Nº 11-522-x au catalogue
ISSN : 1709822-X

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Mesure du sous-dénombrement de deux sources de données dont la couverture est presque parfaite grâce à la capture et à la recapture en présence d'erreurs de couplage

par A. Dasylva, A. Goussanou et C.O. Nambeu

Date de diffusion : le 22 octobre 2021

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2021 de Statistique Canada
Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société

Mesure du sous-dénombrement de deux sources de données dont la couverture est presque parfaite grâce à la capture et à la recapture en présence d'erreurs de couplage

A. Dasylva, A., A. Goussanou et C.O. Nambeu[1]

Résumé

Dans le contexte de son paradigme « données administratives d'abord », Statistique Canada donne la priorité à l'utilisation de sources autres que les enquêtes pour produire des statistiques officielles. Ce paradigme repose de façon capitale sur des sources autres que les enquêtes pouvant fournir une couverture quasi parfaite de certaines populations cibles, y compris des fichiers administratifs ou des sources de mégadonnées. Toutefois, cette couverture doit être mesurée, en appliquant par exemple la méthode de capture-recapture, selon laquelle les données sont comparées à d'autres sources présentant une bonne couverture des mêmes populations, y compris un recensement. Cependant, il s'agit d'un exercice difficile en présence d'erreurs de couplage, qui surviennent inévitablement lorsque le couplage se fonde sur des quasi-identificateurs, comme cela est généralement le cas. Pour faire face à cet enjeu, une nouvelle méthodologie est décrite, selon laquelle la méthode de capture-recapture est améliorée grâce à un nouveau modèle d'erreur fondé sur le nombre de couplages contigus à un enregistrement donné. Elle est appliquée dans le cadre d'une expérience avec des données publiques de recensement.

Mots clés : estimation de système dual; appariement de données; couplage d'enregistrements; qualité; intégration des données; mégadonnées.

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

1. Introduction

Les statistiques officielles reposent de plus en plus sur des données ne provenant pas d'enquêtes, comme des données administratives et des mégadonnées. Il est toutefois essentiel de mesurer la couverture de ces sources, par exemple quand on fournit des indicateurs de qualité pour une source donnée ou quand on estime des tailles de population à partir de données administratives. À cette fin, on peut utiliser la méthode de capture-recapture, qui s'appuie sur de nombreuses hypothèses, y compris celle d'un couplage sans erreur. Dans la pratique, cette hypothèse doit être assouplie, malgré la difficulté à essayer de prendre en compte les erreurs de couplage, y compris les faux négatifs et les faux positifs, qui font qu'un faux négatif ne permet pas de coupler les enregistrements d'une même unité de population et qu'un faux positif couple des enregistrements de différentes unités.

Dans leurs travaux, Ding et Fienberg (1994) ont abordé le problème dans le contexte d'une estimation de système dual fondée sur un recensement et une enquête sur la couverture, quand on suppose que les faux positifs n'impliquent pas les unités (c.-à-d. leurs enregistrements) qui sont sélectionnées dans les deux sources, avec une probabilité connue que l'unité ait au moins un faux positif, étant donné qu'il est inclus seulement dans l'enquête sur la couverture. Par conséquent, un faux positif ne se produit qu'entre les unités sélectionnées soit par le recensement soit par l'enquête sur la couverture, et non par les deux. Di Consiglio et Tuoto (2015) ont étendu cette solution à une estimation de système dual fondée sur deux listes administratives, selon la même hypothèse au sujet des unités incluses dans les deux listes, tout en tenant compte de la probabilité connue que l'unité ait au moins un faux positif, étant donné qu'il

