Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements

par Abel Dasylva et Arthur Goussanou

Date de diffusion : le 6 janvier 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2021
Vol. 47, nº 2, pp. 325-338
Statistique Canada, nº 12-001-X au catalogue

Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements

Abel Dasylva et Arthur Goussanou[1]

Résumé

Dans le couplage d'ensembles de données massifs, on a recours aux pochettes pour sélectionner un sous-ensemble gérable de paires d'enregistrements quitte à perdre quelques paires appariées. Cette perte tient une grande place dans l'erreur de couplage globale, parce que les décisions relatives aux pochettes se prennent tôt dans le processus sans qu'on puisse les réviser par la suite. Mesurer le rôle que joue cette perte demeure un grand défi si on considère la nécessité de modéliser toutes les paires dans le produit cartésien des sources, et non seulement celles qui répondent aux critères des pochettes. Malheureusement, les modèles antérieurs d'erreur ne nous aident guère parce qu'ils ne respectent normalement pas cette exigence. Il sera question ici d'un nouveau modèle de mélange fini, qui ne demande ni vérifications manuelles, ni données d'entraînement, ni hypothèse d'indépendance conditionnelle des variables de couplage. Il s'applique dans le cadre d'une procédure de pochettes typique dans le couplage d'un fichier avec un registre ou un recensement exhaustif lorsque ces deux sources sont exemptes d'enregistrements en double.

Mots-clés : Indexation; ensembles massifs de données; résolution d'entités; intégration des données; apprentissage automatique; classification.

1. Introduction

Le couplage d'enregistrements vise à repérer les enregistrements d'une même personne dans un ou plusieurs fichiers (Fellegi et Sunter, 1969; Christen, 2012; Statistique Canada, 2017a). Il diffère de l'appariement statistique, qui est une méthode d'imputation où on se met à la recherche d'enregistrements de personnes qui se ressemblent (D'Orazio, Di Zio et Scanu, 2006). Il est aujourd'hui une importante méthode d'intégration des données et la création des pochettes en constitue une étape d'importance. Créer des pochettes est sélectionner un sous-ensemble gérable de paires d'enregistrements, qui contient la plupart des paires appariées, c'est-à-dire des paires dont les enregistrements proviennent d'une même personne. Fellegi et Sunter (1969, section 3.4) en donnent une définition abstraite comme la sélection d'un sous-ensemble du produit cartésien de deux sources de données. Herzog, Scheuren et Winkler (2007, page 123, deuxième paragraphe) y vont d'une définition semblable : [traduction] « La création des pochettes est un mode de réduction du nombre de paires d'enregistrements à examiner. » Christen (2012, page 28, troisième paragraphe) parle plutôt d'indexation avec la même signification : [traduction] « Pour réduire le nombre sans doute très grand de paires d'enregistrements à mettre en comparaison, on recourt communément à des techniques d'indexation […] Ces techniques éliminent les paires d'enregistrements qui n'ont guère de chances de correspondre à des paires appariées. » Nous nous servons du terme « création de pochettes » ici pour désigner ce processus qui devient essentiel lors du couplage d'ensembles de données massifs qui comprennent des millions d'enregistrements. En effet, le produit cartésien est juste trop grand. Grâce aux pochettes, on trouve un juste milieu entre les ressources de calcul et de mémoire,

1. Abel Dasylva, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6. Courriel : abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Canada, K1A 0T6. Courriel : arthur.goussanou@statcan.gc.ca.

d'une part, et la perte de quelques paires appariées, d'autre part. Ces paires appariées correspondent à des faux négatifs et représentent une part importante de l'erreur de couplage globale, ne serait-ce que parce que les décisions relatives aux pochettes se prennent habituellement tôt dans le processus de couplage sans possibilité de les modifier par la suite. Il reste qu'on ne dispose là-dessus que de rares données empiriques, puisque ces faux négatifs ne sont jamais déclarés à quelques exceptions près, notamment dans le repérage des doublons dans une base de sondage comme cette opération est décrite par Herzog et coll. (2007, section 12.3). Dans ce rare cas où la base de sondage comprenait 176 000 enregistrements d'entreprises, on a estimé le nombre de paires appariées à 3 219, dont 3 050 avaient été détectées par les critères de pochettes pour une proportion de faux négatifs de $(3\,219 - 3\,050)/3\,219 = 5{,}25\,\%$, phénomène non négligeable si on le compare aux proportions de faux négatifs déclarées dans diverses études de couplage recensées par Bohensky (2016). De nos jours, il est tentant de minimiser les faux négatifs dûs aux pochettes en relâchant les critères de pochettes autant que le permettent les ressources de calcul. Après tout, ces ressources sont déjà considérables et croissent sans cesse en cette ère de mégadonnées. Une conséquence peu souhaitable peut en être que les paramètres de couplage probabiliste deviennent impossibles à estimer parce que la proportion de paires appariées est trop faible (Winkler, 2016, section 2.2.3.2). Ainsi, la question des faux négatifs dûs aux pochettes continue à se poser, quelles que soient les ressources de calcul disponibles. Leur estimation a toutefois présenté tout un défi, parce qu'on se doit de tenir compte de toutes les paires du produit cartésien des deux sources, et non des seules paires répondant aux critères des pochettes. À cet égard, la plupart des modèles antérieurs d'erreur n'aident guère, puisqu'ils ne respectent pas cette exigence; voir notamment à ce sujet Fellegi et Sunter (1969), Armstrong et Mayda (1993), Thibaudeau (1993), Winkler (1993), Belin et Rubin (1995), Sariyar, Borg et Pommerening (2011), Daggy, Xu, Hui, Gamache et Grannis (2013) et Chipperfield, Hansen et Rossiter (2018). Herzog et coll. (2007, chapitre 12.5) ont décrit une technique de capture-recapture qui n'a pas cet inconvénient, mais qui demeure peu pratique parce qu'exigeant des vérifications manuelles et l'indépendance conditionnelle pour un certain nombre de variables de pochettes.

