Journées de méthodologie statistique de l'Insee 2022

PERSPECTIVES FUTURES EN COUPLAGE D'ENREGISTREMENTS

Gautier GISSLER (*) et Abel DASYLVA (*)

(*) Statistique Canada, Direction de la Méthodologie

gautier.gissler@statcan.gc.ca
abel.dasylva@statcan.gc.ca

Mots-clés: classification non supervisée, couplage probabiliste, Fellegi-Sunter, k-moyennes, estimation de seuils, estimation des erreurs.

Domaine concerné : appariement probabiliste, classification

Résumé

En couplage d'enregistrements, beaucoup de défis reposent sur l'absence, en pratique, de données d'entraînement, ce qui nécessite souvent une intervention humaine, généralement d'expertise, non-négligeable. Cet article aborde deux de ces problématiques, sous l'angle des perspectives en couplage visant à diminuer le fardeau des révisions manuelles. Le premier défi est la détermination des seuils dans la méthode probabiliste, pour lequel il est proposé une procédure automatique basée sur la méthode de k-moyennes. Le deuxième défi est l'estimation des erreurs de couplage pour lequel il est proposé d'utiliser un modèle fondé sur le nombre de liens adjacents à un enregistrement donné.

Avertissement: Le contenu de cet article représente le point de vue de ses auteurs et pas nécessairement celui de Statistique Canada. Il décrit des méthodes théoriques qui pourraient ne pas refléter celles implémentées par l'agence.

Introduction

À Statistique Canada, plusieurs couplages sont basés sur la méthode probabiliste et font appel à des vérifications manuelles, c.-à-d. des inspections visuelles de paires d'enregistrements pour déterminer si les enregistrements proviennent de la même unité. Dans un premier temps, les vérifications manuelles servent à déterminer des seuils de similitude pour classer deux enregistrements comme provenant de la même unité, où les seuils choisis sont déterminants pour la qualité des liens résultants. Dans un second temps, les vérifications manuelles servent à mesurer les erreurs de couplage qui sont inévitables lorsque le couplage de deux sources est basé sur des quasi-identificateurs tels que des noms

ou dates. Toutefois, les vérifications manuelles sont coûteuses. Il est donc nécessaire de les réduire et même de les éviter si cela est possible. Pour ce faire, nous proposons de recourir à la méthode de k-moyennes et à la modélisation pour le problème des seuils et celui de l'estimation des erreurs respectivement.

Les paragraphes qui suivent traitent des seuils lors de l'étape de classification des paires en couplage probabiliste, avant d'aborder le problème de l'estimation des erreurs en couplage et de terminer par la conclusion.

Estimation des seuils et classification des paires dans la méthodologie probabiliste : l'exemple de G-Coup.

G-Coup (ou G-Link en anglais) est un logiciel conçu par Statistique Canada afin d'effectuer des couplages probabilistes, fondé sur la théorie de Fellegi-Sunter [1]. Dans G-Coup, après avoir construit un ensemble de paires potentielles (« ensemble lié », noté L) par la méthode du blocage (ou *blocking*), et un ensemble de paires aléatoires (« ensemble non-lié », noté NL), l'utilisateur dresse n règles de comparaisons, composées de plusieurs niveaux d'accords, entre les champs qui composent les enregistrements des paires. C'est à partir de ces règles que G-Coup calcule les scores probabilistes de Fellegi-Sunter des paires, la valeur du degré de la paire (a,b) à la règle k, $\gamma_k(a,b)$ étant défini par le ratio de la probabilité que cet accord arrive lorsque (a,b) est un vrai lien (un *match*) et de la probabilité que cet accord arrive lorsque (a,b) est un faux lien (un *unmatch*). Les ensembles des vrais liens et faux liens n'étant pas connus, ils sont respectivement approximés par les ensembles lié et non-lié :

$$RP(\gamma_k(a,b)) = P\left(\gamma_k(a,b) \middle| (a,b) \in L\right) \Big/ \frac{}{P\left(\gamma_k(a,b) \middle| (a,b) \in NL\right)}$$