[1]Abel Dasylva, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6, abel.dasylva@statcan.gc.ca; Arthur Goussanou, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6; Christian-Olivier Nambeu, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa (Ontario), K1A 0T6

est inclus dans une seule liste. Dans la pratique, il peut être difficile de satisfaire l'hypothèse de l'absence de faux positifs pour les unités sélectionnées dans les deux listes. En outre, il est difficile d'estimer la probabilité qu'une unité ait au moins un faux positif, étant donné qu'il est inclus dans une seule liste. Cette difficulté demeure même quand le rappel et la précision sont donnés. En effet, il n'est pas aisé de calculer la probabilité d'intérêt à partir de ces mesures d'exactitude, car elles s'appliquent aux paires d'enregistrements, alors que cette probabilité concerne toutes les paires formées entre un enregistrement et tous les enregistrements de l'autre liste. Račinskij et coll. (2019) ont décrit une solution pour les situations où les variables de couplage satisfont aux hypothèses d'indépendance conditionnelle et que les erreurs de couplage sont estimées au moyen d'un mélange log-linéaire. Une solution proche s'appuie sur deux décisions de couplage qui sont présumées indépendantes pour les enregistrements provenant des mêmes unités (voir Brown et coll., 2020). Cependant, les estimateurs obtenus peuvent être biaisés, quand les décisions de couplage s'écartent de cette hypothèse. Cela s'explique par problème du biais de corrélation décrit précédemment par Newcombe (1983, chap. E.6, p. 149), Belin et Rubin (1995) et Blakely et Salmond (2002).

Le présent article décrit une nouvelle méthodologie pour résoudre le problème de la capture-recapture avec des erreurs de couplage, tout en assouplissant l'hypothèse selon laquelle il n'y a pas de faux positifs pour les unités capturées dans les deux listes et l'hypothèse selon laquelle les variables de couplage sont conditionnellement indépendantes. La nouvelle méthodologie donne une borne inférieure à la couverture de chaque liste. Elle produit également un estimateur de la couverture réelle si l'on peut supposer qu'il n'y a pas de faux négatifs ou si les deux listes sont couplées par plusieurs décisions de couplage, où le rappel de chaque décision est donné par un modèle paramétrique commun. La méthodologie est une adaptation du modèle d'erreur de Dasylva et Goussanou (2020) et un exemple d'« estimation de système dual sans couplage » (Račinskij et coll., 2019), parce qu'aucun fichier de paires couplées n'est produit.

Les autres sections sont organisées comme suit. La prochaine section donne quelques éléments de contexte. Elle est suivie de sections décrivant la méthodologie proposée et l'expérience sur des données publiques de recensement, dans l'ordre. La dernière section présente une conclusion et les futurs travaux.

2. Contexte

La configuration élémentaire de capture-recapture comprend une population finie U avec une taille de population inconnue N, deux listes indépendantes s_A et s_B tirées de cette population, où la probabilité d'inclusion dans s_A est uniforme. L'objectif est d'estimer la taille de population N et la couverture de chaque liste. Avec la méthodologie de Petersen (1896) et Lincoln (1930), la taille de population et la couverture des deux listes sont estimées comme suit.

$$\begin{aligned} \widehat{N} &= |s_A||s_B|/|s_A \cap s_B|, \\ \hat{P}(i \in s_A) &= |s_A \cap s_B|/|s_B|, \\ \hat{P}(i \in s_B) &= |s_A \cap s_B|/|s_A|. \end{aligned}$$

Cependant, ces estimateurs reposent sur les hypothèses clés suivantes qui doivent être assouplies en pratique.

1. *Population fermée* : ce qui signifie l'absence de naissances, de décès et de migration.

2. *Listes indépendantes* : l'inclusion dans la première liste est indépendante de l'inclusion dans la deuxième liste.

3. *Unités indépendantes* : chaque unité est incluse dans la première ou la deuxième liste indépendamment des autres unités.

4. *Capture homogène* : probabilité d'inclusion uniforme pour au moins une liste.

5. *Absence de surdénombrement* : aucune unité en double ni hors du champ de l'enquête dans chaque liste.

6. *Couplage parfait* : aucune erreur d'identification des unités dans les deux listes.

Il faut assouplir le couplage parfait quand les variables de couplage ne sont pas uniques et comportent des coquilles, de sorte que des erreurs de couplage se produisent, y compris des faux positifs et des faux négatifs. Un faux positif entraîne la surestimation de l'intersection, tandis qu'un faux négatif entraîne la sous-estimation de cette intersection. Ces erreurs sont mesurées par le rappel et la précision, le rappel étant la proportion de couplage entre les paires d'enregistrements qui comprennent des enregistrements de la même unité, tandis que la précision est la proportion de paires d'enregistrements de la même unité dans les couplages.

