Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Surdénombrement dans le recensement canadien de 2011

Abel Dasylva, Robert-Charles Titus et Christian Thibault[1]

Résumé

L'Étude sur le surdénombrement du recensement (ESR) est une étude essentielle de mesure postcensitaire de la couverture. Son principal objectif est de produire des estimations du nombre de personnes dénombrées de façon erronée, selon la province et le territoire, et d'examiner les caractéristiques des personnes dénombrées plus d'une fois, afin de déterminer les raisons possibles de ces erreurs. L'ESR est fondée sur l'échantillonnage et l'examen manuel de groupes d'enregistrements reliés, qui sont le résultat d'un couplage de la base de données des réponses du recensement et d'une base administrative. Dans cette communication, nous décrivons la nouvelle méthodologie de l'ESR de 2011. De nombreuses améliorations ont été apportées à cette méthodologie, y compris une plus grande utilisation du couplage d'enregistrements probabiliste, l'estimation de paramètres de couplage au moyen d'un algorithme espérance-maximisation (EM), et l'utilisation efficace de données sur les ménages pour déceler davantage de cas de surdénombrement.

Mots clés : recensement, couplage d'enregistrements, surdénombrement, espérance-maximisation (EM), sources administratives, groupe d'enregistrements.

1. L'Étude sur le surdénombrement du recensement

Dans le recensement de la population, un surdénombrement a lieu lorsque la même personne est dénombrée plusieurs fois. En 2011, des changements importants ont été apportés au processus du recensement, qui ont fait augmenter la probabilité de surdénombrement. Parmi ces changements figuraient l'adoption de formulaires détaillés optionnels pour l'Enquête nationale auprès des ménages, une méthodologie par vague pour encourager les réponses en ligne et une proportion plus grande de questionnaires envoyés par la poste.

Comme c'est le cas pour le sous-dénombrement, le surdénombrement a des répercussions importantes sur l'exactitude des chiffres du recensement. C'est pourquoi il est mesuré par l'Étude sur le surdénombrement du recensement (ESR), qui résulte de dénombrements multiples dans la population cible. En 2011, la population cible incluait tous les citoyens canadiens et les immigrants reçus qui, le jour du recensement, avaient un lieu de résidence habituel au Canada ou qui vivaient à l'étranger sur une base militaire, dans une mission diplomatique, en mer ou à quai à bord d'un navire marchand enregistré au Canada. La population comprise dans le champ de l'enquête comprenait aussi les résidents non permanents et les membres de leur famille vivant avec eux, si leur lieu de résidence habituel était au Canada et s'ils demandaient le statut de réfugié ou étaient titulaires d'un permis valide d'études ou de travail, pour une période englobant le jour du recensement. Les estimations de l'ESR représentent un élément important du sous-dénombrement net, qui est défini comme la différence entre le sous-dénombrement et le surdénombrement. Il s'agit aussi d'un élément important du Programme des estimations démographiques de Statistique Canada.

L'ESR de 2011 est conçue comme une enquête sur échantillon, le surdénombrement étant estimé au moyen d'un échantillon probabiliste tiré d'une base de sondage de *cas de surdénombrement potentiel*, c'est-à-dire des groupes d'enregistrements du recensement qui sont reliés grâce à un couplage d'enregistrements, même si les enregistrements peuvent ne pas représenter un surdénombrement réel. Par conséquent, l'ESR comprend toutes les étapes des enquêtes types sur échantillon, de l'élaboration de la base de sondage à l'estimation.

[1] Abel Dasylva, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6, Canada (abel.dasylva@statcan.gc.ca); Robert-Charles Titus, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6, Canada (robert-charles.titus@statcan.gc.ca); Christian Thibault, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6, Canada (christian.thibault@statcan.gc.ca)

