployés afin d'obtenir un questionnaire le plus complet possible. Les recenseurs contrôlent les questionnaires et assurent un suivi à l'égard de l'information manquante. Le travail du recenseur est ensuite vérifié par un surveillant et par un technicien du contrôle qualitatif. Malgré tout, il existe toujours un petit nombre de réponses manquantes à la fin de l'étape de la collecte, c'est-à-dire d'**erreurs dues à la non-réponse**. Bien que les réponses manquantes soient éliminées en cours de traitement en remplaçant chacune d'elles par la réponse correspondante figurant dans un enregistrement « similaire », il est possible que certaines erreurs d'imputation s'y glissent. Cela est particulièrement grave lorsque les personnes non répondantes diffèrent des répondants sous certains aspects; en effet, cette procédure introduit un **biais dû à la non-réponse**.

Même lorsqu'une réponse est obtenue, il se peut qu'elle ne soit pas tout à fait exacte. Il est possible que le répondant ait mal interprété la question ou ait donné une réponse au jugé, surtout lorsqu'il répondait pour le compte d'un autre membre du ménage, qui était peut-être absent. Il est aussi possible que le répondant ait inscrit sa réponse au mauvais endroit sur le questionnaire. Ces erreurs sont désignées sous le nom d'**erreurs de réponse**. Bien que ces erreurs surviennent d'ordinaire parce que les répondants ont fourni des renseignements inexacts, elles peuvent aussi résulter d'erreurs commises par les recenseurs qui ont rempli certaines parties du questionnaire, comme le type de construction résidentielle, ou qui ont effectué le suivi pour obtenir une réponse manquante.

Certaines questions du recensement nécessitent une réponse en lettres. Pendant le traitement, on attribue un code numérique à ces réponses. Il est possible que des **erreurs de codage** se produisent lorsque la réponse écrite est ambiguë, incomplète ou difficile à lire, ou lorsque la liste des codes est longue (p. ex. principal domaine d'études, lieu de travail). L'étape formelle du contrôle qualitatif (CQ) permet de cerner et de rectifier les erreurs de codage et d'en réduire le nombre. À l'intérieur de chaque unité de travail, un échantillon des réponses est codé indépendamment une deuxième fois. La résolution des incohérences entre les premier et deuxième codages détermine la nécessité, s'il y a lieu, de coder à nouveau l'unité de travail. Exception faite pour le codage des variables de l'industrie et de la profession, la plupart des tâches de codage du recensement sont maintenant automatisées, ce qui a pour conséquence de réduire le nombre d'erreurs de codage.

Les renseignements figurant dans les questionnaires sont tapés dans un fichier informatique. Deux méthodes de résolution ordonnée sont utilisées pour limiter le nombre d'**erreurs à la saisie des données**. Dans un premier temps, certains contrôles (comme des vérifications d'étendue) sont effectués à mesure que les données sont entrées. Dans un second temps, on tape de nouveau à l'ordinateur un échantillon tiré de chaque lot de documents, puis on compare les entrées résultantes aux entrées initiales. Le travail non satisfaisant est ainsi circonscrit et corrigé et, si cela est nécessaire, le reste du lot est de nouveau saisi.

Une fois saisies, les données font l'objet de vérifications qui consistent à les soumettre à une série de contrôles informatiques visant à relever les réponses manquantes ou incohérentes. À l'étape de l'imputation, on substitue à ces dernières des réponses déduites à partir des autres données de l'enregistrement ou des réponses tirées d'un enregistrement donneur similaire. L'imputation permet d'obtenir une base de données complète dont les données correspondent aux chiffres du recensement et facilitent les analyses multidimensionnelles. Même si des erreurs peuvent être introduites à l'**étape de l'imputation**, les méthodes utilisées ont fait l'objet de tests rigoureux visant à réduire au minimum les erreurs systématiques.

