Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées

par James O. Chipperfield, Glenys R. Bishop
et Paul Campbell

Juin 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2011
Vol. 37, N° 1, pp. 17-30
Statistique Canada, N° 12-001-X au catalogue

Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées

James O. Chipperfield, Glenys R. Bishop et Paul Campbell [1]

Résumé

L'appariement des données consiste à jumeler des enregistrements issus de deux fichiers ou plus que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise). Il s'agit d'un moyen très courant de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des détails. Souvent, le processus d'appariement des données n'est pas exempt d'erreur et peut aboutir à la formation d'une paire d'enregistrements qui n'appartiennent pas à la même unité. Alors que le nombre d'applications d'appariement d'enregistrements croît exponentiellement, peu de travaux ont porté sur la qualité des analyses effectuées en se servant des fichiers de données ainsi appariées. Traiter naïvement ces fichiers comme s'ils ne contenaient pas d'erreurs mène, en général, à des estimations biaisées. Le présent article décrit l'élaboration d'un estimateur du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées. Simple, cette méthode d'estimation est appliquée en utilisant l'algorithme EM bien connu. Dans le contexte qui nous occupe, l'appariement probabiliste des données est une méthode reconnue. Le présent article démontre l'efficacité des estimateurs proposés au moyen d'une étude empirique s'appuyant sur cet appariement probabiliste.

Mots clés : Appariement des données , appariement probabiliste ; maximum de vraisemblance ; tableaux de contingence ; régression logistique.

1. Introduction

L'appariement des données, également appelé appariement ou couplage d'enregistrements, est la tâche consistant à jumeler des enregistrements que l'on pense appartenir à une même unité (par exemple une personne ou une entreprise) et qui sont tirés de deux fichiers ou plus. L'appariement des données est une technique indiquée pour jumeler des ensembles de données en vue de renforcer la dimension temporelle, ou des aspects tels que la portée ou la profondeur des détails. Dans des conditions idéales, l'appariement serait parfait, autrement dit seuls les enregistrements appartenant à la même unité seraient appariés et tous les appariements possibles seraient faits. Malheureusement, très souvent, il n'en est pas ainsi, surtout si l'on se sert pour apparier les enregistrements de champs pouvant contenir des valeurs incorrectes, des valeurs manquantes ou des valeurs légitimement différentes pour une unité particulière.

On recourt souvent à l'appariement probabiliste quand les fichiers contiennent un ensemble commun de variables ou de champs qui fournissent des renseignements d'identification partiels, mais ne constituent pas un identificateur d'unité unique. Dans l'appariement probabiliste (Fellegi et Sunter 1969), un score est attribué à chacun des appariements possibles en se basant sur la probabilité que les enregistrements appartiennent à la même unité. Ce score est calculé en comparant les valeurs des variables d'appariement qui sont communes aux deux fichiers. Un appariement est alors déclaré si le score d'appariement est supérieur à un seuil donné. Un algorithme d'optimisation peut être utilisé pour s'assurer que chaque enregistrement d'un fichier ne soit pas apparié à plus d'un enregistrement d'un autre fichier. Les méthodes probabilistes d'appariement de fichiers sont bien établies aujourd'hui (voir Herzog, Scheuren et Winkler 2007, Winkler 2001 et Winkler 2005), et il existe toute une gamme de progiciels pour les mettre en œuvre.

Cette situation découle de l'importance continue de l'appariement dans divers domaines, particulièrement ceux touchant aux politiques en matière de santé et aux politiques sociales. Les exemples récents d'appariements probabilistes de données effectués par l'Australian Bureau of Statistics (ABS) comprennent l'appariement d'enregistrements provenant du Recensement de la population et du logement de l'Australie de 2006 avec ceux d'un certain nombre d'ensembles de données, dont les enregistrements des décès survenus en Australie (Australian Bureau of Statistics 2008), le Census Dress Rehearsal (Répétition générale du recensement) de 2006 (Solon et Bishop 2009) et l'Australian Migrants Settlements Database (Wright, Bishop et Ayre 2009). Dans le secteur australien de la santé, les méthodes d'appariement probabiliste sont employées par la Western Australian Data Linkage Unit (Holman, Bass, Rouse et Hobbs 1999) et par le New South Wales Centre for Heath Record Linkage. Sur la scène internationale, les méthodes probabilistes sont utilisées par Statistique Canada (Fair 2004), le Census Bureau des États-Unis (voir Winkler

1. James O. Chipperfield, Australian Bureau of Statistics. Courriel : james.chipperfield@abs.gov.au ; Glenys R. Bishop, The Australian National University ; Paul Campbell, Australian Bureau of Statistics.

2001), le National Center for Health Statistics des États-Unis (National Center for Health Statistics 2009) et l'Office fédéral de la statistique de la Suisse dans le cadre de son étude longitudinale des personnes vivant en Suisse.

L'appariement des données rend possibles de nouveaux produits et analyses statistiques. En général, traiter naïvement le fichier de données appariées de manière probabiliste comme s'il s'agissait d'un appariement parfait produit des estimations biaisées. Lahiri et Larsen (2005), ainsi que Scheuren et Winkler (1993) ont proposé des méthodes permettant d'estimer sans biais les coefficients d'un modèle de régression linéaire sous appariement probabiliste d'enregistrements. Plus récemment, Chambers et coll. (2009) et Chambers (2008) ont élargi la portée de ces travaux à un vaste ensemble de modèles en utilisant des équations d'estimation généralisées et, dans le cas de l'appariement de deux fichiers, en permettant que l'un des fichiers soit un sous-ensemble de l'autre.

Le présent article décrit l'élaboration d'une approche d'estimation du maximum de vraisemblance (MV) pour analyser les données appariées de manière probabiliste. La technique d'estimation est simple et mise en œuvre en se servant de l'algorithme EM bien connu. L'approche consiste à remplacer les statistiques qui seraient observées dans le cas de données parfaitement appariées par leur espérance conditionnellement aux données appariées. En supposant que l'espérance est spécifiée correctement, cette approche permet d'éviter les deux limites des travaux antérieurs décrites ci-après.

Premièrement, alors que les méthodes antérieures reposaient sur un appariement exécuté en un seul passage, l'appariement probabiliste comporte habituellement des passages multiples. Dans ce cas, seuls les enregistrements non appariés au moment du premier passage peuvent faire l'objet d'un appariement durant le deuxième passage, puis seuls les enregistrements non appariés au cours des deux premiers passages peuvent être appariés lors du troisième, et ainsi de suite. Chaque passage est conçu en vue d'apparier les enregistrements présentant un ensemble commun particulier de caractéristiques. Par exemple, le premier passage peut être conçu de manière à apparier les enregistrements appartenant aux personnes qui n'ont pas changé d'adresse entre les dates de référence des deux fichiers. Le deuxième passage peut être conçu pour tenir compte des changements d'adresse. Un exemple de ce genre d'approche est donné au tableau 1 à la section 5.

Deuxièmement, les méthodes antérieures reposaient sur l'hypothèse que les deux fichiers contenaient des enregistrements appartenant exactement aux mêmes unités ou que l'ensemble d'unités figurant dans un fichier était un sous-ensemble de celles présentes dans l'autre fichier. L'approche que nous proposons ne requiert pas que l'un des fichiers à apparier soit un sous-ensemble de l'autre fichier. Fréquente en pratique, cette situation s'est présentée dans tous les exemples susmentionnés d'appariements à l'ABS. Il convient aussi de mentionner que les fichiers à apparier ne doivent pas nécessairement être reliés par un mécanisme d'échantillonnage, comme dans le cas où le fichier le plus petit est un sous-échantillon aléatoire des unités comprises dans le plus grand fichier. L'élimination de cette restriction signifie que les deux fichiers peuvent être des ensembles de données administratives.

Considérons l'appariement de deux fichiers désignés par X et Y. Le fichier Y contient la variable y de la population d'individus U_y comprenant n_y enregistrements. Le fichier X contient un vecteur, $\mathbf{x}$, de variables sur la population d'individus U_x comprenant n_x enregistrements. L'inférence a pour cible la population de n_{xy} individus, désignée par $U_{xy} = U_x \cap U_y$, qui sont communs au fichier X et au fichier Y. Les fichiers X et Y contiennent également un vecteur de champs, désigné par $\mathbf{z}$, qui sont utilisés pour apparier les fichiers en utilisant un algorithme d'appariement probabiliste. Naturellement, puisque nous considérons un appariement probabiliste ici, la variable $\mathbf{z}$ ne constitue pas un identificateur d'unité unique.

L'appariement des fichiers X et Y permet d'analyser la distribution conjointe de $\mathbf{x}$ et de y. Deux sources d'erreur peuvent avoir une incidence sur l'analyse de la distribution conjointe en se servant du fichier de données appariées. Ces erreurs correspondent aux *appariements incorrects* et aux *enregistrements non appariés*.

Un appariement est correct si les deux enregistrements appariés appartiennent à une même personne. Un appariement est incorrect si les deux enregistrements appariés n'appartiennent pas à la même personne. Les appariements incorrects peuvent accroître ou réduire artificiellement la corrélation entre $\mathbf{x}$ et y. Un exemple du second cas est l'appariement aléatoire, dans lequel les enregistrements du fichier X sont appariés aléatoirement aux enregistrements du fichier Y.

