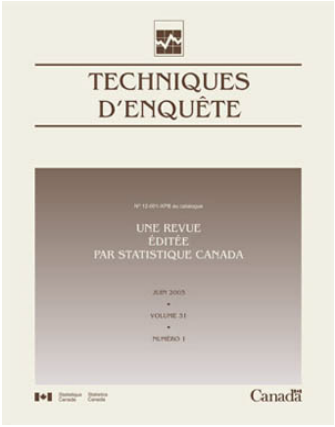
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

Modèles d’erreur de couplage pour l’estimation par capture-recapture sans vérifications manuelles



par Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu

Date de diffusion : le 20 décembre 2024

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2024
Vol. 50, nº 2, pp. 407-441
Statistique Canada, nº 12-001-X au catalogue

Modèles d'erreur de couplage pour l'estimation par capture-recapture sans vérifications manuelles

Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu[1]

Résumé

Il est possible d'appliquer la méthode de capture-recapture pour mesurer la couverture des sources de données administratives et de mégadonnées dans les statistiques officielles. Dans sa forme de base, elle comporte le couplage de deux sources tout en supposant un couplage parfait et d'autres hypothèses types. En pratique, des erreurs de couplage surviennent et constituent une source potentielle de biais quand le couplage est fondé sur des quasi-identificateurs. Ces erreurs comprennent des faux positifs et des faux négatifs, où les premiers se produisent quand un lien est établi entre des enregistrements provenant de différentes unités, et les deuxièmes surviennent lorsqu'il n'y a pas de lien entre des enregistrements provenant de la même unité. Jusqu'à présent, les solutions trouvées ont reposé sur des vérifications manuelles coûteuses ou ont posé l'hypothèse restrictive de l'indépendance conditionnelle. Dans le présent article, on assouplit ces exigences en modélisant plutôt le nombre de liens à partir d'un enregistrement. Cette méthode peut aussi être adoptée pour estimer l'exactitude du couplage sans vérifications manuelles, quand on lie deux sources ayant chacune de la sous-couverture.

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

1. Introduction

La méthode de capture-recapture est un outil important pour estimer la couverture des sources de données administratives et des mégadonnées qui sont de plus en plus employées dans les statistiques officielles (Zhang, 2015). Dans sa forme la plus simple, elle permet d'estimer la couverture de deux sources sur la même population finie, en déterminant les unités sélectionnées dans les deux sources, c'est-à-dire leur intersection, selon des hypothèses types qui comprennent un couplage parfait. Alors, la couverture estimée est fondée sur l'estimateur très connu de Petersen (1896) et Lincoln (1930). Il peut toutefois y avoir des erreurs de couplage parce que le couplage est souvent fondé sur des quasi-identificateurs comme les noms et les dates. Ces erreurs peuvent biaiser l'estimation de la couverture, qui doit être corrigée.

Pour ce qui est de l'exactitude, une erreur de couplage est définie comme un *faux négatif* ou un *faux positif*, où un faux négatif correspond à l'absence de lien entre des enregistrements d'une même unité, et un faux positif correspond à la présence d'un lien entre des enregistrements de différentes unités. En relation avec ces concepts, une paire d'enregistrements est dite *appariée* si ses enregistrements proviennent de la même unité (Fellegi et Sunter, 1969; Herzog, Scheuren et Winkler, 2007). Sinon, elle est dite *non appariée*. L'exactitude du couplage peut être mesurée au moyen de la vérification manuelle, d'un modèle statistique ou d'une combinaison des deux méthodes. Les vérifications manuelles consistent à inspecter visuellement un échantillon probabiliste de paires d'enregistrements afin de déterminer si elles sont appariées (Dasylva, Abeysundera, Akpoué, Haddou et Saïdi, 2016). Elles sont très souples et s'appliquent quelles que soient les détails du couplage. Elles sont cependant très coûteuses. La solution de rechange aux vérifications manuelles

1. Abel Dasylva, Arthur Goussanou et Christian-Olivier Nambeu, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario) K1A 0T6. Courriels : abel.dasylva@statcan.gc.ca, arthur.goussanou@statcan.gc.ca, christianolivier.nambeu@statcan.gc.ca.

consiste à ajuster un modèle statistique comme plusieurs études l'ont proposé, notamment par des mélanges log-linéaires (Fellegi et Sunter, 1969; Thibaudeau, 1993; Winkler, 1993; Daggy, Xu, Hui, Gamache et Grannis, 2013; Chipperfield, Hansen et Rossiter, 2018; Winglee, Valliant et Scheuren, 2005; Chipperfield et Chambers, 2015; Haque, Mengersen et Stern, 2021; Haque et Mengersen, 2022), des modèles du poids de couplage probabiliste d'une paire (Belin et Rubin, 1995; Sariyar, Borg et Pommerening, 2011), des modèles bayésiens (Fortini, Liseo, Nuccitelli et Scanu, 2001; Tancredi et Liseo, 2011; Sadinle, 2017; Steorts, Hall et Fienberg, 2016), et des modèles fondés sur le nombre de liens à partir d'un enregistrement donné (Blakely et Salmond, 2002; Dasylva et Goussanou, 2022). La dernière méthode de modélisation mentionnée présente un intérêt particulier dans le présent travail, car elle ne se limite pas aux couplages probabilistes et tient implicitement compte de toutes les interactions entre les variables de couplage. La méthode de modélisation n'est pas aussi coûteuse que les vérifications manuelles, mais elle est moins souple puisqu'elle repose sur des hypothèses concernant la procédure de couplage. Elle peut aussi constituer un défi quand le couplage est soumis à la contrainte d'avoir au plus un lien par enregistrement ou exactement un lien par enregistrement (Lahiri et Larsen, 2005, page 226). Chipperfield et Chambers (2015) et Sadinle (2017) ont traité cette question. Cependant, les méthodes proposées nécessitent d'importantes ressources informatiques et dépendent de l'hypothèse restrictive selon laquelle les variables de couplage sont conditionnellement indépendantes, c'est-à-dire qu'elles sont indépendantes, que la paire soit appariée ou non. En effet, cette hypothèse est une source potentielle de biais, selon Newcombe (1988, chapitre E.6, page 149), Belin et Rubin (1995) et Blakely et Salmond (2002). À la suite de Larsen et Rubin (2001), il est également possible de combiner les examens manuels et la modélisation statistique pour tirer parti de la souplesse des premiers et des faibles coûts de la seconde. Il reste que les coûts globaux dépassent le budget de nombreuses études. Pour ce qui est de la méthode par capture-recapture, les erreurs de couplage sont nuisibles parce qu'elles peuvent biaiser la couverture estimée. En effet, un faux négatif peut entraîner la sous-estimation de la couverture, tandis qu'un faux positif peut produire un biais dans la direction opposée. Bien entendu, il faut éliminer ce biais pour estimer la couverture avec exactitude.

De nombreuses méthodes de correction des erreurs ont été décrites, qui reposent sur les hypothèses types de capture-recapture sauf pour le couplage imparfait, c'est-à-dire une population fermée, des unités indépendantes sélectionnées indépendamment par chaque source, une probabilité de capture homogène par au moins une source et aucune unité en double ou hors du champ de l'enquête dans l'une des sources. Elles comprennent des solutions qui nécessitent des estimations manuelles de l'exactitude du couplage (Ding et Fienberg, 1994; Di Consiglio et Tuoto, 2015; de Wolf, van der Laan et Zult, 2019; Brown, Bycroft, Di Cecco, Elleouet, Powell, Račinskij, Smith, Tam, Tuoto et Zhang, 2020) et d'autres solutions qui reposent sur un modèle statistique selon l'hypothèse d'indépendance conditionnelle (Tancredi et Liseo, 2011; Račinskij, Smith et van der Heijden, 2019). Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et de Wolf et coll. (2019) présentent trois solutions étroitement apparentées du premier type, où ils restreignent le couplage à au plus un lien par enregistrement et ils supposent que la probabilité de faux positifs est négligeable pour les unités capturées par les deux sources. Cependant, ils estiment l'exactitude du couplage au moyen de vérifications manuelles, qui sont coûteuses, mais qui constituent la seule solution pratique, compte tenu des contraintes du couplage. Brown et coll. (2020) exposent une solution différente, qui repose

aussi sur des estimations manuelles de l'exactitude du couplage, et dans laquelle les deux sources doivent être liées deux fois au moyen de procédures de couplage différentes, en partant de l'hypothèse que les indicateurs de couplage associés sont indépendants dans chaque paire appariée. Tancredi et Liseo (2011) et Račinskij et coll. (2019) utilisent des modèles statistiques pour estimer conjointement l'exactitude du couplage et la couverture sans vérifications manuelles. Ils posent toutefois l'hypothèse restrictive d'indépendance conditionnelle.

Le présent article vise à estimer conjointement la couverture et l'exactitude du couplage sans vérification manuelle, tout en relâchant l'hypothèse selon laquelle les variables de couplage sont conditionnellement indépendantes. À cette fin, l'article présente une nouvelle méthodologie, qui constitue une extension d'un modèle antérieur d'erreur de couplage (Dasylva et Goussanou, 2022), selon les hypothèses types de capture-recapture, sauf pour l'hypothèse de couplage parfait. Dans cette méthode fondée sur un modèle, on estime la couverture en couplant les enregistrements avec un rappel suffisamment élevé, ou en précisant les interactions dans les paires appariées, tout en permettant des interactions arbitraires dans les paires non appariées. Les mêmes modèles peuvent servir à estimer le rappel et la précision lors du couplage de deux sources ayant chacune de la sous-couverture.

Les autres sections de l'article porteront, dans cet ordre, sur les notations et les hypothèses, le contexte, la méthodologie proposée, les simulations et la conclusion.

2. Notations et hypothèses

Dans la version de base du problème de capture-recapture, il faut estimer la couverture d'une liste tirée d'une population finie en exploitant une deuxième liste tirée de la même population finie, selon des hypothèses types, qui comprennent une population fermée, des unités indépendantes sélectionnées indépendamment par chaque liste, une capture homogène par au moins une liste, aucune unité en double ou hors du champ de l'enquête dans l'une des listes et un couplage parfait des deux listes. Généralement, ces listes correspondent à des échantillons probabilistes ayant des probabilités de sélection inconnues. Dans ce qui suit, on suppose que les hypothèses types de capture-recapture se vérifient, sauf pour l'hypothèse de couplage parfait qui est assouplie. Chaque liste est modélisée par un échantillon de Bernoulli, où chaque unité incluse est associée à un enregistrement susceptible de contenir des erreurs typographiques. Par exemple, pour ce qui est des listes de personnes, cet enregistrement peut contenir le nom de famille et la date de naissance. On suppose que les valeurs des enregistrements sont indépendantes entre unités, mais aucune hypothèse n'est faite au sujet de la dépendance de ces valeurs pour les enregistrements qui proviennent de la même unité, ou de la dépendance des variables dans ces enregistrements. Pour satisfaire l'hypothèse de capture homogène, on suppose qu'une unité est capturée dans la première liste indépendamment des valeurs d'enregistrement associées (par exemple le nom de famille et la date de naissance enregistrés). La capture dans la deuxième liste peut toutefois dépendre de ces valeurs. Dans ce qui suit, on désigne la cardinalité d'un ensemble s par $|s|$, et pour un uplet $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ on définit $|\mathbf{x}| = |x_1| + \ldots + |x_d|$.

2.1 Population finie et sources de données

Les unités proviennent d'une population finie U ayant N unités étiquetées de 1 à N. Les unités sont sélectionnées selon deux échantillons de Bernoulli qui sont désignés par S_A et S_B et définis par deux sous-ensembles de $\{1, \ldots, N\}$, où l'on suppose que la probabilité d'inclusion $P(i \in S_A)$ ne dépend pas de N. Par exemple, S_A peut être le recensement de la population et S_B peut être une enquête sur la couverture. Dans chaque liste où elle est incluse, l'unité i est associée à un enregistrement dont la valeur est sa caractéristique déterminante. Dans ce qui suit, le terme « enregistrement » signifiera aussi cette valeur quand le contexte permettra de clairement le comprendre. La valeur de l'enregistrement est censée se trouver dans l'espace d'enregistrement $\mathcal{V}_N$, qui est fini mais qui peut être de grande taille. Par exemple, l'espace d'enregistrement peut comprendre toutes les chaînes écrites avec au plus 32 caractères alphabétiques, si le couplage est établi avec le nom de famille. Dans S_B, cet enregistrement est désigné par V_i. Pour S_A, l'étiquetage des enregistrements dépend d'une permutation aléatoire uniforme $\Pi(.)$ de $\{1, \ldots, N\}$, pour modéliser l'absence totale d'informations sur les enregistrements associés à la même unité. Dans cette liste, l'unité i est associée à l'enregistrement $V'_{\Pi(i)}$. L'utilisation d'une permutation aléatoire inconnue est un procédé courant dans la littérature sur le couplage d'enregistrements (Lahiri et Larsen, 2005; Chambers, 2009). Il est commode sur le plan mathématique de visualiser les deux listes sous forme d'échantillons tirés de registres conceptuels A et B, où chaque unité est associée à un enregistrement. Ensuite, on suppose que les processus de création des enregistrements et d'inclusion dans les listes sont tels que les observations

$$\left[\left(I(i \in S_A), I(i \in S_B), V_i, V'_{\Pi(i)}\right)\right]_{1 \le i \le N}$$

sont indépendantes et identiquement distribuées et indépendantes de la permutation aléatoire $\Pi(.)$. Cela signifie que les deux listes sont étiquetées indépendamment et que l'on peut considérer seulement le cas où la permutation $\Pi(.)$ est l'identité (c'est-à-dire conditionner au fait que $\Pi(.)$ soit l'identité dans ce qui suit), sans perte de généralité. Ainsi, l'unité i est associée à V_i dans S_B et à V'_i dans S_A. Dans l'exposé qui suit, tous les arguments sont à la condition que $\Pi(.)$ soit égale à l'identité. On omet toutefois cette information pour simplifier la notation. Afin de satisfaire l'hypothèse de capture homogène, on suppose que la capture dans S_A est indépendante de V_i et V'_i. Toutefois, la capture dans S_B peut dépendre de ces enregistrements, c'est-à-dire que nous pourrions avoir

$$P\left(i \in S_B \mid V_i, V'_i\right) \neq P(i \in S_B).$$

Par exemple, la probabilité de capture peut varier selon des postrates qui sont basées sur V_i.

