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

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Évaluation de l'exactitude lors d'un appariement par vagues

par Abel Dasylva et Arthur Goussanou

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Évaluation de l'exactitude lors d'un appariement par vagues

Abel Dasylva et Arthur Goussanou[1]

Résumé

Au sein de Statistique Canada, de nombreux ensembles de données sont appariés à l'aide de quasi-identificateurs, comme le prénom, le nom de famille ou l'adresse. En pareil cas, les erreurs de couplage peuvent être une préoccupation et doivent donc être mesurées. À cet égard, des études menées antérieurement ont révélé que cette évaluation peut s'appuyer sur la modélisation du nombre de liens d'un enregistrement donné, tout en tenant compte de toutes les interactions qui existent entre les variables de couplage et en se passant de vérifications manuelles, aussi longtemps que la décision de lier deux enregistrements ne fait pas intervenir d'autres enregistrements. Dans le cadre de la présente communication, la méthodologie utilisée est adaptée à une catégorie de stratégies pratiques, lesquelles ne respectent pas cette contrainte en liant les enregistrements par vagues successives, où une vague donnée lie un sous-ensemble d'enregistrements qui ne sont pas liés dans les vagues précédentes. En particulier, le couplage peut être fondé sur une vague déterministe suivie d'une vague probabiliste.

Mots clés : Résolution d'entité; exactitude de couplage; intégration des données; erreur non due à l'échantillonnage.

Avis de non-responsabilité : Le contenu du présent document reflète les opinions des auteurs, mais pas nécessairement celles de Statistique Canada.

1. Introduction

Au sein de Statistique Canada, de nombreux ensembles de données sont appariés à l'aide de quasi-identificateurs, comme le nom de famille, l'adresse ou la date de naissance, qui ne sont pas uniques et qui peuvent présenter des coquilles. Par conséquent, des erreurs de couplage peuvent survenir, qu'il faut évaluer. Ces erreurs comprennent des faux négatifs et des faux positifs, où un faux négatif est l'absence de lien entre des enregistrements qui sont associés à la même unité, et un faux positif est la présence d'un lien entre des enregistrements qui proviennent d'unités différentes. Actuellement, on mesure ces erreurs avec des vérifications manuelles. Cependant, à cause de la pression des coûts, il y a un intérêt grandissant pour le remplacement de ces revues par une procédure d'estimation basée sur un modèle, en utilisant l'un des nombreux modèles accessibles, qui sont examinés par Dasylva et Goussanou (2024). Une approche prometteuse repose sur la modélisation du nombre de liens d'un enregistrement, là où la décision de lier deux enregistrements ne dépend pas d'autres enregistrements (Blakely et Salmond, 2002; Dasylva et Goussanou, 2020). Malheureusement, cette contrainte n'est pas respectée par une stratégie de couplage populaire qui lie les enregistrements par vagues, où chaque vague lie les enregistrements qui n'ont aucun lien dans les vagues précédentes. Par exemple, le couplage peut comprendre une vague déterministe suivie d'une vague probabiliste. Bien que Chipperfield et coll. (2018) proposent un estimateur pour de tels couplages, ils se fondent essentiellement sur l'hypothèse restrictive d'indépendance conditionnelle, qui pourrait ne pas être respectée. Ces travaux cherchent à éviter cette hypothèse en élargissant le modèle que décrivent Dasylva et Goussanou (2020). Les autres sections portent sur la méthodologie, les simulations et la conclusion dans cet ordre.

[1]Abel Dasylva, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A0T6, abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario) K1A0T6.

2. Méthodologie

Parmi les stratégies populaires, il y a le couplage des enregistrements en deux vagues, où la première vague est déterministe (p. ex. un lien où il y a concordance exacte pour chaque variable), et la deuxième vague est un couplage probabiliste des enregistrements, qui n'ont aucun lien dans la première vague. Évidemment, le couplage peut compter plusieurs autres vagues, alors que chaque vague lie les enregistrements qui ne sont pas liés par les vagues précédentes, avec un critère de couplage distinct. Pour évaluer l'exactitude du couplage, une solution naïve repose sur le modèle des voisins univarié (Dasylva et Goussanou, 2020), qui peut produire des estimations biaisées parce que la décision de lier deux enregistrements est une fonction d'autres enregistrements. Une meilleure solution repose sur le modèle des voisins multivarié (Dasylva, Goussanou et Nambeu, 2024).

Les paragraphes suivants présentent les notations et les hypothèses, avant de décrire ces deux approches.

