Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Mesurer la qualité d'un couplage probabiliste par des vérifications manuelles

Abel Dasylva[1], Melanie Abeysundera, Blache Akpoué, Mohammed Haddou et Abdelnasser Saïdi

Résumé

Le couplage probabiliste est susceptible de donner des erreurs d'appariement telles que les faux positifs et les faux négatifs . Dans de nombreux cas, ces erreurs peuvent être mesurées fiablement par des vérifications manuelles, c'est-à-dire l'inspection visuelle d'un échantillon de paires d'enregistrements pour déterminer si elles sont appariées. Nous décrivons un cadre pour la bonne exécution de ces vérifications qui se fonde sur un échantillon probabiliste de paires, des vérifications indépendantes répétées de mêmes paires et une analyse de classes latentes pour tenir compte des erreurs de vérification manuelle.

Mots-clés : couplage d'enregistrements, couplage probabiliste, erreur d'appariement, vérification manuelle.

1. Introduction

En couplage probabiliste, la décision de jumeler deux enregistrements, c.-à-d. de les classer comme étant appariés ou, de façon équivalente, comme se rapportant à une même personne, est basée sur le rapport de cotes des désaccords observés entre les enregistrements (Fellegi et Sunter, 1969). Le couplage probabiliste est sujet à des erreurs qui comprennent des faux positifs et des faux négatifs. Ces erreurs se produisent s'il n'existe pas de clé de couplage unique et pour d'autres raisons, y compris des erreurs typographiques et des différences de format. Les vérifications manuelles représentent une solution viable pour la mesure de ces erreurs dans le cas du couplage de données sociales incluant des noms, des adresses et des dates de naissance, p. ex., dans les études de couverture du recensement (Byrne et coll., 2002, Dasylva et coll., 2014).

La vérification manuelle est l'inspection visuelle d'une paire d'enregistrements par une personne qui décide si les enregistrements sont appariés. Elle est également appelée *résolution manuelle* si l'objectif est d'apparier manuellement certaines paires d'enregistrements. Les vérifications manuelles remplissent de nombreuses fonctions utiles durant le déroulement d'un projet de couplage, dont la mesure des erreurs d'appariement, sur laquelle porte le présent article. Cependant, les vérifications manuelles peuvent être subjectives, sujettes à erreur et coûteuses. Elles soulèvent trois questions importantes qui n'ont pas été abordées pleinement dans la littérature existante, à savoir quel est le meilleur moyen de sélectionner un échantillon pour la vérification manuelle, comment examiner une paire sélectionnée et comment tenir compte des erreurs de vérification manuelle.

La suite de l'exposé est présentée comme il suit. La section 2 décrit le problème des erreurs d'appariement. La section 3 décrit comment échantillonner des paires pour la vérification manuelle. La section 4 décrit comment examiner les paires échantillonnées. La section 5 décrit comment estimer les taux d'erreur. La section 6 décrit comment tenir compte des erreurs de vérification manuelle. La section 7 décrit l'estimation des erreurs d'appariement pour un couplage entre la Base canadienne de données sur la mortalité (BCDM) et les enregistrements de l'Enquête sur la santé dans les collectivités canadiennes (ESCC). La section 8 sert de conclusion.

[1] Auteur correspondant, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6, Canada (abel.dasylva@canada.ca); Melanie Abeysundera, Statistique Canada; Blache Akpoué, Statistique Canada; Mohammed Haddou, Statistique Canada; Abdelnasser Saïdi, Statistique Canada.

2. Le problème des erreurs d'appariement

Le couplage probabiliste peut donner des erreurs d'appariement, telles que des faux positifs et des faux négatifs, dont les sources sont diverses. Cependant, la mesure exacte de ces erreurs est difficile.

2.1 Sources, types et effet des erreurs d'appariement

Les erreurs d'appariement comprennent les faux positifs et les faux négatifs. Un faux positif, ou mauvais appariement, se produit quand un enregistrement est lié à un enregistrement non apparié. Un faux positif est, en outre, catégorisé comme étant un appariement impossible ou incorrect. Un appariement est *impossible* quand il n'existe pas d'enregistrement correspondant dans l'autre ensemble de données. Sinon, il s'agit d'un appariement *incorrect*. Un faux négatif se produit quand un enregistrement n'est pas lié à un enregistrement apparié. On parle alors aussi d'appariement *manquant*. Le taux de faux positifs (TFP) et le taux de faux négatifs (TFN) sont deux mesures de l'erreur d'appariement. Désignons par FP, VP, FN et VN les nombres de faux positifs, de vrais positifs, de faux négatifs et de vrais négatifs, respectivement. Alors, le TFP et le TFN sont calculés respectivement comment étant $FP/(FP + VN)$ et $FN/(FN + VP)$. Les mesures d'erreur supplémentaires comprennent la sensibilité, la spécificité et la précision, qui sont définies respectivement par $1 - TFN$, $1 - TFP$ et $VP/(VP + FP)$.