Le i^{e} enregistrement du fichier X est défini comme un *enregistrement non apparié* si $i \in U_{xy}$ *et* que l'enregistrement i n'a pas été apparié à un enregistrement du fichier Y. Autrement dit, un enregistrement non apparié est un enregistrement du fichier X qui pourrait être apparié correctement, mais n'a pas été apparié du tout (tout au long de l'exposé, nous adoptons la convention de définir les enregistrements non appariés en fonction du fichier X, mais la définition pourrait également être formulée en fonction des enregistrements du fichier Y). Il se pourrait que l'on ne puisse pas toujours apparier un enregistrement particulier du fichier X en ayant la certitude que l'appariement est correct. Cette situation peut se présenter s'il manque dans un enregistrement des champs qui sont utiles pour établir

l'appariement correct. De manière plus générale, des cas d'enregistrements non appariés peuvent avoir lieu quand certaines sous-populations sont relativement difficiles à apparier. Par exemple, des champs tels que l'état matrimonial, la qualification, le domaine d'études et le plus haut niveau de scolarité ne seraient en général pas aussi puissants lorsque l'appariement concerne des enfants que lorsqu'il s'agit d'adultes ayant atteint la maturité. Dans cette situation, l'apparieur des données doit décider s'il doit ou non apparier ce genre d'enregistrements. Nous définissons l'ensemble d'enregistrements appariés par U_I de taille n^* de sorte que $n^* \leq n_x$ et $n^* \leq n_y$.

Le problème que pose l'analyse quand des enregistrements sont non appariés présente des points communs manifestes avec le problème de la non-réponse totale d'une unité. Dans les deux cas, un seul sous-ensemble d'enregistrements légitimes est disponible pour l'analyse. Le mécanisme de non-réponse dans les enquêtes par sondage dépend, en réalité, d'un ensemble inconnu de variables. En revanche, ici, nous avons le léger avantage de savoir que la probabilité qu'un enregistrement demeure non apparié ne peut être qu'une fonction de **z**. Le problème de la non-réponse est souvent résolu par pondération ou au moyen d'un certain argument de conditionnement. Dans le présent article, nous envisageons les deux approches pour résoudre le problème des enregistrements non appariés.

Il existe un compromis naturel entre le nombre d'enregistrements non appariés et le nombre d'appariements incorrects (et par conséquent le biais qu'ils introduisent). Considérons le cas où le fichier X est un sous-échantillon du fichier Y de sorte que $U_{xy} = U_x$. L'appariement de tous les enregistrements du fichier X ne donnera lieu, par définition, à aucun enregistrement non apparié, mais maximisera le nombre d'appariements incorrects. Si nous décidons plutôt de ne former que des appariements dont nous sommes convaincus qu'ils sont corrects, le nombre d'appariements incorrects diminuera, mais le nombre d'enregistrements non appariés augmentera. En pratique, trouver l'équilibre optimal entre les biais dus aux enregistrements non appariés et aux appariements incorrects dépend de l'analyse qui doit être effectuée, de la méthode d'appariement et de leur interaction. Pour une discussion pratique approfondie de cette question, voir Bishop (2009).

Il convient de mentionner que le problème de l'inférence en présence d'appariements incorrects d'enregistrements est semblable au problème de l'inférence en présence de classifications incorrectes de la variable dépendante, qui est une forme d'erreur de mesure (voir Fuller 1987). Dans ce dernier cas, des hypothèses d'identification font la distinction entre le mécanisme d'erreur de classification et les mécanismes du modèle, et sont nécessaires puisqu'habituellement, on ne dispose d'aucune mesure exempte d'erreur. Par exemple, Hausman et coll. (1998) considèrent le cas d'une erreur de classification dans la variable dépendante d'un modèle de régression logistique. Leur hypothèse d'identification est que la valeur de la variable dépendante éventuellement classée incorrectement est une fonction particulière des variables explicatives du modèle. La méthode que nous proposons ne nécessite pas les hypothèses d'identification fortes que requièrent les problèmes d'erreur de mesure, essentiellement parce qu'une mesure sans erreur peut être obtenue au moyen d'un échantillon pour examen manuel qui permet de déterminer les appariements corrects. Les hypothèses que nous faisons dans le présent article sont énoncées à la section 3.

À la section 2, nous résumons l'approche du MV pour les tableaux de contingence et l'analyse de régression sous appariement parfait. À la section 3, nous considérons l'approche du MV en présence d'appariements incorrects. À la section 4, nous examinons l'approche du MV en présence d'appariements incorrects et d'enregistrements non appariés. À la section 5, au moyen d'une étude empirique, nous démontrons l'efficacité de bon nombre des estimateurs proposés. Enfin à la section 6, nous résumons les résultats.

2. Appariement parfait

En vue de présenter la notation, nous discutons dans cette section du cas où l'appariement est parfait. Par conséquent, l'approche d'estimations est classique, puisque, de toute évidence, aucun ajustement spécial n'est nécessaire pour tenir compte des appariements incorrects. À la section 2.1, nous discutons de l'estimation des probabilités dans les cellules d'un tableau de contingence et à la section 2.2, de l'estimation des coefficients de régression d'une régression logistique.

2.1 Tableaux de contingence

En ce qui concerne la notation, il est commode, lorsque l'on envisage l'analyse d'un tableau de contingence, de transformer $\mathbf{x}_i$ en une variable catégorique unique x de sorte que $x = 1, 2, ..., g, ..., G$. Soit y une variable catégorique du fichier Y, où $y = 1, ..., c, ..., C$.

Considérons la factorisation qui suit de la distribution de x et de y

$$p(y,\ x) = p_1(y \mid x\,;\ \mathbf{\Pi})\, p_2(x),$$

où $\mathbf{\Pi} = (\boldsymbol{\pi}_1', ..., \boldsymbol{\pi}_g', ..., \boldsymbol{\pi}_G')'$, $\boldsymbol{\pi}_g = (\pi_{1|g}, ..., \pi_{c|g}, ..., \pi_{C|g})'$, $\pi_{c|g}$ est la probabilité que $y = c$ sachant $x = g$. Nous supposons que, pour chaque valeur de x, il existe C valeurs possibles de y, ce qui implique que la dimension de $\mathbf{\Pi}$ est CG.

Considérons maintenant l'estimation du maximum de vraisemblance du paramètre $\mathbf{\Pi}$, caractérisant p_1, sous

appariement parfait. Un appariement parfait signifie que chacun des enregistrements du fichier X est correctement apparié à l'enregistrement correspondant du fichier Y (c'est-à-dire qu'il n'y a pas d'appariement incorrect ni d'enregistrement non apparié). Sous appariement parfait, $n_{xy} = n_x$ et l'ensemble d'enregistrements appariés est désigné par $\mathbf{d} = \{(y_i,\ x_i) : i = 1, ..., n_{xy}\}$. Sous appariement parfait, la fonction de score pour $\boldsymbol{\pi}_x = (\pi_{1|x}, ..., \pi_{c|x}, ..., \pi_{C|x})'$ caractérisée par la distribution multinomiale est

$$\text{Score}(\boldsymbol{\pi}_x\,;\,\mathbf{d}) =$$

$$(\text{Score}(\pi_{1|x}\,;\,\mathbf{d}),..., \text{Score}(\pi_{c|x}\,;\,\mathbf{d}),..., \text{Score}(\pi_{C-1|x}\,;\,\mathbf{d}))' \quad (1)$$

où

$$\text{Score}(\pi_{c|x}\,;\,d) = \Sigma_i(w_{ic|x}\pi_{ic|x}^{-1} - w_{iC|x}\pi_{iC|x}^{-1})$$

$$= n_{c|x}\pi_{c|x}^{-1} - n_{C|x}\pi_{C|x}^{-1},$$

pour $c = 1, ..., C-1$, où $n_{c|x} = \Sigma_i w_{ic|x}$, $w_{ic|x} = 1$ si $y_i = c$ et $x_i = x$ et $w_{ic|x} = 0$ autrement, et la catégorie correspondant à $y = C$ est choisie arbitrairement comme catégorie de référence. La résolution de $\text{Score}(\boldsymbol{\pi}_x\,;\,\mathbf{d}) = \mathbf{0}_{C-1}$ pour trouver $\boldsymbol{\pi}_x$, où $\mathbf{0}_{C-1}$ est un vecteur colonne de zéros de dimension $C-1$, donne l'estimateur du maximum de vraisemblance (MV)

$$\hat{\pi}_{c|x} = n_{c|x}/n_x, \quad (2)$$

où

$$n_x = \Sigma_c\Sigma_i w_{ic|x}$$

et

$$\hat{\pi}_{C|x} = 1 - \sum_{c=1}^{C-1}\hat{\pi}_{c|x}.$$

2.2 Régression logistique

Considérons le modèle de régression logistique

$$E(y_i) = \upsilon_i \quad (3)$$

$$\upsilon_i = 1/[1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)]. \quad (4)$$

Pour (4), les K éléments de $\mathbf{x}_i$ sont des variables dichotomiques et y_i est maintenant une variable dichotomique disponible dans le fichier Y. Si nous définissons $\mathbf{x} = (\mathbf{x}_1,...,\mathbf{x}_i,...,\mathbf{x}_{n_{xy}})'$, $\mathbf{y} = (y_1,...y_i,...,y_{n_{xy}})'$ et $\boldsymbol{\upsilon} = (\upsilon_1,...\upsilon_i,..., \upsilon_{n_{xy}})'$, la matrice de score pour $\boldsymbol{\beta}$ basée sur des données parfaitement appariées, $\mathbf{d}$, est donnée par

$$\text{Score}(\boldsymbol{\beta}\,;\,\mathbf{d}) = \mathbf{x}'(\mathbf{y} - \boldsymbol{\upsilon}). \quad (5)$$

La résolution de $\text{Score}(\boldsymbol{\beta}\,;\,\mathbf{d}) = \mathbf{0}_K$ pour trouver $\boldsymbol{\beta}$ donne l'estimation du MV, $\hat{\boldsymbol{\beta}}$, que l'on peut obtenir en appliquant la méthode bien connue de Newton-Raphson.

