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

Estimation fondée sur le plan de sondage à partir de fichiers administratifs avec des enregistrements couplés

Abel Dasylva[1]

Résumé

Le couplage d'enregistrements exact est un outil essentiel à l'exploitation des fichiers administratifs, surtout quand on étudie les relations entre de nombreuses variables qui ne sont pas toutes contenues dans un fichier administratif unique. L'objectif est de trouver des paires d'enregistrements associées à une même personne ou entité. Le résultat est un fichier couplé qui peut être utilisé pour estimer les paramètres de population, y compris les totaux et les ratios. Malheureusement, le processus de couplage est complexe et sujet à erreurs parce qu'il s'appuie habituellement sur des variables d'appariement qui ne sont pas uniques et qui peuvent être consignées avec des erreurs. Par conséquent, le fichier couplé contient des erreurs d'appariement, y compris des appariements incorrects d'enregistrements non apparentés et des appariements manquants d'enregistrements apparentés. Ces erreurs peuvent donner lieu à des estimateurs biaisés s'il n'en est pas tenu compte dans le processus d'estimation. Dans le cadre de travaux antérieurs dans ce domaine, ces erreurs ont été prises en considération au moyen d'hypothèses au sujet de leur distribution. En général, la distribution supposée est en fait une approximation très grossière de la distribution réelle, en raison de la complexité intrinsèque du processus de couplage. Donc, les estimateurs résultants peuvent présenter un biais. Un nouveau cadre méthodologique, fondé sur la théorie classique des sondages, est proposé pour obtenir des estimateurs fondés sur le plan de sondage à partir de fichiers administratifs d'enregistrements couplés. Il comprend trois étapes. Pour commencer, on tire un échantillon probabiliste de paires d'enregistrements. Ensuite, on procède à un examen manuel de toutes les paires échantillonnées. Enfin, on calcule des estimateurs fondés sur le plan de sondage en fonction des résultats de l'examen. Cette méthodologie mène à des estimateurs dont l'erreur d'échantillonnage est fondée sur le plan de sondage, même si le processus repose uniquement sur deux fichiers administratifs. Elle s'écarte des travaux antérieurs s'appuyant sur un modèle et fournit des estimateurs plus robustes. Ce résultat est obtenu en plaçant les examens manuels au cœur du processus d'estimation. Le recours aux examens manuels est essentiel, parce qu'il s'agit de fait d'une norme de référence en ce qui a trait à la qualité des décisions au sujet des appariements. Le cadre proposé peut également être appliqué à l'estimation au moyen de données administratives et de données d'enquête couplées.

Mots clés : couplage d'enregistrements, enquête, fichier administratif, plan de sondage, estimateur, modèle, variable auxiliaire, examen manuel, modèle de mélange, algorithme espérance-maximisation (EM), assisté par modèle.

1. Introduction

1.1 Le problème de couplage d'enregistrements

L'objectif du couplage d'enregistrements exact, appelé simplement couplage d'enregistrements dans la suite, consiste à jumeler des enregistrements qui sont associés à une même personne ou entité. Les enregistrements ainsi jumelés sont appelés enregistrements *appariés*. Le couplage d'enregistrements est un exercice simple quand les enregistrements contiennent un identificateur unique. En l'absence d'un tel identificateur, le couplage d'enregistrements doit se fonder sur des pseudo-identificateurs qui ne sont pas uniques et dont l'enregistrement dans différents fichiers présente souvent des variations. Ainsi, dans les fichiers sur les personnes, les pseudo-identificateurs peuvent comprendre le nom de famille, le prénom et la date de naissance. Les variations sont dues à des erreurs typographiques ou à des différences de format pour des variables qui doivent par ailleurs enregistrer la même information. Dans ce contexte, le couplage d'enregistrements est sujet à erreurs, y compris des appariements incorrects entre des enregistrements non apparentés et des appariements manquants entre des enregistrements apparentés qui, accidentellement, paraissent trop différents. Le défi du couplage d'enregistrements

[1]Abel Dasylva, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (ON) K1A 0T6, Canada (abel.dasylva@statcan.gc.ca).

se résume essentiellement à tester l'hypothèse nulle que deux enregistrements donnés sont apparentés, compte tenu de l'information disponible. De ce point de vue, les appariements manquants sont des erreurs de type I, tandis que les appariements incorrects sont des erreurs de type II. Les deux types d'erreurs appauvrissent la qualité d'un couplage, et ces erreurs doivent être contrôlées en se fondant sur les différences observées dans chaque paire d'enregistrements. Dans le cas des petits fichiers, cette information est traitée facilement par des commis chevronnés. Cependant, dans le cas des grands fichiers, cette solution est trop onéreuse étant donné les millions de paires d'enregistrements qui peuvent être générées. Il faut plutôt recourir à une solution informatisée où la décision concernant le couplage est automatisée pour la plupart des paires d'enregistrements.

1.2 Le couplage d'enregistrements probabiliste

Newcombe (1967) et Fellegi et Sunter (1969) ont proposé une méthode probabiliste de couplage d'enregistrements comprenant une base théorique pour les solutions informatisées, où le nombre d'examens manuels est réduit au minimum et où les erreurs de couplage sont contrôlées. Sous l'approche probabiliste, chaque paire d'enregistrements se voit attribuer un poids d'appariement dont la valeur augmente lorsque les enregistrements deviennent plus similaires. Le poids est comparé à des seuils afin de déterminer si une paire devrait être appariée, examinée manuellement ou rejetée (Fellegi et Sunter, 1969). La performance globale du couplage d'enregistrements est déterminée par les poids d'appariements et les seuils.

