QUALITÉ DES COUPLAGES D'ENREGISTREMENTS: DÉFIS ET SOLUTIONS

Abel Dasylva[1]

[1] *Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A0T6, Canada (abel.dasylva@canada.ca)*

Résumé. Cette présentation décrit les applications du couplage d'enregistrements, les défis associés à la mesure précise des erreurs de couplage tels que les faux positifs et les faux négatifs, ainsi que les solutions qui sont en cours d'élaboration pour y faire face à Statistique Canada. Ces solutions incluent des méthodes qui dépendent de vérifications manuelles, et d'autres méthodes qui n'en demandent pas, mais se basent sur des modèles statistiques innovants. L'application de ces méthodes est illustrée dans le cadre du couplage entre les données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et la Base canadienne des données de mortalité (BCDM).

Mots-clés. couplage d'enregistrements, vérification manuelle, erreurs de couplage, faux positifs, faux négatifs, modèle statistique de mélange

1 Le problème des erreurs de couplage

Le but de tout couplage d'enregistrements est d'identifier les enregistrements qui se rapportent à la même entité, ou au même individu. À Statistique Canada comme dans d'autres agences statistiques, le couplage d'enregistrements est un outil essentiel pour exploiter les données administratives. Il est utilisé à des fins analytiques, opérationnelles ou dans le cadre d'une estimation directe telle que l'Étude sur le surdénombrement du recensement (ÉSR) de 2011, voir Dasylva et coll. (2014).

Le problème des erreurs de couplage se pose dès que la décision de lier deux enregistrements est basée sur des pseudo-identifiants, tels que le nom, la date de naissance ou l'adresse. Ces erreurs surviennent principalement pour deux raisons. La première explication est qu'en général les pseudo-identifiants ne sont pas uniques. La deuxième raison est l'occurrence d'erreurs typographiques, de variations orthographiques ou de différences de format. Selon Fellegi (1997), l'utilisation de pseudo-identifiants s'impose souvent à cause de politiques qui empêchent l'usage d'un identifiant unique (tel que le numéro d'assurance social au Canada), afin de protéger la vie privée. Les erreurs qui en résultent incluent les faux positifs et les faux négatifs. Un faux positif est un lien entre des enregistrements associés à des individus distincts. Tandis qu'un faux négatif correspond à l'absence de lien entre deux enregistrements appariés, c.-à-d. qui se rapportent au même individu. Un faux positif est soit un lien impossible lorsqu'il n'y a pas d'enregistrement apparié dans l'autre fichier, soit un lien incorrect.

Les erreurs de couplage affectent la qualité des données couplées et limitent leurs applications dans les statistiques officielles. En effet, plusieurs études ont déjà démontré que ces erreurs sont une source importante de biais lorsqu'elles ne sont pas prises en compte, que ce soit dans les modèles de régression ou dans l'évaluation de la couverture de recensements ou fichiers administratifs par capture-recapture, voir Chambers et Kim (2015), et Abbott et coll. (2015). Il est donc essentiel de mesurer ces erreurs de façon précise et d'ajuster les analyses en conséquence; deux défis de taille. Cependant, la mesure précise des erreurs est un défi plus urgent pour plusieurs raisons. Tout d'abord, elle permet la comparaison objective de différentes méthodes de couplage d'enregistrements. Ensuite, la mesure précise des erreurs est un prérequis pour toute stratégie d'ajustement. Enfin, cette problématique est en phase avec l'utilisation grandissante du couplage d'enregistrements, des

données administratives ainsi que les requêtes des chercheurs concernant la qualité des données couplées, qui sont mises à leur disposition. Toutefois, mesurer les erreurs de couplage est un défi.

2 Méthodologies pour mesurer les erreurs de couplage

En général, la mesure des erreurs de couplage fait intervenir un modèle statistique ou de la vérification manuelle. Winkler (2015) donne un bon aperçu des solutions proposées à date.