Nous décrirons une nouvelle solution où ces deux exigences disparaissent. Il s'agit d'étendre le modèle de Blakely et Salmond (2002) aux situations où les enregistrements sont hétérogènes et où la population finie sous-jacente est nombreuse. Cette solution est d'abord conçue dans le cas idéal où deux sources exemptes de doublons sont liées, soit un fichier avec un registre ou un recensement exhaustif, de sorte que la décision de conserver une paire dans les pochettes dépende uniquement de ses deux enregistrements constitutifs comme pour les procédures typiques de création de pochettes (voir Christen, 2012, chapitre 4.1). Elle est intéressante dans un cadre pratique où les deux sources ont peu de doublons et où le recensement est quasiment exhaustif. Des exemples en sont le couplage entre les données fiscales et le recensement canadien pour le remplacement de questions sur le revenu (Statistique Canada, 2017b), ou encore une étude par cohorte avec couplage entre des données de mortalité et un recensement (Blakely et Salmond, 2002).

Notre exposé est structuré de la manière suivante : la section 2 présente les hypothèses, la notation et la terminologie; la section 3 indique en quoi la distribution des voisins livre une importante information sur les erreurs; les sections 4, 5 et 6 décrivent respectivement le modèle proposé de mélange fini, la procédure d'espérance-maximisation et l'étude empirique; enfin, la section 7 expose les conclusions et les travaux futurs.

2. Définitions, notation et hypothèses

Enregistrements appariés : Dans un couplage d'enregistrements comme pour d'autres problèmes de classification en mode automatique, une nette distinction doit s'établir entre la nature des entités à classer (deux enregistrements sont-ils réellement de la même entité ?) et les décisions à prendre (les enregistrements sont-ils *réputés* appartenir à la même entité ?) d'après les observations qui se font sur ces entités (quel est le degré d'accord entre les enregistrements ?). On ne s'entend toutefois pas sur les termes à employer pour désigner ces concepts clés, car le couplage d'enregistrements est une activité multidisciplinaire au croisement de la statistique, de l'épidémiologie et de l'informatique. Au premier paragraphe de leur résumé, Fellegi et Sunter (1969) écrivent en ce sens : [traduction] « Un modèle mathématique vise à créer un cadre théorique pour une solution informatique au problème consistant à reconnaître les enregistrements de deux fichiers qui représentent les mêmes personnes, objets ou événements (et qui sont donc dits « appariés »). » Le point est donc de savoir si deux enregistrements sont de la même entité. Dans leur livre (2007, page 83, dernier paragraphe), Herzog, Scheuren et Winkler utilisent le terme « appariement réel » pour le même concept. Dans la documentation spécialisée en informatique, le terme « appariement » a cependant un sens tout à fait différent. Il vise la décision de classification, dont Christen (2012) donne le meilleur exemple dans son livre intitulé « Data matching ». Dans son ouvrage, Newcombe (1988, page 105, deuxième paragraphe) déplore aussi le manque de consensus sur le sens à donner au terme « appariement » lorsqu'il écrit : [traduction] « Ce terme est utilisé diversement dans la documentation sur le couplage d'enregistrements. Ici toutefois, il n'a aucun sens technique particulier, désignant seulement un appariement d'enregistrements en fonction d'une certaine similitude (ou dissimilitude) indiquée. »

Dans les lignes qui suivent, le sens du terme sera celui que définissent Fellegi et Sunter (1969) pour les enregistrements d'une même entité (personne, entreprise, ménage, etc.). Il s'emploie aussi à propos d'une paire lorsque les enregistrements constitutifs sont appariés. Deux enregistrements sont dits *non appariés* s'ils proviennent d'entités différentes.