2.2 Solutions existantes pour mesurer les erreurs d'appariement

Les utilisateurs de données couplées ont besoin de mesures exactes des erreurs d'appariement afin d'en tenir compte. Jusqu'à présent, les solutions proposées ont été fondées sur des modèles statistiques ou des vérifications manuelles.
En théorie, les solutions fondées sur un modèle ne nécessitent pas de vérifications manuelles. Une solution de ce genre a été proposée par Fellegi et Sunter (1969) sous des hypothèses d'indépendance conditionnelle. Quoique la solution soit entièrement automatisée et rentable, Belin et Rubin (1995) ont signalé l'inexactitude des estimations résultantes. D'autres solutions comprenant l'incorporation de *données d'apprentissage* ou *données vraies*, éventuellement au moyen de vérifications manuelles, ont été proposées, comme dans Armstrong et Mayda (1993), Thibaudeau (1993) et Belin et Rubin (1995). Larsen et Rubin (2001) ont également décrit une méthode itérative comportant de multiples cycles de vérifications manuelles qui servent de données d'apprentissage, avec une procédure d'espérance-maximisation (E-M).
Les vérifications manuelles permettent de mesurer les erreurs d'appariement par sélection d'un échantillon probabiliste de paires d'enregistrements et l'estimation des mesures erreurs fondée sur le plan de sondage. Une telle solution a été décrite par Heasman (2014) et s'applique également aux couplages déterministes. Guiver (2011) a décrit un cadre de contrôle de la qualité pour la résolution manuelle dans la zone grise, c.-à-d. entre les deux seuils (Fellegi et Sunter, 1969). Ce cadre peut être adapté facilement pour mesurer les erreurs d'appariement. Toutefois, la littérature existante est en grande partie silencieuse en ce qui concerne le processus proprement dit d'examen d'une paire échantillonnée. En effet, ni Fellegi et Sunter (1969), Newcombe (1988), Newcombe et coll. (1983), Guiver (2011) ou Heasman (2014) n'ont fourni des renseignements spécifiques. Une autre question à résoudre est celle des éventuelles erreurs de vérification manuelle qui ont été largement ignorées, à commencer par Fellegi et Sunter (1969). Pourtant, il s'agit d'un problème potentiellement grave, qui diminue la crédibilité des vérifications manuelles.

3. Comment sélectionner les paires pour la vérification?

Dans un couplage probabiliste, une paire se voit attribuer un poids d'appariement et est classée comme étant une paire rejetée, une paire possible ou une paire définitive, en fonction de deux seuils de poids. Une paire rejetée possède un poids d'appariement inférieur à un seuil inférieur, tandis qu'une paire définitive possède un poids supérieur à un seuil supérieur. Une paire possible possède un poids compris entre les deux seuils et doit être résolue manuellement selon Fellegi et Sunter (1969). Dans le cas d'une solution entièrement automatisée, les deux seuils coïncident. Idéalement, le plan d'échantillonnage des paires doit avoir une probabilité d'inclusion positive pour chaque paire possible, rejetée ou définitive. En pratique, il est nécessaire d'exclure les paires rejetées dont le poids est inférieur à un poids limite, en raison du très grand nombre de paires rejetées. La base de sondage est alors constituée des paires dont le poids est supérieur à ce poids limite.