Le score total d'une paire est alors obtenu par celui de son profil d'accords $[\gamma_1,\ldots,\gamma_k,\ldots,\gamma_n]$, en supposant l'indépendance des règles entre elles.

$$RP((a,b)) = RP(\gamma_1(a,b)) \times \ldots \times RP(\gamma_k(a,b)) \times \ldots \times RP(\gamma_n(a,b))$$

Usuellement, par commodité, le ratio de probabilité est transformé en poids w par une opération logarithmique. À partir de ce poids, les paires sont classées en trois zones (rejetées, possibles et définitives), en définissant deux seuils (inférieur T_λ et supérieur T_μ). Ainsi, une paire de poids w sera rejetée si $w < T_\lambda$, possible si $T_\lambda \leq w < T_\mu$, et définitive si $w \geq T_\mu$. La classification dépend donc d'une part de la bonne estimation des poids des paires et d'autre part de la bonne estimation de ces deux seuils. De plus, la première composante dépend de la seconde, comme l'explique la remarque suivante. En pratique, on observe qu'une bonne classification n'est pas obtenue dès le premier coup et qu'il faut itérer le processus de calcul des poids, de choix des seuils et de reclassification à partir des nouveaux seuils pour raffiner le couplage. En effet, le processus est bien itérable puisque le numérateur du ratio de probabilité varie avec l'ensemble lié, ce dernier évoluant avec le raffinement des zones possibles et définitives.

L'étape du choix des seuils est donc essentielle puisque la classification des paires et le calcul de nouveaux poids à l'itération suivante en découlent. Une méthode utilisée en pratique par l'Environnement de couplage des données sociales (ECDS) à Statistique Canada [2], pour sa fiabilité expérimentée au cours de nombreux couplages, s'appuie sur l'examen des paires et profils de concordances et nécessite donc l'intervention d'un expert en couplage. Un lourd volume de profils de concordances à réviser peut être atteint lorsque plusieurs règles (au-delà d'une dizaine généralement) sont définies pour un grand nombre de paires liées (plusieurs millions) à comparer. Par ailleurs, ces règles peuvent être complexes, en

définissant plusieurs niveaux d'accords partiels, ou/et faisant intervenir plusieurs variables, ou/et faisant appel à des fonctions de comparaisons poussées. Pour réduire le fardeau de révision, des filtres sur les règles à conserver dans l'étude des profils et un filtre sur un nombre minimum de paires que représente un profil, peuvent être appliqués. Cette méthode de révision des profils permet à l'utilisateur de bien connaître les paramètres et données de son couplage et de maîtriser sa classification. Cette maîtrise passe par le choix précis des seuils, le raffinement des poids (que ce soit par itération ou par raffinement manuel suite à l'examen des profils), ou les décisions manuelles (par la mise en place de « politiques » de décision sur des profils ou des choix de classification manuels).

Cependant, cette méthode est coûteuse en temps, et demande un utilisateur avancé pour la réaliser, ce qui n'est pas toujours possible pour tous les projets de couplages. Il est donc pertinent de vouloir chercher des alternatives automatisées pour cette étape. À défaut de remplacer l'intervention humaine à l'étape de classification, une estimation automatisée des seuils permettrait au moins d'aiguiller l'utilisateur dans ses choix, particulièrement l'utilisateur débutant, voire d'alléger son fardeau de révision (en utilisant par exemple les estimations comme point de départ à sa classification). On définit donc un estimateur automatique de seuils comme un algorithme permettant de déterminer les seuils inférieurs et supérieurs sans intervention humaine pour un couplage similaire à ceux effectués en pratique (ex. dans le cadre de l'ECDS, avec de nombreuses règles complexes), en ayant donc à disposition uniquement les données du couplage (on se place dans un cadre non-supervisé), c'est-à-dire les paires liées, leur éventuel précédent statut de classification et leur vecteur de poids.

