Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2022 de Statistique Canada : Désagrégation des données : dresser un portrait de données plus représentatif de la société

Proposition pour le problème de l'estimation des probabilités d'appariement dans le couplage d'enregistrements

par Mauro Scanu, Tiziana Tuoto, Marco Fortini et Sara Piombo

Date de diffusion : le 25 mars 2024

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2022 de Statistique Canada
Désagrégation des données : Dresser un portrait de données plus représentatif de la société

Proposition pour le problème de l'estimation des probabilités d'appariement dans le couplage d'enregistrements

Mauro Scanu, Tiziana Tuoto, Marco Fortini et Sara Piombo[1]

Résumé

Le couplage d'enregistrements vise à mettre en évidence les paires d'enregistrements liées à la même unité et observées dans deux ensembles de données différents, disons A et B. Fellegi et Sunter (1969) proposent de mettre à l'essai chaque paire d'enregistrements, qu'elle soit générée à partir de l'ensemble de paires appariées ou non. La fonction de décision est le rapport entre $m(\gamma)$ et $u(\gamma)$, les probabilités d'observer une comparaison γ d'un ensemble de $k>3$ variables d'identification clés dans une paire d'enregistrements, sous l'hypothèse que la paire constitue, respectivement, un appariement ou non. On estime habituellement ces paramètres au moyen de l'algorithme EM en utilisant comme données les comparaisons pour toutes les paires du produit cartésien $\Omega=A\times B$. On émet l'hypothèse que ces observations (sur les comparaisons et sur l'état des paires comme appariement ou non) sont générées indépendamment des autres paires, hypothèse caractérisant la majeure partie de la littérature sur le couplage d'enregistrements et mise en œuvre dans les outils logiciels (p. ex., RELAIS, Cibella et coll. 2012). Au contraire, les comparaisons γ et l'état d'appariement dans Ω sont dépendants de manière déterministe. Par conséquent, les estimations sur $m(\gamma)$ et $u(\gamma)$ fondées sur l'algorithme EM sont généralement mauvaises. Ce fait compromet l'efficacité de l'application de la méthode de Fellegi-Sunter, ainsi que le calcul automatique des mesures de la qualité et la possibilité d'appliquer des méthodes efficaces aux fins d'estimation du modèle sur des données couplées (p. ex. les fonctions de régression), comme dans Chambers et coll. (2015). Nous proposons d'examiner Ω au moyen d'un ensemble d'échantillons, chacun tiré de manière à préserver l'indépendance des comparaisons entre les paires d'enregistrements sélectionnées. Les simulations sont encourageantes.

Mots clés : probabilités d'erreur dans un couplage d'enregistrements; qualité des variables d'appariement; paires d'unités d'échantillonnage.

1. Problème du couplage d'enregistrements et rôle de l'indépendance

1.1 Description et notation

Supposons qu'A et B sont des ensembles de données où il y a chevauchement, du moins en partie, de la liste d'unités observées. Le couplage d'enregistrements vise à reconnaître les enregistrements de ces deux ensembles qui pourraient être couplés (et les variables correspondantes en juxtaposition) en vue de l'obtention d'un nouvel ensemble (ou sous-ensemble) de données d'unités couplées avec un ensemble de renseignements plus riche pouvant faire l'objet d'une étude et d'une analyse statistiques.
Fellegi et Sunter (1969) proposent de fonder la décision sur la question de savoir si une paire est un appariement ou non en comparant un ensemble de *k variables* ($k>3$) qui sont conjointement capables de déterminer les unités (habituellement appelées *variables d'appariement*), mais le problème est qu'elles peuvent présenter des problèmes de qualité ou de stabilité au fil du temps. C'est pourquoi il n'est pas garanti que la même unité déclare les mêmes valeurs de variables d'appariement dans les deux ensembles de données. En supposant que l'état d'une paire d'enregistrements issue des deux ensembles de données est représenté par la variable aléatoire C (avec les valeurs 1 si la paire est un appariement et 0 autrement), tandis que la comparaison entre les *k variables d'appariement* est représentée par un vecteur aléatoire à k valeurs Γ, où chaque composante est soit 1 (si la variable correspondante est la même pour les deux unités) ou 0 (autrement), Fellegi et Sunter examinent les distributions de probabilité suivantes :

$$m(\gamma) = P(\Gamma = \gamma | C = 1), u(\gamma) = P(\Gamma = \gamma | C = 0), \ \forall \gamma. \quad (1)$$