Dans la suite de l'exposé, nous considérons une base de sondage contenant N paires d'enregistrements qui sont étiquetées $i = 1, \dots, N$. Pour la paire i, soit γ_i le vecteur observé de désaccords (également appelé vecteur des résultats de comparaison ou simplement *vecteur des résultats*), $m(\gamma_i)$ la probabilité conditionnelle du vecteur des résultats sachant que la paire est appariée, $u(\gamma_i)$ la probabilité conditionnelle du même vecteur sachant que la paire est non appariée, et $w_i = \log(m(\gamma_i)/u(\gamma_i))$ le poids d'appariement. Notons qu'avec G-LINK, (Chevrette, 2010), le poids d'appariement est plutôt calculé comme étant $10\log_2(m(\gamma_i)/u(\gamma_i))$. Définissons aussi M_i la variable *latente* indiquant le statut d'appariement de la paire i, où $M_i = 1$ si la paire est appariée et $M_i = 0$ autrement. Naturellement, ce statut d'appariement est inconnu pour une paire arbitraire, sauf quand elle est sélectionnée dans l'échantillon pour la vérification manuelle et que le processus de vérification manuelle est infaillible. Enfin, soit L_i la décision concernant l'appariement, où $L_i = 1$ signifie que la paire est appariée.

Pour une taille d'échantillon donnée *n* et une stratification en fonction du poids d'appariement, une répartition de Neyman (Lohr, 1999, p. 108) peut minimiser la variance d'un total qui est relié à une mesure d'erreur. Le plus simple et naturel total de ce genre est le nombre total de paires appariées dans la base de sondage, c.-à-d. $\sum_{i=1}^{N} M_i$. Autrement, la taille d'échantillon peut être minimisée pour une variance cible du même total. Pour une taille d'échantillon fixée et H strates données désignées par $U_1, \dots, U_H$, la répartition de Neyman requiert une estimation de la variance de M_i dans chaque strate. Pour la strate U_h avec la taille N_h, désignons cette variance par S_h^2. Pour l'estimation, supposons que les paires sont indépendantes dans chaque strate. En utilisant la formule de la variance conditionnelle, la variance de strate correspond à la somme de deux termes. Le premier terme est la moyenne de strate de la variance conditionnelle, $var(M_i|\gamma_i)$. Le deuxième terme est la variance de strate de la moyenne conditionnelle, $E[M_i|\gamma_i]$. Notons aussi que, conditionnellement à γ_i, M_i suit une loi de Bernoulli de probabilité $p_i = E[M_i|\gamma_i] = P(M_i = 1|\gamma_i)$ et de variance $var(M_i|\gamma_i) = p_i(1 - p_i)$. Donc, nous obtenons l'expression suivante.

$$\begin{aligned} S_h^2 &= \frac{1}{N_h}\sum_{i\in U_h} p_i(1-p_i) + \frac{1}{N_h - 1}\sum_{i\in U_h}(p_i - \bar{p}_h)^2 \\ \bar{p}_h &= \frac{1}{N_h}\sum_{i\in U_h} p_i \end{aligned}$$

La répartition de Neyman mène à la taille d'échantillon $n_h = (N_h S_h / \sum_{t=1}^{H} N_t S_t)n$ dans la strate U_h. La probabilité p_i est reliée au poids d'appariement quand les paires sont supposées indépendantes et identiquement distribuées (IID) conformément au mélange de distributions $\lambda m(\gamma_i) + (1 - \lambda)u(\gamma_i)$, où λ est une proportion de mélange positive, qui est éventuellement inconnue. Cette relation prend la forme logistique $\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\lambda}{1-\lambda}\right) + w_i$. Une estimation de la proportion de mélange est un sous-produit de toute procédure d'espérance-maximisation (E-M) qui est utilisée pour estimer les poids d'appariement. Cependant, elle est habituellement inconnue quand les poids d'appariement sont fixés au moyen d'une procédure itérative manuelle (Howe et Lindsay, 1981). Dans ce dernier cas, la proportion de mélange peut être estimée pour les paires et leurs poids d'appariement w_i, par maximisation d'une log-vraisemblance partielle en supposant que le poids spécifié est correct pour chaque paire dans les pochettes. Cette log-vraisemblance partielle est simplement calculée comme il suit. Supposons qu'il existe une *fonction de poids de couplage connue* $w(\gamma)$ pour chaque vecteur des résultats possible γ, et une distribution des paires non appariées $u(;\boldsymbol{\theta})$ (également appelée u-distribution pour *unmatched distribution*) paramétrisée par le vecteur $\boldsymbol{\theta}$. Par exemple, $u(;\boldsymbol{\theta})$ peut être basée sur un modèle log-linéaire avec $\boldsymbol{\theta}$ composé de termes d'interaction choisis. Les paires possèdent le mélange de distributions $p(\gamma;\boldsymbol{\psi}) = u(\gamma;\boldsymbol{\theta})\left[\lambda e^{w(\gamma)} + 1 - \lambda\right]$, où $\boldsymbol{\psi} = (\lambda, \boldsymbol{\theta})$ et λ est indépendant de $\boldsymbol{\theta}$. Donc, la log-vraisemblance s'exprime comme il suit.