3. Analyse en présence d'appariements incorrects

À la présente section, nous considérons la situation où le fichier d'enregistrements appariés contient des appariements incorrects, mais aucun enregistrement non apparié. Cette situation se produit quand chacun des enregistrements du fichier X est apparié à un enregistrement du fichier Y (d'où $n_x \le n_y$). Définissons le fichier d'enregistrements appariés par $\mathbf{d}^* = \{\mathbf{d}_i^* = (y_i^*, \mathbf{x}_i) : i = 1, ..., n_x\}$, où y_i^* est la valeur de y qui est *appariée* à l'enregistrement i du fichier X. Pour clarifier, y_i est la valeur vraie de y pour l'enregistrement i dans le fichier X, de sorte que $y_i^* = y_i$ si l'enregistrement i est apparié correctement.

L'estimateur donné par (2), associé à l'hypothèse que $y_i^* = y_i$ pour $i = 1, ..., n_x$, est naïf, puisqu'il traite le fichier d'enregistrements appariés de manière probabiliste comme s'ils étaient parfaitement appariés. En général, l'estimateur naïf contient un biais. À la présente section, nous dérivons les estimateurs du MV qui tiennent compte du fait que les données ont été appariées de manière probabiliste ou appariées imparfaitement d'une certaine façon.

Il est courant, en pratique, de tirer du fichier d'enregistrements appariés un sous-échantillon, désigné par s_c, puis de l'examiner manuellement. Durant l'examen manuel, un appariement, $\mathbf{d}_i$, est classé comme étant correct ou incorrect. Soit $\delta_i = 1$ si l'enregistrement i dans le fichier X est correctement apparié et $\delta_i = 0$ autrement.

La conception du sous-échantillon pour examen manuel est un problème important, surtout parce que l'examen manuel est souvent coûteux. Les utilisations possibles d'un échantillon pour examen manuel comprennent l'estimation de la proportion d'enregistrements correctement appariés et d'enregistrements non appariés, pour pouvoir décider quels enregistrements devraient être appariés et lesquels devraient rester non appariés, pour s'assurer que l'inférence en se servant de $\mathbf{d}^*$ est correcte (c'est-à-dire l'objectif du présent article) ou pour déterminer comment pourraient être améliorée la façon dont les enregistrements sont appariés. (Dans les applications de l'ABS susmentionnées, les échantillons pour examen manuel ont été conçus de manière à s'assurer que chaque appariement ait au moins une probabilité spécifiée d'être correct.) Si l'on veut s'assurer que l'inférence en se servant de $\mathbf{d}^*$ sera correcte, la sélection de l'échantillon pour examen manuel par échantillonnage aléatoire simple est une approche raisonnable. Un sous-échantillon pour examen manuel plus efficace pourrait peut-être être conçu, mais il n'existe aucun moyen évident de le faire, parce que les paramètres que nous devons estimer pour appliquer la méthode du MV décrite dans le présent article dépend de l'analyse particulière (par exemple choix de y et $\mathbf{x}$). Concevoir un échantillon pour examen

manuel convenant pour toutes les analyses possibles serait difficile.

Nous factorisons la distribution conjointe $p(y_i, \mathbf{x}_i, \delta_i)$ par

$$p(y_i \mid \mathbf{x}_i\ ; \boldsymbol{\theta})\, p(\mathbf{x}_i)\, p(\delta_i \mid \mathbf{x}_i), \tag{6}$$

où $\boldsymbol{\theta} = \boldsymbol{\beta}$ dans le cas de la régression, et $\boldsymbol{\theta} = \boldsymbol{\Pi}$ dans le cas du tableau de contingence. La factorisation (6) signifie que les appariements sont incorrects au hasard (IAH), autrement dit, que les distributions $y_i \mid \mathbf{x}_i$ et $\delta_i \mid \mathbf{x}_i$ sont indépendantes. Sous cette hypothèse, il suffit de maximiser la vraisemblance associée au facteur $p(y_i \mid \mathbf{x}_i\ ; \boldsymbol{\theta})$. Tout au long de la présente section, nous faisons l'hypothèse (6). Il importe de souligner que (6) et le développement qui suit ne s'appuient sur aucune hypothèse nécessitant que le fichier X soit un sous-ensemble du fichier Y (par exemple quand les unités du fichier X sont un sous-échantillon des unités du fichier Y) ou que le processus d'appariement ne comporte qu'un seul passage. Nous supposons également que l'exactitude de l'appariement, δ_i, est indépendante d'un enregistrement à l'autre.

Comme nous l'avons mentionné dans l'introduction, un score est attribué à chaque enregistrement apparié en se basant sur la probabilité que les enregistrements appartiennent à la même unité. Désignons le score par r_i. Un examinateur a suggéré d'utiliser r_i pour paramétriser plus exactement la distribution de δ_i. Techniquement, cette suggestion nécessiterait le remplacement de $p(\delta_i \mid \mathbf{x}_i)$ par $p(\delta_i \mid \mathbf{x}_i, r_i)$ dans (6) et réduirait vraisemblablement la variabilité des estimateurs du MV discutés à la section 3. Il serait intéressant d'explorer cette piste dans le cadre de futurs travaux.

3.1 Tableaux de contingence

Soit $w^*_{ic|x} = 1$ si $y^*_i = c$ et $x_i = x$, et $w^*_{ic|x} = 0$ autrement. L'espérance de $w_{ic|x}$ sachant $\mathbf{d}^*_i$ est

$$\begin{aligned} E_{d|d^*}(w_{ic|x} \mid x_i = x, y^*_i = y^*) = \\ w^*_{ic|x} p_{xy^*} + (1 - p_{xy^*})\pi_{c|x} \quad &\text{si } i \notin s_c \\ = w^*_{ic|x} \quad &\text{si } i \in s_c \text{ et } \delta_i = 1 \\ = \pi_{c|x} \quad &\text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned}$$

et p_{xy^*} est la probabilité que le i^{e} appariement soit correct sachant que $x_i = x$ et $y^*_i = y^*$. L'estimateur du MV de $\pi_{c|x}$ en se servant des données appariées de manière probabiliste, $\mathbf{d}^*_i$, est alors

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x} \left(\sum\nolimits_c \tilde{n}_{c|x} \right)^{-1} \tag{7}$$

où

$$\tilde{n}_{c|x} = \Sigma_i \tilde{w}_{ic|x}, \tag{8}$$

$$\begin{aligned} \tilde{w}_{ic|x} &= w^*_{ic|x} \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*})\tilde{\pi}_{c|x} \quad \text{si } i \notin s_c \\ &= w^*_{ic|x} \quad \text{si } i \in s_c \\ &= \tilde{\pi}_{c|x} \quad \text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned} \tag{9}$$

et

$$\hat{p}_{xy^*} = \left(\sum\nolimits_{i \in s_c} w^*_{ic|x} \delta_i \right) \left(\sum\nolimits_{i \in s_c} w^*_{ic|x} \right)^{-1}. \tag{10}$$

La procédure d'estimation consiste à itérer (7), (8) et (9) jusqu'à la convergence. Plus précisément, l'algorithme est :

1. Calculer $\hat{p}_{xy^*}$ d'après (10).
2. Initialiser $\tilde{\pi}^{(0)}_{c|x}$ puis calculer $\tilde{w}^{(0)}_{c|x}$ d'après (9) et ensuite $\tilde{n}^{(0)}_{c|x}$ d'après (8).
3. Calculer $\tilde{\pi}^{(t)}_{c|x}$ d'après (7) en utilisant $\tilde{n}^{(t-1)}_{c|x}$.
4. Calculer $\tilde{w}^{(t)}_{c|x}$ d'après (9) en utilisant $\tilde{\pi}^{(t)}_{c|x}$ puis calculer $\tilde{n}^{(t)}_{c|x}$ d'après (8) en utilisant $\tilde{w}^{(t)}_{c|x}$.
5. Itérer 3 et 4 jusqu'à la convergence.