En outre, ils adoptent le rapport de vraisemblance

[1]Mauro Scanu, Tiziana Tuoto, Sara Piombo, Marco Fortini, Institut national italien de statistique (Istat), via Cesare Balbo 16, Rome, Italie, 00184 (courriel : scanu@istat.it)

$$r(\gamma) = \frac{m(\gamma)}{u(\gamma)}, \forall\, \gamma,$$

comme la fonction de décision pour déclarer chaque paire comme étant un appariement (si $r(\gamma)$ est plus grand qu'un seuil fixe) ou non. Fellegi et Sunter montrent comment estimer les distributions dans (1) dans un cas simple par la méthode des moments. Des situations plus complexes ont été résolues au moyen de l'algorithme EM appliqué au produit cartésien Ω=A×B (Winkler, 1988). De toute façon, l'algorithme EM est appliqué à toutes les paires dans Ω. Par conséquent, la fonction de vraisemblance observée prend la forme suivante :

$\prod_{ab\in\Omega}(m(\gamma_{ab})p)^{c_{ab}}\big(u(\gamma_{ab})(1-p)\big)^{1-c_{ab}}$ (2)

où *p* est la probabilité que *C=1* et l'état de chaque paire c_{ab} n'est pas observé. Le produit implique que les répétitions bivariées (*C*, Γ) sur les paires dans Ω sont indépendantes. Quoi qu'il en soit, si c_{ab} est égal à 1 pour les paires *a'b'*, *a'b''* et *a''b'*, alors la paire *a''b''* n'est pas libre de supposer une valeur autre que 1, de sorte que *C* n'est pas indépendant pour toutes les paires. Il en va de même pour Γ. Par conséquent, le produit de la vraisemblance observée est erroné et les caractéristiques de bonne qualité des estimateurs de paramètres fondés sur la maximisation (2) sont compromises.

Les conséquences sont différentes. Premièrement, la fonction de décision r_{ab} dépend des estimations des distributions *m* et *u*. Par conséquent, on peut se demander si les décisions prises avec une fonction de décision incorrectement estimée sont correctes. Certains résultats (Tuoto, 2016) indiquent que les décisions ne dépendent pas beaucoup des estimations de paramètres obtenues au moyen de l'algorithme EM, quoi qu'il en soit on peut s'attendre à des améliorations. Deuxièmement, compte tenu de l'inexactitude des estimations obtenues par l'algorithme EM, il est impossible de supposer, comme mesure de l'exactitude du couplage d'enregistrements, la probabilité α et β résultant de décisions erronées, où α et β égalent :

$$\alpha = \sum_{ab:\, r_{ab}>\lambda} u(\gamma_{ab}),\ \beta = \sum_{ab:\, r_{ab}<\mu} m(\gamma_{ab}),$$

où λ est un seuil au-dessus duquel toutes les paires avec $r_{ab} > \lambda$ sont déclarées comme des appariements et μ est le seuil sous lequel les paires avec $r_{ab} < \mu$ sont déclarées non appariées. Au final, l'ensemble de données couplées peut être utilisé dans d'autres analyses statistiques dont l'objectif est d'estimer les paramètres de modèle de variables observées distinctement dans les deux fichiers A et B. Chambers et Kim (2015) montrent que la connaissance des distributions *m* et *u* permet de définir des estimateurs plus efficaces que les estimateurs classiques qui ne supposent pas l'existence d'erreurs de couplage.

Pour toutes ces raisons, l'estimation des paramètres de couplage d'enregistrements est une question qui, selon nous, doit être abordée.