2.2 Couplage d'enregistrements et erreurs connexes

L'indicateur d'un lien entre V_i et V'_j est désigné par L_{ij} et appelé *décision de couplage* pour la paire (i, j). Soit n_i le nombre de liens issus de V_i dans S_B, c'est-à-dire

$$n_i = \sum_{j \in S_A} L_{ij}. \tag{2.1}$$

On suppose que le couplage est tel que L_{ij} est seulement une fonction de V_i et V'_j, c'est-à-dire que la décision de lier deux enregistrements ne dépend pas d'autres enregistrements. Dans la configuration actuelle, cette hypothèse signifie précisément que

$$E\left[L_{ij} \middle| \begin{matrix} \left[\left(I(k \in S_A), I(k \in S_B), V_k, V'_k \right) \right]_{k \in \{1, \ldots, N\} \setminus \{i, j\}}, \\ (i, j) \in S_A \times S_B, V_i, V'_j \end{matrix} \right] = E\left[L_{ij} \middle| (i, j) \in S_A \times S_B, V_i, V'_j \right].$$

Cette condition concerne une vaste gamme de stratégies de couplage pratiques qu'il est possible de mettre en œuvre au moyen de méthodes probabilistes, déterministes ou hiérarchiques. Toutefois, elle exclut les couplages qui limitent le nombre de liens par enregistrement (par exemple exactement un seul ou au plus un) même si de tels liens peuvent être établis à partir de couplages plus simples, qui répondent à la condition. Une paire d'enregistrements est désignée par un élément (i, j) de $S_A \times S_B$. Comme cela a été mentionné plus haut, une paire est dite *appariée* si les deux enregistrements proviennent de la même unité. Sinon, elle est dite *non appariée*. Dans l'exposé sur les erreurs de couplage, un *faux négatif* est une paire appariée qui n'est pas liée, un *faux positif* est une paire non appariée qui est liée, et un *vrai positif* est une paire appariée qui est liée. Par souci d'exhaustivité, nous définissions un *vrai négatif* comme une paire non appariée qui n'est pas liée. Par souci de commodité, supposons que FN, FP, VP et VN désignent respectivement le nombre total de faux négatifs, de faux positifs, de vrais positifs et de vrais négatifs. Il est courant de représenter les différents types de paires d'enregistrements dans un tableau 2×2 appelé matrice de confusion où les cellules hors de la diagonale représentent les erreurs, comme le montre le tableau 2.1. L'exactitude du couplage est habituellement mesurée par le rappel et la précision, le *rappel* étant la proportion de paires appariées qui sont liées (c'est-à-dire $\mathrm{VP}/(\mathrm{VP} + \mathrm{FN})$) et la *précision* étant la proportion de paires liées qui sont appariées (c'est-à-dire $\mathrm{VP}/(\mathrm{VP} + \mathrm{FP})$). Elle est également mesurée par le *taux de faux négatifs*, qui est la proportion de paires appariées qui ne sont pas liées (c'est-à-dire $\mathrm{FN}/(\mathrm{VP} + \mathrm{FN})$), et le *taux de faux positifs* (TFP), qui est la proportion de paires non appariées qui sont liées (c'est-à-dire $\mathrm{FP}/(\mathrm{VN} + \mathrm{FP})$). En cas de couplage parfait, la précision et le rappel sont égaux à 1,0 et le TFP est nul. Dans cette situation idéale, la taille de la population est estimée selon Petersen (1896) et Lincoln (1930) par

$$\hat{N} = \frac{|S_A||S_B|}{|S_A \cap S_B|}.$$

Par conséquent, la couverture estimée de S_A est donnée par $|S_A \cap S_B| / |S_B|$, tandis que celle de S_B est donnée par $|S_A \cap S_B| / |S_A|$. En cas d'erreurs de couplage, l'intersection des deux listes n'est pas observée directement. Il faut plutôt inférer la taille de cette intersection à partir des liens observés et de l'exactitude du couplage qui peut être estimée par la modélisation du nombre de liens à partir d'un enregistrement donné.

Tableau 2.1
Matrice de confusion.

	Lien	Non liée
Appariée	VP	FN
Non appariée	FP	VN

3. Contexte

La présente section fournit quelques éléments de contexte sur le modèle d'erreur (Dasylva et Goussanou, 2022), qui doit être adapté au problème qui nous concerne. Ce modèle s'applique quand S_A est un recensement (c'est-à-dire $S_A = U$) et que le couplage est tel que la décision de lier deux enregistrements donnés ne dépend d'aucun autre enregistrement. Dans ce cas, on peut estimer l'exactitude du couplage en modélisant la distribution du nombre de liens à partir d'un enregistrement donné, sans faire de suppositions sur la dépendance des variables de couplage.

3.1 Relation entre les erreurs et le nombre de liens d'un enregistrement

En général, il existe un lien étroit entre le nombre de liens d'un enregistrement donné et les erreurs de couplage qui s'y rapportent. Cette relation est décrite dans le tableau 3.1 quand S_A est un recensement (Dasylva et Goussanou, 2020). Quand $n_i = 0,$ on sait qu'il n'y a pas de faux positif mais un faux négatif parce que S_A est un recensement. Quand $n_i = 1, \ldots, N-1,$ il n'y a pas de faux négatif ou il y en a un et donc n_i ou $n_i - 1$ faux positifs, selon que l'enregistrement est lié à l'enregistrement de recensement apparié ou non, parce que S_A est un recensement et qu'il n'y a pas d'enregistrement en double. Quand $n_i = N,$ on sait qu'il n'y a pas de faux négatifs et qu'il y a $N-1$ faux positifs, pour les mêmes raisons. À titre d'illustration, examinons l'exemple présenté dans le tableau 3.2, où $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\},$ $S_B = \{2, 3\}$ et la nature de chaque paire d'enregistrements est également indiquée. Dans cet exemple, il y a quatre liens, à savoir $(2,1),$ $(2, 2),$ $(3, 2)$ et $(3, 4).$ Selon le tableau 3.2, il est possible de vérifier qu'il y a une relation entre n_i et les erreurs de couplage. Quand $n_i = 1, \ldots, N-1,$ les erreurs ne sont pas entièrement connues et peuvent être prédites au moyen d'un modèle.

Tableau 3.1
Relation entre n_i et les erreurs quand S_A est un recensement.

n_i	Faux négatifs	Faux positifs
0	1	0
$1 \le n_i \le N-1$	0 ou 1	$n_i - 1$ ou n_i
N	0	$N-1$

Tableau 3.2
Exemple, où $N = 5,$ $S_A = U = \{1, 2, 3, 4, 5\}$ et $S_B = \{2, 3\}$ et les liens sont indiqués par les coches.

	$j =$							
	1	2	3	4	5	n_i	N^{bre} FN	N^{bre} FP
$i = 2$	√ FP	√ VP	VN	VN	VN	2	0	1
3	VN	√ FP	FN	√ FP	VN	2	1	2

3.2 Modèle pour enregistrements homogènes

Blakely et Salmond (2002) modélisent n_i par la somme d'une variable de Bernoulli (pour les vrais positifs) avec une variable binomiale indépendante (pour les faux positifs) et ils estiment les paramètres qui s'y rapportent au moyen d'une équation quadratique. Cependant, la distribution de n_i doit être identique pour tous les enregistrements. Sinon, l'estimateur pourrait être biaisé ou ne pas exister (Dasylva et Goussanou, 2022) si l'équation quadratique n'a pas de solution. En pratique, cette question peut se poser quand on apparie avec des noms ou d'autres caractéristiques, qui se présentent à des fréquences différentes dans la population.

3.3 Modèle pour enregistrements hétérogènes

Pour régler le problème, Dasylva et Goussanou (2022) ont étendu le modèle (Blakely et Salmond, 2002) à un mélange fini, qui s'applique quand N devient grand dans des conditions de régularité. Pour décrire ces conditions, supposons que

$$\mathcal{V}_N^* = \left\{ v \in \mathcal{V}_N \text{ s.t. } P\left(V_i = v \middle| \, i \in S_B\right) > 0 \right\}. \tag{3.1}$$

En d'autres termes, $\mathcal{V}_N^*$ est le sous-ensemble des valeurs d'enregistrement qui peuvent être observées dans S_B avec une probabilité positive. À ce stade, il est utile d'examiner le sous-ensemble de toutes les valeurs d'enregistrement (issues de $\mathcal{V}_N$) qui sont liées avec une probabilité positive à une valeur d'enregistrement particulière, ainsi qu'un sur ensemble de cet ensemble, appelé « *voisinage* » et désigné par $\mathcal{B}_N(v)$ pour la valeur $v \in \mathcal{V}_N^*$. Ainsi,

$$\mathcal{B}_N(v) \supset \left\{ v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \middle| \, i \in S_B, \left(V_i, V_j'\right) = (v, v') \right] > 0 \right\}. \tag{3.2}$$

De façon informelle, le voisinage d'une valeur d'enregistrement particulière est un sous-ensemble de valeurs d'enregistrement qui ressemblent à cette valeur selon un certain critère. Par exemple, on envisage de lier les enregistrements en fonction du nom de famille en lettres majuscules et on suppose que deux enregistrements sont liés s'ils correspondent exactement pour cette variable. Dans ce cas, la valeur d'enregistrement (v) « JARO » peut être associée au voisinage singleton $(\mathcal{B}_N(v) = \{v\})$ {« JARO »}. Pour affiner cet exemple, supposons maintenant que deux enregistrements sont liés si les noms de famille sont identiques, ou s'ils ont la même longueur et diffèrent seulement par une lettre. Dans ce cas, la valeur « JARO » peut être associée au voisinage

$$\begin{aligned} &\{\text{“AARO”, “BARO”, …, “ZARO”}\} \cup \{\text{“JARO”, “JBRO”, …, “JZRO”}\} \cup \\ &\quad \{\text{“JAAO”, “JABO”, …, “JAZO”}\} \cup \{\text{“JARA”, “JARB”, …, “JARZ”}\}. \end{aligned}$$

Le concept de voisinage est utile quand on caractérise le pouvoir discriminatif des variables de couplage et quand on définit des conditions de régularité pour l'estimation convergente du rappel et de la précision sans vérifications manuelles. Pour décrire ces conditions, définissons les fonctions $p_N(.)$, $\lambda_N(.)$ et $\lambda_N^{(0)}(.)$ qui

donnent la probabilité de vrais positifs, la probabilité de faux positifs et la probabilité qu'un enregistrement non apparié se trouve dans le voisinage, c'est-à-dire

$$p_N(v) = E\Big[L_{ii} \,\Big|\, i \in S_B, V_i = v \Big], \tag{3.3}$$

$$\lambda_N(v) = E\Big[L_{ij} \,\Big|\, i \in S_B, V_i = v \Big], \; j \neq i, \tag{3.4}$$

$$\lambda_N^{(0)}(v) = P\Big(V_j' \in \mathcal{B}_N\left(V_i\right) \Big|\, i \in S_B, V_i = v \Big). \tag{3.5}$$

Alors, les deux premières conditions de régularité sont données par les équations suivantes, où Λ est positif et fini, F est une distribution bivariée avec support contenu dans $[0,1] \times [0, \Lambda]$ et ni l'un ni l'autre ne dépend de N.

$$\sup_{v \in \mathcal{V}_N^*} (N-1)\, \lambda_N^{(0)}(v) \;\leq\; \Lambda, \tag{3.6}$$

$$\big(p_N(V_i), (N-1)\, \lambda_N(V_i) \big) \,\big|\, \{ i \in S_B \} \;\xrightarrow{d}\; F. \tag{3.7}$$

La première condition implique que le nombre attendu de faux positifs possède une borne supérieure pour chaque enregistrement. Quand la probabilité de faux positifs a comme borne inférieure δ (Dasylva et Goussanou, 2020, équation 6), cela signifie aussi que la précision n'est pas inférieure à $\delta / (\delta + \Lambda)$ en tout et pour chaque poststrate, qui est définie en fonction de V_i (Dasylva et Goussanou, 2020). La deuxième condition signifie que la distribution conjointe du nombre espéré de vrais positifs et du nombre espéré de faux positifs est approximativement donnée par F quand N est de grande taille. Quand F est discret et comporte G atomes, ces deux conditions supposent la convergence en distribution suivante (Dasylva et Goussanou, 2022, Lemme 1).

$$n_i \,\big|\, \{ i \in S_B \} \;\xrightarrow{d}\; \sum_{g=1}^{G} \alpha_g \text{ Bernoulli}(p_g) \,*\, \text{Poisson}(\lambda_g), \tag{3.8}$$

où $*$ désigne l'opérateur de convolution. Cela signifie que, à la limite, un enregistrement appartient à l'une des G classes latentes, où α_g est la probabilité de la classe g, et p_g et λ_g sont les nombres espérés de vrais positifs et de faux positifs pour les enregistrements de cette classe. On peut estimer les paramètres du modèle en maximisant la probabilité composite des n_i. Ceux-ci sont liés à l'exactitude du couplage au moyen des nombres espérés de vrais positifs et de faux positifs par enregistrement dans S_B, qui sont donnés par $\bar{p} = \sum_{g=1}^{G} \alpha_g p_g$ et $\bar{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$. En effet, le rappel et la précision convergent en probabilité respectivement vers $\bar{p}$ et $\bar{p} / (\bar{p} + \bar{\lambda})$, sous les deux conditions de régularité supplémentaires suivantes (Dasylva et Goussanou, 2022, Corollaire 1), où $i \neq i'$ et c est une constante finie positive ne dépendant pas de N.

$$NP\Big(\mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing \,\Big|\, \{i, i'\} \subset S_B \Big) \leq c, \tag{3.9}$$

$$NP\Big(V_i' \in \mathcal{B}_N(V_i) \,\Big|\, \{i, i'\} \subset S_B, \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) \neq \varnothing \Big) \leq c. \tag{3.10}$$

Ces deux conditions signifient que les enregistrements de différentes unités sont très susceptibles d'avoir des voisinages disjoints (3.9) et que les enregistrements appariés ne sont pas très éloignés (3.10). Étant donné les autres conditions de régularité (c'est-à-dire (3.6)-(3.7)), elles impliquent également la convergence des estimateurs composites par le maximum de vraisemblance (Dasylva et Goussanou, 2022, théorème 3).