La valeur initiale $\tilde{\pi}^{(0)}_{c|x}$ pourrait être choisie comme étant l'estimation naïve de $\pi_{c|x}$, décrite plus haut à la section 3. Cependant, nous avons constaté que le choix de la valeur initiale n'est pas important.

3.2 Régression logistique

Nous décrivons ci-après deux méthodes du MV (méthodes 1 et 2) pour estimer $\boldsymbol{\beta}$ en se servant de données appariées de manière probabiliste, $\mathbf{d}^*$. Les deux méthodes donnent des estimations sans biais sous l'hypothèse que les appariements sont incorrects au hasard. Elles se distinguent par le niveau d'agrégation auquel sont estimées les probabilités qu'un appariement soit correct. La méthode 1 requiert l'obtention de ces probabilités à un niveau plus fin d'agrégation, ce qui pourrait signifier qu'elle produit des estimations plus variables que la méthode 2.

3.2.1 Méthode 1

L'espérance de y conditionnellement aux données appariées est

$$\begin{aligned} E_{\text{d}|\text{d}^*}(y_i \mid \mathbf{x}_i = \mathbf{x}, y^*_i = y^*) = \\ y^*_i p_{\mathbf{xy}^*} + (1 - p_{\mathbf{xy}^*})\upsilon_i \quad &\text{si } i \notin s_c \\ = y^*_i \quad &\text{si } i \in s_c \text{ et } \delta_i = 1 \\ = \upsilon_i \quad &\text{si } i \in s_c \text{ et } \delta_i = 0 \end{aligned}$$

et p_{xy^*} est la probabilité que le i^{e} appariement soit correct sachant que $x = x_i$ et $y^*_i = y^*$.

L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\tilde{\boldsymbol{\beta}}$, pour $\boldsymbol{\beta}$ dans (5) avec y_i remplacé par $\tilde{y}_i$, où

$$\begin{aligned}\tilde{y}_i &= y_i^* \hat{p}_{\mathbf{x}y^*} + (1 - \hat{p}_{\mathbf{x}y^*})\tilde{\upsilon}_i \quad \text{si } i \notin s_c \\ &= y_i^* \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ &= \tilde{\upsilon}_i \quad \text{si } i \in s_c \text{ et } \delta_i = 0,\end{aligned} \tag{11}$$

$\tilde{\upsilon}_i$ a la même forme que υ_i excepté que $\boldsymbol{\beta}$ est remplacé par $\tilde{\boldsymbol{\beta}}$, et $\hat{p}_{\mathbf{x}y^*}$ est la proportion estimée d'appariements corrects dans l'échantillon pour examen manuel pour chaque combinaison de $\mathbf{x}$ et y^*.

3.2.2 Méthode 2

Posons que $\mathbf{x}'\mathbf{y}$ dans (5) possède le k^{e} élément

$$r_k = \mathbf{x}_k'\mathbf{y} = \sum_i^n y_i x_{ik} = \sum_i^n r_{ik},$$

où $r_{ik} = y_i x_{ik}$. L'espérance de r_{ik} conditionnellement à $\mathbf{d}^*$ est

$$\begin{aligned}E_{\mathrm{d}|\mathrm{d}^*}(r_{ik} \mid \mathbf{x}_i = \mathbf{x}, y_i^* = y_i) = \\ [y_i^* p_{ky^*} + (1 - p_{ky^*})\upsilon_i]x_{ik} \quad \text{si } i \notin s_c \\ = y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ = \upsilon_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0\end{aligned} \tag{12}$$

et p_{ky^*} est la probabilité qu'un appariement avec $x_{ik} = 1$ soit correct sachant que $y_i^* = y^*$. L'estimateur du MV est alors obtenu par itération en vue de trouver la solution, désignée par $\tilde{\tilde{\boldsymbol{\beta}}}$, pour $\boldsymbol{\beta}$ dans (5) avec r_{ik} remplacé par $\tilde{r}_{ik}$, où

$$\begin{aligned}\tilde{r}_{ik} &= [y_i^* \hat{p}_{ky^*} + (1 - \hat{p}_{ky^*})\tilde{\tilde{\upsilon}}_i]x_{ik} \quad \text{si } i \notin s_c \\ &= y_i^* x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 1 \\ &= \tilde{\tilde{\upsilon}}_i x_{ik} \quad \text{si } i \in s_c \text{ et } \delta_i = 0,\end{aligned} \tag{13}$$

$\tilde{\tilde{\upsilon}}_i$ a la même forme que υ_i excepté que $\boldsymbol{\beta}$ est remplacé par $\tilde{\tilde{\boldsymbol{\beta}}}$, et $\hat{p}_{ky^*}$ est la proportion estimée d'appariements corrects dans l'échantillon pour examen manuel pour chaque combinaison de $\mathbf{x}$ et y^*. Autrement dit, si $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n y_i^* x_{ik} \delta_i\right)\left(\sum_{i \in s_c}^n y_i^* x_{ik}\right)^{-1}$$

et si $y^* = 0$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik} \delta_i\right)\left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\right)^{-1}.$$

Cette approche ne nécessite le calcul que de $2K$ probabilités d'après l'échantillon pour examen manuel et, en ce sens, pourrait être préférable à l'approche décrite à la section 3.2.1 qui requiert le calcul d'un plus grand nombre de probabilités.

3.3 Estimation de la variance par le bootstrap

À la présente section, nous décrivons comment calculer la variance des estimations du MV de la section 3. Désignons le paramètre d'intérêt par $\boldsymbol{\theta}$, présenté plus haut, et l'estimation de son MV par $\tilde{\boldsymbol{\theta}}$. L'estimation par le bootstrap (Rubin et Little 2003) de la variance de $\tilde{\boldsymbol{\theta}}$, dénotée par $\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}})$, s'obtient comme il suit :

1. Tirer un échantillon répété de taille n_x du ficher de données appariées, $\mathbf{d}^*$, par échantillonnage aléatoire simple avec remise. Désigner le r^{e} échantillon répété par $\mathbf{d}^*(r)$. Le r^{e} échantillon pour examen manuel répété est $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculer $\tilde{\boldsymbol{\theta}}(r)$ qui a la même forme que $\tilde{\boldsymbol{\theta}}$ excepté que $\mathbf{d}^*(r)$ est utilisé au lieu de $\mathbf{d}^*$ et que $s_c(r)$ est utilisé au lieu de s_c.
3. Répéter les étapes 1 et 2 R fois, où R est le nombre de répliques.
4. Calculer

$$\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{R}\sum\nolimits_{b=1}^{R}(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})'.$$

4. Analyse en présence d'appariements incorrects et d'enregistrements non appariés

À la présente section, nous discutons de deux moyens d'analyser les données appariées en présence d'appariements incorrects et d'enregistrements non appariés. Comme nous l'avons mentionné dans l'introduction, le problème de l'analyse en présence d'enregistrements non appariés possède des similitudes évidentes avec le problème de la non-réponse d'une unité, ou non-réponse totale. L'existence d'enregistrements non appariés peut entraîner la sur ou la sous-représentation de certaines caractéristiques dans le fichier de données appariées, ce qui peut biaiser l'analyse. Comme nous l'exposons de manière plus détaillée plus bas, nous tirons parti du fait que le mécanisme donnant lieu aux enregistrements non appariés ne peut dépendre que de $\mathbf{z}$.

Nous considérons ici deux méthodes d'inférence en présence d'appariements incorrects et d'enregistrements non appariés, où les enregistrements appariés sont désignés au moyen de l'indice $i = 1, ..., n^*$ (rappelons que le i^{e} enregistrement du fichier X est un *enregistrement non apparié* si $i \in U_{xy}$ *et* que l'enregistrement i n'a été apparié à aucun enregistrement du fichier Y). La méthode consiste à modéliser de façon indépendante les processus qui déterminent quels enregistrements sont appariés incorrectement et lesquels sont non appariés (voir la section 5 pour un exemple). Ces modèles requièrent qu'un sous-échantillon, désigné par s_{xc}, des enregistrements du fichier X soit soumis à un examen manuel. Les enregistrements compris dans le

sous-échantillon seront soit appariés, soit non appariés à des enregistrements du fichier Y. Les enregistrements appariés du sous-échantillon doivent être catégorisés comme étant correctement ou incorrectement appariés durant le processus d'examen manuel. Un enregistrement du sous-échantillon qui n'est pas apparié doit être classé comme étant *non apparié* ou *autre*. *Non apparié* signifie que l'enregistrement correspondant a été trouvé dans le fichier Y, mais qu'il n'y a pas eu d'appariement, tandis que *autre* indique que l'enregistrement correspondant n'a pas été découvert dans le ficher Y et est par conséquent considéré comme non existant. Cette seconde classification pourrait être beaucoup plus difficile et plus longue que la première, parce qu'elle suppose qu'un autre processus exempt d'erreur existe pour vérifier qu'il est correct que certains appariements n'aient pas été faits. De par leur nature, les enregistrements non appariés contiennent peu d'information permettant de déterminer quel est l'appariement correct, même durant l'examen manuel. Ce genre de processus peut ne pas exister, auquel cas la correction pour tenir compte des enregistrements non appariés semble impossible. Cependant, il pourrait comprendre un examen manuel des noms qui figurent dans les deux fichiers à apparier. Par exemple, la personne chargée de l'examen manuel pourrait se rendre compte que les noms *John O. Smith* et *Joh O. Smith* qui figurent dans deux enregistrements distincts pourraient, en fait, faire référence à la même personne (avec un « n » manquant dans le deuxième cas, peut-être à cause d'erreur de lecture optique), tandis que le processus automatisé d'appariement pourrait traiter les deux noms comme étant entièrement différents. L'examinateur peut alors décider que les deux enregistrements susmentionnés correspondent à la même personne et devraient donc être appariés. Bishop (2009) et Wright (2009) discutent des avantages de l'examen manuel.