Un estimateur automatique de seuil idéal permettrait de reproduire exactement la méthodologie itérative du calcul des poids actuelle, c'est-à-dire :

- Calcul des poids à partir des ensembles lié et non-lié
- Estimation automatique des seuils
- Redéfinition de l'ensemble lié par celui des paires possibles et définitives

Avec un estimateur idéal, on observerait, au fur et à mesure de l'itération la convergence de l'ensemble lié $L = Possibles + Définitives$ vers une bonne approximation de l'ensemble des vrais liens.

Plusieurs algorithmes pour estimer les seuils ont été implémentés dans G-Coup. Par exemple, la méthode des valeurs extrêmes est un algorithme semi- automatique provenant du paquet R *RecordLinkage* [3]. Par la visualisation des résidus des moyennes extrêmes (*Mean excee-dances*), les utilisateurs peuvent déterminer une zone (intervalle de poids) où la distribution des valeurs extrêmes est valide et ainsi calculer les seuils. Une autre méthode se base sur le concept de mise en correspondance (*mapping*) : à la fin du couplage, les conflits entre les paires dans l'ensemble lié sont résolus en conservant la paire au score le plus élevé. Le seuil inférieur est obtenu par le maximum du poids des paires dans l'ensemble des paires rejetées manuellement par l'utilisateur. Le seuil supérieur est obtenu en prenant le minimum du poids des paires dans l'ensemble des paires classifiées définitives sans intervention manuelle [4].
Enfin, différentes méthodes d'apprentissage automatique peuvent être utilisées, en particulier dans l'estimateur basé sur la classification en deux étapes suggérée par Christen [5]. Celui-ci se compose d'une partie non-supervisée qui permet de créer un ensemble d'entraînement, puis une partie supervisée qui utilise les données d'entraînement créées par la première étape. Le choix des modèles, ainsi que la question de la représentation des données entrent alors en jeu. Dans G-Coup, la partie non supervisée est effectuée par les k-

moyennes (k=2 pour les deux ensembles Définitifs et Rejetés à identifier) et la partie supervisée par un modèle probit selon les cibles d'erreurs, faux positifs et faux négatifs, indiquées par l'utilisateur. Des paramètres d'entrée permettent également à l'utilisateur de contrôler la représentation des données, comme le taux minimal de valeurs non-nulles et non-manquantes pour considérer une règle. De plus, l'utilisation de la distance de Manhattan à la place de la distance euclidienne pour représenter les vecteurs de poids des paires peut sembler plus pertinente. La question de représenter les paires par le ratio de probabilité (ou la probabilité liée ou non-liée uniquement) à la place du poids se pose aussi. Enfin, l'utilisation des profils de concordance à la place des paires peut s'avérer être un choix judicieux pour diminuer le fardeau calculatoire pour les gros projets.

En pratique, dans le cadre de l'ECDS, bien que d'évaluation expérimentale rigoureuse n'ait pas encore été conduite, il a été observé le caractère très dépendant de ces estimateurs à la distribution des poids. Plus cette dernière est raffinée et le processus de couplage avance, plus les estimateurs seront pertinents.

Estimation des erreurs de couplage

En couplage d'enregistrements, on définit une paire appariée comme une paire d'enregistrements provenant de la même unité. En relation avec ce concept, le statut d'appariement d'une paire est la variable dichotomique indiquant si elle est appariée. En pratique, on voudrait lier toutes les paires appariées et ne lier aucune paire non-appariée. Toutefois, lorsqu'il n'y a pas d'identificateur unique, cela est un vrai défi en raison des erreurs de couplage, qui incluent les faux négatifs (FN) et les faux positifs (FP), où un faux négatif est le fait de ne pas lier une paire appariée et un faux positif est le fait de lier une paire non-appariée. Il est crucial de mesurer ces erreurs pour produire les indicateurs de qualité sur les données couplées, tel que requis par Statistique Canada. Cette information permet aussi d'optimiser les liens dans la méthode probabiliste [1] et d'ajuster les analyses des données couplées pour tenir compte des erreurs. Toutefois, la mesure des erreurs représente encore un défi que ce soit par la vérification manuelle ou par la modélisation.