En théorie, les méthodes fondées sur un modèle ne nécessitent pas de vérification manuelle, voir Fellegi et Sunter (1969) et Winkler et Yancey (2006). Cependant, elles peuvent produire des estimations non convergentes lorsque le modèle est mal spécifié. C'est le cas des solutions qui reposent sur l'hypothèse d'indépendance conditionnelle proposée par Fellegi et Sunter (1969). Pour y remédier, plusieurs auteurs ont proposé des solutions qui exploitent également un échantillon d'apprentissage où le statut d'appariement (c.-à-d. appariée ou non appariée) des paires est connu. C'est le cas de Armstrong et Mayda (1993), Thibaudeau (1993), Larsen et Rubin (2001), et de Belin et Rubin (1995). Les résultats obtenus démontrent que l'usage d'un échantillon d'apprentissage réduit les biais des estimateurs. Cependant des questions demeurent quant à la convergence de ces estimateurs lorsque le modèle statistique est mal spécifié. Il y a aussi le problème de l'identification des modèles proposés, notamment en ce qui concerne les mélanges de distributions multinomiales, voir Kim (1984) et Fienberg et coll. (2009). Selon Raessler (2002), le problème d'identification du modèle se pose aussi dans le contexte de l'appariement statistique.

Les solutions basées sur la vérification manuelle mesurent les erreurs à partir d'un échantillon de paires d'enregistrements. De telles solutions ont été proposées par Guiver (2011), Heasman (2014) et plus récemment par Smith et coll. (2016). Toutefois, la vérification manuelle est potentiellement coûteuse et subjective. D'autant plus que dans la pratique, la qualité de la vérification manuelle est particulièrement discutable dans la zone grise de la règle de Fellegi-Sunter, où se trouvent les paires ayant le moins d'information discriminante. Cependant, les solutions existantes ne tiennent pas compte des erreurs éventuelles dues à la vérification manuelle.

À Statistique Canada, trois nouvelles méthodes sont en cours de développement pour remédier aux insuffisances qui ont été mentionnées plus haut.

La première méthode est basée sur des vérifications répétées pour prendre en compte les erreurs à la vérification manuelle, voir Dasylva et coll. (2016). Elle adapte la méthodologie prescrite par Biemer (2012) pour évaluer les erreurs de mesure dans les enquêtes par sondage, ex. avec des interviews répétées. Cette solution part de l'idée que le statut d'appariement d'une paire ne peut être déterminé de façon certaine et que pour une paire donnée, des vérificateurs indépendants commettent des erreurs qui sont conditionnellement indépendantes, étant donné le statut d'appariement de la paire et d'autres variables explicatives ; par exemple l'expérience du vérificateur, son niveau d'étude. Ainsi, les taux d'erreurs de vérification manuelle peuvent être estimés à partir d'un modèle de classes latentes et un algorithme d'Espérance-Maximisation (E-M), voir Dempster et coll. (1977). Lorsque ces taux sont faibles, il est possible d'obtenir de bonnes approximations en supposant que la décision à la majorité est toujours correcte. La méthodologie proposée recommande aussi un plan de sondage stratifié à partir des poids de couplage des paires lorsque le couplage est probabiliste. En effet, cette méthodologie de couplage (Fellegi et Sunter, 1969) est une application particulière du test du rapport de vraisemblance, dont le caractère optimal est prouvé par le Lemme de Neyman-Pearson (Casella et Berger, 2002, Theorem 8.3.12). On y attribue à chaque paire un poids de couplage basé sur le logarithme de son rapport de vraisemblance. Enfin, des recommandations sont faites quant au protocole à suivre pour vérifier les paires échantillonnées.