Notations et hypothèses : Ici, on considère le couplage imparfait de deux fichiers sans doublons, qui sont des échantillons de Bernoulli indépendants S_A et S_B d'une population finie avec N unités, où N est connu. Le couplage se fonde sur des quasi-identificateurs qui présentent des coquilles dans chaque fichier. Dans une telle situation, il n'y a aucune certitude concernant les paires d'enregistrements qui proviennent de la même unité. Cela signifie que les enregistrements provenant d'une même unité correspondent dans les deux échantillons par une permutation aléatoire $\Pi(.)$ de $\{1, \dots, N\}$, qui est distribuée uniformément et inconnue. Sans perdre de généralité, l'unité i peut être associée à l'enregistrement V_i dans S_B, et l'enregistrement $V_{\Pi(i)}{}'$ dans S_A; les valeurs d'enregistrement se trouvant dans un espace d'enregistrement $\mathcal{V}_N$. On suppose que l'inclusion des unités dans les échantillons et les valeurs de leurs enregistrements (c.-à-d. $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_{\Pi(i)}{}'\right)\right]_{1 \le i \le N}$) sont indépendantes de $\Pi(.)$. En outre, on suppose que les indicateurs d'inclusion dans les échantillons et les valeurs des enregistrements sont mutuellement indépendants et distribués de manière identique parmi les unités, c.-à-d. on suppose que les uplets $\left(I(1 \in S_A), I(1 \in S_B), V_1, V_{\Pi(1)}{}'\right)$,…,$\left(I(N \in S_A), I(N \in S_B), V_N, V_{\Pi(N)}{}'\right)$ sont iid. En raison des hypothèses concernant $\Pi(.)$, aucune généralité n'est perdue en conditionnant sur le fait que c'est la permutation d'identité (c.-à-d. $\Pi(i) = i$ pour chaque i) dans ce qui suit. Cela permet de simplifier la notation en gardant ce conditionnement implicite, c.-à-d. l'espérance conditionnelle $E[\mathcal{E}|\Pi]$ est simplement désignée par $E[\mathcal{E}]$ pour tout événement $\mathcal{E}$ d'intérêt. Le conditionnement signifie également que l'unité i est associée aux enregistrements V_i et $V_i{}'$.

Le couplage repose sur K vagues, qui se fondent sur un ensemble de critères et de méthodes de couplage diversifiés. Par exemple, dans le cas de trois vagues, la première peut être déterministe, tandis que la deuxième est probabiliste et la dernière est fondée sur l'apprentissage automatique, par exemple le partitionnement en k-moyennes. Dans la vague $k = 1, \dots, K$, on désigne l'indicateur d'un lien entre les enregistrements i et j par $L_{ij}^{(1,k)}$, et on suppose que cette décision ne dépend pas d'autres enregistrements que V_i et $V_j{}'$. Les enregistrements sont liés précisément comme suit. Dans la première vague, V_i et $V_j{}'$ sont liés si $L_{ij}^{(1,1)} = 1$. Dans la deuxième vague, les deux enregistrements sont liés si V_i n'est pas lié dans la première vague (impliquant que $L_{ij}^{(1,1)} = 0$) et $L_{ij}^{(1,2)} = 1$. En général, dans la vague $k \ge 2$, V_i et $V_j{}'$ sont liés si V_i n'est pas lié dans toutes les vagues précédentes et $L_{ij}^{(1,k)} = 1$. Pour ce qui suit, il est commode de définir les décisions $L_{ij}^{(2,1)} = L_{ij}^{(1,1)}$, et pour $k > 2$, $L_{ij}^{(2,k)} = L_{ij}^{(1,k)} \prod_{t=1}^{k-1}\left(1 - L_{ij}^{(1,t)}\right)$, où $L_{ij}^{(2,k)}$ ne dépend que de V_i et $V_j{}'$ pour chaque k, et $L_{ij}^{(2,1)}, \dots, L_{ij}^{(2,K)}$ sont mutuellement exclusives, c.-à-d. $L_{ij}^{(2,k)} L_{ij}^{(2,k\prime)} = 0$ si $k \ne k'$. Soit $L_{ij}^{(2)} = L_{ij}^{(1,1)} \vee \dots \vee L_{ij}^{(1,K)}$ ($\vee$ désignant l'opérateur OU binaire) l'indicateur que V_i et $V_j{}'$ sont liés par une règle dans le premier ensemble de règles, qui est également l'indicateur qu'ils sont liés par une règle dans le deuxième ensemble de règles, puisque $L_{ij}^{(1,1)} \vee \dots \vee L_{ij}^{(1,K)} = L_{ij}^{(2,1)} \vee \dots \vee L_{ij}^{(2,K)}$. Enfin, désignons par L_{ij} la décision définitive de couplage pour V_i et $V_j{}'$. On cherche à évaluer les erreurs pour cette dernière décision, lorsque le couplage se fonde sur des quasi-identificateurs, comme les noms ou les dates.

Selon Fellegi et Sunter (1969), une paire d'enregistrements est *appariée* si les enregistrements proviennent de la même unité. Ensuite, une paire d'enregistrements est classée de manière croisée dans une matrice de confusion, selon qu'elle est appariée ou non et qu'elle est liée ou non, comme le montre le tableau 2-1. Un vrai positif (VP) est une paire

appariée qui est liée; un vrai négatif (VN) est une paire non-appariée qui n'est pas liée; un faux négatif (FN) est une paire appariée qui n'est pas liée; un faux positif (FP) est une paire non-appariée qui est liée, où les erreurs comprennent les deux derniers types de paires.