2. Notre proposition

2.1 Comment préserver l'indépendance entre paires (dans la mesure du possible)

Comme cela a été écrit à la Section 1, le problème auquel nous nous attaquons est le manque d'indépendance de C pour les paires dans Ω, ainsi que Γ pour les paires dans Ω, du fait que nous considérons les observations basées sur des *paires* d'unités. En effet, on peut considérer que les observations sur les unités, pour les variables clés $X_1,\ldots X_k$ par exemple, sont générées indépendamment pour des unités différentes. Nous proposons alors d'examiner Ω (estimation de la distribution *m*) paires de l'échantillon dans Ω de manière à conserver, dans la mesure du possible, l'indépendance entre paires. Un échantillonnage à partir de Ω tel que proposé auparavant par Yancey (2004) afin d'augmenter le pourcentage d'appariements dans Ω quand ceux-ci sont peu nombreux, et par Fortini (2020) afin de réduire l'espace de comparaison sans imposer de contraintes comme des blocs ou des filtres.

Une idée fort simple est de choisir l'ensemble de données ayant la plus petite taille, disons A, et de suivre les étapes suivantes :

1. Pour chaque unité dans A, on choisit au hasard une unité dans B jusqu'à ce que tous les enregistrements de A se lient en paire à un enregistrement correspondant dans B. Si la sélection dans B se fait sans répétitions, l'ensemble de variables (*C*, Γ) sur les paires sélectionnées sera indépendant, car les paires sont formées d'unités distinctes.
2. Répéter les étapes 1 un nombre *T* de fois.

La sélection peut s'opérer de bien des manières :

a. Comme nous l'avons indiqué, les enregistrements dans B peuvent être choisis avec ou sans répétition.
b. La sélection dans B peut suivre une distribution uniforme sur les unités de cet ensemble ou on peut privilégier une sélection d'unités telle que *m* soit examiné le plus rapidement et le plus efficacement possible.

Enfin, on peut examiner les échantillons *T* séparément afin de tirer des estimations de *m* qui doivent finalement être agrégées, ou l'on peut composer les échantillons *T* dans un grand échantillon unique qui sera analysé. L'idée est d'utiliser la même procédure que celle décrite à la section 1, à ceci près que la probabilité observée (2) est calculée sur les paires dans : *(i)* chaque échantillon sélectionné *t, t=1 ,… , T*; *(ii)* l'échantillon global obtenu par composition des échantillons *T*.

Chaque méthode a ses avantages et ses inconvénients.

La sélection sans répétition préserve l'indépendance et la forme de la vraisemblance dans chaque échantillon *t*, tandis que l'indépendance est en outre permise dans la composition des échantillons *T* lorsqu'il n'est pas permis que les paires déjà sélectionnées se retrouvent dans un nouvel échantillon. De toute façon, cette méthode est plus fastidieuse du point de vue du calcul, étant donné les restrictions qu'il faut introduire. La sélection avec répétition de B peut introduire certaines dépendances entre les paires qui sont, quoi qu'il en soit, bien inférieures à celles dans Ω. Par conséquent, l'utilisation du produit sur toutes les paires des échantillons sélectionnés dans la fonction de vraisemblance observée est la meilleure approximation de la vraisemblance réelle.
Comme cela a été dit plus haut, l'indépendance sur *C* et sur Γ est assurée pour les paires dans chaque échantillon *t, t*=1, … , *T*. Quoi qu'il en soit, nous savons déjà que le nombre de paires dans Ω n'est pas si grand et que le nombre attendu d'appariements dans les échantillons est très faible. Ce fait peut compromettre l'efficacité de l'algorithme EM à trouver les composantes du mélange entre les distributions *m* et *u*, dans chaque échantillon *t*. Étant donné que *m* est la distribution la plus difficile à estimer, la proposition vise à gonfler la possibilité d'inclure les appariements dans la procédure de sélection. Une possibilité est de sélectionner les unités dans B à l'étape 1 selon une distribution qui dépend des comparaisons Γ : étant donné *a* dans A, plus le nombre d'égalités sur les variables clés est élevé, plus la probabilité de sélectionner *b* dans B est élevée. Nos simulations tiennent compte de toutes ces situations.