Ding et Fienberg (1994) et Di Consiglio et Tuoto (2015) ont décrit des solutions de capture-recapture avec erreurs de couplage, quand il n'y a pas de faux positifs pour les unités sélectionnées dans les deux listes, avec une probabilité connue que l'unité ait au moins un faux positif, étant donné qu'il est inclus dans une seule liste. Toutefois, ces exigences sont difficiles à respecter en pratique. Račinskij et coll. (2019) ont décrit une solution différente, qui suppose l'indépendance conditionnelle des variables de couplage et estime la taille de population et la couverture sans produire un fichier de paires couplées. Toutefois, l'hypothèse d'indépendance conditionnelle peut ne pas s'appliquer, ce qui entraîne un certain biais.

3. Méthodologie

La méthodologie proposée vise à assouplir l'hypothèse de l'absence de faux positifs pour les unités des deux listes et celle de variables de couplage conditionnellement indépendantes. Il s'agit d'un autre exemple d'« estimation de système dual sans couplage » qui adapte le modèle d'erreur de Dasylva et Goussanou (2020). Cette méthodologie donne une borne inférieure à la couverture de chaque liste. Elle produit également des estimateurs convergents de la couverture des listes si l'on peut supposer qu'il n'y a pas de faux négatifs ou si les fichiers sont couplés par plusieurs décisions de couplage, où le rappel de chaque décision est donné par un modèle paramétrique partagé.

Obtenir une borne inférieure sur la couverture : Pour motiver la borne inférieure de la couverture, il faut d'abord observer que la couverture de s_A n'est pas inférieure à la probabilité conjointe conditionnelle que l'unité i soit dans s_A et un vrai positif (VP) étant donné que l'unité est dans s_B, où un vrai positif couple deux enregistrements de la même unité et que le rappel correspond à la probabilité conditionnelle d'un VP étant donné que l'unité est dans les deux listes, c.-à-d. $P(a\,TP\,|i \in s_B \cap s_A)$.

$$P(i \in s_A) \quad = \quad P(i \in s_A\,|i \in s_B) \quad \geq \quad P(i \in s_A \; and \; a\,TP\,|i \in s_B).$$

À partir de cette borne inférieure, on peut aisément calculer une borne supérieure sur la taille de population et une borne inférieure sur la couverture de s_B. On peut estimer la borne inférieure sur la couverture de s_A en réutilisant le modèle de mélange fini décrit par Dasylva et Goussanou (2020), tout en donnant une nouvelle signification aux paramètres du modèle. À cette fin, la première étape consiste à souligner la relation importante entre le nombre de couplages n_i de l'enregistrement $i \in s_B$ et les erreurs de couplage impliquant cet enregistrement, comme le montre le tableau 3-1. Ce tableau diffère du tableau similaire décrit par Dasylva et Goussanou (2020) parce que chaque liste comporte un certain sous-dénombrement. Quand $n_i = 0$, il n'y a aucun ou un faux négatif parce que l'unité peut être à l'extérieur de s_A ou qu'elle peut être sur s_A mais avec un faux négatif, sans qu'on puisse savoir quel cas s'applique. Cependant, on sait qu'il n'y a pas de faux positifs. Quand n_i est positif, il n'y a toujours pas de faux négatif ou un

seul, mais il y a $n_i - 1$ ou n_i faux positifs, ce qui fournit beaucoup d'informations utiles. Il est clair qu'un modèle statistique est nécessaire quand il subsiste une certaine incertitude, c.-à-d. quand n_i est positif.