La deuxième méthode est fondée sur un modèle à partir de pochettes qui sont conditionnellement indépendantes des variables de couplage, voir Dasylva et Sinha (2014). Dans la méthodologie du couplage probabiliste, une pochette est un ensemble d'enregistrements, qui s'accordent sur une clé simple fabriquée à partir de certaines variables. Cette clé est définie par un critère, qui est appelé

critère de pochette et conduit à la formation de plusieurs pochettes. Il est d'usage d'utiliser plusieurs critères de pochette et de sélectionner toutes les paires qui répondent à au moins un de ces critères. Ces paires sont alors appelées paires potentielles. L'utilisation judicieuse de pochettes limite grandement la création des paires. Ainsi, elle accroit l'efficacité du couplage, avec un impact négligeable sur le taux de faux négatifs. L'indépendance conditionelle des pochettes permet d'estimer avec un biais négligeable la distribution non appariée dans les pochettes à partir d'un échantillon de paires aléatoires. Une façon simple et presque sûre de générer une paire aléatoire est de choisir un premier enregistrement uniformément au hasard dans le premier fichier, ainsi qu'un deuxième enregistrement de la même manière dans le deuxième fichier. La méthode des paires aléatoires a déjà été proposée par Jaro (1989). Elle est aussi utilisée dans G-COUP, le système généralisé de couplage de Statistique Canada, voir Chevrette (2010). Cette méthode permet d'estimer les erreurs de couplage sans faire d'hypothèses sur les interactions entre les variables de couplage, lorsque la proportion de mélange (c.-à-d. la proportion de paires appariées dans l'échantillon) est connue ou quand certains vecteurs de comparaison ne sont pas observés parmi les paires appariées. Ces hypothèses conduisent à des modèles qui ne souffrent pas du problème d'identification. Dasylva et Sinha (2014) ont aussi donné des conditions suffisantes pour obtenir des pochettes qui sont conditionnellement indépendantes, incluant le choix d'une clé de pochette qui est indépendante des variables de couplage dans chacun des fichiers. Ces critères éclairent la construction de pochettes qui facilitent aussi l'estimation des taux d'erreurs. Enfin, cette méthodologie s'applique aussi aux couplages déterministes et hiérarchiques.

La troisième méthode a été décrite par Dasylva (2015). Elle estime les taux d'erreur en utilisant un estimateur par calage fondé sur de la vérification manuelle et un modèle logistique. Ce modèle permet d'estimer la probabilité conditionnelle qu'une paire soit appariée étant donné son poids de couplage, qu'elle soit liée ou non. Cette probabilité estimée sert ensuite de variable auxiliaire pour caler les estimations des taux d'erreurs. Les estimateurs obtenus héritent des propriétés des estimateurs par calage. Ils sont notamment convergents sous l'hypothèse que les vérifications manuelles soient infaillibles, indépendamment de la validité du modèle logistique sous-jacent. Ils sont aussi plus efficients lorsque le modèle est valide. À date, les simulations effectuées appuient ces conclusions (Dasylva, 2015). En outre, il est important de noter la différence entre cette approche et les solutions fondées sur un modèle, qui exploitent un échantillon d'apprentissage.

3 Un exemple

Sanmartin et coll. (2015) ont récemment effectué une étude de mortalité à partir d'un couplage probabiliste entre des données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et la Base canadienne des données de mortalité (BCDM), pour la période allant de 2000 à 2011. Cette étude offre une occasion d'éprouver les nouvelles méthodologies d'estimation des erreurs.

Le fichier de l'ESCC comprend 2,3 millions d'enregistrements, tandis que celui de la BCDM compte 3,6 millions d'enregistrements. Le couplage est implémenté avec G-COUP en utilisant comme variables de couplage le prénom, le nom de famille, la date de naissance, le sexe et le code postal.

Pour réduire le nombre de paires, le couplage a recours à des pochettes basées sur le code phonétique du nom de famille et la date de naissance. Au total, ces pochettes contiennent près de 418 millions de paires potentielles. Les poids de couplage sont assignés de façon manuelle et itérative selon la méthode décrite par Howe et Lindsay (1981). La décision de lier une paire donnée est prise en deux temps. Dans un premier temps, deux seuils de poids préliminaires sont utilisés pour déterminer des paires rejetées, possibles (zone grise) et définitives, avec une résolution manuelle pour certaines paires possibles. Dans un deuxième temps, un seuil unique final de 92 est choisi à partir des résultats de la vérification manuelle, qui sont aussi employés pour déterminer les taux d'erreurs de couplage. La décision finale est automatique et basée sur ce seuil pour les paires, qui ne sont pas résolues manuellement.