2.2 Estimation des autres paramètres

La distribution *u* peut être calculée à partir des ensembles de données distincts A et B. En général, si l'on tient compte d'une seule des variables clés et qu'on suppose qu'elle est catégorique avec *k catégories*, on peut poser que :

$$u_j(1) = P\left(\Gamma_{X_j}(a,b) = 1 | C = 0\right) = \sum_{k=1}^{K} P_A(X_j = k) P_B(X_j = k)$$

pour l'indépendance entre unités (a n'est pas b). En supposant une indépendance entre les variables clés sur des unités distinctes (non-appariements), nous pouvons estimer le u global pour tout le vecteur des variables clés *K*.
En ce qui concerne *p*, il s'agit de la proportion d'appariements attendus dans l'ensemble des échantillons *T*. Ce pourcentage peut être gonflé si la sélection de *b* privilégie le domaine de Γ ayant un grand nombre de concordances entre variables clés. Quoi qu'il en soit, il est encore possible de calculer, au moyen du théorème de Bayes, la probabilité d'avoir un appariement étant donné un γ spécifique :

$$P(C = 1 | \Gamma = \gamma) = \frac{pm(\gamma)}{pm(\gamma) + (1-p)u(\gamma)}$$

Cette probabilité peut être appliquée à l'ensemble d'enregistrements avec chaque γ spécifique afin de calculer les appariements dans A et B.
En fait, l'estimation de *m* et *p* au moyen de l'algorithme EM sur les paires sélectionnées, comme cela est proposé à la section 2.1, et de *u* sur les variables clés, telles qu'elles sont estimées respectivement dans A et B, complète l'estimation de la probabilité d'erreurs et de la probabilité a posteriori d'appariement étant donnée une comparaison réalisée par un simple plug-in.

2.3 Interaction avec d'autres problèmes

Comme cela a été constaté par ailleurs (Yancey (2004), Fortini 2020)), un problème important dans l'estimation des paramètres est le fait que Ω consiste en une partition asymétrique dans des paires appariées et non appariées : quand un enregistrement peut être apparié à au plus un enregistrement de l'autre fichier, la fréquence relative des paires

appariées dans Ω ne peut pas être supérieure à l'inverse de la plus grande taille de fichier. Par conséquent, les paires appariées deviennent très rares quand les fichiers sont plus grands, si on les compare à la taille de Ω. En présence d'une telle disproportion dans la partition de Ω entre les appariements et les non-appariements, l'estimation de la distribution *m* peut être extrêmement inexacte. Quelles sont les interactions entre ce problème et l'utilisation de la fonction de vraisemblance (2) touchée par le manque d'indépendance de certaines paires?

3. Exercice sur des données réelles

3.1 Description des données

Nous avons examiné deux fichiers pour lesquels nous connaissons déjà les vraies correspondances exactes : la taille des deux fichiers est de 25 343 et de 24 613. Les variables communes sont de nature différente : il y a des variables de grande qualité avec un pouvoir discriminant limité (genre) et d'autres de qualité moindre mais avec un pouvoir discriminant élevé (nom de famille, année de naissance, jour de naissance) et des variables avec un pouvoir discriminant intermédiaire (mois de naissance).
D'abord A

3.2 Description de la simulation

Nous avons sélectionné des fichiers A et B de taille 50 (pour les deux) et 200 (pour les deux), avec un chevauchement d'unités appariées égal à 20 % pour les deux tailles ou à 80 %. Nous avons ainsi étudié quatre combinaisons au total entre taille et chevauchement de fichiers. Les variables d'appariement sélectionnées suivent le schéma suivant :

- Schéma 1 : nom de famille, sexe, mois et année de naissance;
- Schéma 2 : prénom, nom de famille, sexe et mois de naissance;
- Schéma 3 : prénom, nom de famille, sexe et année de naissance;
- Schéma 4 : prénom, nom de famille, mois et année de naissance.

Ainsi, les 4 schémas peuvent être considérés comme étant ordonnés de l'ensemble le moins discriminant (schéma 1) à l'ensemble le plus discriminant (schéma 4) de variables d'appariement. Cela signifie que le schéma 1 peut admettre plus d'égalités entre les variables d'appariement, ce qui entraîne un plus grand nombre de dépendances dans Ω.