Tableau 2-1
Matrice de confusion

	Liée	Non-liée
Appariée	VP	FN
Non-appariée	FP	VN

Par commodité, désignons également par VP, VN, FN et FP les nombres de paires des différents types, qui servent à calculer le rappel et la précision, où rappel est la proportion des paires correspondantes qui sont liées (c.-à-d. $VP/(VP+FN)$), et la précision est la proportion de paires liées qui sont appariées (c.-à-d. $VP/(VP+FP)$).

Application du modèle des voisins univarié : Il est possible d'évaluer l'exactitude de la décision L_{ij} avec le modèle des voisins univarié (Dasylva et Goussanou, 2020) en se fondant sur le nombre de liens $n_i = \sum_{j=1}^{N} I(j \in S_A) L_{ij}$ de l'enregistrement i de S_B. Selon ce modèle,

$$n_i \quad \sim \quad \sum_{g=1}^{G} \alpha_g Bernoulli(p_g) * Poisson(\lambda_g), \quad (2.1)$$

où $*$ représente l'opération de convolution, G est le nombre de classes d'enregistrements, et p_g et λ_g sont respectivement les nombres espérés de vrais positifs et de faux positifs par enregistrement dans la classe g, avec $\bar{p} = \alpha_1 p_1 + \cdots + \alpha_G p_G$ et $\bar{\lambda} = \alpha_1 \lambda_1 + \cdots + \alpha_G \lambda_G$ désignant les nombres espérés de vrais positifs et de faux positifs de tout enregistrement (c'est-à-dire de toute classe) dans S_B. Alors, le rappel et la précision correspondent respectivement à $P(i \in S_A)^{-1}\bar{p}$ et à $\bar{p}/(\bar{p}+\bar{\lambda})$ (Dasylva, Goussanou et Nambeu, 2024), qui peuvent être estimés en maximisant la vraisemblance des n_i. Cependant, cette approche est naïve et produit des estimateurs biaisés, car la décision de lier deux enregistrements n'est pas indépendante des autres enregistrements dans ce cas-ci. En effet, on peut le voir facilement dans le contre-exemple simple suivant, où $K = 2$, $L_{ij}^{(1,1)}$ se fonde sur la concordance exacte pour toutes les variables, tandis que $L_{ij}^{(1,2)}$ se fonde sur un critère moins rigide. Puis, V_i est lié à $V_j{}'$ si les deux enregistrements sont identiques, ou si $V_i \neq V_{j'}{}'$ pour chaque $j' \in S_A$ et $L_{ij}^{(2,2)} = 1$. De toute évidence, L_{ij} est une fonction de chaque enregistrement dans S_A. Une meilleure solution est d'appliquer la version multivariée du modèle des voisins (Dasylva et coll., 2024).

Application du modèle des voisins multivarié : Le modèle des voisins multivarié (Dasylva, Goussanou et Nambeu, 2024) a été élaboré pour permettre l'estimation conjointe de l'exactitude du couplage de plusieurs critères de couplage mutuellement exclusifs, où chaque critère lie deux enregistrements indépendamment des autres enregistrements. Il s'agit d'un mélange multivarié fini, dans lequel chaque composante se fonde sur la convolution d'une distribution multinomiale avec une distribution de Poisson composée multivariée (Wang, 1996). De manière générale, une distribution de Poisson composée est paramétrée par un paramètre de Poisson positif λ et une distribution de gravité $h(.)$. Lorsque $h(.)$ est discrète, la fonction de masse de probabilité prend la forme

$$P(X = x) \quad = \quad \sum_{t=0}^{\infty} \frac{\lambda^t}{t!} e^{-\lambda} h^{*t}(x),$$

où $h^{*t}(.)$ est la t-ième convolution de $h(.)$ avec elle-même. Par commodité, la distribution composée multivariée avec paramètre de Poisson λ et distribution de gravité $h(.)$ est désignée par $MCP(\lambda, h(.))$. Le mélange fini multivarié découle d'une convergence en loi, lorsque N devient arbitrairement grand sous des conditions de régularité.

Pour décrire comment le modèle des voisins multivarié s'applique au problème en question, on désigne par $n_i^{(k)}$ le nombre de liens de V_i selon $L_{ij}^{(2,\mathrm{k})}$, et on définit $\boldsymbol{n}_i = [n_i^1 \quad ... \quad n_i^K]^\top$ (aussi désigné par $\left[n_i^{(k)}\right]_{1 \le k \le K}$ par commodité).