Pour cette étude, la méthodologie basée sur des vérifications répétées semble le meilleur choix pour mesurer les erreurs de couplage. En effet, la méthodologie basée sur des pochettes conditionnellement indépendantes (voir Dasylva et Sinha (2014)) ne peut être appliquée parce que le nom de famille et la date de naissance servent à la fois à créer les clés pour les pochettes et comme variables de couplage. L'estimateur par calage doit être adapté car les poids de couplage sont déterminés de façon manuelle et itérative, c.-à-d. qu'ils ne sont pas estimés à partir d'un modèle statistique.

Pour effectuer la vérification manuelle, les paires ayant un poids de couplage supérieur ou égal à 1,51 (soit 1,19 million de paires) sont réparties entre huit strates correspondant à des intervalles réguliers de poids. Un échantillon de 1 000 paires est tiré selon une répartition uniforme. Chaque paire tirée est vérifiée par trois vérificateurs indépendants. Les paires ayant un poids inférieur à 1,51 ne sont pas échantillonnées parce qu'elles sont trop nombreuses (418 millions) et très majoritairement non appariées. Le tableau (Tableau 3-1) suivant donne tous les détails de la stratification.

Tableau 3-1: Répartition de l'échantillon pour la vérification manuelle

Strate	Population	Pourcentage	Intervalle de poids	Poids de sondage	Taille d'échan.
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

Codifier la vérification manuelle est un exercice périlleux parce qu'elle est nécessairement subjective et parce que toute instruction ou formation données aux vérificateurs peut biaiser les résultats. Le choix est donc fait de fournir des instructions minimales aux vérificateurs qui doivent décider si chaque paire est appariée ou non appariée, sans la possibilité de coder leur incapacité à faire ce choix. Pour chaque paire échantillonnée, la vérification manuelle est basée sur l'examen des variables de couplage, c.-à-d. prénom, nom de famille, date de naissance, sexe et code postal, sans accès aux poids de couplage des paires.

Les taux de faux négatifs, de faux positifs et la précision sont calculés à partir des résultats de la vérification manuelle. Ces mesures d'erreur sont définies à partir des nombres de vrais positifs (abrégé VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN). Le taux de faux négatifs (abrégé TFN) correspond au ratio $FN/(FN + VP)$, celui de faux positifs (abrégé TFP) correspond au ratio $FP/(FP + VN)$. Tandis que la précision se définit comme le ratio $VP/(FP + VP)$.

Le taux estimé de faux négatifs est de 2,43% tandis que celui de faux positifs est de 0,04%. La précision estimée est de 98,64%. Les taux d'erreurs d'un vérificateur sont estimés en supposant que les vérificateurs sont interchangeables, et que dans chacune des huit strates de poids, ils se trompent de façon conditionnellement indépendante étant donné le statut de la paire en question. En outre, une approximation simple est faite sous l'hypothèse que la décision à la majorité est infaillible. Dans une étude ultérieure, ces taux d'erreurs seront estimés avec plus de précision par un modèle de classes latentes et un algorithme E-M. Les résultats estiment le taux de faux négatifs à 2.97% et celui de faux positifs à 0.15% pour l'ensemble des strates de poids, pour chaque vérificateur.

Le tableau (Tableau 3-2) suivant donne les taux d'erreur par vérificateur dans chaque strate. On observe que ces taux d'erreur varient selon les strates et qu'ils sont plus grands dans les strates 2 à 5, et en particulier dans la strate 2 où le TFN estimé est de 33%. Ce taux élevé est dû au fait qu'une seule paire est déclarée comme appariée par la décision à la majorité, qui n'est pas unanime. C'est aussi une indication que l'approximation d'infaillibilité est très imprécise dans la strate 2.