Tableau 3-1
Voisins et erreurs.

n_i	Faux négatifs	Faux positifs
0	0 ou 1	0
$1 \leq n_i \leq \lvert s_B \rvert - 1$	0 ou 1	$n_i - 1$ ou n_i

Le modèle se présente comme une limite de distribution dans des conditions de régularité légèrement différentes de celles données précédemment par Dasylva et Goussanou (2020). Afin de détailler ces conditions, supposons que les deux listes sont des échantillons tirés de registres théoriques A et B, et supposons que l'unité i est associée à l'enregistrement v_i dans B et à l'enregistrement $v'_{\pi(i)}$ dans A, pour une permutation aléatoire uniforme $\pi(.)$ de $\{1, \dots, N\}$. Les processus d'inclusion de liste et d'enregistrement sont supposés être tels que $\left[\left(I(i \in s_A), , I(i \in s_B), v_i, v'_{\pi(i)}\right)\right]_{1 \leq i \leq N}$ sont indépendants et identiquement distribués et indépendants de la permutation aléatoire $\pi(.)$. On suppose que les enregistrements prennent leurs valeurs à partir d'un ensemble fini mais potentiellement grand $\mathcal{V}$. La décision de coupler deux enregistrements est telle qu'elle est caractérisée par une fonction multivaluée $\mathcal{B}(.)$ à partir de $\mathcal{V}$ dans l'ensemble de puissance $2^{\mathcal{V}}$, c'est-à-dire l'ensemble de tous les sous-ensembles de $\mathcal{V}$, de sorte que v_i soit couplé à v'_j si et seulement si $v'_j \in \mathcal{B}(v_i)$. Pour $v \in \mathcal{V}$, désignons par $\mathcal{B}(v)$ le voisinage de v et définissons les probabilités conditionnelles

$$\begin{aligned} p_N(v) &= P\left(i \in s_A \text{ and } v'_{\pi(i)} \in \mathcal{B}(v_i) | i \in s_B \text{ and } v_i = v\right), \\ \lambda_N(v) &= P\left(i' \in s_A \text{ and } v'_{\pi(i')} \in \mathcal{B}(v_i) | i \in s_B \text{ and } v_i = v\right), \quad i' \neq i, \end{aligned}$$

où les fonctions $p_N(.)$ et $\lambda_N(.)$ incorporent les mécanismes d'inclusion dans les deux listes avec un chevauchement partiel, contrairement aux fonctions connexes dans Dasylva et Goussanou (2020), où les deux sources sont des registres complets, ou Dasylva et Goussanou (2021), où une source est un registre complet et l'autre un fichier. Pour $i \in s_B$, supposons que

$$n_i = \sum_{j \in s_A} I\left(v'_j \in \mathcal{B}(v_i)\right),$$

désigne le nombre de voisins. Avec la notation ci-dessus, un vrai positif correspond à l'événement $\left\{i \in s_A \text{ and } v'_{\pi(i)} \in \mathcal{B}(v_i)\right\} \cap \{i \in s_B\}$, un faux positif correspond à l'événement $\left\{i' \in s_A \text{ and } v'_{\pi(i')} \in \mathcal{B}(v_i)\right\} \cap \{i \in s_B\}$ avec $i' \neq i$, et

$$P(i \in s_A \text{ et a } TP \mid i \in s_B) = P\left(i \in s_A \text{ et } v'_{\pi(i)} \in \mathcal{B}(v_i) | i \in s_B\right).$$

Avec les définitions plus précises de $p_N(.)$, $\lambda_N(.)$ et n_i, on peut supposer les mêmes conditions de régularité que celles de Dasylva et Goussanou (2020), c'est-à-dire un e fonction constante par morceaux $\left(p_N(.), (N-1)\lambda_N(.)\right)$, avec un nombre fini de niveaux, un nombre attendu borné de faux positifs (c.-à-d. $sup_{v \in \mathcal{V}}(N-1)\lambda_N(v) \leq \Lambda$ pour certains Λ positifs) et une distribution conjointe de $p_N(v_i)$ et $(N-1)\lambda_N(v_i)$ qui est invariante par rapport à N. Ensuite, au moyen d'arguments similaires à ceux avancés par Dasylva et Goussanou (2020), on peut montrer que

$$n_i | i \in s_B \xrightarrow{d} \sum_{g=1}^{G} \alpha_g Bernoulli(p_g) * Poisson(\lambda_g),$$

où * désigne l'opération de convolution. Alors $P(i \in s_A \; et \; a \; TP \; | i \in s_B) \rightarrow \bar{p} = \sum_{g=1}^{G} \alpha_g p_g$, où on peut estimer les paramètres $\left[(\alpha_g, p_g, \lambda_g)\right]_{1 \leq g \leq G}$ en maximisant la vraisemblance composite des n_i, avec G déterminé par la minimisation du critère d'information d'Akaike.