Dans l'ensemble, cette méthodologie présente plusieurs avantages par rapport à d'autres solutions fondées sur un modèle, parce qu'elle en prend en compte à la fois les interactions entre les variables de couplage et l'hétérogénéité des enregistrements, de façon transparente. En outre, on peut tester l'adéquation du modèle au moyen de la procédure décrite par Dasylva et Goussanou (2024) pour tenir compte de la corrélation des n_i. Quand S_A est un recensement, la méthodologie peut aussi servir à estimer les faux négatifs générés par la procédure de création des pochettes (Dasylva et Goussanou, 2021). Cependant, il faut l'adapter quand S_A présente de la sous-couverture.

4. Méthodologie

La méthodologie proposée est fondée sur deux extensions du modèle décrit dans la section précédente, qui sera appelé par la suite *modèle de voisinage univarié* ou simplement modèle de voisinage. La première extension tient compte de la sous-couverture de S_A et modifie seulement l'interprétation de certains paramètres. Elle sert à estimer la couverture en cas de couplage lorsque le rappel est suffisamment élevé. La deuxième extension est plus importante, car elle remplace n_i par un vecteur de telles variables, chacune représentant le nombre de liens pour une règle de couplage distincte. Le modèle qui en résulte est appelé *modèle de voisinage multivarié*. On l'utilise pour estimer la couverture en précisant les interactions dans les paires appariées, tout en permettant des interactions arbitraires dans les paires non appariées. Les paragraphes suivants traitent de la stratégie de couplage avant de présenter en détail les différentes extensions et leur emploi aux fins d'estimation de la couverture.

4.1 Couplage des sources

Dans leurs solutions, Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et de Wolf et coll. (2019) limitent chaque enregistrement à un maximum d'un seul lien. Or, cette contrainte complique considérablement la modélisation des erreurs de couplage, comme cela a été indiqué précédemment. Nous proposons plutôt ici de coupler les enregistrements sans cette contrainte, en fonction d'une règle telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement. Ainsi, un enregistrement peut avoir zéro lien, un lien ou plusieurs liens. Cette règle de couplage peut être au moyen des méthodes de couplage d'enregistrements déterministe, hiérarchique ou probabiliste.

4.2 Extension du modèle de voisinage univarié

Quand S_A n'est pas un recensement, la relation entre n_i et les erreurs est conforme au tableau 4.1, qui diffère du tableau 3.1 quand $n_i = 0$ et $n_i = |S_A|$. Quand $n_i = 0$, il n'y a aucune certitude quant à l'occurrence d'un faux négatif, car on ne sait pas si l'unité correspondante est dans S_A, contrairement à ce

qui se passe dans le tableau 3.1. Quand $n_i = |S_A|$, le nombre de faux positifs n'est pas connu avec certitude pour la même raison. Pour tenir compte de la sous-couverture de S_A, redéfinissons $\mathcal{B}_N(v)$ en tant que sous-ensemble de $\mathcal{V}_N$ de sorte que

$$\mathcal{B}_N(v) \supset \left\{v' \in \mathcal{V}_N \text{ s.t. } E\left[L_{ij} \,\middle|\, (i,j) \in S_B \times S_A, \left(V_i, V_j'\right) = (v, v')\right] > 0\right\}. \tag{4.1}$$

Tableau 4.1
Relation entre n_i et les erreurs quand S_A n'est pas un recensement.

n_i	Faux négatifs	Faux positifs
0	0 ou 1	0
$1 \le n_i \le \lvert S_A \rvert - 1$	0 ou 1	$n_i - 1$ ou n_i
$n_i = \lvert S_A \rvert$	0	$\lvert S_A \rvert - 1$ ou $\lvert S_A \rvert$

Redéfinissons aussi $p_N(.)$, $\lambda_N(.)$ et $\lambda_N^{(0)}(.)$ comme étant

$$p_N(v) = E\left[I\left(i \in S_A\right) L_{ii} \,\middle|\, i \in S_B, V_i = v\right], \tag{4.2}$$

$$\lambda_N(v) = E\left[I\left(j \in S_A\right) L_{ij} \,\middle|\, i \in S_B, V_i = v\right], \; j \neq i, \tag{4.3}$$

$$\lambda_N^{(0)}(v) = P\left(j \in S_A, V_j' \in \mathcal{B}_N\left(V_i\right) \,\middle|\, i \in S_B, V_i = v\right), \tag{4.4}$$

où $p_N(v)$ est la probabilité conjointe d'inclure i dans S_A et d'avoir un vrai positif, $\lambda_N(v)$ reste la probabilité de faux positifs et $\lambda_N^{(0)}(v)$ est la probabilité d'avoir un enregistrement non apparié dans le voisinage. Avec cette mise-à-jour de ces définitions, il est facile de démontrer que (3.8) s'applique quand $N \to \infty$ dans les conditions de régularité données par (3.6)-(3.7) et que F est discrète et comporte G atomes. En effet, la démonstration du cas de recensement continue de s'appliquer avec n_i selon (2.1). Voir Dasylva et Goussanou (2022, Lemme 1) pour en savoir plus. Les paramètres α_g et λ_g ont la même interprétation, mais p_g correspond maintenant au produit de la probabilité d'inclusion $P(i \in S_A)$ par la probabilité d'un vrai positif pour un enregistrement dans la classe g. Comme auparavant, soit $\overline{p} = \sum_{g=1}^{G} \alpha_g p_g$ et $\overline{\lambda} = \sum_{g=1}^{G} \alpha_g \lambda_g$, où $\overline{p}$ est le nombre attendu de vrais positifs par enregistrement, qui est aussi égal à $E\left[I(i \in S_A)\, L_{ii} \,\middle|\, i \in S_B\right]$, et où $\overline{\lambda}$ est le nombre attendu de faux positifs par enregistrement. Ainsi, $\overline{p}$ est une borne inférieure utile sur $P(i \in S_A)$. On peut estimer les paramètres du modèle en maximisant la vraisemblance composite des n_i quand G est donné, et en sélectionnant ce dernier paramètre par la minimisation du critère d'information d'Akaike comme dans le cas du recensement (Dasylva et Goussanou, 2022). Supposons que $\hat{\overline{p}}$ et $\hat{\overline{\lambda}}$ désignent les estimateurs par le maximum de vraisemblance qui en résultent. À l'annexe A, il est indiqué que le rappel et la précision (deux paramètres de population finie) convergent respectivement en probabilité vers $P(i \in S_A)^{-1} \overline{p} = E\left[L_{ii} \,\middle|\, i \in S_A \cap S_B\right]$ et $\overline{p} / (\overline{p} + \overline{\lambda})$, dans des conditions de

régularité. Dans les mêmes conditions, $\hat{\bar{p}}$ et $\hat{\bar{\lambda}}$ sont aussi des estimateurs convergents de $\bar{p}$ et $\bar{\lambda}$, de sorte que $P(i \in S_A)^{-1}\hat{\bar{p}}$ et $\hat{\bar{p}}/(\hat{\bar{p}}+\hat{\bar{\lambda}})$ sont respectivement des estimateurs convergents du rappel et de la précision.

4.3 Estimations de la couverture lorsque le rappel est élevé

À partir de l'exposé ci-dessus, on peut obtenir un estimateur convergent de la couverture $P(i \in S_A)$ comme suit :

$$\left(\frac{\text{VP}}{\text{VP + FN}}\right)^{-1}\hat{\bar{p}},$$

si le rappel (c'est-à-dire $\text{VP}/(\text{VP + FN})$) est connu. En particulier, $\hat{\bar{p}}$ est convergent, si l'on sait que le rappel est parfait; autrement dit, $\text{VP}/(\text{VP + FN}) = 1{,}0$, ce qui équivaut à ne pas avoir de faux négatifs. Cependant, il convient de mentionner que le modèle proposé est intéressant uniquement quand le couplage n'est pas parfait, c'est-à-dire si le rappel, la précision ou les deux sont inférieurs à 1,0. Autrement, l'estimateur standard par capture-recapture s'appliquerait, y compris dans la situation idéale où la clé de couplage est un identificateur unique sans erreur, c'est-à-dire une clé de couplage parfaite. En outre, le modèle de voisinage n'est pas conseillé avec une telle clé, car il se peut que certaines hypothèses du modèle ne se vérifient pas.

Pour appliquer la méthode ci-dessus, il faudrait concevoir la règle de couplage de sorte qu'elle génère très peu de faux négatifs, voire aucun, et idéalement en présentant un taux de faux négatifs inférieur à $\min\left(P(i \in S_A), 1-P(i \in S_A)\right)$ d'un ordre de grandeur. Cela peut s'inspirer des procédures de création des pochettes, qui sont utilisées dans les couplages probabilistes pour sélectionner un petit sous-ensemble du produit cartésien avec la majorité des paires appariées. Christen (2012) donne un bon aperçu de ces procédures, qui sont indispensables quand les sources sont de grande taille. Cependant, l'obtention d'un rappel suffisamment élevé peut se faire au détriment de la précision et nécessiter de tolérer une précision très faible, ce qui peut empêcher l'estimation de la couverture avec l'exactitude requise. Dans de tels cas, il est proposé d'estimer la couverture en précisant les interactions dans les paires appariées. Cela nécessite toutefois une extension multivariée du modèle de voisinage.

4.4 Modèle de voisinage multivarié

L'extension multivariée concerne les ensembles finis de règles de couplage simples qui sont également mutuellement exclusives, c'est-à-dire que pour chaque règle, la décision de lier deux enregistrements ne dépend d'aucun autre enregistrement, et chaque paire est liée par au plus une règle. La meilleure façon d'expliquer la nécessité de cette extension est de donner un exemple. Par souci de simplicité, supposons que S_A est connu pour être un recensement complet et que les deux sources doivent être liées au prénom, au nom de famille et à la date de naissance. Pour ce faire, il faut évaluer sept règles, identifiées dans le tableau 4.2, où γ se trouve dans l'ensemble fini $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$, et les composantes de γ indiquent

s'il y a une concordance exacte pour ce qui est du nom de famille, du prénom et de la date de naissance, dans cet ordre. Pour la règle γ, on désigne par $n_i^{(\gamma)}$ le nombre correspondant de liens pour l'enregistrement de l'échantillon i, par exemple $n_i^{(0,0,1)}$ est le nombre de liens quand le couplage est fondé sur le fait d'avoir les mêmes noms, mais une date de naissance différente. Une façon simple d'évaluer les différentes règles consiste à ajuster un modèle de la forme

$$n_i^{(\gamma)} \Big| \{i \in S_B\} \sim \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \text{Bernoulli}\left(p_g^{(\gamma)}\right) * \text{Poisson}\left(\lambda_g^{(\gamma)}\right), \tag{4.5}$$

séparément pour chaque γ, où les nombres attendus de vrais positifs et de faux positifs par enregistrement sont respectivement $\overline{p}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} p_g^{(\gamma)}$ et $\overline{\lambda}^{(\gamma)} = \sum_{g=1}^{G^{(\gamma)}} \alpha_g^{(\gamma)} \lambda_g^{(\gamma)}$. Notons que nous avons nécessairement la contrainte $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1$, parce que les règles sont mutuellement exclusives. Cependant, il se peut que les estimateurs du rappel et de la précision qui en résultent soient inefficaces parce que des informations importantes ne sont pas pris en compte, comme la contrainte $\sum_{\gamma \in \Gamma} \overline{p}^{(\gamma)} \leq 1$ ou la corrélation entre les nombres de liens des différentes règles pour le même enregistrement de l'échantillon, qui n'est pas exploitée non plus. De plus, quand on choisit le nombre de classes $G^{(\gamma)}$ en fonction du critère d'information d'Akaike, le résultat $\hat{G}^{(\gamma)}$ peut varier selon les différentes règles, ce qui est contre-intuitif. Ajoutons que, même dans le meilleur des cas où $\hat{G}^{(\gamma)}$ est identique pour toutes les règles, les classes peuvent correspondre à différentes partitions latentes des enregistrements de l'échantillon selon les différentes règles, ce qui est également contre-intuitif et non souhaitable. De plus, les limites ci-dessus s'appliquent dans la situation plus générale où S_A n'est pas un recensement et où sa couverture est inconnue.

Tableau 4.2
Règles mutuellement exclusives basées sur le prénom, le nom de famille et la date de naissance.