$$\log L = \underbrace{\sum_i \log\left((\lambda e^{w_i} + 1 - \lambda)\right)}_{I} + \underbrace{\sum_i \log\left(u(\gamma_i;\boldsymbol{\theta})\right)}_{II}$$

La log-vraisemblance se divise en deux parties qui peuvent être maximisées indépendamment. La maximisation de la première partie (partie I) donne l'estimateur du maximum de vraisemblance de la proportion de mélange λ. Cette procédure d'estimation est non paramétrique parce que les paramètres de la u-distribution ne jouent aucun rôle. Cependant, elle s'appuie sur deux hypothèses importantes. La première hypothèse est que les paires possibles, c.-à-d. les paires qui satisfont les critères des pochettes, se comportent comme des paires IID possédant une proportion de mélange constante. En pratique, cette hypothèse peut ne pas être vérifiée, parce que la proportion de mélange peut

varier d'une pochette à l'autre. La deuxième hypothèse est que les poids de couplage spécifiés ne s'écartent pas des vrais poids de couplage. En pratique, des écarts importants sont attendus quand les poids de couplage sont modifiés manuellement. La conséquence de ces écarts peut être une proportion de mélange estimée qui se situe en dehors de l'intervalle $[0,1]$. Un tel résultat fournit aussi un diagnostic sur les poids de couplage spécifiés. Le cas échéant, une solution simple consiste à fixer manuellement la proportion de mélange à une valeur « raisonnable » comprise entre 0 et 1.

Afin d'optimiser davantage la répartition de l'échantillon, les limites de strates peuvent être optimisées en utilisant différents scénarios, y compris la règle cum-$\sqrt{f}$ de Dalenius (Dalenius et Hodges, 1959; Dalenius, 1957) basée sur les valeurs p_i distinctes.

4. Comment examiner les paires sélectionnées?

Codifier les vérifications manuelles est un exercice difficile, parce que ces vérifications sont intrinsèquement subjectives. Cependant, pour fournir leur plein potentiel, elles doivent être assujetties aux lignes directrices simples suivantes. Premièrement, un examinateur doit posséder aussi peu d'information que possible ou même n'en posséder aucune au sujet du couplage. Cette information comprend notamment les poids de couplage des paires, les vecteurs des résultats, les seuils de poids ou les statistiques agrégées au sujet des différents types de paires (définitives, rejetées, ou possibles). Deuxièmement, certaines paires, voire toutes, doivent faire l'objet de vérifications répétées par au moins trois examinateurs indépendants, chacun de ces examinateurs étant tenu de prendre une décision affirmative ou négative concernant le statut d'appariement, et les conflits étant éventuellement résolus par décision selon la règle de la majorité. Bien que ces conditions nécessitent des ressources humaines supplémentaires, elles offrent de plusieurs avantages importants. Le premier est l'amélioration de la qualité des décisions fondées sur la vérification manuelle pour les paires qui font l'objet de vérifications répétées. Le deuxième est la capacité d'évaluer les erreurs de vérification manuelle en se basant sur les conflits observés. Le troisième est l'existence d'un mécanisme intégré pour déceler les paires qui ne fournissent pas suffisamment de données pour prendre des décisions fiables, sans devoir recourir à une catégorie NE SAIS PAS. En fait, l'utilisation d'une telle catégorie est vivement déconseillée, parce que les examinateurs individuels pourraient l'utiliser hâtivement. Troisièmement, chaque examinateur doit consigner toute référence à une source externe durant le traitement d'une paire. Une source externe peut être une ressource publique en ligne, par exemple une notice nécrologique, un annuaire téléphonique ou une carte en ligne. Quand le coût pose problème, les vérifications répétées peuvent être appliquées à un sous-échantillon tiré avec soin de l'échantillon original.