La vérification manuelle consiste à inspecter visuellement des paires d'enregistrements pour déterminer si leurs enregistrements constitutifs proviennent de la même unité. Toutefois, la vérification manuelle est coûteuse.

L'autre solution est la modélisation, par exemple avec les modèles de mélange log-linéaire qui ont été longtemps étudiés [1][6][7][8][9]. La plupart de ces solutions modélisent la distribution des accords qui sont observés dans une paire, par un mélange de deux distributions conditionnelles, qui représentent la distribution des paires appariées et celle des paires non-appariées. Souvent, l'hypothèse est faite que les accords sur les différentes variables sont conditionnellement indépendants, étant donné qu'une paire est appariée ou non-appariée [1][9]. Toutefois, c'est une hypothèse qui peut s'avérer fausse, surtout quand on applique le blocage (aussi appelé pochettes). Dans ce cas, elle produit des estimations biaisées. Pour remédier à ce problème, il est bien possible d'inclure des interactions. Toutefois, les modèles résultants sont plus complexes à ajuster. En outre, la mise en œuvre de cette méthodologie peut nécessiter un recours à des paires dont le statut d'appariement est connu pour la sélection des interactions, comme dans l'étude de Armstrong et Mayda [6] et celle de Thibaudeau [7]. À la différence de ces études, Winkler [8] n'utilise pas de telles paires mais des contraintes ad-hoc sur les paramètres du modèle, bien que la spécification de ces contraintes soit un enjeu.

Il est aussi possible de modéliser les erreurs à partir de la distribution du nombre de liens adjacents à un enregistrement donné [10][11][12], c.-à-d. le nombre de liens à partir dudit enregistrement. L'avantage principal de ces modèles est qu'ils prennent en compte toutes les interactions entre les variables de couplage de façon implicite. Ils sont fondés sur la relation entre le nombre de liens adjacents à un enregistrement donné et les erreurs qui impliquent cet enregistrement. Afin de décrire cette relation, considérons une population finie de N unités et le couplage d'un fichier de taille $O(N)$ et d'un recensement exhaustif à propos de cette population. Considérons aussi que chaque source ne contient aucun doublon et le couplage est tel que la décision de lier deux enregistrements donnés ne fait pas intervenir d'autres enregistrements. Soit n_i le nombre de liens adjacents à l'enregistrement i du fichier, c.-à-d. le nombre de liens à partir de cet enregistrement. Dans ce cas, la relation entre n_i et les erreurs, qui impliquent l'enregistrement i, est décrite dans la table suivante,.

Table 1 : Relation entre n_i et les erreurs

n_i	FN	FP
0	1	0
$1 \leq n_i \leq N-1$	0 ou 1	$n_i - 1$ ou n_i
N	0	$N-1$

Lorsque $n_i = 0$ ou $n_i = N$, il n'y a pas d'incertitude sur les erreurs. Autrement, il y a de l'incertitude même si celle-ci est très réduite pour les faux positifs dont le nombre est de $n_i - 1$ ou n_i. Dans ce cas, la méthode de Blakely et Salmond [10] permet de prédire les erreurs en tenant compte du fait que n_i est la somme des vrais positifs (VP) et des faux positifs (FP), et en supposant que ces deux contributions sont indépendantes et telles que $VP \sim Bernoulli(p)$ et $FP \sim Binomiale\big(N-1, \lambda/(N-1)\big)$, où p et λ sont les paramètres du modèle qui représentent les nombres espérés de vrais positifs et de faux positifs pour chaque enregistrement du fichier. En somme

$$n_i \sim Bernoulli(p) * Binomiale\big(N-1, \lambda/(N-1)\big),$$

où ¿ dénote l'opération de convolution. À partir des paramètres p et λ, il est possible d'estimer le rappel et la précision, c.-à-d. la proportion des paires appariées qui sont liées et la proportion des paires liées qui sont appariées. En effet, le rappel s'estime par p, tandis que la précision s'estime par $p/(p+\lambda)$, où les paramètres p et λ sont estimés par une méthode des moments, à partir des nombres d'enregistrements où $n_i = 0,1,2$.