Indice de la règle $\gamma = (\gamma_1, \gamma_2, \gamma_3)$	Nom de famille identique	Prénom identique	Date de naissance identique
(0,0,1)	x	x	√
(0,1,0)	x	√	x
(0,1,1)	x	√	√
(1,0,0)	√	x	x
(1,0,1)	√	x	√
(1,1,0)	√	√	x
(1,1,1)	√	√	√

La solution consiste à modéliser la distribution conjointe du vecteur de fréquences $\left[n_i^{(\gamma)}\right]_{\gamma \in \Gamma}$ au moyen d'une extension multivariée du modèle de voisinage, comme suit (des précisions sont données à l'annexe B). Pour décrire cette extension, il est commode de définir les distributions multivariées suivantes. La première distribution est la distribution conjointe de variables mutuellement indépendantes qui sont indexées sur un ensemble fini Γ, où la variable γ suit la distribution $\text{Poisson}\left(\lambda^{(\gamma)}\right)$. La distribution conjointe est alors simplement la distribution du produit. Pour faciliter la notation, nous définissons $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma \in \Gamma}$ et désignons cette distribution par $\text{PPoisson}(\boldsymbol{\lambda})$, où le premier « P » signifie produit. La deuxième distribution correspond à la distribution conjointe du nombre de cellules dans une expérience multinomiale avec n essais, où la

dernière cellule est exclue, les autres cellules sont indexées sur Γ, et la probabilité d'observer la cellule γ est indiquée par $p^{(\gamma)}$, de sorte que $\sum_{\gamma\in\Gamma} p^{(\gamma)} \le 1$. Dans ce cas, nous définissons $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma\in\Gamma}$ et nous désignons la distribution conjointe par $\text{IMultinomial}(n, \mathbf{p})$, où le « I » signifie « incomplète ». De façon générale, on peut envisager l'extension multivariée pour modéliser la distribution conjointe du nombre de liens, qui sont le résultat de l'application de règles de couplage simples mutuellement exclusives, indexées sur un ensemble fini Γ, où $n_i^{(\gamma)}$ désigne le nombre de liens de la règle γ pour l'enregistrement de l'échantillon i et $\mathbf{n}_i = \left[n_i^{(\gamma)} \right]_{\gamma\in\Gamma}$. Le modèle multivarié est un mélange fini de distri-butions discrètes à $|\Gamma|$ dimensions, où chaque composante est la convolution d'une distribution multi-nomiale incomplète avec un produit de distributions de Poisson indépendantes, à savoir

$$\mathbf{n}_i \big| \{i \in S_B\} \sim \sum_{g=1}^{G} \alpha_g \, \text{IMultinomial}(1, \mathbf{p}_g) * \text{PPoisson}\left(\boldsymbol{\lambda}_g\right), \tag{4.6}$$

où G est le nombre de classes d'enregistrements, α_g est la probabilité qu'un enregistrement d'échantillon provienne de la classe g, et $\mathbf{p}_g = \left[p_g^{(\gamma)} \right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda}_g = \left[\lambda_g^{(\gamma)} \right]_{\gamma\in\Gamma}$ sont les vecteurs des nombres espérés de vrais positifs et de faux positifs pour un enregistrement dans la classe. De plus, $p_g^{(\gamma)}$ est le nombre espéré de vrais positifs et $\lambda_g^{(\gamma)}$ est le nombre espéré de faux positifs, selon la règle γ. Ensuite, étant donné G, le modèle est paramétré par $\left[\left(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g \right) \right]_{1\le g\le G}$. Quand les enregistrements sont homogènes,

$$\mathbf{n}_i \,\big|\, \{i \in S_B\} \sim \text{IMultinomial}(1, \mathbf{p}) * \text{PPoisson}\left(\boldsymbol{\lambda}\right), \tag{4.7}$$

où $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)} \right]_{\gamma\in\Gamma}$. De plus, si $\min_{\gamma\in\Gamma} \lambda^{(\gamma)} > 0$ et $\mathbf{t} = \left[t^{(\gamma)} \right]_{\gamma\in\Gamma}$, nous avons

$$\begin{aligned} P\left(\mathbf{n}_i = \mathbf{t} \,\middle|\, i \in S_B\right) \;&= I\left(|\mathbf{t}| = 0\right)\left(1 - |\mathbf{p}|\right) e^{-|\boldsymbol{\lambda}|} \\ &\quad + I\left(|\mathbf{t}| > 1\right) \Bigg(\left(1 - |\mathbf{p}|\right) \prod_{\gamma\in\Gamma} \frac{e^{-\lambda^{(\gamma)}} \left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}}}{t^{(\gamma)}!} \\ &\qquad + \sum_{\gamma\in\Gamma;\, t^{(\gamma)} > 0} p^{(\gamma)} \, \frac{e^{-\lambda^{(\gamma)}} \left(\lambda^{(\gamma)}\right)^{t^{(\gamma)}-1}}{\left(t^{(\gamma)} - 1\right)!} \prod_{\gamma'\in\Gamma\setminus\{\gamma\}} \frac{e^{-\lambda^{(\gamma')}} \left(\lambda^{(\gamma')}\right)^{t^{(\gamma')}}}{t^{(\gamma')}!} \Bigg). \end{aligned} \tag{4.8}$$

Comme auparavant, le modèle est motivé par la convergence de la distribution du vecteur des fréquences $\mathbf{n}_i = \left[n_i^{(\gamma)} \right]_{\gamma\in\Gamma}$, quand $N \to \infty$, comme l'indique le lemme 2 de l'annexe B. À partir du mélange multivarié, il s'ensuit que la distribution marginale de $n_i^{(\gamma)}$ est toujours donnée par (4.5), sauf que $G^{(\gamma)}$ et $\alpha_g^{(\gamma)}$ sont identiques pour toutes les règles, comme cela était souhaité. De plus, les catégories d'enregistrements correspondent maintenant pour toutes les règles. On obtient un modèle restreint quand $\mathbf{p}_g = \varrho\left(\boldsymbol{\beta}_g\right)$ pour chaque classe, où $\varrho(.)$ est une fonction injective connue et $\boldsymbol{\beta}_g$ est un vecteur de coefficients de régression de dimension inférieure à $|\Gamma|$. Cela signifie que les $|\Gamma|$ probabilités de vrais positifs pour les différentes règles ne sont pas libres, mais liées par moins de coefficients de régression, pour chaque classe d'enregistrements. Une telle restriction est utile quand on exploite l'information sur les interactions des variables dans les paires appariées, au moyen d'une spécification log-linéaire. Ensuite, un mélange multivarié ayant G composantes est paramétré par $\left[\left(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g \right) \right]_{1\le g\le G}$. Selon le théorème 2 de l'annexe B, les paramètres du

modèle (pour chaque paramétrisation proposée) sont estimés de manière convergente en maximisant la probabilité composite des vecteurs de fréquences observés $\left[\mathbf{n}_i\right]_{i\in S_B}$, quand G est connu ou inconnu. Dans ce dernier cas, on peut sélectionner G selon le critère d'information minimum d'Akaike comme auparavant.

4.5 Estimer la couverture au moyen de la structure de corrélation

Quand les enregistrements sont liés avec un rappel parfait, la couverture peut être estimée au moyen du modèle de mélange univarié évoqué plus haut. Sinon, on peut estimer la couverture au moyen du modèle de voisinage multivarié, où les probabilités de vrais positifs sont contraintes selon la structure de corrélation des variables de couplage au moyen d'une spécification log-linéaire. Plus précisément, cela signifie que le modèle multivarié proposé est fondé sur un ensemble de règles de couplage simples (c'est-à-dire que chaque règle est telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement), qui sont elles-mêmes fondées sur un premier ensemble de règles de couplage simples divisées en K groupes. Dans le groupe k, il y a H_k règles mutuellement exclusives (c'est-à-dire qu'une paire est liée par au plus une règle à partir du groupe) et $L_{ij}^{(k,h)}$ indique si la paire (i, j) est liée par la règle h. Par exemple, chaque groupe peut être fondé sur une seule variable, et les règles peuvent correspondre à différents niveaux de concordance pour cette variable. À titre d'exemple, pour le nom de famille, ces niveaux peuvent comprendre la concordance exacte, la concordance basée sur une erreur typographique (qui exclut la concordance exacte) et la concordance du code SOUNDEX (qui exclut les concordances exactes et celles basée sur une erreur typographique). Toutefois, un groupe peut comporter plusieurs variables. On obtient un deuxième ensemble de règles en combinant les règles du premier ensemble comme suit. Soit l'ensemble des indices $\Gamma = \{0,\ldots,H_1\}\times\ldots\times\{0,\ldots,H_K\} - \mathbf{0}_K$, et pour $\gamma = (\gamma_1,\ldots,\gamma_K)\in\Gamma$, désignons par $L_{ij}^{(\gamma)}$ l'indicateur que la règle γ lie la paire (i, j), où $L_{ij}^{(\gamma)} = 1$ seulement si $L_{ij}^{(k,\gamma_k)} = 1$ pour chaque k de sorte que $\gamma_k \geq 1$, et $\sum_{h=1}^{H_k} L_{ij}^{(k,h)} = 0$ pour chaque k de sorte que $\gamma_k = 0$. Le modèle proposé est le cas particulier du modèle de voisinage multivarié (4.6), où $p_g^{(\gamma)}$ a la forme suivante, pour un vecteur de covariables $\mathbf{z}^{(\gamma)}$ et des coefficients de régression $\mathbf{u}_g$.

$$p_g^{(\gamma)} = \frac{P(i\in S_A)\exp\left(\mathbf{z}^{(\gamma)\top}\mathbf{u}_g\right)}{1+\sum_{\gamma'\in\Gamma}\exp\left(\mathbf{z}^{(\gamma')\top}\mathbf{u}_g\right)}. \tag{4.9}$$

Pour un nombre donné de classes G, les paramètres du modèle incluent $\left[\left(\alpha_g,\mathbf{u}_g,\boldsymbol{\lambda}_\mathbf{g}\right)\right]_{1\leq g\leq G}$ et $P(i\in S_A)$. La forme particulière de $\mathbf{z}^{(\gamma)}$ et $\mathbf{u}_g$ dépend du modèle. L'exemple suivant en est une illustration, où le modèle comprend tous les termes principaux et aucune interaction. Cela signifie également que les composantes de γ sont indépendantes dans les paires appariées. Dans ce cas, le coefficient correspondant à l'événement $\{\gamma_k = l_k\}$ est indiqué par $u_{g,k(l_k)}$. Selon la convention de codage « dummy », le coefficient est fixé à zéro si $l_k = 0$. La covariable correspondant à ce coefficient est l'indicateur $I(\gamma_k = l_k)$ de sorte que

$$\mathbf{z}^{(\gamma)} = \left[I(\gamma_1=1)\ldots I(\gamma_1=H_1)\ldots I(\gamma_K=1)\ldots I(\gamma_K=H_K)\right]^\top, \tag{4.10}$$

$$\mathbf{u}_g = \left[u_{g,1(1)}\ldots u_{g,1(H_1)}\ldots u_{g,K(1)}\ldots u_{g,K(H_K)}\right]^\top. \tag{4.11}$$

Dans l'exemple suivant, le modèle comprend tous les termes principaux et les interactions du deuxième ordre, mais aucune interaction d'ordre supérieur. Pour $1 \le k_1 < k_2 \le K,$ le coefficient de l'interaction entre les événements $\{\gamma_{k_1} = l_{k_1}\}$ et $\{\gamma_{k_2} = l_{k_2}\}$ est désigné par $u_{g,k_1k_2(l_{k_1}l_{k_2})}$. Selon la même convention de codage, le coefficient est fixé à zéro si $l_{k_1} = 0$ ou $l_{k_2} = 0$. La covariable associée au coefficient est l'indicateur $I\left(\left(\gamma_{k_1}, \gamma_{k_2}\right) = \left(l_{k_1}, l_{k_2}\right)\right)$. Dans ce cas, $\mathbf{z}^{(\gamma)}$ et $\mathbf{u}_g$ ont des expressions plus complexes. Écrivons-les dans une formule claire en définissant

$$\mathbf{z}_{1k}^{(\gamma)} = \left[I\left(\gamma_k = 1\right) \ldots I\left(\gamma_k = H_k\right) \right]^\top.$$

De plus, désignons le deuxième membre de (4.10) par $\mathbf{z}_1^{(\gamma)},$ le deuxième membre de (4.11) par $\mathbf{u}_{g1}^{(\gamma)}$ et, pour $k = 1, \ldots, K-1,$ définissons

$$\begin{aligned}
\mathbf{z}_{2k}^{(\gamma)} &= \left(\left[\mathbf{z}_{1(k+1)}^{(\gamma)\top} \ldots \mathbf{z}_{1K}^{(\gamma)\top} \right] \otimes \mathbf{z}_{1k}^{(\gamma)\top} \right)^\top, \\
\mathbf{u}_{g2k} &= \Bigg[\overbrace{u_{g,k(k+1)(11)} \ldots u_{g,k(k+1)(H_k 1)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau 1 de } \gamma_{k+1} \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \overbrace{u_{g,k(k+1)(1H_{k+1})} \ldots u_{g,k(k+1)(H_k H_{k+1})}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau } H_{k+1} \text{ de } \gamma_{k+1} \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \\
&\qquad \overbrace{u_{g,kK(11)} \ldots u_{g,kK(H_k 1)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau 1 de } \gamma_K \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \ldots \overbrace{u_{g,kK(1H_K)} \ldots u_{g,kK(H_k H_K)}}^{\substack{\text{Termes d'interaction entre} \\ \text{le niveau } H_K \text{ de } \gamma_K \text{ et} \\ \text{tous les niveaux de } \gamma_k}} \Bigg]^\top,
\end{aligned}$$

où $\otimes$ est le produit de Kronecker et $\mathbf{u}_{g2k}$ sont les termes d'interaction entre γ_k et $\gamma_{k+1}, \ldots, \gamma_K$. Nous obtenons alors

$$\begin{aligned}
\mathbf{z}^{(\gamma)} &= \left[\mathbf{z}_1^{(\gamma)\top} \; \mathbf{z}_{21}^{(\gamma)\top} \ldots \mathbf{z}_{2(K-1)}^{(\gamma)\top} \right]^\top, \\
\mathbf{u}_g &= \left[\mathbf{u}_{g1}^\top \, \mathbf{u}_{g21}^\top \ldots \mathbf{u}_{g2(K-1)}^\top \right]^\top.
\end{aligned}$$