L'estimateur de la borne inférieure de la couverture est attrayant parce qu'il s'applique quelle que soit la structure de corrélation des variables de couplage et qu'il n'exige pas d'examens manuels. Il est par conséquent peu coûteux. Il présente de l'intérêt pour les applications nécessitant une ou deux sources pour avoir une couverture minimale, ou si l'on vérifie que l'union de plusieurs listes indépendantes fournit une couverture presque complète de la population cible. Par exemple, quand on établit une base de sondage à partir de L listes indépendantes, où la liste l a une couverture d'au moins ε_l, alors l'union des listes a une couverture d'au moins $1 - \prod_{l=1}^{L} \varepsilon_l$ (p. ex. 99,22 % quand $\varepsilon_l = 1/2$ et $L = 7$), ce qui permet de vérifier qu'il y a suffisamment de listes pour l'objectif souhaité. Bien entendu, les enregistrements en double (c.-à-d. les enregistrements d'une même unité, qu'ils soient identiques ou non) doivent être identifiés et pris en compte lors de la production d'estimations, par exemple au moyen de la technique de comptage partiel de Zhang (2019). Il reste pourtant intéressant d'estimer la couverture réelle de chaque liste.

Estimation de la couverture réelle : La couverture est évidemment un estimateur de la couverture réelle si l'on peut supposer qu'il n'y a pas de faux négatifs, c.-à-d. en supposant que $P(a \; TP \; | i \in s_B \cap s_A) = 1$ ou $p_N(v) = 1$ sur $\mathcal{V}$. La couverture réelle peut également être estimée à partir de plusieurs décisions de couplage et d'un modèle paramétrique commun pour les rappels de ces différentes décisions. Pour être précis, considérons Γ décisions et supposons que $\mathcal{B}^{(\gamma)}(.)$ désigne la fonction multivaluée associée à la décision $\gamma = 1, \ldots, \Gamma$, de sorte que v_i est couplé à v_j' par cette décision, si $v_j' \in \mathcal{B}^{(\gamma)}(v_i)$. Soit $n_i^{(\gamma)}$ le nombre correspondant de voisins. Chaque décision est supposée satisfaire aux conditions de régularité données ci-dessus, de sorte que :

$$n_i^{(\gamma)} | i \in s_B \xrightarrow{d} \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} Bernoulli\left(p_g^{(\gamma)}\right) * Poisson\left(\lambda_g^{(\gamma)}\right)$$

et la borne inférieure de la couverture peut être estimée de la façon décrite ci-dessus. Cette borne inférieure est notée $\bar{p}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} p_g^{(\gamma)}$ et on suppose qu'elle a la forme $\bar{p}^{(\gamma)} = P(i \in s_A \; | i \in s_B) r^{(\gamma)}(\boldsymbol{\beta})$, où $r^{(\gamma)}(\boldsymbol{\beta})$ est le rappel, $r^{(\gamma)}(.)$ étant une fonction connue et $\boldsymbol{\beta}$ un paramètre à d dimensions qui est commun à toutes les décisions, où $d \leq \Gamma - 1$. La couverture peut alors être estimée par la méthode des moments suivante. Soit $\hat{\bar{p}}^{(\gamma)}$ la borne inférieure estimée de la décision γ. Alors, $\boldsymbol{\beta}$ peut être estimé par la solution du système d'équations suivant.

$$\left(\sum_{\gamma=1}^{\Gamma} \hat{\bar{p}}^{(\gamma)}\right)^{-1} \begin{bmatrix} \hat{\bar{p}}^{(1)} \\ \vdots \\ \hat{\bar{p}}^{(\Gamma)} \end{bmatrix} \quad = \quad \left(\sum_{\gamma=1}^{\Gamma} r^{(\gamma)}(\hat{\boldsymbol{\beta}})\right)^{-1} \begin{bmatrix} r^{(1)}(\hat{\boldsymbol{\beta}}) \\ \vdots \\ r^{(\Gamma)}(\hat{\boldsymbol{\beta}}) \end{bmatrix}.$$