Récemment, la méthode de Blakely et Salmond [10] a été généralisée pour tenir compte du caractère hétérogène des enregistrements [11][12]. En effet, on attend un nombre espéré de faux positifs plus petit quand un enregistrement a une caractéristique rare (par ex. un nom de famille rare) que lorsque cette caractéristique est commune. Le modèle proposé s'obtient en observant d'abord que, dans le modèle de Blakely et Salmond [10], la distribution des faux positifs (c.-à-d. la loi $Binomiale\big(N-1, \lambda/(N-1)\big)$) converge en loi vers la loi de Poisson avec le paramètre λ, quand $N \to \infty$, de sorte que l'on obtient la loi $n_i \sim Bernoulli(p) * Poisson(\lambda)$ dans le cadre homogène [12]. Ensuite, le caractère hétérogène des enregistrements est pris en compte en permettant aux paramètres p et λ de varier d'un enregistrement à un autre avec un mélange fini du type

$$n_i \sim \sum_{g=1}^{G} \alpha_g \, Bernoulli(p_g) * Poisson(\lambda_g).$$

Dans ce cas, le rappel est estimé par $\overline{p}=\sum_{g=1}^{G}\alpha_g p_g$. Tandis que la précision est estimée par $\overline{p}/(\overline{p}+\overline{\lambda})$, où $\overline{\lambda}=\sum_{g=1}^{G}\alpha_g \lambda_g$. À la différence de Blakely et Salmond [10], les paramètres du modèle sont estimés en maximisant la vraisemblance composite des n_i, pour chaque choix de G [11][12], où ce dernier paramètre peut être choisi en fonction du critère d'information d'Akaike [13].

Les modèles basés sur les liens adjacents sont commodes pour estimer les taux d'erreur [10] et les faux négatifs dus au blocage [14], lorsqu'un fichier est lié à un recensement d'une population donnée et que les conditions énoncées plus haut s'appliquent.

Les Tables 2 et 3 en donnent un bon exemple, à partir du couplage de deux registres exhaustifs, qui sont basés sur les données du Système de ressourcement de la fonction publique [14]. Dans cet exemple, où la population comprend $N=63555$ unités (personnes) avec un identificateur unique, le blocage est basé sur l'égalité de la date de naissance partielle (c.-à-d. sans l'année de naissance) et sur l'égalité du code phonétique (code SOUNDEX) du prénom ou du nom de famille. La disponibilité de l'identificateur unique permet de calculer les vrais taux d'erreur et de les comparer avec les estimations obtenues avec les différents modèles. Les résultats montrent que les estimations sont proches du vrai rappel avec tous les modèles ; l'erreur absolue relative ne dépassant 100(0.9718-0.9697)/0.9718=0.22%. Pour la précision, les estimations sont aussi proches de la vrai valeur ; la plus grande erreur absolue relative étant obtenue avec le modèle de Blakely et Salmond (2002) où elle est égale à 100(0.9864-0.8837)/0.9864=10.41%. Toutefois, avec l'alternative proposée, l'erreur absolue relative est beaucoup plus petite ; étant égale à 100(0.9864-0.9862)/0.9864=0.02%. Ces résultats suggèrent aussi qu'une classe suffit avec le modèle de mélange fini.

Table 2 : Table de fréquence des n_i dans l'exemple.

n_i	0	1	2	3	4	5
Fréquence	1659	53951	6875	603	62	5

Table 3 : Taux d'erreur estimés dans l'exemple.