En général, pour $t \le K,$ le coefficient de l'interaction entre les événements $\{\gamma_{k_1} = l_{k_1}\}, \ldots, \{\gamma_{k_t} = l_{k_t}\}$ est désigné par $u_{g,k_1 \ldots k_t(l_{k_1} \ldots l_{k_t})}$. Comme auparavant, le coefficient est fixé à zéro par convention si $\min\left(l_1, \ldots, l_t\right) = 0.$ La covariable associée au coefficient est l'indicateur $I\left(\left(\gamma_{k_1}, \ldots, \gamma_{k_t}\right) = \left(l_{k_1}, \ldots, l_{k_t}\right)\right)$. Quand le modèle comprend tous les principaux termes et interactions d'ordre d ou d'un ordre inférieur, nous avons

$$\begin{aligned}
\mathbf{z}^{(\gamma)\top} \mathbf{u}_g &= \sum_{t=1}^{d} \sum_{1 \le k_1 < \ldots < k_t \le K} \sum_{l_{k_1}=1}^{H_{k_1}} \ldots \sum_{l_{k_t}=1}^{H_{k_t}} \\
&\qquad I\left(\left(\gamma_{k_1}, \ldots, \gamma_{k_t}\right) = \left(l_{k_1}, \ldots, l_{k_t}\right)\right) u_{g,k_1 \ldots k_t(l_{k_1} \ldots l_{k_t})}.
\end{aligned}$$

Selon le théorème 2 de l'annexe B, cela signifie que l'on peut estimer les paramètres du modèle de mélange asymptotique (dont la couverture $P(i \in S_A)$) de manière convergente en maximisant la probabilité des $\mathbf{n}_i,$ dans les conditions énoncées. À titre d'exemple, cette méthodologie peut être intéressante dans la configuration simple suivante, où le couplage est fondé sur des comparaisons exactes du nom de famille (premier groupe), du prénom (deuxième groupe) et de la date de naissance (troisième groupe), avec $K = 3,$ $H_1 = H_2 = H_3 = 1,$ $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$ et $|\Gamma| = 7.$ On peut estimer la couverture en maximisant la

probabilité des $\mathbf{n}_i$, si les différentes concordances n'ont pas d'interactions du troisième ordre dans les paires appariées pour chaque valeur possible d'un enregistrement dans S_B, c'est-à-dire que tous les termes principaux et les interactions du deuxième ordre peuvent être inclus (cela donne au total six paramètres en plus de la couverture $P(i \in S_A)$). Cela est particulièrement vrai si les concordances sont indépendantes dans les paires appariées. Dans ce cas, la solution est liée à celle décrite par Račinskij et coll. (2019). Elle est également liée à la solution décrite par Brown et coll. (2020), sauf qu'elle ne repose pas sur des examens manuels. Au-delà de ce cas particulier, la distribution des vrais positifs est associée à sept paramètres inconnus ($P(i \in S_A)$ et les six paramètres log-linéaires) et à sept équations (une pour $p^{(\gamma)}$ pour chaque γ), pour chaque composante du mélange.

Ici, quelques remarques sont nécessaires. Premièrement, le modèle proposé tient implicitement compte de toutes les interactions entre les variables de couplage dans les paires non appariées, tout en tenant compte de toutes les interactions d'ordre $K-1$ ou d'ordre inférieur dans les paires appariées, dans chaque classe d'enregistrement. Il offre donc une plus grande flexibilité de modélisation que les mélanges log-linéaires classiques, tout en conservant la propriété d'identification (voir les lemmes 3 et 4) et la capacité d'estimer de manière convergente les paramètres connexes (voir le théorème 2). Cela s'observe le mieux dans le cas plus simple où la distribution des vrais positifs est homogène entre les enregistrements et $H_1 = \ldots = H_K = 1$, c'est-à-dire K comparaisons dichotomiques. Dans ce cas, il est clair que l'on ne peut pas utiliser un mélange log-linéaire à K dimensions à deux composantes pour modéliser les paires d'enregistrements, tout en tenant compte de toutes les interactions d'ordre $K-1$ ou d'ordre inférieur dans les paires appariées. En effet, cela signifie qu'il y a au moins 2^K paramètres libres, dont $2^K - 2$ paramètres pour les paires appariées, un paramètre pour la proportion de mélange et au moins un paramètre pour les paires non appariées. Cependant, il n'y a que 2^K combinaisons observables et, par conséquent, seulement $2^K - 1$ équations pour déterminer les paramètres. Le même problème se produit quand $K = 2$, même en l'absence d'interactions dans la distribution des paires appariées. Deuxièmement, la souplesse de modélisation supplémentaire facilite grandement la conception des règles de couplage, qui répondent aux conditions de l'estimation convergente de la couverture (voir le théorème 2 et le lemme 4).

4.6 Capture hétérogène et enregistrements incomplets

La méthodologie est applicable quand la probabilité de capture varie selon la poststrate en fonction des covariables qui sont enregistrées sans erreur dans chaque échantillon, dans la mesure où les hypothèses posées (voir les sections 2.1, 3.3 et l'annexe) se vérifient dans chaque poststrate. Dans ce cas, S_A et S_B correspondent aux sous-ensembles d'enregistrements d'une poststrate, où la probabilité de capture peut être estimée au moyen d'un des modèles de voisinage. Bien entendu, la construction des poststrates est une question pratique importante, qui sera étudiée dans de futurs travaux.

Une autre préoccupation d'ordre pratique est la présence d'enregistrements incomplets dans l'un des échantillons. Pour traiter la question sans encombrer la notation, supposons que S_A et S_B désignent maintenant les deux échantillons dans une poststrate, que S'_A et S'_B désignent les sous-échantillons correspondants des enregistrements complets, et que $P(i \in S'_A)$ désigne la couverture de S'_A dans la poststrate,

où i est une unité qui s'y trouve. Little et Rubin (1987, chapitre 1.2) ont décrit différentes stratégies pour effectuer une analyse statistique en présence d'enregistrements incomplets. Une première solution consiste à utiliser uniquement les enregistrements complets, avec ou sans pondération visant à tenir compte des enregistrements incomplets. Les deux autres solutions sont l'imputation des valeurs manquantes et la méthode fondée sur un modèle, qui consiste à maximiser la probabilité que les données soient incomplètes. Dans ce cas, on peut envisager la première solution sans repondérer les enregistrements complets, quand les hypothèses énoncées s'appliquent dans chaque poststrate, notamment le fait que l'inclusion dans S'_A et celle dans S'_B sont indépendantes et que la probabilité d'inclusion dans S'_A est uniforme. Essentiellement, l'idée consiste à traiter les données manquantes comme une deuxième étape de la sélection et à estimer la couverture en deux étapes comme suit. Premièrement, on applique l'une des deux méthodes proposées dans la poststrate pour obtenir une estimation de $\hat{P}(i \in S'_A)$ de la couverture de S'_A. Deuxièmement, on estime la couverture de S_A (dans la poststrate) par $|S_A| / (|S'_A| / \hat{P}(i \in S'_A))$.

5. Simulations

La méthodologie proposée est évaluée au moyen de simulations Monte Carlo comprenant 100 répétitions. Dans une répétition, on génère une population finie comportant 100 000 personnes, où chaque personne se voit attribuer un nom de famille et une date de naissance. On tire deux échantillons de Bernoulli de cette population, dans lesquels le nom de famille et la date de naissance ont pu être enregistrés avec des erreurs typographiques. Ensuite, les échantillons sont liés et la couverture est estimée au moyen des modèles proposés et selon Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019), à des fins de comparaison. Différents scénarios sont envisagés en fonction de différentes règles de couplage, y compris certains où s'applique l'hypothèse d'indépendance conditionnelle et où le rappel est parfait. Des précisions sont données dans les paragraphes qui suivent.

5.1 Population finie et sources de données

Pour le nom de famille et la date de naissance, les fréquences sont fondées sur le croisement des répartitions du nom de famille et de l'âge du recensement de la population de 2010 aux États-Unis (US Census Bureau, 2020, 2016). Pour le nom de famille, la fréquence relative est calculée après exclusion de l'observation « tous les autres noms de famille ». Par souci de simplicité, le mois est tiré uniformément de $\{1, \ldots, 12\}$ et le jour est tirée indépendamment et uniformément de $\{1, \ldots, 30\}$. Par conséquent, les composantes du nom de famille et de la date sont mutuellement indépendantes dans la population. Deux registres complets sont créés et les variables sont perturbées dans le deuxième registre. Cette perturbation est décrite du point de vue de la concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance, et d'un critère de référence, qui consiste à avoir le même code SOUNDEX pour le nom de famille, la même année de naissance, ainsi qu'une différence absolue inférieure à 2 pour le jour et le mois. Pour être plus précis, supposons que γ_1, γ_2 et γ_3 désignent les variables indicatrices, qui correspondent à la

satisfaction du critère de référence en plus d'avoir le même nom de famille, le même jour de naissance ou le même mois de naissance, comme le montre le tableau 5.1. Par exemple, quand $\gamma_1 = 1,$ le critère de référence est satisfait et le nom de famille est identique. Quand $\gamma_1 = 0,$ le critère de référence n'est pas satisfait ou il est satisfait et le nom de famille est différent. Dans chaque cas, le jour et le mois de naissance peuvent être identiques ou différents. Pour une personne donnée, les enregistrements associés dans le premier registre et le deuxième registre sont appelés respectivement « premier enregistrement » et « second enregistrement ». On obtient le second enregistrement en tirant d'abord $\gamma = (\gamma_1, \gamma_2, \gamma_3),$ puis en choisissant la valeur de l'enregistrement en fonction de la valeur du premier enregistrement et $\gamma.$ Par exemple, quand $\gamma = (0,1,1),$ le second enregistrement est tel qu'il a la même date de naissance que le premier enregistrement, mais un nom de famille différent ayant le même code SOUNDEX. Au moyen de la convention de codage « dummy », nous pouvons écrire la distribution de γ sous une forme log-linéaire comme suit :

$$P(\gamma) = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1}\gamma_{k_2} u_{k_1 k_2 (11)} + \gamma_1\gamma_2\gamma_3 u_{123(111)} \right), \tag{5.1}$$

où le terme constant u est une fonction des termes principaux et des termes d'interaction parce que $\sum_{\gamma \in \{0,1\}^3} P(\gamma) = 1.$ Par souci de simplicité, nous choisissons les paramètres de sorte que $u_{1(1)} = u_{2(1)} = u_{3(1)}$ et $u_{12(11)} = u_{13(11)} = u_{23(11)}.$ Quand $\gamma_1 = 0,$ le nom de famille dans le second enregistrement est tiré des noms de famille de l'autre recensement ayant le même code SOUNDEX, selon leur fréquence. Quand $\gamma_2 = 0,$ on obtient le jour dans le second enregistrement en augmentant ou en diminuant aléatoirement le jour du premier enregistrement de 1, avec une probabilité de 1/2 pour chaque option, sauf quand le jour est 1 ou 30 dans le premier enregistrement, auquel cas le jour est respectivement augmenté de 1 ou réduit de 1. De même, quand $\gamma_3 = 0,$ on obtient le mois du second enregistrement en augmentant ou en diminuant aléatoirement le mois du premier enregistrement de 1, avec une probabilité de 1/2 pour chaque option, sauf quand le mois est 1 ou 12 dans le premier enregistrement, auquel cas le mois est augmenté de 1 ou réduit de 1, respectivement. On tire un échantillon Bernoulli indépendant de chaque registre, avec une probabilité d'inclusion de 0,9, qui est la couverture réelle.

Tableau 5.1
Indicateurs des perturbations dans un enregistrement.

Indicateur	Critère de référence	Nom de famille identique	Jour de naissance identique	Mois de naissance identique
γ_1	√	√	?	?
γ_2	√	?	√	?
γ_3	√	?	?	√

5.2 Couplage

On considère deux règles de couplage. La première règle permet de coupler les paires qui répondent au critère de référence, ou le sous-ensemble de ces paires dans lequel il y a au moins une concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance, selon le scénario. On utilise les liens qui en résultent pour estimer la couverture au moyen des modèles de voisinage univariés et multivariés et d'un modèle classique de mélange log-linéaire qui intègre l'hypothèse d'indépendance conditionnelle décrite par Račinskij et coll. (2019). Pour une paire donnée, le vecteur des résultats est basé sur les indicateurs de concordance exacte pour le nom de famille, le jour de naissance et le mois de naissance, par exemple $(1,1,1)$ pour une concordance parfaite sur le nom de famille et les deux composantes de la date. Afin d'estimer la couverture au moyen des méthodes proposées par Ding et Fienberg (1994) et Di Consiglio et Tuoto (2015), il faut une deuxième règle de couplage, avec au plus un lien par enregistrement ainsi que des estimations manuelles de l'exactitude du couplage qui en résulte. On dérive cette règle à partir de la première en supprimant un lien, si au moins un enregistrement impliqué comporte plusieurs liens. Les estimations manuelles de l'exactitude du couplage sont fondées sur un échantillon aléatoire simple de 1 000 paires d'enregistrements, qui satisfont au critère de référence, et l'utilisation de la table de vérité. Notons que cette procédure ne tient pas compte des faux négatifs générés par le critère de référence (semblable à un critère de création des pochettes), lorsqu'ils existent.