Toutefois, l'ESR est aussi une enquête inhabituelle parce que sa base de sondage est constituée grâce au couplage probabiliste d'enregistrements (Fellegi et Sunter, 1969). En fait, la base de sondage de l'ESR représente l'union de trois bases se chevauchant; la *base de l'étape 1*, la *base de l'étape 2* et la *base de l'extension*. La base de l'étape 1 est conçue par suite du couplage de la base de données des réponses du recensement et d'un fichier administratif, et de la constitution de groupes d'enregistrements qui sont reliés par les liens en découlant. La base de l'étape 2 est conçue par suite du couplage des enregistrements du recensement qui ne sont pas reliés à l'étape 1 et de l'ensemble de la base de données des réponses du recensement, et de la constitution de groupes d'enregistrements connectés par les liens en résultant. Enfin, la base de l'extension est conçue par suite de l'utilisation d'un identificateur de ménage qui est disponible dans la base de données des réponses du recensement. Deux ménages sont couplés s'ils comprennent des enregistrements du recensement qui ont été couplés à l'étape 1 et à l'étape 2. Pour une paire de ménages couplés, des liens supplémentaires sont créés entre leurs enregistrements de recensement. La base de l'extension est constituée de tous les groupes d'enregistrements qui sont reliés par ces liens.

Une autre différence importante dans le cas d'une enquête type sur échantillon touche l'étape de la collecte des données, dans laquelle un échantillon de paires d'enregistrements fait l'objet d'un examen manuel et à laquelle les répondants ne participent pas. Au cours de cet examen manuel, chaque lien d'un cas possible de surdénombrement est vérifié. Les résultats de ces vérifications sont par la suite traités pour produire les *cas vérifiés de surdénombrement*, c'est-à-dire des groupes d'enregistrements reliés par des liens avec le surdénombrement vérifié.

Dans le cadre de l'ESR de 2011, de nombreuses améliorations ont été apportées à la méthodologie de 2006, y compris un algorithme espérance-maximisation (EM), afin de déterminer les poids de couplage, comme ceux proposés par Winkler (1988) et Jaro (1989), des seuils de pondération provinciaux/territoriaux et le surdénombrement supplémentaire à partir de paires de ménages couplés.

Les sections qui suivent sont organisées ainsi : la section 2 décrit les fichiers d'entrée, la section 3, la base d'échantillonnage, la section 4, le plan d'échantillonnage et la section 5, le traitement et l'estimation. La section 6 présente les résultats.

2. Fichiers d'entrée

La base de données des réponses du recensement (BDRR) et une base administrative (BA) établie à partir de nombreuses sources ont servi de fichiers d'entrée.

2.1 BDRR

La BDRR comprend les réponses des personnes vivant dans des logements privés ou collectifs. En 2011, elle comptait 32 millions d'enregistrements et comprenait les renseignements suivants :

- Noms : prénoms et noms de famille comme deux variables distinctes
- Démographie : date de naissance et sexe
- Géographie : province/territoire, code postal, division de recensement, unité de collecte et numéro de voirie

Les prénoms et noms de famille ont été séparés en composantes et uniformisés.

2.2 Base administrative

La BA est établie pour assurer la plus grande couverture possible de la population cible. En 2011, elle comptait 47 millions d'enregistrements et comprenait les mêmes données sur les noms et les mêmes données géographiques que la BDRR. Toutefois, les données géographiques se limitaient à la province/au territoire, ainsi qu'au code postal de l'adresse postale.

La BA comprenait les enregistrements (ci-après appelés *dossiers administratifs*) provenant des sources suivantes :

- Fichiers maîtres de déclaration de revenus des particuliers T1 (FMPT1) de 2005 à 2009. Ils ont été fournis par l'Agence du revenu du Canada (ARC) et représentaient 58,2 % des dossiers administratifs.
- Fichiers de la Prestation fiscale canadienne pour enfants (PFCE) jusqu'en juillet 2011. Ils ont aussi été fournis par l'ARC et représentaient 15 % des dossiers administratifs.

- Enregistrements de naissance tirés des fichiers de la statistique de l'état civil pour la période de 1974 à 2008. Ils ont été fournis par la Division de la statistique de la santé de Statistique Canada et représentaient 12,2 % des dossiers administratifs.
- Fichiers des immigrants et des résidents non permanents jusqu'en septembre 2011. Ils ont été fournis par Citoyenneté et Immigration Canada (CIC) et représentaient 14,4 % des dossiers administratifs.
- Fichiers territoriaux des soins de santé (FTSS) jusqu'en juillet 2011. Ils ont été fournis par les territoires et représentaient 0,2 % des dossiers administratifs.

À noter que l'ajout de la PFCE en 2011 a considérablement augmenté la couverture de la BA comparativement à 2006.