La première méthode consiste à conditionner l'analyse sur une variable $\zeta_i = \zeta_i(\mathbf{z}_i)$. La variable ζ est définie de façon qu'en présence d'enregistrements non appariés, l'inférence soit sans biais conditionnellement à ζ. Le terme ζ est introduit parce que, dans de nombreux cas, il serait peu commode ou inutile de conditionner sur toute l'information contenue dans $\mathbf{z}$. Il est possible de donner à ζ_i une valeur non manquante, même quand $\mathbf{z}_i$ contient des valeurs manquantes. La forme exacte de la fonction $\zeta(\mathbf{z})$ devrait être justifiée après l'analyse du sous-échantillon, s_{xc}. Par exemple, si les personnes de moins de 20 ans sont sous-représentées dans le fichier de données appariées, ζ indiquerait si une personne a moins de 20 ans. Un moyen d'approcher l'analyse serait d'inclure ζ comme covariable dans le modèle de régression. La méthode décrite à la section 3 s'appliquerait alors directement. Cependant, les analystes pourraient souhaiter procéder à l'intégration sur ζ afin que cette variable ne figure pas dans le modèle logistique ou dans le tableau de contingence. À la section 4.2, nous expliquons comment le faire pour les tableaux de contingence. À la section 4.3, nous discutons d'une approche fondée sur la pseudo-vraisemblance qui consiste à attribuer aux enregistrements non appariés des pondérations visant à tenir compte de toute sur ou sous-représentation de certaines sous-populations dans les données appariées. De nouveau, le choix des pondérations devrait être justifié après l'analyse du sous-échantillon, s_{xc}, qui indique quels sont les enregistrements non appariés. Cet aspect est examiné plus en détail dans le contexte de l'étude empirique.

4.1 Pouvons-nous ignorer les enregistrements non appariés ?

Définissons la variable $\gamma_i = 1$ si l'enregistrement i du fichier X est non apparié et $\gamma_i = 0$ autrement. Soit aussi ζ_i une variable telle que $\zeta_i = 1, 2, ..., h, ...H$, où H est le nombre de catégories pour ζ. Nous pouvons ignorer le fait que des enregistrements sont non appariés si nous sommes prêts à supposer que, conditionnellement à $\mathbf{x}_i$, les distributions de y_i, γ_i et δ_i sont indépendantes. Techniquement, cette hypothèse mène à la factorisation suivante

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto p(y_i \mid \mathbf{x}_i ; \boldsymbol{\theta})\, p(\delta_i \mid \mathbf{x}_i)\, p(\gamma_i \mid \mathbf{x}_i)\, p(\zeta_i)$$

où, de nouveau, $\boldsymbol{\theta} = \boldsymbol{\beta}$ ou $\boldsymbol{\Pi}$. Il vaut la peine de vérifier si cette hypothèse est valide dans le cas du sous-échantillon pour examen manuel. Si l'hypothèse est raisonnable, il n'est pas nécessaire d'appliquer les méthodes décrites aux sections 4.2 et 4.3, et les méthodes de la section 3 suffisent.

Nous pourrions ne pas vouloir émettre l'hypothèse susmentionnée, mais être prêts à supposer que, conditionnellement à $\mathbf{x}$ et ζ, les distributions de y_i, γ_i et δ_i sont indépendantes. Dans ce cas, nous disons que les enregistrements non appariés sont ignorables. Techniquement, cette hypothèse mène à la factorisation suivante,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto p(y_i \mid \mathbf{x}_i, \zeta_i ; \boldsymbol{\Lambda})\, p(\delta_i \mid \mathbf{x}_i ; \boldsymbol{\tau})\, p(\gamma_i \mid \mathbf{x}_i, \zeta_i)\, p(\zeta_i)$$

où $\boldsymbol{\Lambda}$ est le paramètre pour la distribution de $y_i \mid \mathbf{x}_i, \zeta_i$. Si nous nous intéressons à $p(y_i \mid \mathbf{x}_i ; \boldsymbol{\theta})$, mais non à $p(y_i \mid \mathbf{x}_i, \zeta_i ; \boldsymbol{\Lambda})$, une approche consiste à éliminer ζ_i de cette dernière par intégration (c'est-à-dire en prenant la moyenne sur toutes les valeurs possibles).

4.2 Maximum de vraisemblance conditionnel (MVC) pour les tableaux de contingence

Premièrement, paramétrisons la distribution conjointe de y_i, x_i et ζ_i par la distribution multinomiale de paramètre

$\boldsymbol{\Lambda}$. Définissons $\boldsymbol{\Lambda} = (\boldsymbol{\Pi}'_1,...\boldsymbol{\Pi}'_h,..., \boldsymbol{\Pi}'_H)'$, où $\boldsymbol{\Pi}_h = (\boldsymbol{\pi}'_{1h}, ..., \boldsymbol{\pi}'_{gh},..., \boldsymbol{\pi}'_{Gh})'$, $\pi_{gh} = (\pi_{1|gh}, ..., \pi_{c|gh}, ..., \pi_{C|gh})'$ et $\pi_{c|gh}$ est la probabilité que $y_i = c$, $x_i = g$ et $\zeta_i = h$. L'estimateur du MV de $\boldsymbol{\Pi} = (\pi_{c|x})$ de la section 2.1 quand les erreurs d'appariement ne peuvent pas être ignorées est $\tilde{\tilde{\boldsymbol{\Pi}}} = (\tilde{\tilde{\pi}}_{c|x})$, où

$$\tilde{\tilde{\pi}}_{c|x} = \sum_{h=1}^{H} \tilde{\pi}_{c|xh} \hat{\pi}_{h|x}, \qquad (14)$$

où

$$\tilde{\pi}_{c|xh} = \tilde{\mathrm{n}}_{c|xh} \left(\sum\nolimits_c \tilde{\mathrm{n}}_{c|xh} \right)^{-1}, \qquad (15)$$

$\tilde{\mathrm{n}}_{c|xh} = \Sigma_{i \in U_l} \tilde{w}_{ic|xh}$, $\Sigma_{i \in U_l}$ est la somme sur les n^* enregistrements appariés et $\hat{\pi}_{h|x}$ pour $h = 1, ..., H$ est l'estimation classique de la distribution marginale de $\boldsymbol{\zeta}$ sachant x dans le fichier X. En outre, si $i \notin s_c$,

$$\tilde{w}_{ic|xh} = w^*_{ic|xh} \hat{p}_{xy^*h} + (1 - \hat{p}_{xy^*h}) \tilde{\tilde{\pi}}_{c|xh}, \qquad (16)$$

$\hat{p}_{xy^*h}$ est la probabilité que le i^{e} appariement soit correct sachant que $x_i = x$, $\zeta_{\mathrm{i}} = h$ et $y^*_i = y^*$, $w^*_{ic|xh} = 1$ si $x_i = x$, $\zeta_{\mathrm{i}} = h$ et $y^*_i = y^*$, et $w^*_{ic|xh} = 0$ autrement. Si $i \in s_c$, alors $\tilde{w}_{ic|xh} = w^*_{ic|xh} = w_{ic|xh}$ s'il est déterminé que l'appariement est correct et $\tilde{w}_{ic|xh} = \tilde{\tilde{\pi}}_{c|xh}$ s'il est déterminé qu'il est incorrect.

L'estimateur du MV $\tilde{\tilde{\pi}}_{c|x}$ s'obtient par itération entre (14), (15) et (16) jusqu'à la convergence.