Par conséquent, on peut estimer la couverture par

$$\hat{P}(i \in s_A \; | i \in s_B) \quad = \quad \frac{\sum_{\gamma=1}^{\Gamma} \hat{\bar{p}}^{(\gamma)}}{\sum_{\gamma=1}^{\Gamma} r^{(\gamma)}(\hat{\boldsymbol{\beta}})}.$$

En général, les Γ décisions peuvent être construites à partir de K décisions élémentaires basées sur $\mathcal{B}_1(.),\ldots, \mathcal{B}_K(.)$ avec $\mathcal{B}^{(\gamma)}(.)$ caractérisées par un sous-ensemble non vide $S^{(\gamma)}$ de $\{1, \ldots, K\}$, de telle sorte que

$$\mathcal{B}^{(\gamma)}(v) \quad = \quad \left(\bigcap_{k\in S^{(\gamma)}} \mathcal{B}_k(v)\right) \cap \left(\bigcap_{k\in\{1,\ldots,K\}-S^{(\gamma)}} \mathcal{B}_k(v)^c\right),$$

où $\Gamma \leq 2^K - 1$. $\boldsymbol{\beta}$ est ensuite lié aux paramètres du modèle log-linéaire pour $I\left(i' \in s_A \; and \; v'_{\pi(i)} \in \mathcal{B}_1(v_i)\right)$,…, $I\left(i' \in s_A \; and \; v'_{\pi(i)} \in \mathcal{B}_K(v_i)\right)$, en incluant les interactions si les décisions de couplage élémentaire sont corrélées pour deux enregistrements provenant de la même unité. Par exemple, on peut envisager $\Gamma = 3$ décisions, qui sont construites à partir de deux décisions basées sur $\mathcal{B}_1(.)\; et \; \mathcal{B}_2(.)$ qui sont indépendantes pour des enregistrements d'une même unité. Supposons que ces décisions correspondent à $\mathcal{B}^{(1)}(v) = \mathcal{B}_1(v), \mathcal{B}^{(2)}(v) = \mathcal{B}_2(v)$ et $\mathcal{B}^{(3)}(v) = \mathcal{B}_1(v) \cap \mathcal{B}_2(v)$. Supposons que $\bar{p}_1$ et $\bar{p}_2$ désignent les bornes inférieures de la couverture associées à $\mathcal{B}_1(.)\; et \; \mathcal{B}_2(.)$ et $\boldsymbol{\beta} = [\bar{p}_1 \quad \bar{p}_2]^T$. Il est alors facile de démontrer que l'estimateur de la méthode des moments est $\hat{P}(i \in s_A \,|i \in s_B) = \hat{\bar{p}}^{(1)}\hat{\bar{p}}^{(2)}/\hat{\bar{p}}^{(3)}$. En pratique, il peut être difficile de sélectionner les interactions parce que les n_i sont corrélés de sorte que le test du rapport de vraisemblance type ne s'applique pas. Une solution consiste à baser les inférences sur un sous-ensemble de $o\left(\sqrt{N}\right)$ n_i qui sont alors approximativement indépendants, comme l'ont proposé Dasylva et coll. (2019).

Assouplir l'hypothèse de capture homogène : La discussion ci-dessus s'applique aussi quand la capture dans s_A est homogène au sein de post-strates définies en fonction des variables auxiliaires, qui sont disponibles sur la liste s_B. Dans ce cas, un estimateur ponctuel et de la borne inférieure de la couverture est obtenu dans chaque post-strate. On obtient ensuite la couverture globale en agrégeant cette information dans toutes les post-strates.