Population finie et sources de données : Dans l'examen de ce problème, considérons une grande population finie de N individus avec un processus d'enregistrement tel que les enregistrements d'individus différents sont indépendants les uns des autres, tout comme les erreurs d'enregistrement. Soit m la taille du fichier que l'on suppose être une variable aléatoire avec $m \leq N$ et $m \to \infty$ lorsque $N \to \infty$

($m = O(N)$), par exemple). Soit V l'ensemble de valeurs possibles d'enregistrements dans l'une ou l'autre des sources de données, et soit v_i l'enregistrement i du fichier, où $v_i \in V$ par définition. Pour simplifier, posons que V est fini même s'il est habituellement très grand. Simplifions davantage en posant que les deux sources de données sont réellement exemptes de doublons et que le registre est exhaustif. En d'autres termes, chaque enregistrement du fichier correspond à un seul enregistrement du même individu dans le registre. Chaque enregistrement est enfin jugé complet, c'est-à-dire sans valeurs manquantes.

Stratégies de pochettes : Dans le couplage de deux grandes sources de données, la création de pochettes permet d'éliminer la vaste majorité des paires d'enregistrements d'individus différents, tout en conservant les autres paires et en économisant de précieuses ressources de calcul. Toutefois, certaines paires d'enregistrements d'un même individu se perdent inévitablement en cours de processus. Christen (2012, chapitre 4.4) a passé en revue un éventail de procédures de création de pochettes, dont la stratégie la plus simple consistant à sélectionner une paire si les enregistrements s'accordent parfaitement selon une seule clé. C'est une procédure qui est souvent posée au départ dans les études publiées sur l'analyse de données couplées (Chambers et Kim, 2016; Han et Lahiri, 2018). On sélectionne ainsi un sous-ensemble de paires en fonction d'une union de produits cartésiens dans des poststrates disjointes appelées blocs. Dans la pratique, on peut affiner cette méthode en conservant une paire si les enregistrements s'accordent parfaitement pour au moins une clé parmi plusieurs. Dans ce cas, le sous-ensemble de paires sélectionnées n'est plus l'union de produits cartésiens dans des poststrates disjointes. Dans ce qui suit, nous ne regarderons pas de tels détails, mais verrons quelle est notre capacité d'estimer fidèlement la perte attribuable à la procédure de création des pochettes dans un couplage entre un fichier et un registre ou recensement lorsque ces deux sources ont peu de doublons et que le registre ou le recensement est à peu près exhaustif. Le couplage de données fiscales avec le recensement canadien (Statistique Canada, 2017b) et une étude par cohorte dans un couplage entre des données de mortalité et un recensement (Blakely et Salmond, 2002) en sont de parfaits exemples.

Dans la suite de cet exposé, nous posons que la décision de conserver une paire dépend seulement des enregistrements constitutifs, c'est-à-dire que la décision de pochettes est l'équivalent d'une application mathématique de $V \times V$ dans $\{0,1\}$. On a là une grande catégorie de procédures de pochettes comprenant les procédures typiques dans ce domaine (Christen, 2012, section 4.4). On se trouve néanmoins à exclure les stratégies de pochettes comportant une certaine forme de partitionnement, comme le mode dit de « partitionnement de canopée » (Christen, 2012, section 4.8).

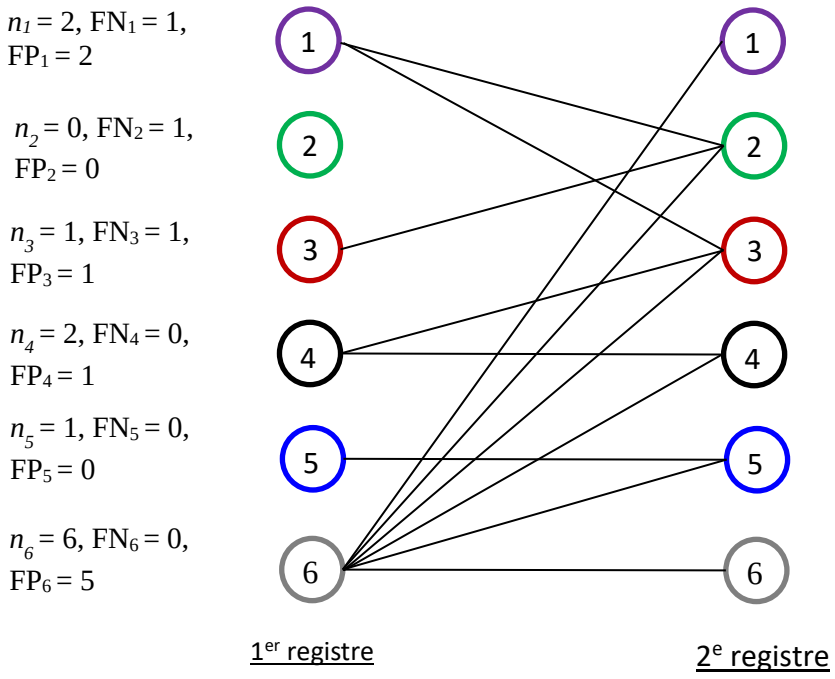
Erreurs : Dans l'application de critères de pochettes, deux types d'erreurs sont possibles sous forme de *faux négatifs* et de *faux positifs*. Il y a faux négatif si une paire appariée est rejetée par les critères applicables. Il y a faux positif si une paire non appariée est acceptée par ces mêmes critères. Ces erreurs se mesurent par le *taux de faux négatifs* (TFN) et le *taux de faux positifs* (TFP), où le premier est la proportion des paires appariées qui sont rejetées et, le second est la proportion des paires non appariées qui sont acceptées.