3.2.1 Résultats expérimentaux

Le tableau 3.2.1-1 montre la distribution *m*, réelle, connue d'avance, pour le schéma 2 dans les différents scénarios. L'algorithme EM appliqué à Ω donne des résultats très éloignés de la vérité, surtout dans le cas des fichiers de taille 200 avec 20 % de chevauchement (c'est le cas quand les appariements sont rares dans Ω, ce qui se traduit par de très mauvaises performances de l'algorithme EM).

Tableau 3.2.1-1
Distribution de $m(\gamma)$ dans les 4 scénarios simulés et estimations correspondantes par l'algorithme EM sur Ω, schéma 2

Taille des fichiers	**50**				**200**			
Chevauchement	**20 %**		**80 %**		**20 %**		**80 %**	
γ	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$	$m(\gamma)$	$\widehat{m_\Omega}(\gamma)$
(0 0 0 0)	0	0,00000	0	0,00001	0	0,00679	0	0,00006
(0 0 0 1)	0	0,00191	0	0,00006	0	0,00157	0	0,00043
(0 0 1 0)	0,05	0,00000	0	0,00035	0 025	0,26015	0,00625	0,00480
(0 0 1 1)	0	0,02493	0	0,00264	0	0,05999	0,0125	0,03222
(0 1 0 0)	0	0,00000	0	0,00007	0	0,00098	0	0,00024
(0 1 0 1)	0	0,00953	0	0,00052	0,05	0,00023	0,00625	0,00164

(0 1 1 0)	0	0,00000	0	0,00283	0	0,03746	0,03125	0,01839
(0 1 1 1)	0,15	0,12425	0,05	0,02136	0,15	0,00864	0,11875	0,12358
(1 0 0 0)	0	0,00000	0	0,00030	0	0,01128	0	0,00029
(1 0 0 1)	0	0,00999	0	0,00223	0	0,00260	0	0,00192
(1 0 1 0)	0	0,00000	0	0,01222	0 025	0,43212	0	0,02165
(1 0 1 1)	0,1	0,13027	0 025	0,09225	0,05	0,09965	0,06875	0,14544
(1 1 0 0)	0	0,00000	0	0,00239	0	0,00162	0	0,00110
(1 1 0 1)	0,05	0,04979	0 025	0,01801	0 025	0,00037	0,0125	0,00738
(1 1 1 0)	0	0,00000	0,1	0,09883	0,05	0,06221	0 075	0,08303
(1 1 1 1)	0,65	0,64932	0,8	0,74594	0 625	0,01435	0,66875	0,55783

La stratégie d'échantillonnage sur Ω – dans laquelle l'unité dans B devant être rattachée à une unité dans A est sélectionnée selon une distribution dont la probabilité augmente avec le nombre de 1 dans γ, et l'unité de B peut être sélectionnée avec remise – donne les estimations des tableaux 3.2.1-2 et 3.2.1-3. L'algorithme EM appliqué à l'échantillon entier obtenu avec l'union de toutes les itérations effectuées donne des résultats satisfaisants quand le chevauchement entre A et B est de 80 %, alors qu'il a de mauvaises performances quand le chevauchement est de 20 %, quel que soit le nombre d'itérations. Ces chiffres peuvent suggérer l'hypothèse que la rareté des appariements dans Ω est plus importante que le manque d'indépendance pour C et Γ.

On a calculé l'indicateur du chi carré comme mesure globale de la divergence entre les estimations et la distribution *m* connue. Avec le schéma 2, pour les deux tailles de fichier, l'estimateur sur un échantillon de Ω paires semble meilleur que celui calculé sur le Ω total, sauf dans la situation où les appariements sont rares dans Ω.