1.3 Les solutions entièrement automatisées et leurs limites

Une solution de couplage probabiliste peut être entièrement automatisée ou semi-automatisée. Newcombe et coll. (1983) ont préconisé l'usage de décisions automatisées en se fondant sur leur expérience avec des fichiers de données de haute qualité sur les personnes, y compris les noms, les dates de naissance ou les adresses. Cependant, la supériorité des solutions entièrement automatisées a été mise en doute en raison de l'inexactitude des taux d'erreur estimés et de la nécessité de les appuyer par des examens manuels (Heasman, 2014) ou des échantillons d'entraînement (Belin, 1995). Le problème est apparenté à celui de l'identifiabilité de mélanges multinomiaux généraux. Kim (1984) a énoncé les conditions suffisantes pour l'identifiabilité de ce genre de mélanges. Ce résultat s'applique au couplage d'enregistrements seulement dans la situation irréaliste où les paires sont classées en groupes connus qui contiennent chacune des paires ayant le même état d'appariement. Cependant, une structure de classement de ce genre n'existe pas en pratique. Par conséquent, la modélisation des paires au moyen d'un mélange multinomial avec interactions peut comprendre des hypothèses non vérifiables. De telles hypothèses sont également formulées pour traiter la non-réponse non ignorable dans les sondages, quand la non-réponse ne touche que quelques unités. Cependant, la difficulté est plus grande dans le cas du couplage d'enregistrements, parce que toute erreur de spécification du modèle affecte toutes les paires.

1.4 Les solutions semi-automatisées

La plupart des solutions pratiques sont semi-automatisées et, en dernière analyse, s'appuient sur des examens manuels ou des échantillons d'entraînement. Dans ces solutions, un petit échantillon de paires est examiné manuellement, tandis que toutes les paires restantes sont appariées automatiquement en fonction de paramètres. Les paramètres peuvent être estimés au moyen du même échantillon. La règle de décision optimale de Fellegi-Sunter (Fellegi et Sunter, 1969) est un bon exemple de solution semi-automatisée, parce qu'elle requiert des examens manuels dans la *zone grise* entre les seuils. Le logiciel G-COUP (anciennement appelé SGCE), qui est le système généralisé de couplage d'enregistrements de Statistique Canada, offre un autre exemple où les poids d'appariements et les seuils sont estimés de façon manuelle et itérative (Howe, 1981). Gill (2001) ainsi que Guiver (2011) ont recommandé d'utiliser les examens manuels pour établir les seuils de pondération. Guiver (2011) remarque aussi que le *spot-checking*, une procédure fréquente pour l'établissement des seuils, ne possède aucun fondement statistique valable. Il recommande plutôt l'examen d'un échantillon probabiliste de paires. Larsen et Rubin (2001) ont décrit une autre solution itérative intégrant des examens manuels pour faciliter la sélection d'un modèle pour les paires. Belin et Rubin (1995) se sont également servis d'un échantillon d'entraînement pour estimer le taux de faux appariements.

1.5 L'estimation fondée sur le plan de sondage avec des données couplées

La présente étude porte sur les questions connexes de l'estimation convergente sous le plan et de l'échantillonnage dans le cas d'échantillons pour examen manuel. Elle comprend l'utilisation de l'information auxiliaire provenant des

résultats des comparaisons et de modèles de leur distribution dans les paires. Särndal et coll. (1992) ont déjà traité le problème général de l'estimation efficace et convergente sous le plan avec des variables auxiliaires, des modèles et des estimateurs connexes. Ils ont proposé des estimateurs par la régression qui sont convergents sous le plan et qui utilisent de façon optimale les données auxiliaires disponibles si le modèle supposé est vérifié. L'application de ces résultats au couplage d'enregistrements a nécessité une certaine adaptation afin de tirer parti de nouvelles possibilités. Les estimateurs résultants sont des estimateurs par la régression dans lesquels les variables auxiliaires sont intégrées au moyen d'une fonction non linéaire. Ils sont construits en deux parties. Dans la première partie, toutes les paires d'enregistrements qui satisfont à des critères de groupage sont utilisées pour ajuster un modèle en vue de prédire l'état d'appariement des paires dans les groupes, ou blocs, qu'elles fassent partie ou non de l'échantillon pour examen manuel. Dans la deuxième partie, un estimateur par la régression est calculé et les inexactitudes possibles du modèle sont corrigées en se fondant sur les données manuelles. Le cadre décrit est également applicable quand l'état d'appariement est déterminé par d'autres moyens que des examens manuels, p. ex. grâce à l'accès limité à des identificateurs uniques ou à de l'information supplémentaire fournie par une tierce partie.

La suite de l'exposé est structurée comme il suit. À la section 2, nous présentons le modèle et la notation. À la section 3, nous décrivons les estimateurs fondés sur un modèle dans le contexte du couplage d'enregistrements. À la section 4, nous discutons des plans de sondage. À la section 5, nous présentons les résultats des simulations. À la section 6, nous présentons les conclusions et les travaux à venir.

2. Modèle et notation

2.1 Registres et population finie de paires d'enregistrements

Considérons deux registres A et B sans doublons, qui contiennent des enregistrements au sujet de N personnes. Le registre A contient K variables d'appariement et la variable d'intérêt x_i pour le i^{e} enregistrement dans A. Le registre B contient les mêmes variables d'appariement que A et la variable d'intérêt y_j pour le j^{e} enregistrement dans B. Soit U la population finie des N^2 paires d'enregistrements dans le produit cartésien des deux fichiers, c.-à-d. de toutes les paires (i, j) où $1 \leq i, j \leq N$.