Lorsqu'on conçoit des critères de pochettes, on peut réduire au minimum le taux de faux positifs, tout en gardant le taux de faux négatifs sous une valeur seuil (1 %, par exemple). Comme il y a habituellement beaucoup plus de paires non appariées qu'appariées dans les pochettes, l'idée est en gros de minimiser le nombre de paires dans les pochettes, tout en gardant le taux de perte de paires appariées sous le seuil établi. Bien sûr, la mise en œuvre d'une telle stratégie exige une estimation fidèle des deux taux d'erreur en question. Le taux de faux positifs est souvent bien plus facile à estimer que le taux de faux négatifs. Soit B le nombre total de paires acceptées par les critères de pochettes. Comme le taux de faux positifs n'est pas inférieur à $(B-m)/m(N-1)$ ni supérieur à $B/m(N-1),$ il est bien approché par B/mN si $B>>m.$ Cet estimateur est lié au *rapport de réduction* défini comme $1-B/mN$ (Christen, 2012, chapitre 7.3). Il est bien plus difficile d'estimer les faux négatifs. Heureusement, le concept de *voisin* jette un bon éclairage sur la question.

3. Voisins et erreurs

Lorsqu'on examine les erreurs possibles, on a intérêt à regarder combien d'enregistrements d'un registre forment une paire acceptée (par les critères de pochettes) avec un enregistrement donné du fichier. Dans ce qui suit, ces enregistrements sont appelés *voisins* de l'enregistrement du fichier et leur nombre est désigné par n_i pour l'enregistrement i du fichier. La distribution empirique de n_i nous renseigne largement sur les erreurs parce que, dans le cadre établi, chaque enregistrement du fichier aurait exactement un voisin si la stratégie de pochettes était exempte d'erreurs, c'est-à-dire sans faux négatifs ni faux positifs. À noter que le respect de cette condition n'implique pas l'absence d'erreurs. Considérons, par exemple, la situation illustrée à la figure 3.1 où les deux sources sont des registres d'une population de $N=6$ individus et où l'individu i a l'enregistrement i dans chacun des registres, c'est-à-dire que la paire d'enregistrements (i,i) est appariée pour $i=1,\ldots,6.$ Avant même de regarder les $n_i,$ nous savons que le nombre de faux négatifs est 0 ou 1, et le nombre de faux positifs, de 0 à 5 pour chaque enregistrement dans le premier registre. Toutefois, les n_i nous éclairent davantage sur les erreurs. Avec $n_i=0$ (avec l'enregistrement 2, par exemple), nous savons avec certitude qu'il existe un faux négatif, mais aucun faux positif. Avec $n_i=1,$ deux cas sont possibles : dans le premier (comme avec l'enregistrement 5), le voisin est l'enregistrement apparié sans erreurs et, dans le second (comme avec l'enregistrement 3), le voisin est un enregistrement non apparié avec deux erreurs, soit un faux négatif et un faux positif. Bref, lorsque $n_i=1,$ le nombre de faux négatifs est 0 ou 1, et il en va de même du nombre de faux positifs. Rien ne nous est connu d'autre sur les faux négatifs; nous savions en effet que leur nombre était 0 ou 1 avant de regarder $n_i.$ En revanche, nous en savons beaucoup plus sur les faux positifs, puisque nous savions qu'ils se trouvaient dans un intervalle plus étendu (0 à 5) avant de regarder $n_i.$ Cette observation confirme la facilité relative avec laquelle les faux positifs peuvent être estimés.

Figure 3.1 Deux registres avec six individus.



Le tableau 3.1 résume le lien général entre le nombre de voisins et les erreurs de couplage pour un certain enregistrement dans le cadre présent où chaque enregistrement du fichier est apparié avec exactement un seul enregistrement du registre.

Tableau 3.1
Voisins et erreurs

Voisins (n_i)	Faux négatifs	Faux positifs	Information complète sur les erreurs (oui/non)
0	1	0	Oui
1	0 ou 1	0 ou 1	Non
$[2, N-1]$	0 ou 1	$n_i - 1$ ou n_i	Non
N	0	$N-1$	Oui

Le tableau qui précède démontre clairement que le nombre de voisins nous renseigne beaucoup sur les erreurs, notamment dans le cas $n_i = 0$ et dans le cas plus improbable $n_i = N$, où l'information est complète. Là où la décision de pochettes concernant deux enregistrements ne dépend d'aucun autre enregistrement, avec une forte probabilité lorsque m et N sont élevés, un certain nombre d'enregistrements du fichier seront inévitablement sans voisin. En théorie, nous pourrions concevoir les critères de pochettes de sorte que n_i soit positif pour chaque enregistrement du fichier, mais ce serait aller

à l'encontre de l'hypothèse formulée à la section 2 d'absence de dépendance de la décision de pochettes concernant deux enregistrements à l'égard de tout autre enregistrement. Selon cette hypothèse, le nombre de voisins nous renseigne utilement sur les erreurs, mais une certaine incertitude subsiste dans le cas $1 \le n_i \le N-1$ où un modèle statistique est nécessaire à la prévision des erreurs en fonction de n_i.