4.3 Pseudo-maximum de vraisemblance (PMV)

À la présente section, nous discutons d'une alternative au MVC décrit à la section 4.2, que nous appelons pseudo-maximum de vraisemblance (voir Chambers et Skinner 2003). Il s'agit essentiellement d'une approche de pondération, qui pourrait être plus facile à mettre en œuvre que le MVC et qui s'appuie sur la factorisation donnée à la section 4.2. Elle comprend la résolution de versions pondérées des fonctions de score, $\mathrm{Score}(\boldsymbol{\pi}_x ; \mathbf{d}) = \mathbf{0}_{C-1}$ et $\mathrm{Score}(\boldsymbol{\beta} ; \mathbf{d}) = \mathbf{0}_{\mathrm{K}}$ pour trouver $\boldsymbol{\pi}_x$ et $\boldsymbol{\beta}$ respectivement, où le poids d'un enregistrement est égal à l'inverse de la probabilité que l'enregistrement demeure non apparié. Nous désignons la probabilité que l'enregistrement i ne restera pas non apparié par $t_i = E(\gamma_i)$, de sorte que les poids unitaires sont donnés par $q_i = t_i^{-1}$, où ici $i = 1, ..., n^*$. Par conséquent, l'estimateur du PMV pour $\pi_{c|x}$ est

$$\tilde{\pi}^{\mathrm{PMV}}_{c|x} = \breve{n}_{c|x} \left(\sum\nolimits_c \breve{n}_{c|x} \right)^{-1}, \qquad (17)$$

où $\breve{n}_{c|xy} = \Sigma_{i \in U_l} q_i \tilde{w}_{ic|x}$. L'estimation de $\tilde{\pi}^{\mathrm{PMV}}_{c|x}$ s'obtient par itération entre la mise à jour de $\tilde{w}_{ic|x}$, donné par (7), et (17) jusqu'à la convergence. L'estimateur du PMV de $\boldsymbol{\beta}$ est le même que l'estimateur du MV, excepté que, maintenant, dans l'équation d'estimation (5) les poids unitaires correspondaient à q_i. Une approche possible pour estimer l'exactitude des estimations du PMV sous appariement parfait consiste à utiliser la méthode du bootstrap décrite plus haut, mais en introduisant maintenant le poids q_i.

Pour illustrer la situation où les enregistrements non appariés ne peuvent pas être ignorés, considérons l'appariement d'une base de données contenant des renseignements personnels sur la situation d'emploi à une autre contenant des renseignements sur le niveau de scolarité. En outre, supposons que l'âge et le sexe, qui sont des variables corrélées à l'emploi et à la scolarité, sont disponibles dans l'une des bases de données. Après avoir effectué un examen manuel, nous pourrions constater que les enregistrements pour les jeunes hommes ont 50 % de chances de plus de rester non appariés que les enregistrements pour les femmes. Cela pourrait tenir au fait que les hommes sont moins susceptibles que les femmes de fournir des renseignements personnels qui sont utiles pour l'appariement. Clairement, dans le fichier de données appariées, il convient de donner aux enregistrements des hommes une pondération double de celle des enregistrements des femmes afin que l'analyse conjointe de la situation d'emploi et du niveau de scolarité soit sans biais.

5. Étude empirique

L'Australian Bureau of Statistics a réalisé une étude de qualité comportant l'appariement des enregistrements du Recensement de la population et du logement de 2006 aux enregistrements de la répétition générale de ce recensement (RGR). Durant la répétition générale du recensement, des renseignements ont été recueillis auprès de 78 349 personnes, un an avant le recensement. Durant le Recensement de 2006, des renseignements ont été recueillis auprès de plus de 19 millions de personnes.

Pendant une brève période, durant laquelle les données du Recensement de 2006 ont été traitées, les noms et adresses étaient disponibles pour le recensement ainsi que pour la répétition générale du recensement. Durant cette période, les deux fichiers d'enregistrements au niveau de la personne ont été appariés en utilisant des normes d'information différentes :

- la *norme d'or* (NO) consistait à utiliser le nom, l'adresse, l'îlot (*mesh block*) et certains éléments de données du recensement. L'îlot est une zone géographique contenant habituellement 50 logements. Tous les noms et adresses ont été détruits à la fin de la période de traitement des données du recensement.

- la *norme de bronze* (NB) consistait à utiliser l'îlot et certains éléments de données du recensement (c'est-à-dire à ne pas se servir du nom et de l'adresse). Il s'agit d'une méthode qu'il est proposé d'utiliser pour les futurs travaux d'appariement de l'ABS.

Une description détaillée de l'étude de la qualité et de la méthode d'appariement figure dans Solon et Bishop (2009). Le rôle de la norme d'or (NO) dans l'étude de la qualité est essentiel. Elle fournit une référence par rapport à laquelle peut être évaluée la fiabilité de la norme de bronze (NB). L'utilité de la NO comme référence est due au fait que le nom et l'adresse sont des variables puissantes pour repérer les personnes figurant dans le recensement et dans la répétition générale du recensement et qu'elles ont été soumises à un examen manuel approfondi. Nous supposons donc que la NO correspond à l'appariement parfait. Par conséquent, les différences entre les estimations fondées sur la NO et sur la NB sont interprétées comme étant des erreurs. Autrement dit, nous nous intéressons à la fiabilité de la NB *relativement* à la NO.

5.1 Méthodologie d'appariement

5.1.1 Variables de groupage et d'appariement et algorithme d'affectation 1 – 1

À la présente sous-section, nous donnons un aperçu de l'appariement des enregistrements de la RGR à ceux du recensement en appliquant la norme de bronze (NB). La méthode d'appariement comprend une série de passages, où chacun est défini par un ensemble de variables de groupage et d'appariement, ainsi qu'un algorithme d'affectation 1-1. Dans le cas de passages multiples, seuls les enregistrements non appariés au premier passage peuvent être appariés au deuxième, seuls les enregistrements non appariés au deuxième passage peuvent être appariés au troisième, et ainsi de suite.

Le tableau 1 donne les variables de groupage, désignées par « G » pour la NB. Par exemple, durant le passage 1, un enregistrement du recensement et un enregistrement de la RGR ne sont considérés comme un appariement possible que s'ils contiennent la même valeur pour l'îlot.

Les variables d'appariement sont utilisées pour mesurer le degré de concordance entre une paire d'enregistrements. Un haut niveau de concordance donne à penser que la probabilité que la paire d'enregistrements constitue un appariement correct est élevée. Le tableau 1 donne les variables d'appariement, désignées par « A », pour la NB. Par exemple, durant le passage 1 pour la NB, une gamme de variables, dont le jour, le mois et l'année de la naissance, le pays de naissance et le plus haut niveau de qualification sont utilisés comme variables de couplage.

Tableau 1
Exemple de variables de groupage (G) et d'appariement (A) utilisées pour apparier les données du Recensement de 2006 avec celles de la répétition générale du recensement. Différentes variables de groupage ont été utilisées lors de chacun des deux passages

Variable	Passage 1	Passage 2
Jour de la naissance	A	G
Mois de la naissance	A	G
Année de la naissance	A	G
Sexe	A	G
Statut d'Autochtone	A	A
Pays de naissance	A	A
Langue parlée	A	A
Année de l'arrivée	A	A
État matrimonial	A	A
Religion	A	A
Domaine d'études pour la qualification la plus élevée	A	A
Niveau de la qualification la plus élevée	A	A
Plus haut niveau de scolarité	A	A
Îlot (*Mesh block*)	G	A

À chaque passage, la sortie comprend un score pour chaque paire d'enregistrements. Le score est une mesure du degré de concordance entre les enregistrements formant la paire. Nous remettons à plus tard la définition formelle du score (pour des détails, voir (3.6), Conn et Bishop 2006), mais nous illustrons plus bas comment il peut être interprété. Considérons la NB au passage 2, pour lequel les enregistrements d'une paire contiennent la même date de naissance complète et le même sexe ; une paire d'enregistrements recevra un score de 23,5 s'il y a concordance pour l'îlot (+17) et l'année d'arrivée (+8) et désaccord pour la religion (-1,5) (dans cet exemple, la situation de concordance pour les autres variables d'appariement contribuerait également au score, mais nous les ignorons pour simplifier l'illustration). La contribution de la concordance pour l'îlot (+17) est plus grande celle de la concordance concernant l'année d'arrivée (+8) parce que la première est moins susceptible que la seconde d'avoir lieu par chance uniquement.

Afin de formaliser la cible de l'algorithme d'appariement, désignons le score pour l'enregistrement *i* de la RGR et l'enregistrement *j* du recensement durant le passage *p* en appliquant la NB par r_{pij}. L'ensemble des scores des paires d'enregistrements r_{pij} et le seuil d'exclusion f_p sont utilisés par le progiciel d'appariement *Febrl* (voir Christen et Churches 2005) pour déterminer l'ensemble optimal d'appariements pour le passage *p*. Le terme f_p est la valeur minimale du score pour qu'une paire d'enregistrements soit considérée comme un appariement durant le passage *p*. L'algorithme *Febrl* cherche à maximiser $\sum_i r_{pij}$, sous la contrainte $r_{pij} > f_p$. De toute évidence, le nombre d'appariements dépend de f_p.

Dans la suite, nous évaluons la NB au moyen de deux ensembles différents de seuils, où un ensemble de seuils est défini par les seuils des passages 1 et 2. Le premier seuil,

que nous disons très faible (TF), est considéré comme un seuil optimal, puisque, pour une gamme de seuils, les estimations naïves obtenues étaient les estimations « les plus proches » des estimations correspondantes produites en appliquant la NO (voir Bishop 2009). Le deuxième seuil, que nous disons extrêmement faible (EF), a effectivement pour objectif de maximiser le nombre d'enregistrements de la RGR qui sont appariés. Ci-après, nous désignons les deux fichiers d'enregistrements appariés sous la NB par les noms de leur seuil, TF et EF.