2.2 Résultats des comparaisons et pochettes

Pour la paire d'enregistrements (i, j) dans le produit cartésien des deux registres, les variables d'appariement peuvent être comparées pour produire un K-uplet $\boldsymbol{\gamma}_{ij} = \left(\gamma_{ij}^{(1)}, \ldots, \gamma_{ij}^{(K)}\right)$ de résultats des comparaisons, également appelé vecteur des résultats des comparaisons. Dans les grands fichiers, certaines variables d'appariement sont également comparées grossièrement pour produire des pochettes de paires qui, réunies, représentent un petit sous-ensemble U^* de U, mais contiennent la plupart des paires appariées. Le sous-ensemble U^* des paires des pochettes est l'union de B sous-ensembles disjoints, $U_1^*, \ldots, U_B^*$, où chaque sous-ensemble représente une pochette distincte. Pour chaque paire, cette information des pochettes est également incluse dans le vecteur des comparaisons γ_{ij}. Ce vecteur des comparaisons $\boldsymbol{\gamma}_{ij}$ sert de fondement au couplage des enregistrements, par exemple en utilisant la règle de couplage optimale de Fellegi et Sunter (1969). Cependant, la méthodologie d'estimation proposée ici ne requiert pas un tel couplage.

2.3 État d'appariement des paires

Soit M_{ij} la variable indicatrice dont la valeur est fixée à 1 si la paire (i, j) est appariée, c.-à-d. si les deux enregistrements sont associés à une même personne. La variable M_{ij} est également appelée état d'appariement de la paire (i, j). Le vecteur des comparaisons $\boldsymbol{\gamma}_{ij}$ est essentiel pour faire une inférence $\widehat{M}_{ij}$ au sujet de l'état d'appariement inconnu M_{ij}. L'état d'appariement inféré $\widehat{M}_{ij}$ peut prendre de nombreuses formes. Par exemple, il peut être établi à la probabilité d'appariement conditionnelle ou *à posteriori* $P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ sachant le vecteur des comparaisons. Il peut aussi être interprété comme la « part de poids » de la paire (i, j), au sens de la méthode généralisée du partage des poids. Voir Lavallée (2002, chap. 9) pour des applications de cette méthode au couplage d'enregistrements.

2.4 Problèmes d'inférence

Dans le cas de l'inférence en population finie, l'objectif est d'estimer un total de la forme suivante :

$$Z \quad = \quad \sum_{(i,j)\in U} M_{ij} z_{ij}$$

Dans l'expression susmentionnée, $z_{ij} = f(x_i, y_j)$ et f est une fonction connue.

Les inférences au sujet de la superpopulation reposent sur l'hypothèse que les variables d'intérêt x_i et y_j dans les paires appariées sont indépendantes et identiquement distribuées (IID). Elles ont pour cible un paramètre de superpopulation θ qui satisfait une équation de score de la forme $E[S(\theta; x_i, y_j)] = 0$, où S est une fonction de score (p. ex., un logarithme du rapport de vraisemblance) et l'espérance est calculée par rapport à la superpopulation. Le paramètre θ peut être estimé au moyen de l'équation d'estimation sans biais qui suit où $z_{ij}(\theta) = S(\theta; x_i, y_j)$.

$$\sum_{(i,j)\in U} M_{ij} z_{ij}(\hat{\theta}) \quad = \quad 0$$

Dans les deux cas, les inférences s'appuient sur les valeurs enregistrées des variables dans les paires appariées, que ces valeurs soient ou non exemptes d'erreurs non dues à l'échantillonnage, telles que des erreurs typographiques, des erreurs de mesure, etc.

2.5 L'échantillon pour examen manuel

Des ressources sont disponibles pour procéder à des examens manuels sans erreurs en vue de déterminer l'état d'appariement. Cependant, ces examens sont coûteux et leur nombre doit être réduit au minimum. L'échantillon pour examen manuel s possède une taille fixe. Il est divisé en une strate des pochettes U^* et une strate hors-pochettes $U \backslash U^*$. Soit s^* l'échantillon de paires des pochettes dans l'échantillon pour examen manuel. Les échantillons dans les différentes strates sont tirés indépendamment et les plans de sondage sont arbitraires.

3. Estimateurs assistés par un modèle

3.1 Une forme générale

Les estimateurs proposés ont la forme générale de différence qui suit :

$$\hat{Z} \quad = \quad \underbrace{\sum_{(i,j)\in U^*} \widehat{M}_{ij} z_{ij} + \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} (M_{ij} - \widehat{M}_{ij})}_{(1)} + \underbrace{\sum_{(i,j)\in s\backslash s^*} \pi_{ij}^{-1} M_{ij} z_{ij}}_{(2)}$$

Cet estimateur est égal à la somme des contributions des deux strates. La première contribution exploite l'état d'appariement inféré pour estimer le total sur la strate des pochettes avec une plus grande précision. La deuxième contribution est simplement un estimateur de Horwitz-Thompson du total sur la strate hors-pochettes. L'estimateur proposé est sans biais si l'état inféré ne tient pas compte de l'information de l'échantillon examiné manuellement :

$$E[\hat{Z}|U] \quad = \quad \sum_{(i,j)\in U} M_{ij} z_{ij} \quad = \quad Z$$

Cela est le cas si $\widehat{M}_{ij}$ est seulement une fonction de z_{ij} et γ_{ij}.