4. Modèle de mélange fini

Le couplage de deux sources est intéressant s'il s'agit d'une possibilité viable, même lorsque N est très grand. Pour traduire l'essentiel dans de telles situations, nous posons les deux conditions suivantes en matière de régularité :

a. deux enregistrements appariés sont voisins avec une probabilité bornée loin de 0, quel que soit N;
b. deux enregistrements non appariés sont des voisins *accidentels* avec une probabilité de $O(1/N)$.

Ces hypothèses impliquent que chaque enregistrement a un nombre espéré de voisins qui est borné et que $O(N)$ paires (plutôt que $O(mN)$ paires et même $O(N^2)$ paires si $m = O(N)$) sont sélectionnées par les critères de pochettes. Une autre implication est que les variables de couplage donnent assez d'information pour permettre de reconnaître les enregistrements appariés avec une probabilité de succès qui est bornée loin de zéro indépendamment de la taille de la population. Une dernière implication avec ces hypothèses est l'existence d'une distribution asymptotique particulière pour le nombre de voisins n_i. Soit $n_i = n_{i|M} + n_{i|U}$, où $n_{i|M}$ est le nombre de voisins appariés et $n_{i|U}$ le nombre de voisins non appariés. À noter que ces dernières variables ne sont pas directement observées sauf si $n_i = 0$ ou $n_i = N$ (voir le tableau 3.1). Elles sont conditionnellement indépendantes étant donné v_i de sorte que $n_{i|M} \,|\, v_i \sim \text{Bernoulli}(p(v_i))$, $n_{i|U} \,|\, v_i \sim \text{Binomial}(N-1, \lambda(v_i)/(N-1))$, si un enregistrement non apparié est un voisin avec la probabilité $\lambda(v_i)/(N-1)$ indépendamment des autres enregistrements non appariés. Là où les fonctions $p(.)$ et $\lambda(.)$ ne dépendent pas de N et où N est élevé, nous avons $n_{i|U} \,|\, v_i \dot{\sim} \text{Poisson}(\lambda(v_i))$ (Billingsley, 1995), où $\dot{\sim}$ signifie « approximativement distribué comme ». Dans ce cas, $n_i \,|\, v_i \dot{\sim} \text{Bernoulli}(p(v_i)) * \text{Poisson}(\lambda(v_i))$, où $*$ est l'opérateur de convolution. À noter qu'en général, les fonctions $p(.)$ et $\lambda(.)$ sont des paramètres inconnus de haute dimension. Pour simplifier, posons également que $(p(.), \lambda(.))$ est représenté (bien approché) par une fonction constante par morceaux avec G niveaux, ce qui nous donne le modèle demélange fini $n_i \sim \sum_{g=1}^{G} \alpha_g (\text{Bernoulli}(p_g) * \text{Poisson}(\lambda_g))$ qui se vérifie approximativement. Avec G fixe, les paramètres inconnus du modèle sont donnés par le vecteur $\psi = \left[(\alpha_g, p_g, \lambda_g)\right]_{1 \le g \le G}$ qui peut être estimé par la procédure d'espérance-maximisation (EM) à la prochaine section.

Le lien entre les taux d'erreur et les paramètres du modèle s'établit si on note d'abord que les définitions du TFN et du TFP impliquent ce qui suit :

$$\begin{aligned} \text{TFN} &= \frac{1}{m}\sum_{i=1}^{m}(1-n_{i|M}), \\ (N-1)\,\text{TFP} &= \frac{1}{m}\sum_{i=1}^{m} n_{i|U}. \end{aligned} \tag{4.1}$$

Si $m = N$ presque sûrement, les équations qui précèdent impliquent que

$$\begin{aligned} E[\text{TFN}] &= 1-E[n_{i|M}], \\ &= 1-E[p(v_i)], \\ (N-1)\,E[\text{TFP}] &= E[n_{i|U}] \\ &= E[\lambda(v_i)], \end{aligned} \tag{4.2}$$

où $E[p(v_i)] = \sum_{g=1}^{G}\alpha_g p_g$ et $E[\lambda(v_i)] = \sum_{g=1}^{G}\alpha_g\lambda_g$ avec le modèle de mélange fini. Si m est aléatoire de sorte que

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m} n_{i|M} &\xrightarrow{p} E[n_{i|M}], \\ \frac{1}{m}\sum_{i=1}^{m} n_{i|U} &\xrightarrow{p} E[n_{i|U}], \end{aligned} \tag{4.3}$$

et si $N \to \infty$, les taux d'erreur et les paramètres du modèle sont liés de la manière suivante :

$$\begin{aligned} \text{TFN} &\xrightarrow{p} 1-E[p(v_i)], \\ (N-1)\,\text{TFP} &\xrightarrow{p} \quad E[\lambda(v_i)]. \end{aligned} \tag{4.4}$$