5.3 Scénarios

Cinq scénarios sont envisagés. Dans le premier scénario, l'hypothèse d'indépendance conditionnelle est satisfaite à partir de $u_{1(1)} = 1, u_{12(11)} = 0$ et $u_{123(111)} = 0,$ et la première règle de couplage est fondée sur le critère de référence et le rappel est parfait. Dans le deuxième scénario, il y a un écart par rapport à l'indépendance conditionnelle en raison des interactions du deuxième ordre, à partir de $u_{1(1)} = 1$ et $u_{12(11)} = 1$ et $u_{123(111)} = 0,$ mais la première règle de couplage reste fondée sur le critère de référence. Dans le troisième scénario, il y a aussi un écart par rapport à l'indépendance conditionnelle en raison des interactions des deuxième et troisième ordres, à partir de $u_{1(1)} = 1, u_{12(11)} = 1$ et $u_{123(111)} = 1/4,$ mais aucun changement n'est apporté à la première règle de couplage. Le quatrième scénario est identique au deuxième scénario, sauf que la première règle de couplage permet de lier maintenant les paires qui répondent au critère de référence et qui ont au moins une concordance exacte sur le nom de famille, le jour de naissance ou le mois de naissance. Enfin, le cinquième scénario est identique au quatrième, sauf que l'interaction du troisième ordre est ajoutée à partir de $u_{123(111)} = 1/4.$ Les caractéristiques des différents scénarios sont résumées dans les tableaux 5.2 et 5.3. Dans ce dernier tableau, les chiffres correspondent aux moyennes des répétitions.

Tableau 5.2
Scénarios de simulation.

Scénario	Paramètres log-linéaires			Indépendance conditionnelle
	$u_{k(1)}$	$u_{k_1k_2(11)}$	$u_{123(111)}$	
1	1	0	0	√
2 et 4	1	1	0	X
3 et 5	1	1	1-4	X

Tableau 5.3
Moyennes empiriques des taux d'erreur de couplage.

Scénario	Couplage	Rappel	Précision	Taux de faux positifs (TFP) $\times 10^{-9}$	Rappel parfait
1	1	1,000	0,952	498,92	√
	2	0,944	1,000	4,15	X
2 et 3	1	1,000	0,953	495,03	√
	2	0,950	1,000	4,17	X
4 et 5	1	0,996	0,964	371,84	X
	2	0,957	1,000	3,48	X

5.4 Estimateurs

Les modèles de voisinage sont appliqués selon une distribution homogène des vrais positifs, pour traduire la configuration actuelle et la situation en pratique, où l'on s'attend à ce que l'hétérogénéité de la distribution des faux positifs soit la source dominante d'hétérogénéité pour la distribution des n_i (Dasylva et Goussanou, 2022). Cela signifie que la probabilité p_g et le vecteur $\mathbf{p}_g = \left[p_g^{(\gamma)} \right]_{\gamma \in \Gamma}$ sont identiques pour toutes les classes. Cela signifie également que le paramètre $\boldsymbol{\beta}_g$ est identique dans toutes les classes, si $\mathbf{p}_g$ a la forme $\varrho\left(\boldsymbol{\beta}_g\right)$ pour une fonction connue $\varrho(.)$. Par souci de commodité, supposons que p, $\mathbf{p}$ et $\boldsymbol{\beta}$ désignent respectivement les valeurs communes de p_g, $\mathbf{p}_g$ et $\boldsymbol{\beta}_g$. Supposons aussi que

$$r^{(\gamma)} = \exp\left(u + \sum_{k=1}^{3} \gamma_k u_{k(1)} + \sum_{1 \le k_1 < k_2 \le 3} \gamma_{k_1} \gamma_{k_2} u_{k_1 k_2 (11)} \right), \gamma \in \Gamma = \{0,1\}^3, \tag{5.2}$$

où $\sum_{\gamma \in \{0,1\}^3} r^{(\gamma)} = 1$, le terme constant u est une fonction des autres paramètres, et le deuxième membre inclut seulement les interactions du deuxième ordre contrairement à celui de (5.1). Ensuite

$$\boldsymbol{\beta} = \left(u_{1(1)}, u_{2(1)}, u_{3(1)}, u_{12(11)}, u_{13(11)}, u_{23(11)} \right),$$

et la fonction $\varrho(.)$ est caractérisée par

$$p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}, \gamma \in \Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}.$$

On calcule les estimations en maximisant la probabilité numériquement dans R, où le nombre de classes est choisi en minimisant le critère d'information d'Akaike. Dans le modèle univarié, les estimations sont fondées sur le plafonnement des n_i par 10 (c'est-à-dire le remplacement de n_i par $\min(10, n_i)$) et la maximisation de la probabilité des observations qui en résultent, comme le décrivent Dasylva et Goussanou (2022). Lorsque l'on utilise le modèle multivarié, on estime la couverture quand seuls les termes principaux sont inclus et aussi quand les termes d'interaction du deuxième ordre sont inclus et qu'on ignore que les termes principaux sont égaux, et que les termes d'interaction du deuxième ordre sont égaux. En général, la maximisation numérique de la probabilité est plus difficile qu'avec le modèle univarié, et les estimations qui en résultent perdent en précision quand la précision du couplage diminue. C'est pourquoi il faut une bonne procédure d'initialisation, décrite à l'annexe C. À des fins de comparaison, nous calculons aussi les estimateurs proposés par Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019) ainsi que l'estimateur naïf de capture-recapture, qui ne tient pas compte des erreurs de couplage. Notons

que ce dernier estimateur est calculé comme étant le rapport du nombre de liens par la deuxième règle de couplage sur $|S_B|$.

5.5 Résultats

Les résultats de la simulation sont présentés dans le tableau 5.4. Dans le scénario 1, où l'hypothèse d'indépendance conditionnelle s'applique, la meilleure performance est obtenue par les estimateurs de Račinskij et coll. (2019) et les estimateurs de voisinage, tant pour ce qui est du biais relatif que de l'erreur quadratique moyenne, avec un avantage pour les estimateurs de voisinage quand on observe cette dernière mesure de performance. Parmi les estimateurs de voisinage, le modèle univarié affiche les meilleures performances pour ce qui est du biais, de la variance et de l'erreur quadratique moyenne. Sans surprise, l'estimateur naïf affiche les pires performances, tandis que les estimateurs de Ding et Fienberg (1994), et de Di Consiglio et Tuoto (2015) donnent de meilleurs résultats, mais une grande variance, parce qu'ils intègrent des estimations manuelles de l'exactitude du couplage. Notons que le mélange log-linéaire proposé par Račinskij et coll. (2019) présente un biais, une variance et une erreur quadratique moyenne plus grands que les estimateurs basés sur des modèles de voisinage multivariés, à une petite exception près. (Dans le tableau 5.4, le mélange log-linéaire comporte un biais relatif légèrement plus petit que celui du modèle de voisinage multivarié avec interactions.) En effet, tous ces estimateurs visent à estimer la couverture en exploitant la structure de corrélation des variables de couplage, ce que fait pleinement le mélange log-linéaire en intégrant l'indépendance des variables de couplage dans les paires appariées et dans les paires non appariées. Toutefois, les modèles de voisinage multivariés n'exploitent que l'information sur la structure de corrélation dans les paires appariées sans contrainte sur les paires non appariées. Pourtant, ils donnent des estimateurs qui sont nettement plus précis que le mélange log-linéaire pour ce qui est de l'erreur quadratique moyenne. Cela illustre la différence importante entre les mélanges log-linéaires classiques et les modèles de voisinage multivariés, quand ces derniers intègrent une spécification log-linéaire de la structure de corrélation dans les paires appariées.

Dans le scénario 2, les estimateurs proposés continuent de donner les plus petites erreurs quadratiques moyennes. Comme auparavant, le modèle univarié affiche les meilleures performances globales pour ce qui est du biais, de la variance et de l'erreur quadratique moyenne. Parmi les deux modèles multivariés, celui qui comprend les interactions donne de meilleurs résultats, comme on peut s'y attendre, étant donné un biais environ quarante fois plus petit et une erreur quadratique moyenne environ cinq fois plus petite. Sans surprise, l'estimateur de Račinskij et coll. (2019) donne de moins bons résultats que dans le scénario précédent, en raison de la violation de l'hypothèse d'indépendance conditionnelle dans ce scénario. Cette dégradation est toutefois telle que ses performances sont pires que celles de l'estimateur naïf pour chaque mesure de performance. De façon intéressante, nous observons que ses performances sont aussi nettement moins bonnes que celles de l'estimateur basé sur le modèle de voisinage multivarié sans interactions, étant donné qu'il comporte un biais relatif et une erreur quadratique moyenne plus grands de plus d'un ordre de grandeur, tandis que ce dernier estimateur donne de meilleurs résultats que l'estimateur naïf pour chaque mesure de performance. Entre autres explications possibles, cela est attribuable au fait que le modèle de

voisinage multivarié tient implicitement compte de toutes les interactions dans la distribution des paires non appariées, tandis que le modèle de Račinskij et coll (2019) ignore ces interactions. Cela est une illustration supplémentaire de la différence entre les mélanges log-linéaires classiques et les modèles de voisinage multivariés. Comme auparavant, les estimateurs de Ding et Fienberg (1994) et de Di Consiglio et Tuoto (2015) affichent de moins bons résultats que les modèles de voisinage, en raison de la variance de l'exactitude estimée du couplage. Cependant, pour ce qui est du biais et de l'erreur quadratique moyenne, ils donnent de meilleurs résultats que l'estimateur naïf et que celui de Račinskij et coll. (2019). Le scénario 3 diffère du scénario 2 par l'ajout d'une interaction du troisième ordre avec le coefficient 1/4. Ce changement a toutefois un effet négligeable sur les résultats obtenus et les tendances observées.

Dans le scénario 4, la première règle de couplage a un faible taux de faux négatifs d'environ 0,4 % (c'est-à-dire un rappel imparfait), car elle ne permet pas de lier les paires qui n'ont pas de concordance exacte pour le nom de famille, le jour de naissance ou le mois de naissance. Cela a un effet direct sur l'estimateur de voisinage univarié, qui a maintenant la troisième plus petite erreur quadratique moyenne, derrière les deux estimateurs de voisinage multivariés, celui qui présente les interactions ayant les meilleures performances. L'exclusion des paires sans concordance exacte dégrade davantage les performances de l'estimateur de mélange log-linaire (comparativement aux scénarios 3 et 4), qui présente toujours la plus grande erreur quadratique moyenne et de moins bonnes performances que l'estimateur naïf. Toutefois, ce changement a un effet limité sur les performances des estimateurs de Ding et Fienberg (1994) et de Di Consiglio et Tuoto (2015). Le scénario 5 diffère du scénario 4 par l'ajout d'une interaction du troisième ordre avec le coefficient 1/4. Ce changement a toutefois un effet négligeable sur les résultats.

En résumé, les résultats de simulation démontrent que les estimateurs proposés peuvent servir à estimer la couverture en ayant un petit biais relatif et une erreur quadratique moyenne plus petite que les autres estimateurs proposés par Ding et Fienberg (1994), Di Consiglio et Tuoto (2015) et Račinskij et coll. (2019) quand les faux négatifs sont négligeables ou que les probabilités de vrais positifs sont contraintes par une spécification log-linéaire.

Tableau 5.4
Résultats de simulation.

Scénario	Estimateur	Biais relatif (%)	Variance $\times 10^{-7}$	Erreur quadratique moyenne $\times 10^{-7}$
1	Naïf	-5,522	12,90	24 711,31
	R	-0,034	824,62	817,32
	DF	1,618	2 377,74	4 475,68
	DT	1,159	2 550,13	3 613,51
	VUNI	-0,003	8,43	8,35
	VMULTI sans interactions	-0,023	28,25	28,41
	VMULTI avec interactions du deuxième ordre	-0,119	27,06	38,25

DF : estimateur de Ding et Fienberg (1994)
DT : estimateur de Di Consiglio et Tuoto (2015)
VMULTI : estimateur basé sur le modèle de voisinage multivarié
Naïf : estimateur de Lincoln-Petersen qui ignore les erreurs de couplage
R : estimateur de Račinskij et coll. (2019)
VUNI : estimateur basé sur le modèle de voisinage univarié

Tableau 5.4(suite)
Résultats de simulation.