Dans le cas de $v \in \mathcal{V}_N$, définissons le voisinage de v comme l'ensemble $\mathcal{B}_N(v)$ des valeurs d'enregistrements de $\mathcal{V}_N$, qui sont liés à v dans une vague donnée, avec une probabilité positive, c.-à-d.

$$\mathcal{B}_N(v) \quad \supset \quad \left\{ v' \in \mathcal{V}_N \; s.t. \; E\left[\sum_{k=1}^{K} L_{ii}^{(2,k)} \middle| (i,j) \in S_B \times S_A, (V_i, V_i{}') = (v, v') \right] > 0 \right\}.$$

Enfin, désignons par $\mathcal{V}_N^*$ le sous-ensemble de valeurs d'enregistrements qui peuvent être observées dans S_B avec une probabilité positive, et étant donné i et j distincts, définissons

$$\begin{aligned} p_N^{(k)}(v) &= E\left[I(i \in S_A) L_{ii}^{(2,k)} \middle| i \in S_B, V_i = v \right], \\ \lambda_N^{(k)}(v) &= E\left[I(j \in S_A) L_{ij}^{(2,k)} \middle| i \in S_B, V_i = v \right], \\ \lambda_N^{(0)}(v) &= P\left(j \in S_A \; et \; V_j{}' \in \mathcal{B}_N(V_i) \middle| i \in S_B, V_i = v \right), \end{aligned}$$

$\boldsymbol{p}_N(v) = \left[p_N^{(k)}(v) \right]_{1 \le k \le K}$, et $\boldsymbol{\lambda}_N(v) = \left[\lambda_N^{(k)}(v) \right]_{1 \le k \le K}$. Le modèle élargi se fonde sur la convergence en loi de $\boldsymbol{n}_i$ dans les conditions suivantes (Dasylva, Goussanou et Nambeu, 2024, Lemma 2).

$$sup_{v \in \mathcal{V}_N^*} (N-1) \lambda_N^{(0)}(v) \quad \le \quad \Lambda, \quad (2.2)$$

$$\left(\boldsymbol{p}_N(V_i), (N-1) \boldsymbol{\lambda}_N(V_i) \right) \quad \overset{d}{\to} \quad F, \quad (2.3)$$

où Λ et F sont indépendants de N, et F est la distribution atomique, qui place la masse de probabilité α_g sur $\left(\boldsymbol{p}_g, \boldsymbol{\lambda}_g \right)$, avec $\boldsymbol{p}_g(v) = \left[p_g^{(k)}(v) \right]_{1 \le k \le K}$, $\boldsymbol{\lambda}_g(v) = \left[\lambda_g^{(k)}(v) \right]_{1 \le k \le K}$ et $g = 1, \dots, G$. Pour détailler cette loi asymptotique, désignons par $\mathbf{0}_K$ le vecteur colonne de dimension K composé de zéros, désignons par $\boldsymbol{e}_k$ le vecteur colonne de dimension K avec un 1 à la position k et des zéros ailleurs et, étant donné $\boldsymbol{z} \in \mathbb{N}^K$, désignons par $\delta_{\boldsymbol{z}}(.)$ la distribution qui place l'ensemble de la masse à $\boldsymbol{z}$. Alors

$$\boldsymbol{n}_i \quad \overset{d}{\to} \quad \sum_{g=1}^{G} \alpha_g p_g(.) * MCP\left(\lambda_g, h_g(.) \right), \quad (2.4)$$

$$p_g(.) \quad = \quad \left(1 - \sum_{k=1}^{K} p_g^{(k)} \right) \delta_{\mathbf{0}_K}(.) + \sum_{k=1}^{K} p_g^{(k)} \delta_{\boldsymbol{e}_k}(.), \quad (2.5)$$

$$h_g(.) \quad = \quad \sum_{k=1}^{K} \left(\frac{\lambda_g^{(k)}}{\lambda_g} \right) \delta_{\boldsymbol{e}_k}(.), \quad (2.6)$$

$$\lambda_g \quad = \quad \lambda_g^{(1)} + \cdots + \lambda_g^{(K)}. \quad (2.7)$$

Il faut souligner que la distribution composée $MCP\left(\lambda_g, h_g(.) \right)$ correspond au produit de K distributions de Poisson indépendantes, tandis que la distribution $p_g(.)$ porte également le nom de *distribution multinomiale incomplète* dans des travaux antérieurs (Dasylva, Goussanou et Nambeu, 2024). Les paramètres du modèle comprennent α_g, $\boldsymbol{p}_g = \left[p_g^{(k)} \right]_{1 \le k \le K}$ et $\boldsymbol{\lambda}_g = \left[\lambda_g^{(k)} \right]_{1 \le k \le K}$ pour chaque g, et ils sont estimés de manière convergente en maximisant la vraisemblance des $\boldsymbol{n}_i$ (Dasylva, Goussanou et Nambeu, 2024, Theorem 2), y compris la sélection de G, qui est choisi selon le minimum du critère d'information d'Akaike.

Avec ces paramètres, le rappel et la précision de $L_{ij}^{(2,k)}$ prennent respectivement la forme de $P(i \in S_A)^{-1} \sum_{g=1}^{G} \alpha_g p_g^{(k)}$ et $\sum_{g=1}^{G} \alpha_g \lambda_g^{(k)}$, pour chacun des K critères de couplage mutuellement exclusifs (c.-à-d. $L_{ij}^{(2,1)}, \dots, L_{ij}^{(2,K)}$), tout en tenant compte de l'hétérogénéité de V_i d'une manière cohérente pour l'ensemble de ces critères. C'est l'avantage principal du modèle multivarié lorsqu'on établit une comparaison avec l'application distincte du modèle univarié (Dasylva et Goussanou, 2020) à $L_{ij}^{(2,k)}$ pour chaque k. Il faut ensuite dériver le rappel et la précision lorsque les enregistrements sont liés dans K vagues successives, comme décrit ci-dessus.