			Rappel	Précision
Vrais taux			0.9718	0.9864
Estimations fondées sur un modèle	Blakely et Salmond (2002)		0.9701	0.8837
	Dasylva et Goussanou (2020)	G=1	0.9699	0.9862
		G=2	0.9702	0.9862
		G=3	0.9697	0.9862

À la différence des mélanges log-linéaires, les modèles fondés sur le nombre de liens adjacents prennent en compte toutes les interactions entre les variables de couplage de façon implicite, c.-à-d. sans avoir à les spécifier, et sans nécessiter de paires dont le statut d'appariement est connu. Ils sont donc faciles à employer et peu coûteux. En outre, ils s'appliquent à plusieurs méthodes de couplage incluant les méthodes probabiliste, déterministe et hiérarchique. Toutefois, ces modèles doivent être généralisés pour prendre en compte les doublons, la sous-couverture [15] et les couplages complexes, où la décision de lier deux enregistrements donnés peut impliquer d'autres enregistrements. Il y a aussi la question ardue de l'inférence statistique, parce que la structure de corrélation des n_i

complique le calcul des variances, intervalles de confiance et niveaux critiques pour les tests d'hypothèse. À ce sujet, il a été démontré qu'il est possible de traiter un échantillon aléatoire simple sans remise des n_i de taille $o\left(N^{1/2}\right)$ comme étant approximativement indépendant et identiquement distribué ; la taille optimale d'échantillon étant de l'ordre $O\left(N^{2/5}\right)$ sous des conditions de régularité [11].

Conclusion

La détermination des seuils par la méthode de k-moyennes et la modélisation des erreurs par le biais des liens adjacents sont des solutions pratiques au problème du recours à la vérification manuelle, en couplage d'enregistrements. Toutefois, les recherches continuent pour améliorer ces méthodes et les rendre plus robustes.

Bibliographie

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

[2] Babyak, C. and Saidi, A. (2017) "Record Linkage Methodology for the Social Data Linkage Environment at Statistics Canada", IJPDS (2017) Issue 1, Vol 1:032, *Proceedings of the IPDLN Conference (August 2016)*, International Journal of Population Data Science, 1(1). doi: 10.23889/ijpds.v1i1.49.

[3] Sariyar, M., Borg, A. and Pommerening, K. (2011). "Controlling False Match Rates in Record Linkage using Extreme Value Theory", *Journal of Biomedical Informatics*, 44(2011), pp. 648 – 654.

[4] "Guide d'utilisateur de G-Coup 3.5", Statistique Canada, 2021.

[5] Christen, P. (2007). "A Two-step Classification Approach to Unsupervised Record Linkage", appeared at the Sixth Australasian Data Mining Conference.

[6] Armstrong, M., et Mayda, J. (1993). "Model-based estimation of record linkage error rates", *Techniques d'enquête*, 19, 137-147.

[7] Thibaudeau, Y. (1993). "The discrimination power of dependency structures in record linkage", *Techniques d'enquête*, 19, 1-16.

[8] Winkler, W. (1993). "Improved decision rules in the Fellegi-Sunter Model of Record Linkage", *dans les Actes de la Section des méthodes d'enquête*, Association américaine de statistique, 274-279.

[9] Chipperfield, J., Hansen, N., et Rossiter, P. (2018). "Estimating Precision and Recall for Deterministic and Probabilistic Record Linkage'', *International Statistical Review*, 86, 219-236.

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

[11] Dasylva, A., Goussanou, A., Ajavon, D., et Abousaleh, H. (2019). "Revisiting the probabilistic method of record linkage", https://arxiv.org/abs/1911.01874.

[12] Dasylva, A., et Goussanou, A. (2020). "Estimating linkage errors under regularity conditions", *dans les Actes de la Section des méthodes d'enquête*, Association américaine de statistique.

[13] Akaike, H. (1974). "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19, 716-723.

[14] Dasylva, A., et Goussanou, A. (2021). "Estimating the false negatives due to blocking in record linkage", *Survey Methodology*, 47, 299-311.

[15] Dasylva, A., Goussanou, A., et Nambeu, C.-O. (2021). Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors. https://www.statcan.gc.ca/en/conferences/symposium2021/program. (Présentation donnée au Symposium international de Statistique Canada de 2021)