Chaque source administrative est unique. Toutefois, la BA comportait des doubles entre les différentes sources, parce qu'aucun identificateur unique ne couvrait toutes les sources. Par conséquent, la BA comptait un plus grand nombre d'enregistrements que la BDRR. Pour résoudre ce problème, les liens entre un enregistrement de recensement donné et la BA ont été classés par priorité, de la façon suivante. Pour un adulte (c'est-à-dire une personne âgée de plus de 18 ans le jour du recensement) d'une province (c'est-à-dire ne vivant pas dans un territoire), les liens avec les enregistrements des FMPT1 avaient préséance, suivis par les liens avec les enregistrements de la PFCE et les liens avec les enregistrements de CIC. Selon ce plan, un lien comportant une préséance donnée a été laissé de côté en faveur d'un lien comportant une préséance plus grande. Des priorités différentes ont été utilisées pour un enfant dans une province; les liens avec les enregistrements de la PFCE ont eu la priorité la plus grande, suivis par les liens avec les enregistrements de naissance. Enfin, tous les enregistrements territoriaux du recensement ont été couplés uniquement aux enregistrements des FTSS.

3. Bases d'échantillonnage des cas de surdénombrement potentiel

Dans le cadre de l'ESR de 2011, on a élaboré trois bases d'échantillonnage des cas de surdénombrement potentiel, y compris la base de l'étape 1, fondée sur le couplage de la BDRR et de la BA, la base de l'étape 2, fondée sur le couplage de la BDRR résiduelle et de l'ensemble de la BDRR, et la base de l'extension. Chaque base comprenait des groupes d'enregistrements du recensement qui ont été reliés au moyen de liens.

3.1 Base de l'étape 1

À cette étape, la BDRR et la BA ont été couplées grâce au couplage d'enregistrements probabiliste au moyen de G-COUP, le système généralisé de couplage d'enregistrements de Statistique Canada. Ce couplage a permis la détermination des cas les plus probables de surdénombrement, pour lesquels deux enregistrements ou plus de la BDRR ont été couplés au même dossier administratif, selon l'ordre de priorité décrit à la section 2.2. Il a aussi permis de déterminer les pseudo-doublons. Il s'agit d'enregistrements du recensement qui concordent pour de nombreuses variables de couplage, mais représentent des personnes différentes.

3.1.1 Critères des pochettes

La création de pochettes était nécessaire, compte tenu de la taille des fichiers d'entrée, soit 32 et 47 millions respectivement pour la BDRR et la BA. Il a été fondé sur le code Soundex des noms, les composantes de la date de naissance, y compris les transpositions du mois et du jour de naissance, et le code postal.

3.1.2 Règles et leurs poids de résultats

Six règles ont servi à la comparaison détaillée des enregistrements compris dans des paires. Il s'agissait des suivantes :

- *Deux premiers noms de famille* : règle matricielle avec concordances partielles (y compris exacte et erreur typographique dans cet ordre) et transpositions;
- *Deux premiers prénoms* : règle matricielle avec concordances partielles (y compris exacte, erreur typographique et surnom dans cet ordre) et transpositions;
- *Jour et mois de naissance* : règle matricielle avec comparaisons exactes et transpositions;
- *Année de naissance* : comparaison exacte;

- *Code postal* : comparaison exacte;
- *Sexe* : comparaison exacte.

Les poids de résultats des règles ont été estimés au moyen d'un algorithme EM fondé sur des hypothèses d'indépendance conditionnelle, conformément à la proposition de Winkler (1988) et Jaro (1989). L'algorithme comprenait trois améliorations. Tout d'abord, l'algorithme a servi à estimer directement la distribution en U à partir d'un échantillon de paires aléatoires, à l'extérieur de l'algorithme EM. Cela signifie que seules les m-probabilités et la proportion de mélange ont été estimées de façon itérative au moyen de l'algorithme EM. En deuxième lieu, la u-distribution estimée a tenu compte des interactions entre les paires non appariées satisfaisant aux critères des pochettes. En troisième lieu, l'algorithme EM a tenu compte des variables de couplage manquantes, comme il est expliqué dans la section suivante.

3.1.3 Variables de couplage manquantes

Des variables de couplage étaient manquantes dans certaines paires. Ces données manquantes peuvent être considérées comme une forme de non-réponse partielle, qui a compliqué l'estimation des poids de couplage. L'EM a été amélioré, afin de tenir compte des valeurs manquantes, selon l'hypothèse de données manquantes entièrement au hasard. Pour les règles simples, qui faisaient intervenir l'année de naissance, le code postal ou le sexe, ce choix a mené à un poids de résultat nul, chaque fois que la variable correspondante était manquante dans un des enregistrements. La situation était plus complexe pour les règles matricielles, les poids de résultats n'étant pas toujours nuls selon le modèle particulier de données manquantes. La solution représentait une amélioration par rapport aux heuristiques précédentes, qui attribuaient simplement la même valeur par défaut (c'est-à-dire nulle) à chaque règle comportant une valeur manquante (Samuels, 2011).