5. Comment estimer les taux d'erreurs d'appariement?

Soit s l'échantillon probabiliste et π_i la probabilité d'inclusion pour l'unité i. Alors, des estimations ponctuelles pour les différentes erreurs de mesure peuvent être calculées au moyen des estimateurs par le ratio simples suivants :

$$\widehat{TFN} = \frac{\sum_{i\in s} \pi_i^{-1} M_i (1 - L_i)}{\sum_{i\in s} \pi_i^{-1} M_i}$$

$$\widehat{TFP} = \frac{\sum_{i\in s} \pi_i^{-1} (1 - M_i) L_i}{\sum_{i\in s} \pi_i^{-1} (1 - M_i)}$$

La variance et les intervalles de confiance peuvent être calculés par rééchantillonnage ou par linéarisation. Le bootstrap de Rao-Wu (Rao et Wu, 1988) convient bien à cette application, étant donné le plan d'échantillonnage à un seul degré et la fraction d'échantillonnage habituellement faible dans chaque strate.

6. Quelle est la fiabilité des vérifications manuelles?

Il est probable que des erreurs de vérification manuelle aient lieu et qu'elles causent des *conflits* quand la même paire est évaluée par de nombreux examinateurs indépendants. Parallèlement, des décisions contradictoires indiquent

clairement l'existence d'une erreur. Une stratégie simple de *résolution de conflit* est fondée sur la *règle de la majorité* en faisant appel à un nombre impair d'examinateurs. Cependant, des scénarios plus élaborés peuvent tenir compte de la performance de chaque examinateur. La fiabilité des vérifications manuelles peut être mesurée en supposant que le statut d'appariement d'une paire n'est jamais vraiment connu mais que les erreurs des examinateurs sont conditionnellement indépendantes sachant le statut d'appariement de la paire et d'autres covariables. Certaines de ces covariables peuvent être propres à l'examinateur, comme le niveau d'études ou l'expérience. L'idée des vérifications répétées remonte à Newcombe et coll. (1983). Dans les enquêtes classiques, les interviews répétées ont rempli la même fonction et ont été utilisées pour évaluer les erreurs de mesure (Biemer, 2011).

Considérons trois examinateurs indépendants pour chaque paire. Par souci de commodité, définissons $s_h = s \cap U_h$ le sous-échantillon tiré de la strate h. En outre, pour $h = 1, \dots, H$, et $j = 1,2,3$, désignons par TFN_{hj} et TFP_{hj} respectivement le TFN et le TFP pour l'examinateur j dans la strate h. Nous supposons que les examinateurs commettent des erreurs conditionnellement indépendantes sachant le statut d'appariement d'une paire et sa strate. Dans le cas le plus simple, chaque strate est un intervalle de poids. Cependant, elle peut aussi comprendre des renseignements supplémentaires propres aux paires.

Désignons par TFN_h et TFP_h respectivement le TFN et TFP pour la décision selon la règle de la majorité dans la strate h. Pour une paire échantillonnée, soit C_i la décision selon la règle de la majorité au sujet du statut d'appariement de la paire i, où $C_i = 1$ si la paire est déclarée appariée et $C_i = 0$ autrement. De façon similaire, soit C_{ij} la décision de l'examinateur j au sujet du statut d'appariement de la paire i, où $C_{ij} = 1$ si la paire est déclarée appariée et $C_{ij} = 0$ autrement. Quand le TFN et le TFP sont faibles pour tous les examinateurs dans chaque strate, et que l'on peut supposer sans risque que la décision selon la règle de la majorité est infaillible (c.-à-d. $C_i = M_i$), le TFN (défini comme étant $P(C_{ij} = 0|M_i = 1)$) et le TFP (défini comme étant $P(C_{ij} = 1|M_i = 0)$) de l'examinateur j peuvent être estimés simplement comme il suit.

$$\widehat{TFN}_{hj} = \frac{\sum_{i \in s_h} \pi_i^{-1} C_i (1 - C_{ij})}{\sum_{i \in s_h} \pi_i^{-1} C_i}$$

$$\widehat{TFP}_{hj} = \frac{\sum_{i \in s_h} \pi_i^{-1} C_{ij} (1 - C_i)}{\sum_{i \in s_h} \pi_i^{-1} (1 - C_i)}$$

La performance de l'examinateur j sur l'ensemble des strates est estimée comme il suit.

$$\widehat{TFN}_{.j} = \frac{\sum_{h=1}^{H} \left(\sum_{i \in s_h} \pi_i^{-1} C_i\right) \widehat{TFN}_{hj}}{\sum_{h=1}^{H} \sum_{i \in s_h} \pi_i^{-1} C_i}$$

$$\widehat{TFP}_{.j} = \frac{\sum_{h=1}^{H} \left[\sum_{i \in s_h} \pi_i^{-1} (1 - C_i)\right] \widehat{TFP}_{hj}}{\sum_{h=1}^{H} \sum_{i \in s_h} \pi_i^{-1} (1 - C_i)}$$