Pour y parvenir, désignons par $TP_i^{(k)}$ et $FP_i^{(k)}$ les nombres de vrais positifs et de faux positifs pour V_i dans la vague k, et désignons par $TP_i = TP_i^{(1)} + \cdots + TP_i^{(K)}$ et $FP_i = FP_i^{(1)} + \cdots + FP_i^{(K)}$ les nombres totaux de vrais positifs et de faux positifs pour le même enregistrement dans toutes les vagues. Il faut noter que $TP_i^{(k)}$ et $FP_i^{(k)}$ ne sont pas les nombre de vrais positifs et de faux positifs sous $L_{ij}^{(2,k)}$. Par commodité, redéfinissons $\bar{p}$ et $\bar{\lambda}$ comme $\bar{p} = \lim_{N\to\infty} E\left[TP_i^{(k)} \middle| i \in S_B\right]$ et $\bar{\lambda} = \lim_{N\to\infty} E\left[FP_i^{(k)} \middle| i \in S_B\right]$, c.-à-d. les valeurs asymptotiques des nombres espérés de vrais positifs et de faux positifs par enregistrement dans S_B dans toutes les vagues. Le lemme suivant donne les expressions pour ces paramètres. La démonstration se trouve dans l'annexe.

Lemme 1 :

$$\bar{p} = \sum_{k=1}^{K}\sum_{g=1}^{G} \alpha_g \left(I(k=1) p_g^{(1)} + I(k>1) p_g^{(k)} \exp\left(-\sum_{t=1}^{k-1} \lambda_g^{(t)} \right) \right), \quad (2.8)$$

$$\bar{\lambda} = \sum_{k=1}^{K}\sum_{g=1}^{G} \alpha_g \left(I(k=1) \lambda_g^{(1)} + I(k>1) \lambda_g^{(k)} \left(1 - \sum_{t=1}^{k-1} p_g^{(t)} \right) \exp\left(-\sum_{t=1}^{k-1} \lambda_g^{(t)} \right) \right). \quad (2.9)$$

Le rappel et la précision de L_{ij} convergent en probabilité vers $P(i \in S_A)^{-1}\bar{p}$ et $\bar{p}/\left(\bar{p} + \bar{\lambda}\right)$, respectivement, sous les conditions de régularité additionnelles suivantes (Dasylva, Goussanou et Nambeu, 2024, Theorem 1 and Corollary 1) pour i et i' distincts de $\{1, \ldots, N\}$.

$$NP(\mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \emptyset | \{i, i'\} \subset S_B) \leq c, \quad (2.10)$$

$$NP(V_{i'} \in \mathcal{B}_N(V_i) | \{i, i'\} \subset S_B, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \emptyset\ ) \leq c, \quad (2.11)$$

où c est une constante positive indépendante de N. L'équation (2.10) signifie que les enregistrements non liés dans S_B ont des voisinages disjoints, avec une probabilité élevée. Cela indique également que les deux enregistrements sont très peu similaires. L'équation (2.11) signifie que les enregistrements liés sont suffisamment similaires. Ainsi, il est possible d'estimer le rappel et la précision de manière convergente à l'aide de $P(i \in S_A)^{-1}\hat{\bar{p}}$ et de $\hat{\bar{p}}/\left(\hat{\bar{p}} + \hat{\bar{\lambda}}\right)$, où $\hat{\bar{p}}$ et $\hat{\bar{\lambda}}$ se fondent sur le maximum de vraisemblance (Dasylva, Goussanou et Nambeu, 2024, Theorem 2).

Jusqu'à maintenant, on a supposé que $L_{ij}^{(1,k)}$ ne dépend que de V_i et de V_j', au lieu de dépendre d'autres enregistrements. Cependant, cette condition n'est pas respectée si la décision de couplage est une fonction de paramètres de décision qui dépendent de tous les enregistrements des deux sources. Par exemple, cette situation se produit lorsque $L_{ij}^{(1,k)}$ se fonde sur la méthode probabiliste, les paramètres de la décision comprenant les poids de couplage et un seuil de poids, lorsqu'il n'y a pas de ressources de vérifications manuelles. C'est également le cas si la décision se fonde sur un partitionnement en deux moyennes, lorsque les paramètres de la décision comprennent les centres des deux grappes. Néanmoins, exactitude du couplage peut être estimée de façon convergente si les paramètres de décision sont estimés de façon convergente.