Cette méthode a été répétée huit fois, afin de déterminer si la variabilité supplémentaire due à l'échantillonnage à partir de Ω a une incidence sur les résultats. Le tableau 3.2.1-5 présente une distance du chi carré entre les distributions de *m* réelles et estimées pour tous les schémas, quand la taille du fichier est de 200 et que 80 % des enregistrements dans A ou B forment un appariement. Cet exemple montre que l'estimation dans tout Ω est généralement bonne. En effet, celle basée sur un échantillon de paires indépendantes a besoin d'un plus grand nombre d'itérations pour donner de meilleurs résultats que celle calculée sur le Ω en entier.

Tableau 3.2.1-2
Estimations de l'algorithme EM sur un échantillon de paires indépendantes, taille de fichiers 50, schéma 2

Chevauchement	20 %				80 %			
Nombre T d'échantillons	30	50	80	$m(\gamma)$	30	50	80	$m(\gamma)$
γ	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$		$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	$\widehat{m_1}(\gamma)$	
(0 0 0 0)	0,00000	0,00000	0,00132	0	0,00000	0,00009	0,00002	0
(0 0 0 1)	0,00519	0,00000	0,00248	0	0,00000	0,00065	0,00013	0
(0 0 1 0)	0,00000	0,00000	0,01603	0,05	0,00008	0,00284	0,00034	0,00625
(0 0 1 1)	0,02158	0,00000	0,03009	0	0,00176	0,02017	0,00273	0,0125
(0 1 0 0)	0,00000	0,00000	0,00099	0	0,00000	0,00034	0,00047	0
(0 1 0 1)	0,04630	0,04178	0,00185	0	0,00000	0,00241	0,00380	0,00625
(0 1 1 0)	0,00000	0,00000	0,01198	0	0,00101	0,01048	0,00970	0,03125
(0 1 1 1)	0,19261	0,15062	0,02248	0,15	0,02266	0,07452	0,07788	0,11875
(1 0 0 0)	0,00000	0,00000	0,01384	0	0,00000	0,00073	0,00016	0
(1 0 0 1)	0,01434	0,00000	0,02597	0	0,00000	0,00521	0,00127	0
(1 0 1 0)	0,00000	0,00000	0,16775	0	0,00301	0,02259	0,00323	0
(1 0 1 1)	0,05966	0,00000	0,31482	0,1	0,06731	0,16074	0,02597	0,06875
(1 1 0 0)	0,00000	0,00000	0,01034	0	0,00000	0,00270	0,00451	0
(1 1 0 1)	0,12796	0,17538	0,01941	0,05	0,00000	0,01924	0,03620	0,0125
(1 1 1 0)	0,00000	0,00000	0,12536	0	0,03869	0,08347	0,09230	0 075
(1 1 1 1)	0,53236	0,63222	0,23527	0,65	0,86548	0,59382	0,74129	0,66875

Tableau 3.2.1-3
Estimations de l'algorithme EM sur un échantillon de paires indépendantes, taille de fichiers 200, schéma 2

Chevauchement	20 %				80 %			
Nombre T d'échantillons	50	100	300	$m(\gamma)$	50	100	300	$m(\gamma)$
γ	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$		$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	
(0 0 0 0)	0,00000	0,00000	0,00000	0	0,00018	0,00006	0,00006	0
(0 0 0 1)	0,00000	0,00000	0,00000	0	0,00061	0,00043	0,00066	0
(0 0 1 0)	0,54281	0,36497	0,32086	0 025	0,00831	0,00397	0,00199	0,00625
(0 0 1 1)	0,14359	0,07971	0,06513	0	0,02809	0,02814	0,02244	0,0125
(0 1 0 0)	0,00000	0,00000	0,00000	0	0,00052	0,00031	0,00033	0
(0 1 0 1)	0,00000	0,00000	0,00000	0,05	0,00174	0,00217	0,00369	0,00625
(0 1 1 0)	0,05219	0,07755	0,04227	0	0,02385	0,02008	0,01111	0,03125
(0 1 1 1)	0,01381	0,01694	0,00858	0,15	0,08063	0,14220	0,12522	0,11875
(1 0 0 0)	0,00000	0,00000	0,00000	0	0,00107	0,00025	0,00030	0
(1 0 0 1)	0,00000	0,00000	0,00000	0	0,00361	0,00174	0,00334	0
(1 0 1 0)	0,17864	0,31194	0,41363	0 025	0,04942	0,01616	0,01004	0
(1 0 1 1)	0,04725	0,06813	0,08396	0,05	0,16708	0,11444	0,11314	0,06875
(1 1 0 0)	0,00000	0,00000	0,00000	0	0,00306	0,00124	0,00165	0
(1 1 0 1)	0,00000	0,00000	0,00000	0 025	0,01036	0,00881	0,01861	0,0125
(1 1 1 0)	0,01718	0,06628	0,05450	0,05	0,14186	0,08168	0,05602	0 075
(1 1 1 1)	0,00454	0,01448	0,01106	0 625	0,47963	0,57832	0,63140	0,66875