Scénario	Estimateur	Biais relatif (%)	Variance $\times 10^{-7}$	Erreur quadratique moyenne $\times 10^{-7}$
2 et 3	Naïf	-4,994	15,10	20 216,56
	R	-7,784	324,65	49 403,40
	DF	1,667	3 381,80	5 598,43
	DT	0,961	3 560,11	4 272,01
	VUNI	-0,004	8,31	8,24
	VMULTI sans interactions	-0,423	21,05	165,44
	VMULTI avec interactions du deuxième ordre	-0,091	23,70	30,12
4 et 5	Naïf	-4,292	15,46	14 934,57
	R	-3,497	280 414,54	287 515,73
	DF	1,760	2 255,34	4 740,96
	DT	1,159	2 550,13	3 613,51
	VUNI	-0,393	9,30	134,50
	VMULTI sans interactions	-0,423	21,05	165,44
	VMULTI avec interactions du deuxième ordre	-0,091	23,70	30,12

DF : estimateur de Ding et Fienberg (1994)
DT : estimateur de Di Consiglio et Tuoto (2015)
VMULTI : estimateur basé sur le modèle de voisinage multivarié
Naïf : estimateur de Lincoln-Petersen qui ignore les erreurs de couplage
R : estimateur de Račinskij et coll. (2019)
VUNI : estimateur basé sur le modèle de voisinage univarié

6. Conclusion

Nous avons décrit une nouvelle méthodologie aux fins d'estimation par la capture-recapture comportant des erreurs de couplage, qui est fondée sur la modélisation du nombre de liens à partir d'un enregistrement, sans examens manuels, notamment un modèle univarié et un modèle multivarié connexe. Dans le modèle univarié, on estime la couverture en liant les enregistrements dont le rappel est suffisamment élevé. Dans le modèle multivarié, on estime la couverture en contraignant les interactions dans les paires appariées au moyen d'une spécification log-linéaire, tout en permettant des interactions arbitraires dans les paires non appariées, ce qui tranche nettement avec les mélanges log-linéaires classiques. Dans ce dernier cas, il faut lier les enregistrements avec une grande précision pour obtenir une estimation fiable de la couverture. Les simulations à partir des données de recensement publiques démontrent les bonnes performances des estimateurs proposés, comparativement aux solutions précédentes. De futurs travaux porteront sur l'obtention des variances et d'intervalles de confiance, ainsi que sur la validation de la spécification log-linéaire quand le modèle multivarié est utilisé.

Remerciements

Le présent article expose les opinions des auteurs qui ne sont pas nécessairement celles de Statistique Canada.

Annexe

A. Extension pour la sous-couverture

La présente annexe vise à étendre les résultats de Dasylva et Goussanou (2022) pour démontrer que

$$\begin{aligned}
\frac{\text{VP}}{\text{VP + FN}} &\xrightarrow{p} P(i \in S_A)^{-1}\, \overline{p}, \\
\frac{\text{VP}}{\text{VP + FP}} &\xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}}, \\
\hat{\overline{p}} &\xrightarrow{p} \overline{p}, \\
\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} &\xrightarrow{p} \frac{\overline{p}}{\overline{p} + \overline{\lambda}}.
\end{aligned}$$

Par conséquent, $P(i \in S_A)^{-1}\, \hat{\overline{p}}$ et $\hat{\overline{p}} \Big/ \left(\hat{\overline{p}} + \hat{\overline{\lambda}}\right)$ estiment le rappel et la précision de façon convergente, en ce sens que

$$\begin{aligned}
P(i \in S_A)^{-1}\, \hat{\overline{p}} - \frac{\text{VP}}{\text{VP + FN}} &\xrightarrow{p} 0, \\
\frac{\hat{\overline{p}}}{\hat{\overline{p}} + \hat{\overline{\lambda}}} - \frac{\text{VP}}{\text{VP + FP}} &\xrightarrow{p} 0.
\end{aligned}$$

L'extension consiste à tenir compte de la sous-couverture dans S_A. Pour continuer, il faut des notations supplémentaires. Appelons V_j' un *voisin* de V_i, si l'unité j est incluse dans S_A et V_j' est contenu dans le voisinage de V_i, c'est-à-dire $\mathcal{B}_N(V_i)$. Le voisin est considéré comme apparié si les deux enregistrements proviennent de la même unité. Sinon, il est considéré comme non apparié. De plus, définissons la notation supplémentaire suivante. Pour $i, i' \in S_B$ de sorte que $i \neq i'$, soit

$$n_i^{(0)} = \sum_{j \in S_A} I\left(V_j' \in \mathcal{B}_N(V_i)\right), \tag{A.1}$$

$$\left(n_{i|U}^{(0)}, n_{i|U}\right) = \left(n_i^{(0)}, n_i\right) - \left(I\left(\{i \in S_A\} \cap \left\{V_i' \in \mathcal{B}_N(V_i)\right\}\right), I(i \in S_A)\, L_{ii}\right), \tag{A.2}$$

$$\left(n_{ii'|U}^{(0)}\right) = \sum_{t \in S_A : t \neq i, i'} I\left(V_t' \in \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'})\right), \tag{A.3}$$

où $n_i^{(0)}$ est le nombre de voisins de V_i, $n_{i|U}^{(0)}$ est le nombre de voisins non appariés de cet enregistrement, $n_{i|U}$ est le nombre d'enregistrements non appariés, qui sont liés au même enregistrement, et $n_{ii'|U}^{(0)}$ est le nombre de voisins non appariés, qui sont communs à V_i et $V_{i'}$. En ce qui concerne $n_i^{(0)}$ et $n_{i|U}^{(0)}$, soit $S_{Ai} = \left\{t \in S_A \text{ s.t. } V_t' \in \mathcal{B}_N\left(V_i\right)\right\}$ le sous-ensemble d'unités, qui sont incluses dans S_A et ont leur enregistrement dans le voisinage, et soit $S_{Ai|U} = S_{Ai} - \{i\}$ le sous-ensemble de ces unités qui sont différentes de l'unité i. Ces dernières unités sont associées à des voisins non appariés. Enfin, pour les variables aléatoires (ou vecteurs) X, Y et Z, notons l'indépendance de X et Y par $X \perp\!\!\!\perp Y$, et leur indépendance conditionnelle étant donné Z par $X \perp\!\!\!\perp Y \mid Z$.

Le lemme suivant est une extension du lemme 2 présenté dans Dasylva et Goussanou (2022). Toutes les démonstrations se trouvent dans la version longue de l'article (Dasylva, Goussanou et Nambeu, 2024).

Lemme 1. *Supposons que* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_i'\right)\right]_{1 \le i \le N}$ *sont indépendants et identiquement distribués et que* $Z_1, \ldots, Z_N$ *désignent des variables aléatoires identiquement distribuées, de sorte qu'elles soient conditionnellement indépendantes étant donné* $\left[\left(I(i \in S_A), I(i \in S_B), V_i, V_i'\right)\right]_{1 \le i \le N}$ *avec une distribution marginale conditionnelle de* Z_i *qui est seulement une fonction de* $I(i \in S_A)$, $I(i \in S_B)$, V_i, $S_{Ai|U}$, $\left[V_t'\right]_{t \in S_{Ai|U}}$ *et* V_i'. *Alors*

$$\left(I(i' \in S_A), V_{i'}', \left[V_t'\right]_{t \in S_{Ai'|U}}\right) \perp\!\!\!\perp \left(I(i \in S_A), V_i', \left[V_t'\right]_{t \in S_{Ai'|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V_i', S_{Ai'|U}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), \\ V_{i'}' \notin \mathcal{B}_N(V_i), \end{array} \right. \tag{A.4}$$

$$S_{Ai'|U} \perp\!\!\!\perp \left(I(i \in S_A), V_i', \left[V_t'\right]_{t \in S_{Ai|U}}\right) \left| \begin{array}{l} i \in S_B, V_i, S_{Ai|U}, \\ i' \in S_B, V_{i'}, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), \\ V_{i'}' \notin \mathcal{B}_N(V_i). \end{array} \right. \tag{A.5}$$

Aussi, pour tous $(v_i, v_{i'}) \in \mathcal{V}_N^* \times \mathcal{V}_N^*$, $a_i \in \{0,1\}$, $w_i \notin \mathcal{B}_N(v_{i'})$, $s_{Ai|U} \subset \{1, \ldots, N\} \setminus \{i, i'\}$ *et* $\left[w_t\right]_{t \in s_{Ai|U}}$ *fixes de sorte que* $\mathcal{B}_N(v_i) \cap \mathcal{B}_N(v_{i'}) = \varnothing$ *et* $w_t \in \mathcal{B}_N(v_i)$ *pour tous les* $t \in s_{Ai|U}$

$$P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & i \in S_B, V_i = v_i, \\ V_i' = w_i, & S_{Ai|U} = s_{Ai|U}, \\ \left[V_t'\right]_{t \in s_{Ai|U}} = & i' \in S_B, V_{i'} = v_{i'} \\ \left[w_t\right]_{t \in s_{Ai|U}} & \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ & V_i' \notin \mathcal{B}_N(V_{i'}), \\ & V_{i'}' \notin \mathcal{B}_N(V_i) \end{array}\right) = P\left(\begin{array}{l|l} I(i \in S_A) = a_i, & \\ V_i' = w_i, & i \in S_B, V_i = v_i, \\ \left[V_t'\right]_{t \in s_{Ai|U}} = & S_{Ai|U} = s_{Ai|U}, \\ \left[w_t\right]_{t \in s_{Ai|U}} & V_i' \notin \mathcal{B}_N(v_{i'}) \end{array}\right). \tag{A.6}$$

Donc

$$E\left[Z_i Z_{i'} \left| \begin{array}{l} i \in S_B, V_i = v_i, n_{i|U}^{(0)} = k, \\ i' \in S_B, V_{i'} = v_{i'}, n_{i'|U}^{(0)} = \ell, \\ \mathcal{B}_N(V_i) \cap \mathcal{B}_N(V_{i'}) = \varnothing, \\ V_i' \notin \mathcal{B}_N(V_{i'}), V_{i'}' \notin \mathcal{B}_N(V_i) \end{array} \right. \right] = g(v_i, k; v_{i'})\, g(v_{i'}, \ell; v_i), \tag{A.7}$$

où

$$g(v_i, k; v_{i'}) = E\left[Z_i \left| \begin{array}{l} i \in S_B, V_i = v_i, \\ n_{i|U}^{(0)} = k, V_i' \notin \mathcal{B}_N(v_{i'}) \end{array} \right. \right]. \tag{A.8}$$

Le lemme ci-dessus mène au théorème suivant, qui prolonge le théorème 1 présenté dans Dasylva et Goussanou (2022).

Théorème 1. *Considérons que* $V_N,\ \mathcal{B}_N(.),\ S_A,\ S_B$ *et* $[Z_i]_{1\le i\le N}$ *sont des variables aléatoires identiquement distribuées, de sorte que les conditions suivantes s'appliquent.*

(C.1) $\lim_{N\to\infty} P(i\in S_B)=\tau$ *pour un certain* τ *positif.*

(C.2) Pour $\Lambda\ge 0$ *ne dépendant pas de* N, $\sup_{v\in V_N^*}(N-1)\,\lambda_N^{(0)}(v)\le\Lambda$.

(C.3) Pour $c\ge 0$ *ne dépendant pas de* N

$$NP\left(\mathcal{B}_N(V_i)\cap\mathcal{B}_N(V_{i'})\ne\varnothing\mid\{i,i'\}\subset S_B\right)\le c,$$

$$NP\left(V_i'\in\mathcal{B}_N(V_{i'})\,\middle|\,\{i,i'\}\subset S_B,\ \mathcal{B}_N(V_i)\cap\mathcal{B}_N(V_{i'})=\varnothing\right)\le c.$$

(C.4) $Z_1,\ldots,Z_N$ *sont conditionnellement indépendants étant donné* $\left[\left(I(i\in S_A),I(i\in S_B),V_i,V_i'\right)\right]_{1\le i\le N}$ *de sorte que la distribution marginale conditionnelle de* Z_i *est seulement une fonction de* $I(i\in S_A)$, $I(i\in S_B)$, V_i, $S_{Ai|U}$, $[V_t']_{t\in S_{Ai|U}}$ *et* V_i'.

(C.5) $|Z_i|\le R_N\left(n_{i|U}^{(0)}\right)$ *où* $R_N(.)$ *est un polynôme de degré fini ne dépendant pas de* N *et des coefficients non négatifs de* $O\left((\log N)^d\right)$, *où* d *ne dépend pas non plus de* N.

(C.6) $\lim_{N\to\infty} E\left[Z_i\,\middle|\,i\in S_B\right]=\mu$, *où* $|\mu|<\infty$.

Alors

$$\frac{1}{|S_B|}\sum_{i\in S_B} Z_i\ \xrightarrow{p}\ \mu. \qquad (A.9)$$

Le résultat suivant montre la convergence du rappel et de la précision respectivement vers $P(i\in S_A)^{-1}\,\overline{p}$ et $\overline{p}/(\overline{p}+\overline{\lambda})$. Il prolonge le corollaire 1 dans Dasylva et Goussanou (2022) et est une conséquence directe du théorème 1.

Corollaire 1. *Supposons que les hypothèses C.1 à C.3 se vérifient et que le couplage respecte les conditions suivantes :*

(C.7) $[L_{1j}]_{1\le j\le N},\ldots,[L_{Nj}]_{1\le j\le N}$ *sont conditionnellement indépendants étant donné que* $[(I(i\in S_A),I(i\in S_B),V_i,V_i')]_{1\le i\le N}$, *quand la distribution conditionnelle de* $[L_{ij}]_{1\le j\le N}$ *est seulement une fonction de* $I(i\in S_A)$, $I(i\in S_B)$, V_i, $S_{Ai|U}$, $[V_t']_{t\in S_{Ai|U}}$ *et* V_i'.