3.1.4 Poids de fréquence provinciaux/territoriaux des noms

Les poids de fréquence rendaient compte des fréquences relatives des noms. Ils ont été ajoutés à un poids de couplage de paires lorsque le résultat était une concordance pour la règle matricielle correspondante. Ces poids ont été calculés selon la fréquence provinciale/territoriale F du nom correspondant, à partir de la formule $-10\log_2 F$. La province/le territoire d'une paire a été déterminé au moyen de l'enregistrement du recensement. Les poids de fréquence provinciaux/territoriaux étaient essentiels parce que les distributions de noms différaient au Canada, et particulièrement au Québec.

3.1.5 Décision de couplage fondée sur les seuils provinciaux/territoriaux supérieurs

Pour chaque province et territoire, un seuil de pondération supérieur distinct a été calculé. Il était fondé sur une probabilité de non-appariement conditionnel de 1 %, à condition qu'une paire eût satisfait aux critères des pochettes et que ses poids se situaient au-dessus du seuil. Le seuil a été calculé en se fondant sur le résultat de l'algorithme EM. Des paires d'enregistrements de la BDRR-BA comportant un poids de couplage supérieur à leur seuil provincial/territorial supérieur ont été couplées. À noter que la province/le territoire d'une paire a été déterminé à partir de l'enregistrement du recensement. Enfin, les liens redondants ont été supprimés selon l'ordre de priorité établi à la section 2.2.

3.1.6 Groupes d'enregistrements

Des groupes mutuellement exclusifs d'enregistrements du recensement et de la BA reliés ont été formés à partir des paires restantes se situant au-dessus du seuil supérieur. La grande majorité des groupes était des groupes un-à-un, c'est-à-dire des paires d'enregistrements de la BDRR et de la BA. Les cas un-à-un et un-à-plusieurs (un enregistrement de la BDRR couplé à de nombreux enregistrements de la BA) ne représentaient pas le surdénombrement. Les groupes restants représentaient des cas de surdénombrement potentiel de la base de l'étape 1.

3.2 Base de l'étape 2

La majorité des enregistrements du recensement ont été couplés à la BA à l'étape 1. Les enregistrements restants de la BDRR non couplés ont constitué la *BDRR résiduelle*. Le surdénombrement de cette partie de la BDRR a été laissé de côté à l'étape 1, peut-être parce que la personne correspondante n'était pas couverte par la BA. À l'étape 2, on a décelé une partie de ce surdénombrement en couplant la BDRR résiduelle et l'ensemble de la BDRR. La méthode de couplage d'enregistrements de cette étape était similaire à celle de l'étape 1, sauf pour des changements mineurs.

3.2.1 Poids de fréquence provinciaux/territoriaux des noms

On a aussi utilisé les poids provinciaux/territoriaux de fréquence des noms. Toutefois, les fréquences nationales ont été utilisées dans des paires comportant des enregistrements de provinces ou de territoires différents.

3.2.2 Décision de couplage fondée sur les seuils provinciaux/territoriaux inférieurs

Pour chaque province/territoire, un seuil de pondération inférieur distinct a été calculé. Il était fondé sur une probabilité conditionnelle de 1 %, à condition que les poids de la paire soient inférieurs au seuil, sous réserve que celle-ci soit appariée. Comme auparavant, le seuil a été calculé en se fondant sur le résultat de l'algorithme EM. Un seuil inférieur unique a été calculé pour toutes les paires dans le cas des enregistrements du recensement provenant de provinces ou de territoires différents. Les paires comportant un poids de couplage supérieur à leur seuil provincial/territorial inférieur ont été couplées.

3.2.3 Groupes d'enregistrements

Des groupes mutuellement exclusifs d'enregistrements reliés ont été constitués à partir de paires dont le poids était supérieur à leur seuil de pondération inférieur. Ces groupes comprenaient une grande majorité de paires. Tous les groupes représentaient des cas de surdénombrement potentiel de la base de l'étape 2.