3.2 Inférence de l'état d'appariement

L'état inféré peut être considéré comme étant la probabilité conditionnelle d'appariement sachant le vecteur des résultats des comparaisons et les variables x_i, y_j, c.-à-d.

$$\widehat{M}_{ij} = P(M_{ij} = 1 | x_i, y_j, \boldsymbol{\gamma}_{ij})$$

Cette stratégie particulière d'inférence minimiserait l'erreur quadratique moyenne (sur la superpopulation) entre le total prédit $\sum_{(i,j)\in U^*} \widehat{M}_{ij} z_{ij}$ et le total réel $\sum_{(i,j)\in U^*} M_{ij} z_{ij}$ sur la strate des pochettes, parmi toutes les stratégies d'inférence où $\widehat{M}_{ij}$ est seulement une fonction de x_i, y_j et $\boldsymbol{\gamma}_{ij}$, si les paires d'enregistrements sont IID. Sous un plan à échantillonnage aléatoire simple (EAS) dans la strate des pochettes, l'estimateur résultant serait également plus efficace que l'estimateur de Horwitz-Thompson, si les paires étaient IID.

La probabilité conditionnelle d'appariement peut être estimée sous l'hypothèse que les paires sont IID selon un modèle de mélange de lois à deux composantes, où les différents résultats des comparaisons et les variables x_i, y_j sont supposés être conditionnellement indépendants sachant l'état d'appariement, et où $\tau = 0{,}1$:

$$P\left(x_i, y_j, \gamma_{ij} \middle| M_{ij} = \tau\right) \quad = \quad P\left(x_i, y_j \middle| M_{ij} = \tau\right) \prod_{k=1}^{K} P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = \tau\right)$$

Les paramètres $\boldsymbol{\psi}$ de ce mélange englobent la proportion de mélange $\lambda = P\left(M_{ij} = 1\right)$, les probabilités marginales pour les paires appariées $P\left(x_i, y_j \middle| M_{ij} = 1\right)$ et $P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = 1\right)$, et les probabilités marginales pour celles non-appariées $P\left(x_i, y_j \middle| M_{ij} = 0\right)$ et $P\left(\gamma_{ij}^{(k)} \middle| M_{ij} = 0\right)$, sous l'hypothèse de paires IID. Elles peuvent être estimées au moyen d'un algorithme d'espérance-maximisation (EM). Voir Jaro (1989) ou Winkler (1988) pour des applications de l'algorithme EM au couplage d'enregistrements, et Dempster et coll. (1977) pour une référence générale à l'algorithme EM. Une caractéristique importante de ce modèle de mélange est l'utilisation de x_i et y_j comme variables d'appariement supplémentaires. Le modèle de mélange devient plus simple quand les variables x_i et y_j sont fortement corrélées aux variables d'appariement. Dans ce cas, x_i et y_j n'apportent aucune information nouvelle au sujet de l'état d'appariement, sachant $\boldsymbol{\gamma}_{ij}$. Mathématiquement, cela s'exprime par l'indépendance conditionnelle de $\left(x_i, y_j\right)$ et l'état d'appariement sachant les résultats des comparaisons :

$$P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}\right) \quad = \quad P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$$

La stratégie d'inférence peut être inefficace si le modèle de mélange présumé n'est pas vérifié. Par exemple, ce problème peut se poser si le couple $\left(x_i, y_j\right)$ contient de l'information supplémentaire au sujet de l'état d'appariement, mais que l'inférence $\widehat{M}_{ij} = P\left(M_{ij} = 1 \middle| \boldsymbol{\gamma}_{ij}\right)$ est utilisée à la place. L'estimateur est également moins efficace si les variables d'appariement sont corrélées, mais que leur indépendance conditionnelle est supposée.

Soit $P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$ une estimation préliminaire de la probabilité conditionnelle d'appariement sous le modèle de mélange. Cette estimation est calculée à l'étape E de l'algorithme EM et elle n'utilise pas les résultats des examens manuels. Dans la plupart des cas, ce modèle de mélange estimera la probabilité conditionnelle d'appariement avec un certain biais, même s'il tient compte de certaines interactions entre les différentes variables. Pour corriger ce biais, l'état d'appariement peut être inféré en utilisant une fonction linéaire $\beta_0 + \beta_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$ de la probabilité conditionnelle estimée, où les coefficients de régression β_0 et β_1 sont estimés pour l'échantillon examiné manuellement. Dans ce cas, l'état d'appariement inféré est calculé comme il suit :

$$\widehat{M}_{ij} \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$$

Un cas particulier est celui où un estimateur par le ratio estime le total sur la strate des pochettes. C'est-à-dire,

$$\hat{Z} \quad = \quad \frac{\sum_{(i,j)\in U^*} z_{ij}\, P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)} \sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}$$

Dans ce cas, $\hat{\beta}_0 = 0$ et $\hat{\beta}_1$ est calculé comme il suit :

$$\hat{\beta}_1 \quad = \quad \frac{\sum_{(i,j)\in U^*} z_{ij}\, P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)}{\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij} P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)}$$

L'estimateur peut aussi s'écrire en fonction de poids g uniformes $\left[g_{ij}\right]_{ij}$, où $g_{ij} = \hat{\beta}_1$:

$$\hat{Z} \quad = \quad \sum_{(i,j)\in s^*} g_{ij}\pi_{ij}^{-1} z_{ij} M_{ij} + \sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}$$