Table 3-2: Taux d'erreurs par vérificateur dans chaque strate.

Strate	Intervalle de poids	TFP par vérificateur (%)	TFN par vérificateur (%)
1	1,51 – 23,51	0,00	0,00
2	23,52 – 49,51	0,54	33,33
3	49,52 – 73,51	4,01	5,88
4	73,52 – 97,51	6,86	5,49
5	97,52 – 123,51	11,11	1,09
6	123,52 – 149,51	0,00	0,27
7	149,52 – 163,51	0,00	0,53
8	163,51 – 194,52	0,00	0,27

Toutefois, ces résultats démontrent que la vérification manuelle permet une mesure fiable des erreurs de couplage lorsqu'il y a assez d'informations; en l'occurrence, les noms, la date de naissance, le sexe et le code postal.

Cette stratégie d'estimation des erreurs peut être améliorée de plusieurs façons. En effet, il faut d'abord noter que les taux d'erreur estimés sont caractérisés par une variance, qui n'a pas été mesurée, mais qui peut l'être par un processus de rééchantillonnage, tel que la méthode de Rao et Wu (1988). En outre la variance des estimations n'est pas minimisée par la répartition uniforme de l'échantillon et les fractions de sondage disparates, qui en résultent. La décision a été prise de garder cette répartition par manque de ressources affectées à la vérification manuelle et parce qu'un vérificateur avait déjà traité l'échantillon produit selon la répartition uniforme. Ensuite, l'estimation des erreurs de vérification manuelle peut être raffinée. En effet, ces erreurs sont estimées sous les hypothèses que la décision à la majorité est infaillible et que les vérificateurs sont interchangeables dans chaque strate. Bien entendu, ce sont des hypothèses simplistes, qui introduisent potentiellement des biais, notamment dans la zone grise. Une meilleure solution est d'estimer les taux d'erreur avec un algorithme E-M tel que suggéré par Biemer (2012), sans supposer que la décision à la majorité est sans erreur. Il est aussi probable que les vérificateurs ne sont pas interchangeables et qu'il faille modéliser l'effet de vérificateur et pour ce faire introduire des variables explicatives pertinentes, ex. l'éducation, l'expérience du vérificateur. Enfin, il faut aussi ajuster les mesures d'erreurs de couplage pour prendre en compte les erreurs de vérification manuelle. Toutes ces améliorations font l'objet de travaux, qui sont en cours.

4 Conclusions

L'estimation des erreurs de couplage demeure le plus grand défi pour les applications du couplage d'enregistrements dans les statistiques officielles. À Statistique Canada, les solutions qui sont en cours de développement prennent en compte une diversité de situations selon la richesse de l'information disponible, la possibilité d'effectuer des vérifications manuelles, la méthode de couplage et la technique employée pour estimer les poids de couplage dans le cas probabiliste. Bien que les premiers résultats soient encourageants, l'effort de recherche se poursuit pour mieux évaluer les solutions existantes et élargir leur champ d'application.

Bibliographie

[1] Abbott, O., P., Jones and M., Ralphs (2015). "Large-scale linkage for total populations in official statistics" in Methodological Developments in Data Linkage, pp. 8-35, UK:John Wiley.
[2] Armstrong, J., and Mayda, J. (1993), "Model-based estimation of record linkage error rates", *Survey Methodology*, 19, pp. 137-147, 1993.
[3] Belin, T.R. and Rubin, D.B.. (1995). "A Method for calibrating false-match rates in record