Diverses études sont réalisées afin d'évaluer la qualité des réponses obtenues dans le cadre du recensement de 2001. Ainsi, on a calculé les taux de non-réponse et les taux de rejet au contrôle pour chaque question. Ces taux peuvent permettre de déterminer le potentiel d'erreurs dues à la non-réponse et d'autres types d'erreurs. De même, les totalisations établies à partir des données du recensement de 2001 ont été ou seront comparées avec les estimations correspondantes obtenues à partir des données des recensements précédents, des enquêtes-échantillon (comme l'Enquête sur la population active) et de divers dossiers administratifs (comme les registres des naissances et le cadastre municipal). Ces comparaisons peuvent permettre de cerner les problèmes de qualité éventuels ou, à tout le moins, de relever les divergences entre les sources.

Outre ces comparaisons entre données agrégées, certaines études de couplage de microdonnées sont actuellement menées afin de comparer les réponses de certains particuliers obtenues au recensement à celles d'une autre source de renseignements. Pour un certain nombre de caractéristiques « stables » (comme l'âge, le sexe, la langue maternelle et le lieu de naissance), on compare les réponses obtenues auprès d'un échantillon de personnes à l'occasion du recensement de 2001 aux réponses obtenues des mêmes personnes à l'occasion du recensement de 1996.

Erreurs d'échantillonnage

Les estimations obtenues en pondérant les réponses recueillies auprès d'un échantillon sont susceptibles d'être entachées d'erreurs en raison de la répartition des caractéristiques au sein de l'échantillon, qui n'est généralement pas identique à la répartition correspondante au sein de la population dans laquelle l'échantillon a été prélevé.

L'erreur susceptible d'être introduite par l'échantillonnage variera en fonction de la rareté relative de la caractéristique étudiée au sein de la population. Lorsque la valeur contenue dans la case est élevée, cette erreur sera relativement faible proportionnellement à cette valeur. Lorsque la valeur contenue dans la case est faible, cette erreur sera relativement importante proportionnellement à cette valeur.

L'erreur susceptible d'être introduite par l'échantillonnage est d'ordinaire exprimée sous forme d'« erreur type ». Il s'agit de la racine carrée de la moyenne, calculée pour l'ensemble des échantillons de même taille prélevés selon le même plan d'échantillonnage, des carrés de l'écart de l'estimation obtenue à partir de l'échantillon par rapport à la valeur pour l'ensemble de la population.

Le tableau ci-dessous fournit des mesures approximatives de l'erreur type due à l'échantillonnage. Ces mesures sont données uniquement à titre indicatif.

Erreur type approximative due à l'échantillonnage pour les données-échantillon du recensement de 2001

Valeur contenue dans la case	Erreur type approximative
50 ou moins	15
100	20
200	30
500	45
1 000	65
2 000	90
5 000	140
10 000	200
20 000	280
50 000	450
100 000	630
500 000	1 400

Les utilisateurs souhaitant déterminer l'erreur d'échantillonnage approximative pour une case de données dont la valeur a été obtenue à partir de l'échantillon de 20 % doivent choisir l'erreur type correspondant à l'entrée dans la colonne « Valeur contenue dans la case » ci-dessus qui se rapproche le plus de celle qui figure dans la case de données de la totalisation en cause. En utilisant la valeur ainsi obtenue pour l'erreur type, l'utilisateur peut, en général et à juste titre, être certain que la valeur réelle pour la population dénombrée (ne tenant pas compte des formes d'erreurs autres que l'erreur d'échantillonnage) ne s'écarte pas de la valeur contenue dans la case dans une proportion supérieure ou inférieure à trois fois l'erreur type (p. ex., si la valeur contenue dans la case est 1 000, la fourchette à l'intérieur de laquelle se situe la valeur réelle serait de 1 000 ± [3 x 65] ou de 1 000 ± 195).

Les erreurs types données dans le tableau ci-dessus ne s'appliquent pas aux chiffres de population, de logements, de ménages ou de familles pour la région géographique étudiée (voir Échantillonnage et pondération ci-dessous). On peut déterminer l'effet de l'échantillonnage pour ces valeurs en les comparant à celles des produits correspondants contenant des données intégrales.

Il est à noter que l'effet du plan d'échantillonnage et de la méthode de pondération utilisés dans le cadre du recensement de 2001 variera d'une caractéristique à l'autre et d'une région géographique à l'autre. Il est donc possible que les valeurs de l'erreur type données dans le tableau ci-dessus sous-estiment ou surestiment l'erreur attribuable à l'échantillonnage.