Tableau 3.2.1-4
Divergence entre les distributions *m* vraies et estimées avec le schéma 2, quand les tailles de fichier sont 50 (nombre d'échantillons tirés 80 pour $\widehat{m_{\mathrm{I}}}(\gamma)$) et 200 (nombre d'itérations 300 pour $\widehat{m_{\mathrm{I}}}(\gamma)$)

Taille des fichiers	**50**				**200**			
Chevauchement	**20 %**		**80 %**		**20 %**		**80 %**	
Estimateur	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$	$\widehat{m_{\Omega}}(\gamma)$	$\widehat{m_{\mathrm{I}}}(\gamma)$
χ 2	0,058732	1,70963	0,11769	0,04513	100,6169	37,17019	0,12716	0,09658

Afin d'avoir une idée préliminaire de la variance induite par la sélection des paires dans $\widehat{m_{\mathrm{I}}}(\gamma)$, nous avons également effectué huit répétitions différentes de l'estimateur quand la taille du fichier est de 200, avec le schéma 2, pour un nombre différent d'itérations : 50, 100, 200, 300, 400, 450. L'augmentation du nombre d'itérations semble atténuer la variabilité. Quoi qu'il en soit, la plage des valeurs de la distance du chi carré par rapport à la distribution vraie pour $\widehat{m_{\mathrm{I}}}(\gamma)$ est généralement assez grande, et comprend la distance de la distribution vraie avec l'estimateur basé sur Ω. Il semblerait que l'estimateur basé sur toutes les paires dans Ω puisse être amélioré, même si les stratégies adoptées jusqu'à présent ne peuvent pas donner constamment de meilleurs résultats.

Tableau 3.2.1-5
Divergence entre les distributions *m* vraie et estimée, quand la taille des fichiers est de 200 et que le chevauchement entre A et B est de 80 %, avec le schéma 2, pour différentes répétitions de $\widehat{m_{\mathrm{I}}}(\gamma)$ estimées avec 50 et 300 tirages d'échantillons à partir de Ω (la divergence pour $\widehat{m_{\Omega}}(\gamma)$ est de 0,12716037)

T	χ 2							
	Répétitions de la méthode basée sur des tirages à partir de Ω							
	n° 1	n° 2	n° 3	n° 4	n° 5	n° 6	n° 7	n° 8
50	0,135127	0,264034	0,278951	0,424961	0,149647	0,106711	0,046630	0,152528
300	0,096584	0,136924	0,163529	0,162267	0,112410	0,079939	0,137170	0,100193

4. Commentaires et questions en suspens

Le présent article traite du problème de l'estimation de paramètres dans un problème de couplage d'enregistrements. Les paires d'enregistrements pour deux fichiers peuvent être un appariement ou un non-appariement. Les connaissances se limitent à un ensemble de variables d'appariement dans les deux fichiers qui peuvent comporter des erreurs ou des fautes. On suppose que la comparaison des variables d'appariement est générée par une variable aléatoire dont la distribution est *m* pour les appariements et *u* pour les non-appariements. On suppose généralement que les paires sont indépendantes et la fonction de vraisemblance est construite en conséquence. Cependant, en général, l'indépendance entre les paires n'est pas vraie. C'est pourquoi étant donné que le produit cartésien des deux fichiers Ω est la seule source d'information sur l'état des paires, nous proposons d'estimer l'échantillonnage *m* à partir de Ω pour que les paires de l'échantillon puissent être indépendantes.