linkage", *Journal of the American Statistical Association*, 90(430):694–707.
[4] Biemer, P. (2012). Latent Class Analysis of Survey Error, New Jersey:John Wiley.
[5] Casella, G. and Berger, R.L. (2002). Statistical Inference, 2nd edition, Canada: Duxbury.
[6] Chambers W.E., and G., Kim (2015). "Secondary analysis of linked data", in Methodological Developments in Data Linkage, pp.83-108, UK:John Wiley.
[7] Chevrette A (2010). "G-Link: Constructing an Avatar". in *Proceedings of the International Symposium on Statistical Methodology*, Ottawa:Statistics Canada..
[8] Dasylva, A., and S. , Sinha (2014), "Reducing the Structure of Statistical Models for Probabilistic Record Linkage", poster presented at the *Joint Statistical Meetings*.
[9] Dasylva, A., Titus, R.-C., and Thibault, C. (2015), "Overcoverage in the 2011 Canadian Census", in *Proceedings of the International Symposium on Statistical Methodology*, Ottawa:Statistics Canada.
[10] Dasylva, A (2015), "Design-based Estimation with Record-Linked Administrative Files", in *Proceedings of the International Symposium on Statistical Methodology*.
[11] Dasylva, A., M. Abeysundera, B. Akpoué, M. Haddou and A. Saïdi (2016). "Measuring the Quality of a Probabilistic Linkage through Clerical Reviews", presentation given at the 2016 *International Symposium on Survey Methodology*.
[12] Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society Series B*, 39, pp. 1-38.
[13] Fellegi, I.P., and Sunter, A.B. (1969), "A Theory of Record Linkage", *JASA*, 64, pp. 1183-1210.
[14] Fellegi, I. (1997). "Record linkage and Public Policy – A Dynamic Evolution", in *Proceedings of the 1997 International Workshop on Record Linkage Techniques*, Arlington VA, pp. 3-12.
[15] Fienberg, S., Hersh, P., Rinaldo, A., and Zhou, Y. (2009), "Maximum likelihood in latent class models for contingency table data", in *Algebraic and Geometric Methods in Statistics*, Cambridge University Press, pp. 27-62.
[16] Guiver, T. (2011), "Sampling-Based Clerical Review Methods in Probabilistic Linking", unpublished report, Australia: Australia Bureau of Statistics.
[17] Heasman, D. (2014), "Sampling a matching project to establish the linking quality", *Survey Methodology Bulletin*, Office of National Statistics, 72, pp. 1-16.
[18] Howe, G.R., Lindsay, J.A. (1981), "A generalized iterative record linkage computer system for use in medical follow-up studies", *Computers and Biomedical Research*, 14, pp. 327-340.
[19] Jaro, M. A. (1989). "Advances in record linkage methodology to matching the 1985 census of Tampa, Florida", *Journal of the American Statistical Association*, 84, pp. 414-420.
[20] Kim, B.S. (1984), "Studies of multinomial mixture models", PhD thesis, University of North Carolina , Chapel Hill.
[21] Larsen, M., and Rubin, D. (2001), "Iterated automated record linkage using mixture models", *JASA*, 96, pp. 32-41.
[22] Raessler, S. (2002) Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York: Springer-Verlag.
[23] Rao, J.N.K., and Wu, C.F.J. (1988). "Resampling inference with complex survey data", JASA, 83, pp. 231-241.
[24] Sanmartin, C., Y., Decady, R.,Trudeau, A., Dasylva, M., Tjepkema, P., Fines, R., Burnett, N., Ross, and D., Manuel (2015)."Linking the Canadian Community Health Survey to the Canadian Mortality Database: A national resource to study mortality in Canada", submitted to Health Reports.
[25] Smith, P., S. Gammon, S. Cummins, C. Chatzoglou and D. Heasman (2016). "Sampling procedures for assessing accuracy of record linkage", Presentation given at *the 2016 International Symposium on Survey Methodology*.
[26] Thibaudeau, Y. (1993), "The discrimination power of dependency structures in record linkage", *Survey Methodology*, vol. 19, pp. 31-38, June 1993.
[27] Winkler, W.E., and Yancey, W.E. (2006), "Record-linkage error-rate estimation without training data", in *Proceedings of the Section on Survey Research Methods*, ASA.
[28] Winkler W.E. (2015). "Probabilistic linkage" in Methodological Developments in Data Linkage, pp.8-35, UK:John Wiley.