Échantillonnage et pondération

Les données du recensement de 2001 sont soit des données intégrales (c'est-à-dire recueillies auprès de l'ensemble des ménages), soit des données-échantillon (c'est-à-dire recueillies auprès d'un échantillon aléatoire comprenant un ménage sur cinq) que l'on a pondérées pour obtenir des estimations pour l'ensemble de la population. Les données ont été recueillies auprès d'un échantillon de 20 % et pondérées pour compenser pour l'échantillonnage. Tous les en-têtes de tableaux sont annotés en conséquence. On notera que, dans les réserves indiennes et les régions éloignées, toutes les données ont été recueillies auprès de l'ensemble de la population.

Il est possible que, pour une région géographique donnée, le total ou le total partiel pondéré de la population, des ménages, des logements ou des familles diffère du chiffre correspondant figurant dans les publications contenant des données intégrales. Ces variations sont attribuables à l'échantillonnage et au fait que les données intégrales n'excluent pas les pensionnaires d'établissements institutionnels, contrairement aux données-échantillon.

Confidentialité et arrondissement aléatoire

Afin de protéger le caractère confidentiel des renseignements fournis, les chiffres indiqués aux tableaux ont fait l'objet d'un **arrondissement aléatoire** qui supprime toute possibilité d'associer des données statistiques à une personne facilement reconnaissable. Selon cette méthode, tous les chiffres, y compris les totaux et les marges, sont arrondis de façon aléatoire (vers le haut ou vers le bas) jusqu'à un multiple de « 5 » et, dans certains cas, de « 10 ». Cette technique assure une protection efficace contre la divulgation sans ajouter d'erreur significative dans les données du recensement. Les utilisateurs doivent savoir que les totaux et les marges sont arrondis séparément et qu'ils ne correspondent pas nécessairement à la somme des chiffres arrondis séparément dans les répartitions. De plus, il faut s'attendre à ce que les totaux et les autres chiffres correspondants dans diverses totalisations du recensement présentent quelques légères différences. De même, la somme des pourcentages, qui sont calculés à partir de chiffres arrondis, ne correspond pas forcément à 100 %. Les statistiques d'ordre (médiane, quartiles, percentiles, etc.) ainsi que les mesures de dispersion comme l'erreur type sont calculées de la façon habituelle. Lorsqu'une statistique est définie comme le quotient de deux nombres (c'est le cas pour des moyennes, des pourcentages et des proportions), les deux nombres sont arrondis avant d'effectuer la division. S'il s'agit de revenu, de dépenses de propriété, de valeur du logement, d'heures travaillées, de semaines travaillées ou d'âge, la somme est définie comme le produit de la moyenne par la fréquence pondérée arrondie. Sinon, c'est la somme pondérée qui est arrondie. La distorsion importante pouvant résulter de l'arrondissement aléatoire dans le cas des cases de faible valeur mérite aussi d'être signalée. Cette distorsion peut entraîner une perte de précision pour les cases de données renfermant des chiffres peu élevés. De plus, une statistique est supprimée si le nombre actuel d'enregistrements ayant servi au calcul est inférieur à 4 ou si la somme du poids de ces enregistrements est inférieure à 10. En outre, dans le cas de valeurs exprimées en dollars, d'autres règles

s'ajoutent. Ainsi, pour les produits normalisés, si toutes les valeurs sont égales, la statistique est supprimée. Pour tous les autres produits, la statistique est supprimée si l'étendue des valeurs est trop petite ou si toutes les valeurs sont inférieures, en valeur absolue, à un certain seuil.

Les utilisateurs devraient, lors de l'agrégation des données arrondies, être conscients de cette distorsion. Les erreurs dues à l'arrondissement ont tendance à s'annuler lorsque les chiffres contenus dans les cases sont agrégés de nouveau. Cependant, il est possible de réduire les distorsions en intégrant dans la mesure du possible les totaux partiels appropriés dans les totalisations.

Les utilisateurs désirant obtenir un maximum de précision peuvent aussi choisir de demander des totalisations personnalisées. Dans le cas de produits personnalisés, l'agrégation se fait à partir des enregistrements dans la base de données du recensement se rapportant aux particuliers. L'arrondissement aléatoire a lieu uniquement après que les cases de données ont été agrégées, ce qui réduit la distorsion au minimum.