5. Procédure d'estimation

Nous pouvons estimer les paramètres du modèle en maximisant la vraisemblance *composite* (Varin, Reid et Firth, 2011) de l'échantillon $n_1, \ldots, n_m$. Dans ce qui suit, nous emploierons le terme « vraisemblance » tout court pour désigner la vraisemblance composite. Dans l'application de la procédure EM (Dempster, Laird et Rubin, 1977), il est commode de calculer d'abord les équations de maximum de vraisemblance pour les *données complètes* de la manière suivante : $n_{i|M}$, $n_{i|U}$ et $(c_{i1}, \ldots, c_{iG})$ pour chaque i; c_{ig} est l'indicateur d'appartenance de l'enregistrement i à la classe g.

Avec un peu d'algèbre, voici les équations de maximum de vraisemblance pour les données complètes :

$$\begin{aligned} \hat{p}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|M}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} c_{ig} n_{i|U}}{\sum_{i=1}^{m} c_{ig}} \\ \hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} c_{ig}. \end{aligned} \tag{5.1}$$

Par conséquent, les équations applicables aux données observées (les n_i) sont les suivantes :

$$\begin{aligned}\hat{p}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|M} \mid n_i;\psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi]} \\ \hat{\lambda}_g &= \frac{\sum_{i=1}^{m} E[c_{ig} n_{i|U} \mid n_i;\psi]}{\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi]} \\ \hat{\alpha}_g &= \frac{1}{m}\sum_{i=1}^{m} E[c_{ig} \mid n_i;\psi].\end{aligned} \tag{5.2}$$

La procédure EM alterne entre l'étape M à l'équation (5.2) et les équations de l'étape E à l'annexe A.

La procédure en question peut produire des estimateurs ponctuels convergents même si elle traite l'échantillon $n_1, \ldots, n_m$ comme s'il était indépendant et identiquement distribué. Il peut cependant en résulter un certain biais lors de l'estimation de la variance et des valeurs critiques de test d'hypothèse.

6. Étude empirique

L'étude empirique est fondée sur les données de dotation du Système de ressourcement de la fonction publique (SRFP) au service des demandeurs d'emploi dans la fonction publique fédérale au Canada. Un utilisateur peut ouvrir autant de comptes et solliciter autant d'emplois qu'il le désire avec le même compte; chaque compte est lié à une adresse électronique distincte. Dans l'exécution de son mandat, la Commission de la fonction publique doit recenser tous les comptes d'un même demandeur. La chose est difficile, parce qu'il n'y a pas d'identificateur unique, sauf pour une minorité de demandeurs. Pour la plupart des enregistrements, le couplage doit plutôt se faire avec le prénom, le nom et une date partielle de naissance qui peuvent être tirés de tous les enregistrements. La date de naissance partielle comprend le jour et le mois et le dernier chiffre de l'année de naissance.

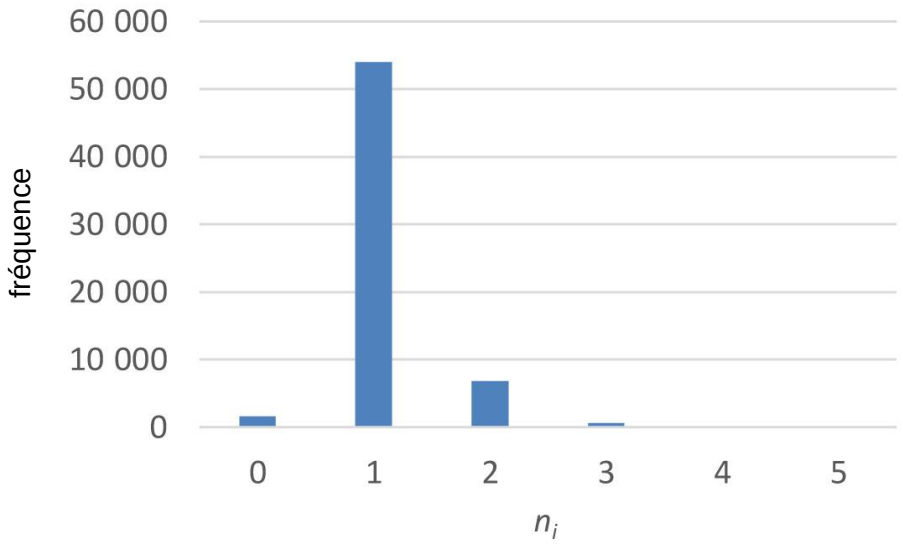
L'étude empirique fait appel à un sous-ensemble de 126 330 enregistrements prélevés sur les données SRFP depuis 2006. Les critères de sélection sont les suivants :

- identificateur unique non manquant;
- prénom, nom et date de naissance partielle non manquants;
- deux enregistrements pour chaque valeur sélectionnée de l'identificateur unique.