3.3 Base de l'extension

Par le passé, le dénombrement en double de ménages entiers a été à l'origine d'une partie importante du surdénombrement. Pour déterminer une part plus grande de ce surdénombrement, des liens supplémentaires ont été créés entre les enregistrements du recensement sur la base d'identificateurs des ménages présents dans la BDRR. Ces liens ont été créés en deux étapes. La première étape a permis la création d'un lien entre deux ménages, s'ils étaient associés à deux enregistrements du recensement couplés à l'étape 1 ou à l'étape 2. La deuxième étape a permis la création de nouveaux liens entre les autres enregistrements associés aux ménages couplés, après les avoir comparés sur la base du sexe et de la date de naissance. Ces nouveaux liens ont servi à créer des groupes reliés d'enregistrements du recensement. Ces enregistrements représentaient des cas de surdénombrement potentiel de la base de l'extension.

4. Plan d'échantillonnage

Trois échantillons stratifiés indépendants ont été tirés des trois bases. Ces échantillons ont fait l'objet d'un examen manuel, afin de vérifier l'occurrence du surdénombrement. L'examen manuel d'un cas possible a pris la forme d'un examen des liens le constituant. Pour chacun de ces liens, l'examen comprenait la comparaison des enregistrements sélectionnés et de ceux des ménages correspondants, en vue de prendre des décisions manuelles plus appropriées.

La construction séquentielle des différentes bases et les échantillons indépendants ont donné lieu à un certain chevauchement, car certains enregistrements du recensement étaient inclus dans plusieurs cas potentiels provenant de différentes bases. Ce problème a été résolu à l'étape de l'estimation de la façon suivante. Tout d'abord, chaque cas possible a été inclus dans un groupe d'enregistrements plus important, appelé *groupe de chevauchement*. À l'intérieur du même groupe de chevauchement, les enregistrements peuvent être reliés au moyen des liens de l'étape 1, de l'étape 2 ou de l'extension. En deuxième lieu, les poids d'échantillonnage ont été corrigés à partir du partage des poids, en considérant l'union des trois échantillons comme un échantillon indirect des groupes de chevauchement. Lavallée (2002) décrit la méthode généralisée de partage des poids pour l'échantillonnage indirect.

4.1 Strates

Dans chaque base de cas de surdénombrement potentiel, les groupes d'enregistrements ont été stratifiés selon leur nombre d'enregistrements et la province/le territoire de chaque enregistrement. La mesure de taille $\hat{P}(M|\gamma)\left[1-\hat{P}(M|\gamma)\right]$ a servi à stratifier à nouveau les paires de l'étape 2, où $\hat{P}(M|\gamma)$ représente la probabilité d'appariement conditionnel estimée $\hat{P}(M|\gamma)$ et γ, le vecteur des résultats observés dans la paire. La probabilité d'appariement conditionnel a été calculée au moyen de l'algorithme EM.

4.2 Sélection de l'échantillon

À l'intérieur des strates de chaque base, les groupes d'enregistrements ont d'abord été triés selon le sexe et la date de naissance. Puis, on a tiré un échantillon systématique.

4.3 Répartition de l'échantillon

Aux étapes 1 et 2, la répartition a été optimisée sous réserve d'un coefficient de variation maximum pour l'estimation provinciale/territoriale du surdénombrement. Les strates comportant des groupes de grande taille ont été désignées comme des strates à tirage complet. À l'étape 2, la répartition a utilisé la mesure de taille $\hat{P}(M|\gamma)\left[1-\hat{P}(M|\gamma)\right]$ pour les strates comprenant des paires.

Dans la base de l'extension, une taille d'échantillon égale a été attribuée à chaque strate.

5. Traitement et estimation

Deux ensembles d'estimations ont été produits, y compris des estimations précédant le rajustement et des estimations finales. Les estimations finales incluaient le surdénombrement laissé de côté par l'ESR, mais déterminé au moyen de l'Étude par appariement automatisé (EAA).

5.1 Traitement

Les résultats de l'examen manuel ont été traités en cas vérifiés de surdénombrement, c'est-à-dire des groupes d'enregistrements du recensement qui ont été reliés grâce à des liens comportant un surdénombrement vérifié. Dans un cas vérifié de surdénombrement, le surdénombrement a été calculé comme correspondant au nombre d'enregistrements moins un. Le surdénombrement d'un cas possible échantillonné a été établi comme le surdénombrement total pour l'ensemble des cas vérifiés inclus.