Dans le cas d'examinateurs homogènes, il n'y a pas d'*effet d'examinateur* si bien que $TFN_{hj} = TFN_{h.}$ et $TFP_{hj} = TFP_{h.}$ pour tout j. Alors, les taux communs d'erreurs des examinateurs peuvent être estimés par $\widehat{TFN}_{h.} = \left(\sum_{j=1}^{3} \widehat{TFN}_{hj}\right)/3$ et $\widehat{TFP}_{h.} = \left(\sum_{j=1}^{3} \widehat{TFP}_{hj}\right)/3$. Les TFN et TFP communs des examinateurs sur l'ensemble des strates sont estimés par $\widehat{TFN}_{..} = \left(\sum_{j=1}^{3} \widehat{TFN}_{.j}\right)/3$ et $\widehat{TFP}_{..} = \left(\sum_{j=1}^{3} \widehat{TFP}_{.j}\right)/3$.

L'infaillibilité de la décision selon la règle de la majorité est contestable dans les strates où les paires fournissent peu d'information discriminante, par exemple dans le milieu de la zone grise. De meilleures estimations peuvent être calculées en analysant les résultats des vérifications manuelles au moyen d'un modèle de classes latentes et d'une procédure E-M distincte dans chaque strate, comme l'a proposé Biemer (2011). Dans ce cas, un vecteur des résultats concernant les paires est composé des décisions affirmatives/négatives prises par les examinateurs individuels.

7. Couplage ESCC-BCDM

La méthodologie susmentionnée a été appliquée à un couplage entre les enregistrements de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et la Base canadienne de données sur la mortalité (BCDM), voir Sanmartin et coll. (2015). L'ensemble de données de l'ESCC est composé de 2,3 millions d'enregistrements pour les périodes de collecte débutant en 2000, 2003, 2005, et entre 2007 et 2011. L'ensemble de données de la BCDM est quant à lui composé de 3,6 millions d'enregistrements pour la période allant de 2000 à 2011. Les deux fichiers ont été couplés par la méthode probabiliste en se servant de G-LINK et en utilisant comme variables la date de naissance, le sexe, le nom de famille, le prénom et le code postal. L'application de critères de pochettes a produit 418 millions de paires incluant 114 000 paires définitives et 22 000 paires possibles en se servant d'un premier ensemble de seuils. Certaines des paires possibles ont été résolues manuellement. Pour toutes les paires qui n'ont pas été résolues manuellement, la décision finale concernant le couplage s'est appuyée sur un seuil unique fixé à 92. Pour la vérification manuelle, la base de sondage était composée de paires ayant un poids de couplage non inférieur à 1. Cette base a été stratifiée en huit intervalles de pondération uniformément espacés, et une taille d'échantillon de 1 000 a été répartie uniformément entre les strates. La répartition de l'échantillon est résumée au tableau 7-1.

Tableau 7-1
Plan d'échantillonnage pour les vérifications manuelles

Strate	Fréquence	Pourcentage	Intervalle de pondération	Poids d'échantillonnage	Taille de l'échantillon
1	880 515	73,58	1,51 – 23,51	7 044,12	125
2	277 757	23,21	23,52 – 49,51	2 222,06	125
3	5 447	0,46	49,52 – 73,51	43,57	125
4	2 405	0,20	73,52 – 97,51	19,24	125
5	3 274	0,27	97,52 – 123,51	26,19	125
6	2 699	0,23	123,52 – 149,51	21,59	125
7	21 198	1,77	149,52 – 163,51	169,58	125
8	3 347	0,28	163,51 – 194,52	26,77	125

Chaque paire échantillonnée a été évaluée par trois examinateurs indépendants en appliquant la décision selon la règle de la majorité. Les taux d'erreurs estimés sont donnés au tableau 7-2, y compris un TFN de 2,43 %, un TFP de 0,04 % et une précision de 98,64 %. Ces résultats donnent la preuve de la bonne qualité du couplage et sont en harmonie avec ceux d'études antérieures, voir Da Silveira et Artmann (2009).