Les enregistrements sélectionnés correspondent à 63 155 valeurs distinctes de l'identificateur et à autant d'individus à raison de deux enregistrements appariés par individu. Ces enregistrements sont répartis entre deux registres complets et exempts de doublons qui sont liés et soumis aux critères de pochettes qui suivent, et ce, sans l'identificateur unique. Une paire est sélectionnée si la date de naissance partielle est la même, tout comme le code SOUNDEX (Herzog et coll., 2007, chapitre 11) pour le prénom ou le nom. Les taux d'erreur espérés s'estiment avec le modèle et se comparent aux valeurs réelles basées sur l'identificateur unique.

L'histogramme de la figure 6.1 indique que, dans leur vaste majorité, les enregistrements ont exactement un seul voisin. Néanmoins, 1 659 enregistrements sont sans voisin et cinq en ont cinq, soit le maximum de voisins pour tout enregistrement.

Figure 6.1 Histogramme du nombre de voisins.



Le tableau 6.1 croise les enregistrements selon le nombre de voisins et les erreurs de couplage conformément au tableau 3.1.

Tableau 6.1
Nombre de voisins et d'erreurs

Voisins (n_i)	Faux négatifs	Faux positifs	Fréquence
0	1	0	1 659
1	1	1	116
1	0	0	53 835
2	1	2	8
2	0	1	6 867
3	1	3	1
3	0	2	602
4	0	3	62
5	0	4	5

La matrice de confusion est la suivante :

Tableau 6.2
Matrice de confusion

	Lien	Non-lien	Total
Paires appariées	61 371	1 784	63 155
Paires non appariées	8 412	3,99E9	3,99E9
Total	69 783	3,988E9	3,989E9

Dans cette matrice, $\text{TFN} = 1\,784 / 63155 = 0{,}0282$ et $\text{TFP} = 8\,412 / 3{,}99\text{E}9 = 2{,}11\text{E} - 6$. Les deux mesures peuvent être considérées comme les estimateurs $\hat{E}[\text{TFN}]$ et $\hat{E}[\text{TFP}]$ des espérances correspondantes respectives. Comme le taux de faux négatifs est la sommation de variables aléatoires indépendantes et identiquement distribuées, sa variance peut s'estimer par

$$\hat{\text{var}}(\text{TFN}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (1 - n_{i|M} - \text{TFN})^2,$$

en fonction des variables latentes $n_{1|M}, \ldots, n_{m|M}$, qui ne sont pas directement observées dans la pratique. Ainsi, la variance estimée du TFN est $\hat{\text{var}}(\text{TFN}) = 4{,}35\text{E} - 7$. L'erreur-type estimée est donc $\hat{\text{SE}}(\hat{E}[\text{TFN}]) = 6{,}6\text{E} - 4$ pour l'estimateur $\hat{E}[\text{TFN}]$ et l'intervalle normal de confiance à 95 % est $\hat{E}[\text{TFN}] \mp z_{\alpha/2} \hat{\text{SE}}(\hat{E}[\text{TFN}]) = (2{,}82\text{E} - 2 \mp 1{,}3\text{E} - 3)$ pour le TFN espéré, où $\alpha = 0{,}05$ et $z_{\alpha/2} = 1{,}96$. L'intervalle de confiance à 99 % correspondant est $(2{,}82\text{E} - 2 \mp 1{,}71\text{E} - 3)$. Il est plus difficile d'estimer la variance du TFP, puisque ce taux fait intervenir une statistique U du second ordre (Hoeffding, 1948; Lee, 1990). En fait, le tableau 6.1 ne donne pas suffisamment d'information pour l'estimation de cette statistique. Il est tout autant difficile d'estimer la variance des estimateurs par modèle parce que les n_i sont corrélés. Toutes les estimations ponctuelles figurent au tableau 6.3 où la première ligne donne les valeurs réelles du TFN et du TFP.

Tableau 6.3
Estimations ponctuelles

		$\hat{E}$[TFN]	$\hat{E}$[TFP]
Identificateur unique		0,0282	2,11E-6
Modèle	G = 1	0,0301	2,14E-6
	G = 2	0,0298	2,13E-6
	G = 3	0,0303	2,14E-6

Les résultats indiquent que les estimations par le modèle sont très proches des valeurs réelles du TFN et du TFP si on emploie une, deux ou trois classes. Pour le taux de faux négatifs, l'erreur relative est $100 \times |0{,}0303 - 0{,}0282| \,/\, 0{,}0282 = 7{,}45\,\%$, et, pour le taux de faux positifs, $100 \times |2{,}11 - 2{,}14| \,/\, 2{,}11 = 1{,}42\,\%$. La faiblesse des erreurs relatives est encourageante pour ce qui est de la précision des estimateurs proposés, bien que les estimations par le modèle du TFN espéré se situent quelque peu en dehors de l'intervalle de confiance à 95 %. Il reste que l'estimation se trouve bel et bien dans l'intervalle de confiance à 99 % lorsque deux classes sont utilisées. Le choix de deux classes paraît optimal, puisque l'estimation résultante accuse l'erreur relative la plus basse par rapport à la valeur réelle du TFN.