3. Simulations

On évalue la méthodologie à l'aide de simulations qui se fondent sur 100 répétitions. Lors de chaque répétition, une population fictive est produite, comprenant un nom de famille et une date de naissance pour chaque personne, alors que les distributions de ces variables se fondent sur les chiffres du recensement américain de 2010 (Bureau du recensement américain, 2016, 2020). À partir de cette population finie, deux fichiers sont créés en tirant des échantillons de Bernoulli indépendants avec la même probabilité d'inclusion de 0,9, dans lesquels le nom de famille et la date de naissance sont perturbés, comme suit. La chaîne de caractère du nom de famille est modifiée par plusieurs coquilles, où le nombre de coquilles respecte la loi de Poisson avec une moyenne de 0,1. Étant donné plusieurs coquilles, les différentes coquilles sont appliquées en séquence, c'est-à-dire qu'une coquille est appliquée au résultat de la modification du nom de famille par les coquilles précédentes. Chaque coquille est stochastique avec une probabilité égale d'être une suppression ou une insertion, à une position uniformément aléatoire dans la chaîne. Dans

le cas d'une insertion, la lettre insérée est tirée uniformément de l'alphabet. Dans le cas de la date de naissance, des perturbations indépendantes sont appliquées à l'année d'une part et au jour et au mois d'une autre part. L'année est inscrite sans coquilles avec une probabilité de 0,9. Sinon, elle est augmentée ou diminuée de 1 avec une probabilité égale pour chaque option. Dans le même ordre d'idées, le jour et le mois sont inscrits sans coquilles avec une probabilité de 0,9. Sinon, ces deux composantes de la date sont transposées.

Le couplage est également réalisé dans SAS. Il se fonde sur deux vagues, où la première vague est basée sur le fait d'avoir le même nom de famille et la même date de naissance, et la deuxième vague se fonde sur la méthode probabiliste. Pour cette deuxième vague, un simple couplage probabiliste est réalisé, qui tient compte des pratiques actuelles, y compris l'hypothèse de travail de l'indépendance conditionnelle (c'est-à-dire l'hypothèse selon laquelle les concordances pour différentes variables sont indépendantes étant donné qu'une paire est appariée ou non). Dans le cas de ce couplage, les pochettes se fondent sur le fait d'avoir la même année et de satisfaire au moins un des critères suivants : le même nom de famille et le même jour, le même nom de famille et le même mois, la même première lettre du nom de famille et le même jour et le même mois, ou la même première lettre du nom de famille et une concordance croisée pour le jour et le mois. Par concordance croisée pour le jour et le mois, on entend que le jour du premier enregistrement concorde avec le mois du deuxième enregistrement, et que le mois du premier enregistrement concorde avec le jour du deuxième enregistrement. Le couplage probabiliste est simple et se fonde sur trois comparaisons, à savoir si la similarité Jaro-Winkler du nom de famille est supérieure à 0,8, si le jour est le même et si le mois est le même. Les paramètres de couplage sont estimés numériquement avec SAS OPTMODEL en fonction de l'hypothèse d'indépendance conditionnelle. Enfin, une paire d'enregistrements est liée lorsque la probabilité d'appariement conditionnelle estimée (c.-à-d. la probabilité conditionnelle qu'une paire est appariée en raison des résultats de comparaison observés) est égale ou supérieure à $1/2$, afin de minimiser une combinaison linéaire des faux positifs et des faux négatifs. Pourtant, la décision qui en découle est susceptible d'être sous-optimale parce que l'hypothèse relative à l'indépendance conditionnelle n'est pas respectée et que la probabilité d'appariement conditionnelle est estimée avec un certain biais. Toutefois, cela ne nuit pas à l'atteinte de notre objectif principal, qui est d'évaluer notre capacité à mesurer avec précision l'exactitude du couplage réalisé, indépendamment de son optimalité. En outre, la méthodologie proposée ne repose pas sur l'hypothèse d'indépendance conditionnelle. La procédure d'estimation des erreurs est exécutée dans R, en fonction de la version restreinte du modèle multivarié, où $p_1^{(k)} = \cdots = p_G^{(k)} = p^{(k)}$. À titre de comparaison, l'exactitude du couplage est également estimée naïvement avec le modèle univarié restreint, en fonction de la contrainte $p_1 = \cdots = p_G = p$. Dans les simulations, la taille de la population passe de 100 000 à 200 000. En augmentant la taille de la population, on augmente le nombre de faux positifs et on diminue la précision. Les résultats figurent dans le tableau 3-1. Bien que la variance soit inférieure dans le modèle univarié, le biais relatif et l'erreur quadratique moyenne (EQM) sont beaucoup plus petits avec l'estimateur proposé. La baisse de l'EQM est plutôt considérable lorsqu'on passe du modèle univarié au modèle multivarié. En outre, les résultats montrent que l'EQM augmente avec N dans le cas de chaque estimateur. Dans l'ensemble, le modèle multivarié affiche une meilleure performance que le modèle univarié, comme on s'y attend.