Tableau 7-2
Taux d'erreurs d'appariement

	Couplés (L)	Non couplés (NL)	
Appariements	34 298	855,09	**TFN = 2,43 %**
Non-appariements	473,57	1 161 015	**TFP = 0,04 %**

Pr = 98,64 %

Les taux d'erreurs de vérification manuelle ont également été estimés sous l'hypothèse que la décision selon la règle de la majorité est infaillible et qu'il n'existe pas d'effet d'examinateur. Globalement, les taux estimés d'erreurs d'examinateur sont $\overline{TFN}_{..} = 2{,}97$ % et $\overline{TFP}_{..} = 0{,}15$ %. Ces résultats appuient la thèse selon laquelle les vérifications manuelles représentent une option viable pour l'estimation des erreurs d'appariement. Le plan d'échantillonnage utilisé pour sélectionner l'échantillon pour la vérification manuelle aurait pu comporter une stratification ou une répartition d'échantillon plus optimale. Néanmoins, il a été décidé de l'utiliser parce que l'échantillon résultant avait déjà été traité par un premier examinateur.

8. Conclusion

La vérification manuelle est une option viable pour mesurer les erreurs d'appariement dans le cas du couplage de données sociales en se servant du nom, de la date de naissance et de l'adresse. Cependant, des solutions fondées sur un modèle entièrement automatisé sont également nécessaires quand les vérifications manuelles sont impossibles, comme dans le cas du couplage de données anonymisées.

Bibliographie

Armstrong, J., and Mayda, J. (1993), "Model-based estimation of record linkage error rates", *Survey Methodology*, 19, pp. 137-147, 1993.

Belin, T.R. and Rubin, D.B. (1995). "A Method for calibrating false-match rates in record linkage", *JASA*, 90, pp. 694–707.

Byrne, R., Beaghen, M., and Mulry, M., (2002). "Clerical review of census duplicates", Accuracy and Coverage Evaluation (ACE) Revision II Report #PP-43, Washington:US Census Bureau.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New Jersey:John Wiley.

Chevrette A. G-Link (2010). "Constructing an Avatar". *In Proceedings of the 2010 International Methodology Symposium*, Ottawa: Statistics Canada.

Dalenius, T. (1957). *Sampling in Sweden*. Stockholm: Almquist and Wiksell.

Dalenius, T., and Hodges, J.L. (1959), "Minimum variance stratification", *JASA*, 54, pp. 88-101.

Dasylva A, Titus RC, Thibault C. (2014). "Overcoverage in the 2011 Canadian Census". *In Proceedings of the 2014 International Methodology Symposium*, 29-31 October 2014, Ottawa: Statistics Canada. Available at http://www.statcan.gc.ca/sites/default/files/media/14269-eng.pdf

Da Silveira DP, Artmann E (2009). "Accuracy of probabilistic record linkage applied to health databases: systematic review", *Rev Saude Publica*, 43, pp. 875-882.

Fellegi, I.P., and A.B. Sunter (1969). "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.

Guiver, T. (2011). "Sampling-Based Clerical Review Methods in Probabilistic Linking", unpublished report. Australia:Australian Bureau of Statistics. Available at http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.034

Heasman, D. (2014). "Sampling a matching project to establish the linking quality". *Survey Methodology Bulletin*, 72, pp. 1-16.

Howe GR, Lindsay JA (1981). "A generalized iterative record linkage computer system for use in medical follow-up studies". *Computers and Biomedical Research*, 14, pp. 327-340.

Larsen, M., and Rubin, D. (2001). "Iterated automated record linkage using mixture models", *JASA*, 96, pp. 32-41.

Lohr, S (1999). *Sampling: Design and Analysis*. New York: Duxbury.

Newcombe, H.B., Smith, M.E., and Howe, G.R. (1983). "Reliability of computerized versus manual death searches in a study of the health of eldorado uranium workers", *Computers in Biology and Medecine*, 13, pp. 157-169.

Newcombe, H.B. (1988). *Handbook of Record Linkage*. New York: Oxford University Press.

Rao, J.N.K., and Wu, C.F.J. (1988). "Resampling inference with complex survey data", *JASA*, 83, pp. 231-241.

Sanmartin C, Y Decady, R Trudeau, A Dasylva, M Tjepkema, P Fines, R Burnett, N Ross, D Manuel (2015). "Linking the Canadian Community Health Survey to the Canadian Mortality Database: A national resource to study mortality in Canada", to appear in *Health Reports*.

Thibaudeau Y (1993). "The discrimination power of dependency structures in record linkage", *Survey Methodology*, 19, pp. 31-38.