Outre l'arrondissement aléatoire, on a adopté la technique de la **suppression des régions**, afin d'assurer encore mieux la confidentialité des réponses des particuliers.

Dans le cadre de la **suppression des régions**, toutes les données caractéristiques se rapportant aux régions géographiques dont la population est inférieure à une taille donnée sont supprimées. L'importance de la suppression est fonction des facteurs suivants :

– Si les données sont totalisées à partir de la base de données intégrales, elles sont supprimées si la population totale de la région est inférieure à 40 personnes.

– Si les données sont totalisées à partir de la base de données-échantillon, elles sont supprimées si la population totale de la région, à l'exclusion des pensionnaires d'un établissement institutionnel, est inférieure à 40 personnes, selon la base de données intégrales ou la base de données-échantillon.

Il y a quelques exceptions à ces règles :

– Les données renfermant une répartition du revenu et les statistiques connexes sont supprimées si la population de la région, à l'exclusion des pensionnaires d'un établissement institutionnel, est inférieure à 250 personnes selon la base de données intégrales ou la base de données-échantillon, ou encore si le nombre de ménages privés est inférieur à 40, selon la base de données-échantillon.

– Les données renfermant une répartition du lieu du travail et les statistiques connexes sont supprimées si le nombre de personnes occupées dans la région est inférieur à 40, selon la base de données-échantillon. Si ces données incluent, en plus, une répartition du revenu, le seuil est changé à 250 personnes, toujours selon la base de données-échantillon.

– Les totalisations traitant à la fois du lieu de travail et du lieu de résidence ainsi que les statistiques connexes sont supprimées si le nombre de personnes occupées dans la région est inférieur à 40 selon la base de données-échantillon ou si la population totale de la région, à l'exclusion des pensionnaires d'un établissement institutionnel, selon la base de données intégrales ou la base de données-échantillon est inférieure à 40 personnes. Si ces totalisations incluent, en plus, une répartition du revenu, le seuil est changé à 250 personnes dans tous les cas et les totalisations sont supprimées si le nombre de ménages privés dans la région du lieu de résidence est inférieur à 40.

– Les données renfermant une répartition sur les couples de même sexe et les statistiques connexes sont supprimées si la population de la région dans les ménages privés est inférieure à 5 000 personnes, selon la base de données-échantillon.

– Si les données sont totalisées à partir de la base de données intégrales et se réfèrent aux codes postaux de six caractères ou encore à des regroupements d'îlots ou de côtés d'îlots, elles sont supprimées si la population totale de la région est inférieure à 100 personnes.

– Si les données sont totalisées à partir de la base de données-échantillon et se réfèrent aux codes postaux de six caractères ou encore à des regroupements d'îlots ou de côtés d'îlots, elles sont supprimées si la population totale de la région, à l'exclusion des pensionnaires d'un établissement institutionnel, et selon la base de données intégrales ou la base de données-échantillon, est inférieure à 100 personnes.

– Si les données se réfèrent à des regroupements d'îlots ou de côtés d'îlots, et renferment une répartition du lieu de travail, elles sont supprimées si le nombre de personnes occupées dans la région est inférieur à 100 selon la base de données-échantillon.

– Si les données se réfèrent à des regroupements d'îlots ou de côtés d'îlots, et renferment, à la fois, une répartition du lieu de travail et du lieu de résidence, elles sont supprimées si le nombre total de personnes occupées dans la région est inférieur à 100 selon la base de données-échantillon ou si la population totale de la région, à l'exclusion des pensionnaires d'un établissement institutionnel, selon la base de données intégrales ou la base de données-échantillon, est inférieure à 100 personnes.

Dans tous les cas, les données supprimées sont incluses dans les totaux ou totaux partiels du niveau d'agrégation supérieur approprié.

La technique de suppression est appliquée à tous les produits renfermant des données infraprovinciales (c'est-à-dire la série des Profils, les tableaux croisés de base, les produits personnalisés et semi-personnalisés), qu'il s'agisse de données intégrales ou de données-échantillon.

Pour obtenir de plus amples renseignements sur la qualité des données du recensement, veuillez communiquer avec la Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6, ou en composant le (613) 951-4783.