Tableau 3-1
Résultats de la simulation

		$N = 100,000$			$N = 200,000$		
Mesure	**Estimateur**	**Biais relatif (en %)**	**Variance ($\times 10^{-6}$)**	**EQM ($\times 10^{-3}$)**	**Biais relatif (en %)**	**Variance ($\times 10^{-6}$)**	**EQM ($\times 10^{-3}$)**
Précision	Univarié	7,92	1,42	4,79	14,11	1,38	11,96
	Multivarié	-1,79	3,90	0,25	-3,63	2,55	0,79
Rappel	Univarié	7,92	2,87	2,83	14,11	1,85	8,96
	Multivarié	-1,92	5,11	0,17	-3,89	3,44	0,68

4. Conclusions

Pour les applications de données couplées aux statistiques officielles, il est essentiel de mesurer l'exactitude du couplage. Cela est également vrai lorsqu'on utilise une stratégie populaire, qui consiste à lier les enregistrements par vagues, chaque vague liant un sous-ensemble des enregistrements qui ne sont pas liés dans toutes les vagues précédentes. Cependant, l'évaluation de l'exactitude résultante a été un défi jusqu'à ce jour, car la décision de lier deux

enregistrements dépend d'autres enregistrements. Dans cet article, ce problème est finalement résolu en modélisant le nombre de liens d'un enregistrement, sans supposer l'indépendance conditionnelle et sans nécessiter de vérifications manuelles. Dans de futurs travaux, cette approche sera étendue à d'autres stratégies de couplage.

Bibliographie

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

Chipperfield, J., Hansen, N., et Rossiter, P. (2018), "Estimating precision and recall for deterministic and probabilistic record linkage", *International Statistical Review*, 86, p. 219-236.

Dasylva, A. et Goussanou, A. (2020), "Estimating linkage errors under regularity conditions", *Proceedings of the Survey Methods Section, American Statistical Association*, p. 687-692. Accessible à l'adresse http://www.asasrms.org/Proceedings/y2020/files/1505346.pdf.

Dasylva, A., Goussanou, A., et Nambeu, C.-O. (2024), "Models of linkage error for capture-recapture estimation without clerical reviews", *Survey Methodology*, 50, p. 375-408.

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

US Census Bureau. (2016), File b: surnames occurring 100 or more times. https://www2.census.gov/topics/genealogy/2010surnames/names.zip. , Consulté : 2020-10-17.

US Census Bureau. (2020), Annual state resident population estimates for 6 race groups (5 race alone groups and two or more races) by age, sex, and Hispanic origin: April 1, 2010 to July 1, 2019. https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/asrh/sc-est2019-alldata6.csv. , Consulté : 2020-10-17.

Wang, Y.H. (1996), "On the multivariate compound distributions", *Journal of Multivariate Analysis*, 59, p. 13-21.

Annexe

Démonstration du lemme 1 : Dans la première vague, un vrai positif se produit si V_i et V_i' sont liés, et un faux positif se produit si V_i et V_j' sont liés étant donné $j \neq i$. Par conséquent,

$$\begin{aligned} E\left[TP_i^{(1)} \middle| i \in S_B, V_i = v\right] &= p_N^{(1)}(v), \\ E\left[FP_i^{(1)} \middle| i \in S_B, V_i = v\right] &= (N-1)\lambda_N^{(1)}(v), \end{aligned}$$

et

$$\begin{aligned} E\left[TP_i^{(1)} \middle| i \in S_B\right] &= E\left[p_N^{(1)}(V_i) \middle| i \in S_B\right] \\ &= \int p^{(1)} dF(\boldsymbol{p}, \boldsymbol{\lambda}) + \int p^{(1)} d\left(F_N(\boldsymbol{p}, \boldsymbol{\lambda}) - F(\boldsymbol{p}, \boldsymbol{\lambda})\right) \\ &= \sum_{g=1}^{G} \alpha_g p_g^{(1)} + \underbrace{\int p^{(1)} d\left(F_N(\boldsymbol{p}, \boldsymbol{\lambda}) - F(\boldsymbol{p}, \boldsymbol{\lambda})\right)}_{o(1)}. \end{aligned}$$

D'une manière similaire

$$
\begin{aligned}
E\left[FP_i^{(1)} \middle| i \in S_B\right] &= E\left[(N-1)\lambda_N^{(1)}(V_i) \middle| i \in S_B\right] \\
&= \int \lambda^{(1)} dF(\boldsymbol{p}, \boldsymbol{\lambda}) + \int \lambda^{(1)} d\left(F_N(\boldsymbol{p}, \boldsymbol{\lambda}) - F(\boldsymbol{p}, \boldsymbol{\lambda})\right) \\
&= \sum_{g=1}^{G} \alpha_g \lambda_g^{(1)} + \underbrace{\int \lambda^{(1)} d\left(F_N(\boldsymbol{p}, \boldsymbol{\lambda}) - F(\boldsymbol{p}, \boldsymbol{\lambda})\right)}_{o(1)}.
\end{aligned}
$$