Le modèle qui suit sert de fondement à l'obtention de meilleurs estimateurs par les moindres carrés pondérés :

$$E\left[M_{ij}\middle|x_i, y_j, \gamma_{ij}\right] \quad = \quad \beta_0 + \beta_1 P\left(M_{ij} = 1\middle|x_i, y_j, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)$$

$$var\left(M_{ij}\middle|x_i, y_j, \gamma_{ij}\right) \quad \propto \quad P\left(M_{ij} = 1\middle|z_{ij}, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij} = 1\middle|x_i, y_j, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)\right]$$

Dans ce cas, les coefficients de régression estimés minimisent la fonction quadratique qui suit :

$$Q\left(\beta_0, \beta_1; \widehat{\boldsymbol{\psi}}\right) \quad = \quad \sum_{(i,j)\in s^*} \pi_{ij}^{-1} \frac{\left[M_{ij} - \beta_0 + \beta_1 P\left(M_{ij} = 1\middle|x_i, y_j, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)\right]^2}{P\left(M_{ij} = 1\middle|x_i, y_j, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)\left[1 - P\left(M_{ij} = 1\middle|x_i, y_j, \gamma_{ij}; \widehat{\boldsymbol{\psi}}\right)\right]}$$

L'estimateur résultant peut s'écrire en fonction de poids g non uniformes dans lesquels est incorporé l'état d'appariement inféré. Cet estimateur est amélioré en affinant la structure de variance par la méthode des équations d'estimation généralisées (Jiang, 2007).

Les estimateurs proposés ne sont plus sans biais parce que les résultats de l'examen manuel sont utilisés pour faire des inférences au sujet de l'état d'appariement des paires. Cependant, à l'instar des estimateurs par la régression classiques (Särndal et coll., 1992), ils sont convergents sous le plan quels que soient les modèles présumés.

4. Plan de sondage

L'échantillonnage stratifié fondé sur un modèle a été utilisé pour minimiser approximativement la variance des estimateurs par la régression (Särndal et coll., 1992). Dans ce plan, les strates sont définies par la variance de l'erreur dans le modèle linéaire supposé. Cette stratégie s'applique aussi au contexte actuel où un seul total est estimé. Précisément, la variance sous le plan de sondage $var\left(\hat{Z}|U\right)$ de l'estimateur assisté par modèle est égale à la somme de deux termes :

$$var\left(\hat{Z}|U\right) \quad = \quad var\left(\sum_{(i,j)\in s^*} \pi_{ij}^{-1} z_{ij}\left(M_{ij} - \widehat{M}_{ij}\right)\middle| U\right) + var\left(\sum_{(i,j)\in s\setminus s^*} \pi_{ij}^{-1} M_{ij} z_{ij}\middle| U\right)$$

Le premier terme peut être minimisé approximativement par une répartition de Neyman selon laquelle les paires sont stratifiées en fonction de la variance conditionnelle fondée sur le modèle $var\left(z_{ij}\left(M_{ij} - \widehat{M}_{ij}\right)\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$. En effet, supposons les probabilités conditionnelles d'appariement connues données par $P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ et la meilleure inférence possible $\widehat{M}_{ij} = P\left(M_{ij} = 1\middle|x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ pour chaque paire dans les pochettes. Supposons que les paires sont stratifiées en fonction de $\left(x_i, y_j\right)$, ou d'une certaine approximation discrète exacte de ces dernières, et $\boldsymbol{\gamma}_{ij}$. La variance dans la strate de $\left(x_i, y_j, \boldsymbol{\gamma}_{ij}\right)$ serait donnée par $z_{ij}^2 \widehat{M}_{ij}\left(1 - \widehat{M}_{ij}\right)$, si les paires étaient IID. Dans la répartition de Neyman correspondante, la taille de l'échantillon est proportionnelle à la variance dans la strate. Un estimateur possédant la même variance minimale est obtenu au moyen d'une répartition de Neyman, où les strates sont fondées sur l'estimation $z_{ij}^2 \widehat{M}_{ij}\left(1 - \widehat{M}_{ij}\right)$ de la variance conditionnelle de l'erreur.

5. Simulations

Deux scénarios de simulation ont servi à évaluer les différents estimateurs et à mesurer leur sensibilité aux hypothèses sous-jacentes. Les deux scénarios ont trait à un couplage un à un entre deux registres. Dans chaque

registre, les enregistrements sont partitionnés en pochettes parfaites de taille égale. Par conséquent, deux enregistrements appariés se trouvent toujours dans la même pochette.

5.1 Premier scénario

Dans le premier scénario, les variables d'appariement sont binaires, IID et de même loi pour les erreurs typographiques. Cette loi est donnée par les probabilités de transition suivantes :

$$\begin{aligned} P\left(c_i^{(k)}, c_j^{(k)} \middle| \zeta_i^{(k)}, M_{ij} = 1\right) &= P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) P\left(c_j^{(k)} \middle| \zeta_i^{(k)}\right) \\ P\left(c_i^{(k)} \middle| \zeta_i^{(k)}\right) &= (1-\alpha) I\left(c_i^{(k)} = \zeta_i^{(k)}\right) + \alpha I\left(c_i^{(k)} \neq \zeta_i^{(k)}\right) \end{aligned}$$

Dans les expressions susmentionnées, $c_i^{(k)}$ est la k^{e} variable d'appariement pour l'enregistrement i dans le registre A, $\zeta_i^{(k)}$ est la vraie valeur (c.-à-d. exempte d'erreurs d'enregistrement) latente de la variable pour la personne associée, avec $c_j^{(k)}$ et $\zeta_j^{(k)}$ désignant les variables correspondantes dans le registre B. Notons que, par définition $\zeta_i^{(k)} = \zeta_j^{(k)}$ dans une paire appariée (i, j). Pour chaque enregistrement i, les variables latentes $\zeta_i^{(k)}$ sont IID. Les résultats des comparaisons sont fondés sur des comparaisons exactes avec $\gamma_{ij}^{(k)} = I\left(c_i^{(k)} = c_j^{(k)}\right)$.