5.1.2 Résultats d'appariement

Sous la norme d'or (NO), 70 274 des 78 349 enregistrements de la RGR ont été appariés. Sous l'hypothèse que la NO correspond à l'appariement parfait, 8 075 personnes possédaient un enregistrement dans la RGR, mais non dans le recensement. En réalité, la NO n'est pas parfaite. Pour une discussion à ce sujet, voir Bishop, 2009.

Sous la norme de bronze avec seuil très faible (TF), 57 790 enregistrements de la RGR ont été appariés. Des 70 274 enregistrements de la RGR qui ont été appariés en appliquant la NO, 13 784 sont demeurés non appariés, 700 ont été appariés incorrectement et 55 790 ont été appariés correctement an appliquant le seuil TF. En outre, 1 300 enregistrements ont été appariés en appliquant le seuil TF, mais non en appliquant la NO et sont également des appariements incorrects. Donc, il y a eu, en tout, 2 000 (= 700 + 1 300) appariements incorrects.

Sous la norme de bronze avec seuil extrêmement faible (EF), 74 350 enregistrements de la RGR ont été appariés. Des 70 274 enregistrements de la RGR qui ont été appariés en appliquant la NO, 2 811 sont demeurés non appariés, 9 793 ont été appariés incorrectement et 57 670 ont été appariés correctement en appliquant la norme de bronze avec le seuil EF. En outre, 6 887 enregistrement de la RGR ont été appariés sous la norme de bronze EF, mais non sous la NO.

En résumé, 97 % des appariements du fichier TF sont corrects et 20 % (=13 784/70 274) des enregistrements de la RGR appariés en appliquant la NO demeurent non appariés. Les chiffres correspondants pour le fichier EF sont de 78 % et 4 % (=2 811/70 274).

5.1.3 Modélisation de la probabilité qu'un appariement soit correct

Pour *tous* les appariements dans les conditions EF et TF, nous savions s'il s'agissait d'un appariement correct ou incorrect (par exemple, si un appariement EF est également produit par la NO, cet appariement est correct. Sinon, l'appariement EF est incorrect). Par conséquent, p_{xy^*} de la section 3.1 était connue pour la NO. Toutefois, pour simuler la réalité, nous avons estimé p_{xy^*} d'après un échantillon pour examen manuel de taille 1 000 qui a été tiré des fichiers de données appariées par échantillonnage aléatoire simple.

5.1.4 Modélisation de la probabilité qu'un enregistrement demeure non apparié

À chaque enregistrement de la RGR apparié en appliquant la NO, nous avons attribué une variable indiquant si l'enregistrement était non apparié en appliquant la NB. Autrement dit, si l'enregistrement demeurait non apparié sous la NB, la variable indicatrice prenait la valeur « 1 » et autrement, la valeur « 0 ». Nous avons ajusté un modèle logistique en utilisant la NO, où la variable dépendante était la variable indicatrice susmentionnée et les variables explicatives provenaient de la RGR. Le modèle contenait plus de 20 variables explicatives qui ont été choisies selon la méthode classique de sélection ascendante/descendante. Les variables explicatives comprennent le niveau de scolarité, la langue, la naissance outre-mer, le statut d'autochtone, et les indicateurs de variable clé manquante, tels que l'*îlot* (*meshblock*). La prédiction résultante a donné t_i qui est utilisé plus bas pour mettre en œuvre la méthode du pseudo-MV pour les tableaux de contingence ainsi que la régression logistique.

5.2 Résultats de l'analyse tabulaire

Le tableau 2 donne les résultats du recoupement des situations d'emploi des personnes autochtones déclarées à la RGR et au recensement. Le tableau 2a montre que en appliquant la NO, l'estimation de la proportion d'Autochtones occupés au recensement, sachant qu'ils étaient occupés au moment de la RGR, est de 78,3 %. L'estimation naïve correspondante dans les conditions TF, qui reposent sur l'hypothèse que les données sont parfaitement appariées, est de 86,7 %. Même après avoir remplacé chacun des 700 appariements TF incorrects par l'appariement correct correspondant et écarté les 1 300 enregistrements non appariés pour lesquels aucun appariement correct n'existe, l'estimation naïve demeure pratiquement inchangée, soit 86,0 % (*appariements or* dans le tableau 2a). nous voyons donc que la différence entre les estimations TF et NO ne résultent pas tellement d'appariements incorrects, et qu'elle est due principalement à des enregistrements non appariés. Cela explique partiellement pourquoi l'estimation du MV (86,4 %) dans les conditions EF (voir la section 3.1), qui comporte une correction pour les appariements incorrects seulement, ne produit qu'une faible amélioration. Le MV conditionnel (MVC) (voir la section 4) a été examiné pour essayer de réduire l'erreur due aux enregistrements non appariés risquant de donner lieu à une représentation incorrecte, pour ce qui est de l'âge et du sexe, dans le fichier de données appariées. L'estimation de l'emploi par le MVC

était de 86,6 %. L'amélioration produite par le MVC n'était pas importante, ce qui indique que le mécanisme sous-jacent qui donne des enregistrements non appariés ne dépendait pas de l'âge ni du sexe. Les estimations du PMV (voir la section 4) n'ont pas non plus amélioré beaucoup la situation, ce qui indique que le modèle logistique décrit à la section 5.1.4 n'expliquait pas le mécanisme générant les enregistrements non appariés. Curieusement, l'estimation du MV en utilisant le fichier EF était de 81,8 %, c'est-à-dire l'estimation de loin la plus proche de l'estimation de 78,3 % sous la NO. Or, dans le cas du seuil EF, les appariements incorrects sont la principale source d'erreur, c'est-à-dire le type d'erreur d'appariement que corrige l'estimateur du MV. Par conséquent, la correction des erreurs dues à des appariements incorrects a été beaucoup plus fructueuse que celle des erreurs dues à des enregistrements non appariés.

Les erreurs types des estimations NO, naïves et MV sont indiquées entre parenthèses dans le tableau 2a. Dans le cas de TF et de EF, les erreurs-types de l'estimation MV sont environ 25 % et 75 % plus grandes, respectivement, que les erreurs-types de l'estimation naïve correspondantes. En outre, les erreurs-types de l'estimation MV pour EF sont légèrement plus faibles que pour TF, ce qui signifie que les appariements supplémentaires effectués en appliquant le seuil EF valaient la peine. Clairement, l'inférence naïve dans les conditions EF surestime le niveau de confiance des estimations. Pour les conditions TF, les estimations naïve et MV et leurs erreurs-types sont très proches.

Quel que soit le seuil utilisé, les estimations MV des tableaux 2a, b et c sont systématiquement plus proches de l'estimation NO que l'estimation naïve correspondante. Par exemple, dans le tableau 2b, l'estimation MV pour le seuil TF est de 36,9 %, c'est-à-dire appréciablement plus proche de l'estimation NO de 37,9 % que l'estimation naïve de 33,3 %. En fonction des estimations du tableau 2, nous pourrions soutenir que la décision d'utiliser le seuil TF ou EF n'est pas tellement importante, à condition d'utiliser l'estimateur du MV.

Le tableau 3 ressemble au tableau 2, à part le fait qu'il décrit les analyses des enregistrements appariés pour toutes les personnes de 15 ans et plus, plutôt que pour les Autochtones seulement. De nouveau, l'estimateur du MV produit systématiquement une amélioration dans le cas du seuil EF, mais non dans le cas du seuil TF. Le tableau 4 donne la situation d'étudiant en 2006 pour les personnes qui étaient étudiantes en 2005. Encore une fois, le MV donne généralement des estimations plus proches de l'estimation or correspondante, en particulier pour le seuil EF.

Tableau 2
Pourcentages d'Autochtones dans les diverses catégories d'emploi en 2006 sachant leur catégorie d'emploi en 2005. Pour chaque ensemble de données appariées, soit Seuil très faible ou extrêmement faible, les méthodes d'estimation peuvent être comparées à la norme d'or

	Estimations pour divers ensembles de donnée appariées et méthodes							
a : Autochtones occupés en 2005								
Situation en 2006	**or**			**Seuil très faible**			**Seuil extrêmement faible**	
		Naïve	**Appariements or**	**MV**	**PMV**	**MVC**	**Naïve**	**MV**
Occupé	78,3	86,7	86,0	86,4	86,6	86,1	71,9	81,8
	(1,7)	(2,4)		(3,0)			(1,7)	(2,9)
En chômage	3,7	4,2	4,3	4,1	4,1	4,2	6,3	3,3
	(0,84)	(1,2)		(2,5)			(0,82)	(2,1)
Inactif	17,8	9,0	9,6	9,3	9,1	9,6	21,6	14,7
	(1,6)	(2,4)		(3,1)			(1,6)	(2,8)

b : Autochtones en chômage en 2005					
Situation en 2006	**or**	**Très faible**		**Extrêmement faible**	
		Naive	**ML**	**Naive**	**ML**
Occupé	27,5	27,7	27,2	35,2	23,8
En chômage	34,4	38,9	36,4	32,3	38,0
Inactif	37,9	33,3	36,3	32,3	38,0
c : Autochtones inactifs en 2005					
Occupé	13,7	10,8	10,7	24,3	10,5
En chômage	5,8	7,6	7,4	6,3	5,8
Inactif	80,4	81,5	81,8	69,2	83,5