Dans la vague $k \geq 2$, un vrai positif se produit si $I(i \in S_A) L_{ii}^{(2,k)} = 1$ et $I(j \in S_A) \sum_{t=1}^{k-1} L_{ij}^{(2,t)} = 0$ pour chaque $j \neq i$. En outre, il existe une indépendance mutuelle entre les variables qui composent $I(i \in S_A) L_{ii}^{(2,k)}$ et $I(j \in S_A) \sum_{t=1}^{k-1} L_{ij}^{(2,t)}$ pour chaque $j \neq i$, en conditionnant par l'événement $\{\mathrm{i} \in S_B\} \cap \{V_i = v\}$. Ainsi

$$
E\left[TP_i^{(k)} \middle| i \in S_B, V_i = v\right] = p_N^{(k)}(v) \left(1 - \sum_{t=1}^{k-1} \lambda_N^{(t)}(v)\right)^{N-1}.
$$

et

$$
\begin{aligned}
E\left[TP_i^{(k)} \middle| i \in S_B\right] &= E\left[p_N^{(k)}(V_i) \left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} (N-1) \lambda_N^{(t)}(V_i)\right)^{N-1} \middle| i \in S_B \right] \\
&= \int p^{(k)} \exp\left(-\sum_{t=1}^{k-1} \lambda^{(t)}\right) dF(\boldsymbol{p}, \boldsymbol{\lambda}) + \\
&\quad \int p^{(k)} \underbrace{\left(\left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)}\right)^{N-1} - \exp\left(-\sum_{t=1}^{k-1} \lambda^{(t)}\right) \right)}_{O(1/N)} dF(\boldsymbol{p}, \boldsymbol{\lambda}) + \\
&\quad \underbrace{\int p^{(k)} \left(1 - \frac{1}{N-1} \sum_{t=1}^{k-1} \lambda^{(t)}\right)^{N-1} d\left(F_N(\boldsymbol{p}, \boldsymbol{\lambda}) - F(\boldsymbol{p}, \boldsymbol{\lambda})\right)}_{o(1)} \\
&= \sum_{g=1}^{G} \alpha_g p_g^{(k)} \exp\left(-\sum_{t=1}^{k-1} \lambda_g^{(t)}\right) + o(1).
\end{aligned}
$$

Dans la même vague, pour i et j distincts, un faux positif entre V_i et $V_j{}'$ se produit si $I(j \in S_A) L_{ij}^{(2,k)} = 1$, $I(i \in S_A) \sum_{t=1}^{k-1} L_{ii}^{(2,t)} = 0$ et $I(j' \in S_A) \sum_{t=1}^{k-1} L_{ij'}^{(2,t)} = 0$, pour chaque $j' \neq i, j$, avec l'indépendance mutuelle parmi les variables qui comprennent $I(j \in S_A) L_{ij}^{(2,k)}$, $I(i \in S_A) \sum_{t=1}^{k-1} L_{ii}^{(2,t)}$ et $I(j' \in S_A) \sum_{t=1}^{k-1} L_{ij'}^{(2,t)} = 0$, pour chaque $j' \neq i, j$, en conditionnant par l'événement $\{\mathrm{i} \in S_B\} \cap \{V_i = v\}$. Donc

$$
E\left[FP_i^{(k)} \middle| i \in S_B, V_i = v\right] = (N-1) \lambda_N^{(k)}(v) \left(1 - \sum_{t=1}^{k-1} p_N^{(t)}(v)\right) \left(1 - \sum_{t=1}^{k-1} \lambda_N^{(t)}(v)\right)^{N-2},
$$

et

$$
\begin{aligned}
E\left[FP_i^{(k)} \middle| i \in S_B\right] &= E\left[(N-1)\lambda_N^{(k)}(v)\left(1-\sum_{t=1}^{k-1} p_N^{(t)}(v)\right)\left(1-\sum_{t=1}^{k-1}\lambda_N^{(t)}(v)\right)^{N-2} \middle| i \in S_B\right] \\
&= \int \lambda^{(k)}\left(1-\sum_{t=1}^{k-1} p^{(t)}\right)\exp\left(-\sum_{t=1}^{k-1}\lambda^{(t)}\right)dF(\boldsymbol{p},\boldsymbol{\lambda}) + \\
&\quad \int \lambda^{(k)}\left(1-\sum_{t=1}^{k-1} p^{(t)}\right)\underbrace{\left(\left(1-\frac{1}{N-1}\sum_{t=1}^{k-1}\lambda^{(t)}\right)^{N-2} - \exp\left(-\sum_{t=1}^{k-1}\lambda^{(t)}\right)\right)}_{O(1/N)} dF(\boldsymbol{p},\boldsymbol{\lambda}) + \\
&\quad \underbrace{\int \lambda^{(k)}\left(1-\sum_{t=1}^{k-1} p^{(t)}\right)\left(1-\frac{1}{N-1}\sum_{t=1}^{k-1}\lambda^{(t)}\right)^{N-2} d\left(F_N(\boldsymbol{p},\boldsymbol{\lambda}) - F(\boldsymbol{p},\boldsymbol{\lambda})\right)}_{o(1)} \\
&= \sum_{g=1}^{G}\alpha_g\lambda_g^{(k)}\left(1-\sum_{t=1}^{k-1} p_g^{(t)}\right)\exp\left(-\sum_{t=1}^{k-1}\lambda_g^{(t)}\right) + o(1).
\end{aligned}
$$

QED