Les variables d'intérêt x_i et y_j sont également binaires et mutuellement indépendantes des variables d'appariement dans chaque registre et chaque paire appariée. Les fichiers sont couplés pour étudier la distribution conjointe de ces deux variables, c.-à-d. pour estimer les fréquences des différentes cellules dans un tableau de contingence à double entrée. Dans ce cas, $z_{ij} = I\left(\left(x_i, y_j\right) = (x, y)\right)$ où $x, y = 0{,}1$. Cette configuration est similaire à celle décrite par Chipperfield et coll. (2011). Cependant, ici, l'objectif est une inférence en population finie sur une seule population finie.

Différents échantillons IID sont tirés de la population finie en utilisant l'un des deux plans de sondage. Pour chaque échantillon résultant, on calcule trois estimateurs du nombre de paires appariées dans chaque cellule du tableau de contingence à double entrée. Il s'agit de l'estimateur de Horwitz-Thompson, d'un deuxième estimateur assisté par modèle (appelé simplement ci-après 2e estimateur) en utilisant l'inférence $\widehat{M}_{ij} = P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$, et d'un troisième estimateur (appelé simplement ci-après 3e estimateur) en utilisant l'inférence $\widehat{M}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 P\left(M_{ij} = 1 \middle| x_i, y_j, \boldsymbol{\gamma}_{ij}; \widehat{\boldsymbol{\psi}}\right)$.

Le premier plan de sondage est stratifié en fonction des paires de valeurs x-y. Un échantillon à EAS de taille fixe est tiré dans chaque strate. Le deuxième plan de sondage est également stratifié en fonction des paires de valeurs x-y, mais il comprend des sous-strates, qui sont fondées sur la variance conditionnelle de l'erreur de prédiction. Chaque strate x-y contient le même nombre de sous-strates, mais les limites sont choisies de manière à obtenir des tailles de sous-strates quasi égales, après avoir trié les paires en fonction de leur variance conditionnelle dans chaque strate. Par conséquent, les limites des sous-strates peuvent différer d'une strate x-y à la suivante. La taille d'échantillon de strate x-y utilisée est la même que dans le premier plan de sondage. Cependant, dans le deuxième plan de sondage, cette taille d'échantillon est répartie de manière optimale entre les sous-strates en utilisant une répartition de Neyman, où la variance d'une sous-strate est estimée comme étant la variance conditionnelle moyenne de l'erreur sur l'ensemble des paires correspondantes. Un échantillon de sous-strate issu de la répartition est en outre contraint de contenir au moins deux unités et de ne pas excéder la taille de la sous-strate.

Le premier scénario est choisi pour évaluer les deux estimateurs assistés par modèle dans le meilleur cas, où le modèle correct est utilisé pour la distribution des résultats. Cette situation devrait maximiser leur avantage relatif par rapport à l'estimateur de Horwitz-Thompson naïf.

5.2 Deuxième scénario

Le deuxième scénario est identique au premier, sauf en ce qui concerne la corrélation des variables latentes $\zeta_i^{(k)}$. Cette corrélation est produite en générant les $\zeta_i^{(k)}$ conformément à un modèle de mélange avec indépendance conditionnelle fondée sur une classe latente binaire ξ_i. Cependant, la probabilité conditionnelle d'appariement

$P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}}\right)$ est estimée sous l'hypothèse d'indépendance conditionnelle entre toutes les variables d'appariement.

Le deuxième scénario est choisi pour évaluer les deux estimateurs assistés par modèle, quand la distribution des résultats est spécifiée incorrectement. Il s'agit d'une situation moins favorable, mais plus réaliste. Dans ce cas, les avantages relatifs devraient être réduits comparativement à l'estimateur de Horwitz-Thompson, qui en principe doit donner les mêmes résultats que dans le premier scénario.

5.3 Résultats

Les paramètres de simulation autres que ceux déjà décrits sont présentés au tableau 5-1. Les résultats de la simulation figurent au tableau 5-2. Dans chaque cellule x-y de ce tableau, la 1re colonne de chiffres correspond à $E\left[\left(1-\hat{Z}/Z\right)^2\middle|U\right]^{1/2}$, qui est la racine carrée de l'erreur quadratique moyenne relative, pour l'estimateur de Horwitz-Thompson, et les 2e et 3e estimateurs, respectivement, pour le premier plan de sondage. Dans la même cellule x-y, la 2e colonne contient les valeurs des mêmes estimateurs sous le deuxième plan de sondage.