4. Expérience sur des données

La méthodologie proposée est évaluée au moyen de simulations dans deux scénarios, avec 100 répétitions Monte-Carlo dans chaque scénario. Dans chaque répétition, une population synthétique est générée, comprenant un million de personnes dont le nom de famille et la date de naissance sont tirés des données publiques du recensement américain de 2010. Les listes sont créées en tirant des échantillons bernoulliens indépendants avec une probabilité d'inclusion de 0,95 et en injectant des coquilles dans les variables selon Copas et Hilton (1990). Pour le nom de famille, les erreurs sont générées en tirant un nombre t de coquilles selon une distribution de Poisson avec intensité ε, où $\varepsilon = 0.01$ dans le premier scénario et $\varepsilon = 0.0025$ dans le second scénario. Le nom de famille est ensuite transformé en appliquant t étapes d'un automate fini, où l'état initial est le nom de famille original et l'état à l'étape $i \leq t$ est le nom de famille modifié par les i premières coquilles. À une étape donnée, une coquille aléatoire est générée indépendamment des coquilles des étapes précédentes, de sorte qu'il est tout aussi probable qu'il s'agisse d'une suppression ou d'une insertion, à une position aléatoire uniforme dans la chaîne de caractères. Quand la coquille est une insertion, le caractère inséré est tiré uniformément de l'alphabet. Pour la date de naissance, chaque composante est modifiée par l'ajout d'une erreur aléatoire indépendante, qui est égale à 0, -1 ou 1 avec une probabilité $1 - \varepsilon$, $\varepsilon/2$ et $\varepsilon/2$, respectivement. Par souci de simplicité, la date modifiée est enregistrée même si elle n'est plus légitime, par exemple en raison d'une composante de jour ou de mois nulle. Les coquilles sont générées indépendamment pour le nom de famille, d'une part, et le jour et le mois de naissance, d'autre part. Ensuite, les fichiers sont couplés au moyen des décisions décrites au tableau 4-1, où l'on peut constater que les décisions 1 et 2 ne sont pas indépendantes (pour les enregistrements d'une même unité) parce qu'elles comportent toutes deux le critère de pochettes. Ce serait le cas s'il n'y avait pas de faux négatifs attribuables à la création de pochettes, c'est-à-dire un rappel de 1,0 pour la décision 0.

On estime la borne inférieure de la couverture en ajustant le modèle proposé avec la contrainte $p_1 = \cdots = p_G$, après avoir remplacé chaque n_i sélectionné par $\min(\tau, n_i)$ pour se protéger contre les valeurs aberrantes, avec $\tau = 10$. Les estimations des paramètres sont calculées par optimisation non linéaire avec la procédure R constrOptim() sous les contraintes $p_1 \in [0{,}1[$, $\lambda_g \geq 0$, $\alpha_g \geq 0$ et $\sum_{g=1}^{G-1} \alpha_g \leq 1$. On peut estimer la couverture en utilisant les bornes inférieures des décisions 1, 2 et 3, décrites ci-dessus, et en supposant que les décisions 1 et 2 sont indépendantes.

Les performances de la borne inférieure estimée (par rapport à la borne inférieure réelle) sont indiquées dans les tableaux 4-2 et 4-3, tandis que celles de la couverture estimée (par rapport à la couverture réelle) figurent dans le tableau 4-4. Elles indiquent que la borne inférieure de la couverture est estimée avec un petit biais relatif et une petite variance. Elles montrent aussi que la couverture est estimée avec un biais plus petit dans le deuxième scénario, où il y a moins de corrélation entre les décisions 1 et 2 parce qu'il y a moins de faux négatifs attribuables à la création de pochettes, ce qui signifie un rappel plus élevé pour la décision 0.

Tableau 4-1
Décisions de couplage.

Décision	Description
0	Critère de pochettes basé sur la même année de naissance, le même nom de famille SOUNDEX et une concordance directe ou croisée sur le jour et le mois de naissance
1	Critère de pochettes et concordance parfaite sur le nom de famille
2	Critère de pochettes et concordance parfaite sur le jour et le mois de naissance
3	Concordance parfaite sur toutes les variables (implique la satisfaction du critère de pochettes)

Tableau 4-2
Borne inférieure estimée pour le premier scénario.