La première question est donc simple : existe-t-il un moyen formel de représenter la probabilité réelle à partir de l'ensemble entier de paires dans Ω?

Les exemples présentés dans l'article montrent que les estimations obtenues par l'algorithme EM dans tout Ω peuvent être très éloignées des estimations réelles. Cependant, cela ne semble pas toujours vrai. La première question consisterait à déterminer dans quelles conditions l'estimateur classique $\widehat{m_\Omega}(\gamma)$ est fiable (en suivant également les indications données dans Yancey (2004) et Fortini (2020) pour l'améliorer).

Les expériences indiquent que certains aspects doivent être approfondis pour pouvoir être confirmés. Dans un avenir proche, nous aimerions étudier les deux affirmations suivantes.

- Plus les variables utilisées comme variables d'appariement sont identifiables et plus le chevauchement entre les deux fichiers est élevé, le meilleur estt $\widehat{m_\Omega}(\gamma)$.
- Moins les variables d'appariement sont identifiables, plus l'estimateur $\widehat{m_1}(\gamma)$ (comme pour le schéma 1) donne des résultats moins éloignés de la vérité. Cela pourrait être l'effet d'un plus grand nombre d'équivalences entre les variables d'appariement, ce qui entraîne une plus grande dépendance entre les observations de paires.
- Plus on a un grand nombre d'itérations dans l'estimateur basé sur un échantillonnage à partir de Ω, $\widehat{m_1}(\gamma)$, moins l'estimation semble éloignée de la vérité. Est-il possible d'obtenir des résultats constamment meilleurs que $\widehat{m_\Omega}(\gamma)$ pour toute répétition du même nombre d'itérations pour $\widehat{m_1}(\gamma)$?

D'autres aspects mériteraient aussi d'être abordés.

- L'échantillonnage à partir de Ω a été effectué dans un seul cas, celui de la sélection de paires indépendantes pour chaque *a* dans A, de l'itération de cette sélection et des analyses communes de toutes les paires obtenues. Il importerait d'évaluer ce qui se produira si l'estimation se fait dans chaque échantillon et qu'une estimation unique est donnée.
- Il s'agit de procéder à la simulation pour différents nombres d'itérations, des nombres bien plus grands que ceux calculés dans notre article.
- Il s'agit aussi de procéder à des simulations pour des nombres différents de tailles de fichiers pour A et B.
- Nous avons supposé que les comparaisons entre variables d'appariement sont indépendantes tant pour les appariements que pour les non-appariements. Comme il y a des zéros dans la distribution *m* présentée dans les différents tableaux, cette affirmation est généralement fausse. D'autres modèles peuvent être envisagés.

Ces travaux constituent un simple point de départ : les commentaires et idées nouvelles sont les bienvenus.

Bibliographie

Chambers, R. et G. Kim (2015), « Secondary analysis on linked data », dans K. Harron, H. Goldstein, C. Dibben (éd.) *Methodological developments in record linkage*, New York: Wiley, p. 83 à 108.

Cibella N., Scannapieco M., Tosco L., Tuoto T. et L. Valentino (2012), « Record Linkage with RELAIS: Experiences and Challenges », *Revista Estadistica Española*, 179, p. 311 à 328.

Fellegi, I. P. et A. B. Sunter (1969), « A Theory for Record Linkage », *Journal of the American Statistical Association*, 64, p. 1183 à 1210.

Fortini, M. (2020), « An Improved Fellegi-Sunter Framework for Probabilistic Record Linkage Between Large Data Sets », *Journal of Official Statistics*, vol. 36(4), p. 803 à 825.

Tuoto T. (2016), « Newproposal for linkage error estimation », *Statistical Journal of the IAOS* 32, p. 413 à 420.

Winkler, W. 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, American Statistical Association*, p. 667 à 671.

Yancey, W.E. (2004), « Improving EM Algorithm Estimates for Record Linkage Parameters », Research Report Series, Statistics #2004-01 U.S. Census Bureau, Washington, U.S.A.