Tableau 5-1
Paramètres de simulation

Paramètre	***Valeur***
α	0,1
K	7
N	10 000
Nombre de pochettes	1 000
Taille de pochette	10
$P\left(\zeta_i^{(k)}=1\right)$	0,5

Paramètre	***Valeur***
$P(\xi_i=1)^*$	0,5
$P\left(\zeta_i^{(k)}=1\middle\|\xi_i=0\right)^*$	0,3
$P\left(\zeta_i^{(k)}=1\middle\|\xi_i=1\right)^*$	0,7
$P(x=1)$	0,5
$P(y=1\|x=0)$	0,4
$P(y=1\|x=1)$	0,7

Paramètre	***Valeur***
Itérations E-M	100
Taille de l'échantillon de strate x-y	1 000
Sous-strates par strate x-y	10
Nombre d'échantillons	100

* utilisé dans le 2e scénario

Dans le premier scénario, et le premier plan de sondage, le 3e estimateur est de loin supérieur au 2e estimateur, qui est lui-même supérieur à l'estimateur de Horwitz-Thompson. L'importante différence de performance entre les deux estimateurs assistés par modèle est étonnante, étant donné l'utilisation d'un modèle paramétrique correct pour les paires. Sous le deuxième plan de sondage, les trois estimateurs donnent des résultats tout aussi bons les uns que les autres et nettement meilleurs que sous le premier plan de sondage. Cette bonne performance est due au fait qu'on s'intéresse à une seule variable. Donc, il est possible d'optimiser la répartition de l'échantillon entre les sous-strates pour cette variable particulière. Il est plus difficile d'obtenir cet avantage quand on s'intéresse de manière égale à de nombreuses variables.

Sous le deuxième scénario, les résultats révèlent une tendance similaire, excepté deux différences notables. Premièrement, les estimateurs assistés par modèle donnent de moins bons résultats que dans le premier scénario, sous les mêmes plans de sondage. Deuxièmement, la dégradation de la performance est la plus importante sous le premier plan de sondage pour le 2e estimateur en utilisant l'inférence $\widehat{M}_{ij}=P\left(M_{ij}=1\middle|x_i,y_j,\boldsymbol{\gamma}_{ij};\widehat{\boldsymbol{\psi}}\right)$, où le modèle supposé n'est pas ajusté en se fondant sur l'échantillon examiné manuellement. Dans ce cas, cet estimateur donne effectivement des résultats nettement moins bons que l'estimateur de Horwitz-Thompson. Sous le deuxième plan de sondage optimisé, la dégradation de la performance est à peine percevable pour les deux estimateurs assistés par modèle. En outre, ce deuxième scénario établit clairement le 3e estimateur comme une solution pratique, parce que sa performance demeure bonne même si le modèle de mélange est spécifié incorrectement.

Dans les deux scénarios, l'estimateur de Horwitz-Thompson donne de nettement meilleurs résultats sous le deuxième plan de sondage que sous le premier. En effet, sa performance passe de lamentable à excellente entre le premier et le deuxième plans de sondage, y compris dans le deuxième scénario où le modèle sous-jacent est spécifié incorrectement. Cette différence frappante souligne encore davantage l'importance d'utiliser des variables auxiliaires qui tirent parti des résultats des comparaisons.

La performance des estimateurs assistés par modèle est assez sensible à la proportion de mélange estimée au moyen de l'algorithme EM. Contrairement aux probabilités m- et u-, ce paramètre converge assez lentement vers son estimation du maximum de vraisemblance, d'où la nécessité d'un grand nombre (100) d'itérations EM. Les études

antérieures, qui étaient axées sur les probabilités m- et u-, ont négligé cette question essentielle quand la probabilité conditionnelle d'appariement doit être estimée. C'est une autre manifestation du taux de convergence EM linéaire qui a été déploré dans des études antérieures. Le problème est résolu en utilisant une procédure numérique de Newton-Raphson, où le taux de convergence est quadratique.

Bien que l'on considère dans la présente étude un couplage un à un, cette hypothèse ne joue pas un rôle important dans la procédure d'estimation. Donc, la méthodologie proposée s'applique aussi à un couplage incomplet à condition que les examens manuels demeurent exempts d'erreur. Cependant, les estimateurs assistés par modèle résultants pourraient être moins efficaces si la distribution des enregistrements non appariables diffère considérablement de celle des autres enregistrements. Le cas échéant, les résultats des paires sont mieux modélisés par un mélange à trois composantes incluant deux classes de paires non appariées. Dans ce cas, la spécification d'un bon modèle peut être plus compliquée.

Un rapport complémentaire contient un plus grand ensemble de résultats de simulation, ainsi que des détails supplémentaires sur la méthodologie utilisée (Dasylva, 2015).

Tableau 5-2
Racine carrée de l'erreur quadratique moyenne relative (%) pour différents estimateurs et scénarios

	1er scénario (%)			
	Y=0		***Y=1***	
X=0	(91)	(5)	(96)	(7)
	(61)	(5)	(83)	(7)
	(14)	(5)	(24)	(7)
X=1	(95)	(7)	(96)	(6)
	(95)	(7)	(69)	(6)
	(32)	(7)	(20)	(6)

	2e scénario (%)			
	Y=0		***Y=1***	
X=0	(96)	(7)	(96)	(8)
	(92)	(7)	(148)	(8)
	(16)	(7)	(44)	(8)
X=1	(95)	(8)	(97)	(5)
	(147)	(7)	(94)	(5)
	(42)	(7)	(17)	(5)

6. Conclusion et travaux à venir

6.1 Importance des variables auxiliaires, des modèles et des examens manuels dans l'estimation

La présente étude décrit le problème de l'estimation fondée sur le plan de sondage à partir de fichiers administratifs avec des enregistrements couplés, dans le cadre de la méthodologie d'enquête classique. Elle propose aussi une nouvelle méthode d'estimation fondée sur des estimateurs assistés par modèle et des plans de sondage qui sont évalués au moyen de simulations. Les simulations montrent clairement l'importance égale de l'utilisation de variables auxiliaires fondée sur les variables d'appariement et d'examens manuels de haute qualité. L'efficacité des estimateurs résultants dépend aussi de la spécification de bons modèles. Cependant, l'utilisation du modèle correct n'est pas obligatoire, car, comme les estimateurs assistés par modèle antérieurs (Särndal et coll., 1992), les estimateurs proposés demeurent convergents sous le plan même quand le modèle est spécifié incorrectement.