Tableau 3
Pourcentages de l'ensemble des personnes de plus de 15 ans dans les diverses catégories d'emploi en 2006 sachant leur catégorie d'emploi en 2005. Pour chaque ensemble de données appariées, soit seuil Très Faible ou Extrêmement Faible, les méthodes d'estimation peuvent être comparées à la norme d'or

	Estimations pour divers ensembles de donnée appariées et méthodes				
Situation en 2006	**or**	**Très faible**		**Extrêmement faible**	
		Naïve	**MV**	**Naïve**	**MV**
a : Personnes occupées en 2005					
Occupé	91,8	92,2	92,6	89,7	92,4
En chômage	1,8	1,7	1,6	1,9	1,6
Inactif	6,2	6,1	5,6	8,3	5,8
b : Personnes en chômage en 2005					
Occupé	44,5	44,3	44,0	49,4	43,8
En chômage	26,8	26,6	27,5	22,8	27,6
Inactif	28,6	28,7	28,4	27,6	28,5
c : Personnes inactives en 2005					
Occupé	12,1	12,3	11,1	16,8	11,0
En chômage	3,1	3,1	3,0	3,0	3,0
Inactif	84,7	84,5	85,7	80,1	85,9

Tableau 4
Situation d'étudiant en 2006 pour les étudiants du secondaire en 2005

Situation d'étudiant en 2006	**Or**	**Très faible**		**Extrêmement faible**	
		Naïve	**MV**	**Naïve**	**MV**
Élève du secondaire	79,3	79,3	79,6	77,4	79,6
Études secondaires terminées	14,0	14,3	13,7	14,7	14,1
Études secondaires non terminées	6,6	6,3	6,6	7,8	6,2

5.3 Simulation

L'étude par simulation qui suit illustre les problèmes que pose l'analyse naïve et les avantages de l'utilisation de la méthode décrite dans le présent article. Dans la simulation, les fichiers X et Y, contenant chacun 2 000 enregistrements, sont générés indépendamment 400 fois, chaque fichier généré étant désigné par X(r) et Y(r), respectivement, avec $r = 1, ..., 400$. En particulier, sur X(r), x_i est tiré aléatoirement de la loi de Bernoulli de paramètre 0,5. Sur Y(r), y_i est tiré aléatoirement de la loi de Bernoulli de paramètre υ_i, où $\upsilon_i = 1/[1+\exp(\beta_0 + \beta_1 x_i)]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)', \beta_0 = -0,5$, $\beta_1 = 1,5$. Le r^{e} ensemble de données imparfaitement appariées, $\mathbf{d}^*(r)$, est généré en appariant correctement chaque enregistrement du fichier Y(r) à un enregistrement du fichier X(r) avec la probabilité $p = 0,8, 0,90, 0,95$ et 1. Pour chaque r^{e} ensemble de données appariées, un échantillon pour examen manuel de 300 appariements est sélectionné. Chaque appariement contenu dans l'échantillon pour examen manuel est classé comme étant correct ou incorrect. Nous résumons la performance de l'estimateur du MV décrit à la section 3.2.2 et de la méthode naïve, qui repose sur l'hypothèse qu'il n'existe aucune erreur d'appariement, par leur taux de couverture à 95 % et leur erreur quadratique moyenne (EQM). Les taux de couverture sont fondés sur les erreurs-types calculées par la méthode du bootstrap décrite à la section 3.3 avec $R = 40$ répliques. L'EQM de $\tilde{\boldsymbol{\beta}}$ est calculée par

$$\text{EQM}(\tilde{\boldsymbol{\beta}}) = \frac{1}{400}\sum_{r=1}^{400}(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})'$$

où $\tilde{\boldsymbol{\beta}}_r$ est l'estimation du MV de $\boldsymbol{\beta}$ d'après $\mathbf{d}^*(r)$.

Le tableau 5 montre que l'approche naïve produit de mauvais taux de couverture, à cause de son biais important en présence d'erreurs d'appariement et, par conséquent, son EQM relativement élevée. Pour le MV-méthode 1, les taux de couverture sont très proches des taux nominaux. Les résultats montrent que, si le pourcentage d'appariements corrects passe de 100 % à 80 %, l'EQM pour l'estimation MV augmente d'un facteur 3 environ pour β_0 et β_1 (les taux de couverture et l'EQM des méthodes 1 et 2 d'estimation du MV étant fort semblables, seuls les résultats pour la première sont présentés ici).

Tableau 5
Erreur quadratique moyenne et taux de couverture pour les données simulées appariées, avec la probabilité p qu'un appariement soit correct

		Erreur quadratique moyenne				**Taux de couverture à 95 %**		
		0,8	0,9	0,95	1	0,8	0,9	0,95
Naïve	β_0	0,024	0,010	0,0056	0,0043*	0,35	0,80	0,93
	β_1	0,11	0,038	0,016	0,011*	0,05	0,62	0,88
MV-Méthode 1	β_0	0,013	0,0078	0,0055	0,0043*	93,0	94,25	93,5
	β_1	0,031	0,018	0,013	0,011*	96,0	94,5	96,25

* Quand $p = 1$, les estimateurs naïf et MV sont identiques par définition.

6. Discussion

L'appariement des données est une méthode appropriée quand des ensembles de données doivent être jumelés en vue de renforcer la dimension temporelle ou des aspects tels que la portée ou la profondeur des renseignements. L'appariement des données est utilisé de plus en plus fréquemment par les organismes statistiques partout dans le monde. Il est bien connu que des erreurs peuvent se produire durant l'appariement des fichiers, par exemple quand des méthodes probabilistes sont appliquées. Toutefois, peu d'études traitant de la façon de faire des inférences valides en présence de ce genre d'erreurs ont été publiées. Le présent article offre des conseils méthodologiques et pratiques en vue d'aider les analystes dans ce domaine.

En général, traiter naïvement un fichier de données appariées comme si l'appariement était parfait donne lieu à des estimations biaisées. L'analyste ne devrait adopter l'approche naïve que si le nombre d'enregistrements non appariés, définis comme étant des enregistrements qui pourraient être appariés correctement, mais qui ne l'ont pas été du tout, ainsi que le nombre d'appariements incorrects sont négligeables. Le présent article décrit une approche fondée sur le maximum de vraisemblance en vue de faire de inférences valides en présence des deux sources d'erreur. Cette approche s'appuie sur l'algorithme EM bien connu et est facile à appliquer en pratique. La méthode peut être utilisée quand l'un des fichiers n'est pas nécessairement un sous-ensemble de l'autre et que l'appariement comporte des passages multiples. Ces situations se présentent fréquemment en pratique, comme en témoignent de nombreux exemples récents à l'Australian Bureau of Statistics. L'étude empirique montre que l'approche du MV améliore de manière significative les estimations fondées sur des données appariées.

Dans le cas particulier où le fichier X est obtenu par tirage d'un échantillon aléatoire du fichier Y, la procédure d'estimation décrite n'est pas « complètement » celle du maximum de vraisemblance, parce qu'elle ne s'appuie pas sur le fait que les totaux de population pour le fichier Y sont connus. Bien que l'inférence selon la méthode décrite ici demeure valide dans ce cas, il y aurait peut-être moyen de la rendre plus efficace (voir Scott et Wild 1997).

Remerciements

Les auteurs remercient Raymond Chambers et deux examinateurs de *Techniques d'enquête* de leur contribution au présent article.

Bibliographie

Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study, 2006-07. Document d'information n° de catalogue 4723.0.

Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Chambers, R., Chipperfield, J.O., Davis, W. et Kovačević, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Soumis pour publication.

Chambers, R.L., et Skinner, C.J. (2003). *Analysis of Survey Data*. New York : John Wiley & Sons, Inc.

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4, http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Christen, P., et Churches, T. (2005). Febrl – Freely extensible biomedical record linkage. Version 0.3.1, vue le 17 novembre 2008, http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html.

Conn, L., et Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Articles du comité consultative sur la méthodologie, n° de catalogue 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 et 2), 37-53.

Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fuller, W.A. (1987). *Measurement Error Models*. New York : John Wiley & Sons, Inc.

Hausman, J.A., Abrevaya, J. et Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York : Springer.

Holman, C.D.J., Bass, A.J., Rouse, I.L. et Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.

Lahiri, P., et Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Rapport de méthodologie, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.

Rubin, D.B., et Little, R.J.A. (2003). *Statistical analysis of missing data*, 2e Édition. New York : John Wiley & Sons, Inc.

Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

Scott, A.J., et Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.

Solon, R., et Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Collection de rapports de recherche statistique, n° RR2001/03, Bureau of the Census.

Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Collection de rapports de recherche statistique, n° RRS2005/02, Bureau of the Census.

Wright, J., Bishop, G. et Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Travaux de recherche de méthodologie, n° de catalogue 1351.0.55.027, Australian Bureau of Statistics, Canberra.