Décision	Rappel	Précision	Borne inférieure réelle	Borne inférieure estimée	
				Biais relatif (%)	Variance ($\times 10^{-7}$)
0	0,971	0,207	0,922	-0,540	2,58
1	0,964	0,369	0,915	-0,478	1,58
2	0,933	0,940	0,886	-0,029	1,30
3	0,926	0,972	0,880	-0,030	1,73

Tableau 4-3
Borne inférieure estimée pour le deuxième scénario.

Décision	Rappel	Précision	Borne inférieure réelle	Borne inférieure estimée	
				Biais relatif (%)	Variance ($\times 10^{-7}$)
0	0,993	0,210	0,943	-0,135	1,98
1	0,991	0,372	0,941	-0,119	0,886
2	0,983	0,942	0,934	-0,008	0,693
3	0,981	0,973	0,932	-0,007	0,634

Tableau 4-4
Couverture estimée des deux scénarios.

Scénario	Biais relatif (%)	Variance ($\times 10^{-7}$)
1	-3,34	8,73
2	-0,845	1,02

5. Conclusion

Nous proposons une nouvelle méthodologie aux fins d'estimation de la capture-recapture avec erreurs de couplage, sans examens manuels, tout en assouplissant l'hypothèse d'absence de faux positifs pour les unités incluses dans les deux listes et l'hypothèse selon laquelle les variables de couplage sont conditionnellement indépendantes. La méthode donne une borne inférieure sur la couverture de chaque liste. Elle produit également un estimateur de la couverture réelle si l'on peut supposer qu'il n'y a pas de faux négatifs ou si les deux listes sont couplées par plusieurs décisions de couplage, avec un modèle paramétrique commun pour le rappel de ces décisions. Il s'agit d'un exemple de solution « sans couplage » parce qu'elle estime la couverture en exploitant la relation entre ce paramètre et les mesures de l'exactitude du couplage, y compris le rappel et la précision, sans produire de fichier de paires couplées. Cette étude indiquerait aussi qu'il faudrait donner la priorité à un rappel élevé plutôt qu'à une grande précision lors du couplage aux fins d'estimation de la capture-recapture, contrairement aux priorités des couplages qui ont un but analytique.

Bibliographie

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

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

Brown, J., C. Bycroft, D. Di Cecco, J. Elleouet, G. Powell, V. Račinskij, P. Smith, S.-M. Tam, T. Tuoto, et L.-C. Zhang (2020), « Exploring developments in population size estimation », *Survey Statisticain*, 82, p. 27-39.

Copas, J., et F. Hilton (1990), « Record linkage : Statistical models for matching computer records », *Journal of the Royal Statistical Society A*, 153, 287-320.

Dasylva, A., A. Goussanou, A. Ajavon, et H. Abousaleh (2019), « Revisiting the probabilistic method of record linkage », arXiv :1911.01874.

Dasylva, A. et A. Goussanou (2020), « Estimating linkage errors under regularity conditions », *Proceedings of the Survey Methods Section, American Statistical Association*, p. 687-692.

Dasylva, A. et Goussanou, A. (2021), « Estimation des faux négatifs attribuables à la création des pochettes dans le couplage d'enregistrements », *Techniques d'enquête*, 47.

Ding, Y., et S.E. Fienberg (1994), « Estimation de système dual du sous-dénombrement dans le recensement lorsqu'il y a erreur d'appariement », *Techniques d'enquête,* 20, p. 149-158.

Di Consiglio, L., et T. Tuoto (2015), « Coverage Evaluation on Probabilistically Linked Data », *Journal of Official Statistics*, 31, p. 415-429.

Lincoln, F.C. (1930). « Calculating Waterflow Abundance on the Basis of Banding Returns », *United States Department of Agriculture Circular*, 118 : 1-4.

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

Petersen, C.G. J. (1896). « he Yearly Immigration of Young Plaice Into the Limfjord From The German Sea » *Report of the Danish Biological Station (1895)* **6**, 5-84.

Račinskij, V., P. A. Smith, et P. van der Heijden (2019), « Linkage Free Dual System Estimation », arXiv :1903.10894.

Zhang, L. C. (2019), On provision of UK neighbourhood population statistics beyond 2021. Report for the ONS.