6.2 Défis persistants

Les examens manuels peuvent poser deux problèmes, qui ont trait à la qualité de l'information complémentaire et à la qualité du processus d'examen. Il est évidemment impossible de procéder à des examens manuels valables à moins de disposer d'information complémentaire suffisante et fiable. Même lorsqu'on dispose d'une telle information, de nombreuses questions persistent quant à la qualité du processus d'examen et aux moyens de la mesurer objectivement. En effet, les études sur le sujet sont rares, outre celle de Newcombe et coll. (1983). De surcroît, les études de ce genre pourraient être difficiles à répéter, soit parce que d'importants détails méthodologiques n'ont pas été divulgués, soit parce que leurs résultats dépendent fortement des jeux de données utilisés et que ces derniers ne sont pas disponibles.

Un deuxième défi concerne l'élaboration de techniques d'anonymisation. Ces techniques empêchent les examens manuels et ont un effet indésirable sur l'efficacité du couplage. Des solutions fondées sur le couplage d'enregistrements préservant la confidentialité sont recherchées activement pour résoudre ces problèmes (Schnell et

coll., 2009). Toutefois, dans des situations où les examens manuels ont été efficaces (p. ex., avec des noms, des dates de naissance et des adresses disponibles dans les fichiers originaux), il reste encore à établir clairement que ces solutions offrent des options de rechange, qui peuvent concurrencer les examens manuels tout en préservant la confidentialité.

Un troisième défi concerne les valeurs manquantes dans les fichiers couplés. Le problème découle du fait que, comme les examens manuels sont coûteux, il est souhaitable d'éviter d'échantillonner des paires pour lesquelles les données manquent pour certaines variables d'intérêt. Ces variables pour lesquelles des données manquent représentent une forme inhabituelle de non-réponse partielle, parce qu'elle est connue avant la sélection de l'échantillon. Concevoir des solutions en vue de sélectionner un échantillon optimal représente une nouvelle piste de recherche prometteuse.

Bibliographie

BELIN, Thomas R. and RUBIN, Donald B. (1995). «A Method for calibrating false-match rates in record linkage», *Journal of the American Statistical Assocation*, vol. 90, nº 430, p. 694 à 707.

CHIPPERFIELD, James O., Glenys R. BISHOP et Paul CAMPBELL. 2011. « Estimation du maximum de vraisemblance pour les tableaux de contingence et la régression logistique en présence de données incorrectement appariées », *Techniques d'enquête*, vol. 37, nº 1, p. 17 à 36.

DASYLVA, Abel. 2015. « Estimation fondée sur le plan de sondage à partir de fichiers administratifs d'enregistrements couplés », rapport inédit, Statistique Canada.

DEMPSTER, Arthur, Nan LAIRD et Donald RUBIN. 1977. « Maximum Likelihood from Incomplete Data via the EM Algorithm », *Journal of the Royal Statistical Society: Series B*, vol. 39, nº 1, p. 1 à 38.

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.

GILL, Leicester. 2001. « Methods for Automatic Record Matching and Linkage and their Use in National Statistics », *National Statistics Methodological Series*, vol. 25.

GUIVER, Tenniel. 2011. « Sampling-Based Clerical Review Methods in Probabilistic Linking », rapport inédit, Australia Bureau of Statistics.

HEASMAN, Dick. 2014. « Sampling a matching project to establish the linking quality », *Survey Methodology Bulletin*, nº 72, p. 1 à 16, Office for National Statistics.

HOWE, G.R. 1981. « A generalized iterative record linkage computer system for use in medical follow-up studies », *Computers and Biomedical Research*, vol. 14, nº 4, p. 327 à 340.

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.

JIANG, Jiming. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

KIM, B.S. 1994. « Studies of multinomial mixture models », thèse de doctorat, University of North Carolina, Chapel Hill.

LARSEN, Michael D., et Donald B. RUBIN. 2001. « Iterated automated record linkage using mixture models », *Journal of the American Statistical Association*, vol. 96, nº 453, p. 32 à 41.

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

NEWCOMBE, Howard B. 1967. « Record linking: the design of efficient systems for linking records into individual and family histories », *Amercian Journal of Human Genetics*, vol. 19, nº 3, partie I.

NEWCOMBE, Howard B., M.E. SMITH, G.R. HOWE, J. MINGAY, A. STRUGNELL et J.D. ABBATT. 1983. « Reliability of computerized versus manual death searches in a study of the health of eldorado uranium workers », *Computers in Biology and Medecine*, vol. 13, nº 3 p. 157 à 169.

SÄRNDALL, Carl-Erik, Bengt SWENSSON et Jan WRETMAN. 1992. *Model Assisted Survey Sampling*, New-York: Springer.

SCHNELL, Rainer, Tobias BACHTELER et Jörg REIHER. 2009. « Privacy-preserving Record Linkage using Bloom Filters », *BMC Medical Informatics and Decision Making*, vol. 9, nº 41.

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.