7. Conclusions et travaux futurs

Nous avons proposé un nouveau modèle de mélange fini pour l'estimation des faux négatifs dans une procédure de création de pochettes typique lors du couplage entre un fichier et un registre ou un recensement exhaustif et là où les deux sources sont exemptes de doublons. Une étude empirique avec des données sociales donne des résultats encourageants. Il faudra cependant poursuivre les travaux pour étudier les questions d'estimation de variance et d'inférence statistique sur le nombre de classes. Il faudra également pousser le traitement du sous-dénombrement et des doublons.

Avertissement

Cet article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada. Il décrit des méthodes théoriques qui pourraient ne pas correspondre à celles qu'emploie actuellement l'organisme.

Remerciements

Les auteurs expriment leur gratitude envers le Dr. Jonnagada Rao pour son éclairage et envers la Commission de la fonction publique du Canada pour l'accès aux données.

Annexe A

Pour l'étape E, les équations sont les suivantes :

$$
\begin{aligned}
P(n_i \mid c_{ig}=1) &= I(n_i=0)(1-p_g)\,e^{-\lambda_g} + I(n_i>0)\left(p_g+(1-p_g)\frac{\lambda_g}{n_i}\right)\frac{e^{-\lambda_g}\lambda_g^{n_i-1}}{(n_i-1)!}\\
P(c_{ig}=1 \mid n_i) &= \frac{\alpha_g P(n_i \mid c_{ig}=1)}{\sum_{g'=1}^{G}\alpha_{g'}P(n_i \mid c_{ig'}=1)}\\
P(n_{i\mid M}=1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i+(1-p_g)\lambda_g}\\
P(n_{i\mid U}=n_i \mid n_i, c_{ig}=1) &= I(n_i=0)+I(n_i>0)\frac{(1-p_g)\lambda_g}{p_g n_i+(1-p_g)\lambda_g}\\
P(n_{i\mid U}=n_i-1 \mid n_i, c_{ig}=1) &= \frac{p_g n_i}{p_g n_i+(1-p_g)\lambda_g}
\end{aligned}
$$

et

$$
\begin{aligned}
E\left[c_{ig} n_{i|M} \mid n_i\right] &= P\left(c_{ig}=1 \mid n_i\right) P\left(n_{i|M}=1 \mid n_i, c_{ig}=1\right) \\
E\left[c_{ig} n_{i|U} \mid n_i\right] &= P\left(c_{ig}=1 \mid n_i\right) E\left[n_{i|U} \mid n_i, c_{ig}=1\right] \\
E\left[n_{i|U} \mid n_i, c_{ig}=1\right] &= \left(\frac{p_g\left(n_i-1\right)+\left(1-p_g\right)\lambda_g}{p_g n_i+\left(1-p_g\right)\lambda_g}\right) n_i .
\end{aligned}
$$

Bibliographie

Armstrong, M., et Mayda, J. (1993). Estimation modéliste des taux d'erreur liés au couplage d'enregistrements. *Techniques d'enquête*, 19, 2, 147-158. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14459-fra.pdf.

Belin, T., et Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Billingsley, P. (1995). *Probability and Measure*, Wiley.

Blakely, T., et Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predicted value. *Journal of Epidemiology*, 31, 1246-1252.

Bohensky, M. (2016). Bias in data linkage studies. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 63-82.

Chambers, R., et Kim, G. (2016). Secondary analysis of linked data. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 83-108.

Chipperfield, J., Hansen, N. et Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *Revue Internationale de Statistique*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*, Springer.

Daggy, J., Xu, H., Hui, S., Gamache, R. et Grannis, S. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

Dempster, A., Laird, N. et Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39, 1-38.

D'Orazio, M., Di Zio, M. et Scanu, M. (2006). *Statistical Matching, Theory and Practice*, Wiley.

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

Han, Y., et Lahiri, P. (2018). Statistical analysis with linked data. *Revue Internationale de Statistique*, 87, numéro spécial, 139-157, 1C19 doi:10.1111/insr.12295.

Herzog, T., Scheuren, F. et Winkler, W. (2007). *Data Quality and Record Linkage Techniques*, Springer.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19, 293-325.

Lee, A.J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker.

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Sariyar, M., Borg, A. et Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Statistique Canada (2017a). *Modèle du processus d'un projet de couplage d'enregistrements*, nº 12-605-X au catalogue.

Statistique Canada (2017b). *Guide de référence sur le revenu, Recensement de la population, 2016*, nº 98-500-X2016004 au catalogue.

Thibaudeau, Y. (1993). Le pouvoir discriminant des strructures de dépendance dans le couplage d'enregistrements. *Techniques d'enquête*, 19, 1, 35-43. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993001/article/14477-fra.pdf.

Varin, C., Reid, N. et Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-42.

Winkler, W. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. Dans *Proceedings of the 1993 Joint Statistical Meetings*, American Statistical Association, 274-279.

Winkler, W.E. (2016). Probabilistic linkage. Dans *Methodological Developments in Data Linkage*, (Éds., K. Harron, H. Goldstein et C. Dibben), Chichester: Wiley, 7-35.