5.2 Estimations précédant le rajustement

Les estimations du surdénombrement ont été fondées sur l'estimateur de Horwitz-Thompson, avec une certaine repondération pour tenir compte de tout chevauchement.

5.3 Rajustement

L'EAA est une étude d'évaluation qui est fondée sur les ménages. Une petite proportion du surdénombrement peut être déterminée par l'EAA, mais oubliée par l'ESR. En 2011, pour chaque province et territoire, un rajustement distinct a permis d'ajouter le surdénombrement oublié au moment de l'estimation précédant le rajustement pour produire l'estimation finale de l'ESR.

6. Résultats

En 2011, dans le cadre de l'ESR, on a estimé le surdénombrement à 632 846 à l'échelle du Canada, avec une erreur type de 6 675 (Dasylva, 2013). Le taux de surdénombrement a été estimé à 1,85 %. Il s'agit d'une augmentation par rapport à l'estimation de 1,59 % observée en 2006 (Statistique Canada, 2006).

Le tableau 6-1 montre le surdénombrement selon l'étape (Dasylva, 2013). La majorité du surdénombrement a été détectée à l'étape 1 et à l'étape 2. L'extension a permis de détecter 5,05 % du surdénombrement. Quant au rajustement, il représentait seulement 1,90 % de l'estimation finale du surdénombrement.

Le tableau 6-2 montre la répartition du surdénombrement selon le type, ainsi que selon le scénario, pour les cas faisant intervenir des ménages différents (Dasylva, 2013). Le surdénombrement entre des ménages identiques a représenté la majorité du surdénombrement, soit 51 %. Le surdénombrement restant (48 %) faisait intervenir des ménages différents, 29 % de ce dénombrement représentant des enfants en garde partagée.

Tableau 6-1
Surdénombrement selon l'étape

Étape	*Total*	*%*
Étapes 1 et 2	588 856	93,05
Extension	31 939	5,05
Rajustement	12 051	1,90
Total	632 846	100

Tableau 6-2
Surdénombrement selon le type et le scénario

Type	*%*
Identique, une personne, loin	1
Identique, une personne, près	4
Identique, plusieurs personnes, loin	7
Identique, plusieurs personnes, près	39
Non identique, plusieurs personnes, une en commun	28
Non-identique, plusieurs personnes, ≥ 2 en commun	20
Valeur manquante	<1

Scénario du surdénombrement entre des ménages différents	*%*
Enfant(s) de parents dans des ménages séparés	29
Adulte avec d'autres parents	17
Étudiant/jeune adulte récemment sorti du domicile familial	15
Enfant(s) vivant avec deux parents/adultes	6
Adulte récemment entré dans une relation maritale ou une union civile ou récemment sorti d'une telle relation ou union	5
Adulte avec des adultes sans lien de parenté	5
Jeune adulte récemment sorti du domicile familial pour une relation maritale ou une union civile	4
Un ménage collectif	3
Autre	16
Valeur manquante	1

Bibliographie

DASYLVA, Abel, et Robert-Charles TITUS. 2013. « 2011 Census Overcoverage Survey (COS) Methodology Report », rapport interne, Ottawa, Statistique Canada.

FELLEGI, Ivan P., et Alan B. SUNTER. 1969. « A Theory of Record Linkage », *Journal of the American Statistical Association*, vol. 64, nº 328, p. 1183 à 1210.

JARO, Matthew A. 1989. « Advances in record linkage methodology to matching the 1985 census of Tampa, Florida », *Journal of the American Statistical Association*, vol. 84, nº 406, p. 414 à 420.

LAVALLÉE, Pierre. 2002. *Le sondage indirect ou la méthode du partage des poids*, Bruxelles: Éditions de l'Université de Bruxelles.

SAMUELS, C. 2011. « Using the EM algorithm to estimate the parameters of the Fellegi-Sunter model for data linking », numéro de rapport 1352.0.55.120, Australian Bureau of Statistics.

STATISTIQUE CANADA. 2010. *Rapport technique du Recensement de 2006 : Couverture*, numéro 92-567-X au catalogue de Statistique Canada.

WINKLER, William E. 1988. « Using the EM algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage », *Proceedings of the Section on Survey Research Methods*, p. 667 à 671, American Statistical Association.