(C.8)

$$\lim_{N\to\infty} E\left[\left(p_N(V_i),(N-1)\,\lambda_N(V_i)\right)\ \middle|\ i\in S_B\right]=\left(\overline{p},\overline{\lambda}\right). \qquad (A.10)$$

Alors

$$\left(\frac{\text{VP}}{\text{VP}+\text{FN}},\ \frac{\text{VP}}{\text{VP}+\text{FP}}\right) \xrightarrow{p} \left(\frac{\overline{p}}{P(i\in S_A)},\ \frac{\overline{p}}{\overline{p}+\overline{\lambda}}\right). \tag{A.11}$$

En particulier, (A.11) se vérifie si C.8 est remplacé par la condition

$$\left(p_N(V_i),\,(N-1)\,\lambda_N(V_i)\right)\ \Big|\ \{i\in S_B\} \xrightarrow{d} F(.,.)$$

avec $\overline{p}=\int p dF(p,\lambda)$ *et* $\overline{\lambda}=\int \lambda dF(p,\lambda).$

D'autres résultats présentés dans Dasylva et Goussanou (2022) demeurent valides. Ils sont fondés sur le théorème 1 et le corollaire 1, comme le lemme 2 (convergence dans la distribution de n_i vers un mélange comme dans le deuxième membre de (3.8)), le théorème 2 (convergence de l'estimateur de Blakely et Salmond (2002) dans le cas homogène) et le théorème 3 (convergence de l'estimateur par le maximum de vraisemblance). On peut facilement le constater en examinant les démonstrations connexes dans Dasylva et Goussanou (2022). Ainsi, les estimateurs $\hat{\overline{p}}$ et $\hat{\overline{\lambda}}$ sont convergents, et $P(i\in S_A)^{-1}\hat{\overline{p}}$ et $\hat{\overline{p}}\Big/\left(\hat{\overline{p}}+\hat{\overline{\lambda}}\right)$ sont des estimateurs convergents du rappel et de la précision, respectivement.

B. Extension multivariée

Pour décrire la version multivariée du modèle de voisinage, supposons que Γ désigne l'ensemble d'indices d'un ensemble de règles et que $L_{ij}^{(\gamma)}$ indique si la paire (i,j) est couplée par la règle $\gamma\in\Gamma$. (Pour éviter tout conflit avec la notation définie précédemment, on suppose que les règles sont indexées de sorte que Γ ne contienne pas 0.) L'ensemble Γ peut prendre diverses formes, comme un sous-ensemble de nombres entiers consécutifs à partir de 1, ou il peut être un sous-ensemble de $\{0,1\}^K$ si l'on lie les enregistrements avec les K variables de couplage et si l'on effectue une comparaison exacte pour chaque variable. Pour $\gamma\in\Gamma$, soit $n_i^{(\gamma)}=\sum_{j\in S_A} L_{ij}^{(\gamma)}$, $\mathbf{n}_i=\left[n_i^{(\gamma)}\right]_{\gamma\in\Gamma}$, et redéfinissons $\mathcal{B}_N(v)$ comme un sous-ensemble de $\mathcal{V}_N$, qui satisfait à la condition

$$\mathcal{B}_N(v)\ \supset\ \left\{v'\in\mathcal{V}_N\ \text{s.t.}\ E\left[\sum_{\gamma\in\Gamma} L_{ij}^{(\gamma)}\ \middle|\ (i,j)\in S_B\times S_A, \left(V_i,\,V_j'\right)=(v,v')\right]\ >\ 0\right\}.$$

En d'autres termes, $\mathcal{B}_N(v)$ est un surensemble de valeurs d'enregistrement, qui sont liées par au moins une règle de l'ensemble avec une probabilité positive, étant donné que $i\in S_B$ et $V_i=v$. La fonction $\lambda_N^{(0)}(.)$ reste définie par (4.4), tandis que $p_N(.)$ et $\lambda_N(.)$ sont remplacées par les vecteurs $\mathbf{p}_N(v)=\left[p_N^{(\gamma)}(v)\right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda}_N(v)=\left[\lambda_N^{(\gamma)}(v)\right]_{\gamma\in\Gamma}$, avec

$$p_N^{(\gamma)}(v)\ =\ E\left[I\left(i\in S_A\right)L_{ii}^{(\gamma)}\ \middle|\ i\in S_B, V_i=v\right], \tag{B.1}$$

$$\lambda_N^{(\gamma)}(v)\ =\ E\left[I\left(j\in S_A\right)L_{ij}^{(\gamma)}\ \middle|\ i\in S_B, V_i=v\right],\ j\neq i, \tag{B.2}$$

et $\sum_{\gamma\in\Gamma} p_N^{(\gamma)}(v) \le 1$ pour tous les $v \in \mathcal{V}_N^*$. En d'autres termes, $p_N^{(\gamma)}(v)$ et $\lambda_N^{(\gamma)}(v)$ sont les nombres attendus de vrais positifs et de faux positifs pour la règle γ, étant donné que $i \in S_B$ et $V_i = v$. La condition de régularité de (3.7) est remplacée par la condition plus générale suivante :

$$\left(\mathbf{p}_N(V_i),\ (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} F. \tag{B.3}$$

Nous avons un cas particulièrement intéressant quand $\mathbf{p}_N(v)$ a la forme $\varrho\left(\boldsymbol{\beta}_N(v)\right)$, pour une certaine fonction $\boldsymbol{\beta}_N : \mathcal{V}^* \to \mathbb{R}^m$, et une certaine *injection* $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$ indépendante de N, où $m < |\Gamma|$. Dans ce cas, (3.7) est remplacée par la condition suivante :

$$\left(\boldsymbol{\beta}_N(V_i), (N-1)\,\boldsymbol{\lambda}_N(V_i)\right) \,\middle|\, \{i \in S_B\} \xrightarrow{d} H, \tag{B.4}$$

où H ne dépend pas de N.

Le lemme suivant indique la convergence de $\mathbf{n}_i$ vers un mélange multivarié, quand $N \to \infty$ dans les conditions données par (3.6) et (B.3) ou (B.4). La distribution latente est donnée par F ou H selon que (B.3) ou (B.4) s'applique. Dans les deux cas, les distributions des composantes proviennent de familles de distributions discrètes à $|\Gamma|$ dimensions, qui correspondent à la convolution d'une distribution multinomiale avec un produit de distributions de Poisson indépendantes. Pour décrire plus en détail les distributions limites, supposons que $\mathcal{F}$ désigne la famille des distributions des composantes selon (B.3), où chaque membre a la forme

$$\text{IMultinomial}(1, \mathbf{p}) \ * \ \text{PPoisson}(\boldsymbol{\lambda}),$$

pour $\mathbf{p} = \left[p^{(\gamma)}\right]_{\gamma\in\Gamma}$ et $\boldsymbol{\lambda} = \left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$. Quand (B.4) s'applique, les distributions des composantes proviennent du sous-ensemble $\mathcal{F}_\varrho$ de $\mathcal{F}$, où $\mathbf{p} = \varrho(\boldsymbol{\beta})$ pour un certain $\boldsymbol{\beta} \in \mathbb{R}^m$. Un membre de cette famille est une distribution paramétrique dont les paramètres sont $\boldsymbol{\beta} \in \mathbb{R}^m$ et $\boldsymbol{\lambda} \in (0, +\infty)^{|\Gamma|}$. Comme auparavant, toutes les démonstrations se trouvent dans la version longue de l'article (Dasylva et coll., 2024).

Lemme 2. *Supposons que (3.6) s'applique. Si (B.3) s'applique aussi,* $\mathbf{n}_i$ *converge en distribution vers le mélange des distributions provenant de* $\mathcal{F}$*, avec la distribution latente* F*. Si (B.4) s'applique aussi, alors* $\mathbf{n}_i$ *converge en distribution vers le mélange des distributions provenant de* $\mathcal{F}_\varrho$*, avec la distribution latente* H*.*

Le lemme suivant est une extension du lemme 4 présenté dans Dasylva et Goussanou (2022). Il fournit des conditions suffisantes pour l'identification de mélanges finis sur $\mathcal{F}$ (ou $\mathcal{F}_\varrho$); une propriété essentielle pour prouver la convergence des estimateurs par le maximum de vraisemblance. Le lemme nécessite un ordre lexicographique sur $(0,\infty)^{|\Gamma|}$. Pour ce faire, ordonnons les éléments de Γ en fonction d'une certaine application bijective provenant de $\{1,\ldots,|\Gamma|\}$, dans Γ, ce qui est désigné par $\gamma(.)$ avec un léger abus de la notation. Ensuite, désignons un uplet $\left[\lambda^{(\gamma)}\right]_{\gamma\in\Gamma}$ de manière équivalente par $\left[\lambda^{(\gamma(t))}\right]_{1\le t\le|\Gamma|}$, et écrivons $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ (c'est-à-dire $\boldsymbol{\lambda}$ est supérieur à $\boldsymbol{\lambda}'$), si $\lambda^{(\gamma(1))} > \lambda'^{(\gamma(1))}$ ou s'il existe un $t_0 = 2,\ldots,|\Gamma|$ tel que

$\lambda^{(\gamma(t))} = \lambda'^{(\gamma(t))}$ pour $t < t_0$ et $\lambda^{(\gamma(t_0))} > \lambda'^{(\gamma(t_0))}$. Aussi, convenons que $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}$ si $\boldsymbol{\lambda} \succ \boldsymbol{\lambda}'$ sinon $\max(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \boldsymbol{\lambda}'$.

Lemme 3. *Pour les nombres entiers naturels G et G', soit $\boldsymbol{\lambda}_1 \succ \ldots \succ \boldsymbol{\lambda}_G$, et $\boldsymbol{\lambda}'_1 \succ \ldots \succ \boldsymbol{\lambda}'_G$, et désignons par h_g et h'_g les membres de $\mathcal{F}$ dont les paramètres sont $\left(\mathbf{p}_g, \boldsymbol{\lambda}_g\right)$ et $\left(\mathbf{p}'_g, \boldsymbol{\lambda}'_g\right)$, respectivement, et supposons que les mélanges $h = \sum_{g=1}^{G} \alpha_g h_g$ et $h' = \sum_{g=1}^{G'} \alpha'_g h'_g$ sont égaux, où α_g et α'_g sont positifs pour chaque g. Alors $G = G'$, $\alpha_g = \alpha'_g$ et $\left(\mathbf{p}_g, \boldsymbol{\lambda}_g\right) = \left(\mathbf{p}'_g, \boldsymbol{\lambda}'_g\right)$, pour $g = 1, \ldots, G$. De plus, s'il existe une injection $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$ telle que $\mathbf{p}_g = \varrho(\boldsymbol{\beta}_g)$ et $\mathbf{p}'_g = \varrho(\boldsymbol{\beta}'_g)$ pour chaque g, nous avons aussi $\boldsymbol{\beta}_g = \boldsymbol{\beta}'_g$ pour chaque g.*

Le théorème suivant étend le théorème 3 de Dasylva et Goussanou (2022), qui concerne la convergence des estimateurs par le maximum de vraisemblance. Des notations supplémentaires sont nécessaires pour préciser cette extension. Pour $G \geq 1$, considérons le mélange fini de G distributions provenant de $\mathcal{F}$, où la g^{e} composante composant a une probabilité α_g et des paramètres $\mathbf{p}_g$ et $\boldsymbol{\lambda}_g$. Aussi, désignons par $\boldsymbol{\theta} = \left[\left(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$ les paramètres de mélange associés et par $q(.;\boldsymbol{\theta})$, la fonction de masse de probabilité correspondante.

$$
\begin{aligned}
q(\mathbf{n};\boldsymbol{\theta}) \quad = \quad & \sum_{g=1}^{G} \alpha_g \left(I(|\,\mathbf{n}\,| = 0)\left(1 - |\mathbf{p}_g|\right) e^{-|\boldsymbol{\lambda}_g|} + I(|\,\mathbf{n}\,| > 1)\left(\left(1 - |\mathbf{p}_g|\right) \prod_{\gamma \in \Gamma} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)}}}{n^{(\gamma)}!} \right.\right. \\
& \left.\left. + \sum_{\gamma \in \Gamma:\, n^{(\gamma)} > 0} p_g^{(\gamma)} \frac{e^{-\lambda_g^{(\gamma)}} \left(\lambda_g^{(\gamma)}\right)^{n^{(\gamma)}-1}}{\left(n^{(\gamma)} - 1\right)!} \prod_{\gamma' \in \Gamma \setminus \{\gamma\}} \frac{e^{-\lambda_g^{(\gamma')}} \left(\lambda_g^{(\gamma')}\right)^{n^{(\gamma')}}}{n^{(\gamma')}!} \right)\right), \quad \mathbf{n} = \left[n^{(\gamma)}\right]_{\gamma \in \Gamma} \in \mathbb{N}^{|\Gamma|},
\end{aligned}
\tag{B.5}
$$

Définissons

$$
M_N(\boldsymbol{\theta}) = \frac{1}{|\,S_B\,|} \sum_{i \in S_B} \log q(\mathbf{n}_i;\boldsymbol{\theta}), \tag{B.6}
$$

et pour un nombre naturel $\tau_N > 0$, définissons

$$
\begin{aligned}
M_N(\boldsymbol{\theta};\tau_N) \quad = \quad & \frac{1}{|\,S_B\,|} \sum_{i \in S_B} \left(I\left(|\,\mathbf{n}_i\,| \leq \tau_N\right) \log q(\mathbf{n}_i;\boldsymbol{\theta}) \right. \\
& \left. + I\left(|\,\mathbf{n}_i\,| \geq \tau_N + 1\right) \log\left(\sum_{\mathbf{n} \in \mathbb{N}^{|\Gamma|}:\, |\mathbf{n}| \geq \tau_N + 1} q(\mathbf{n};\boldsymbol{\theta}) \right) \right).
\end{aligned}
\tag{B.7}
$$

Pour $\boldsymbol{\theta}_\varrho = \left[\left(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$, soit $\boldsymbol{\theta}\left(\boldsymbol{\theta}_\varrho\right) = \left[\left(\alpha_g, \varrho\left(\boldsymbol{\beta}_g\right), \boldsymbol{\lambda}_g\right)\right]_{1 \leq g \leq G}$. De même pour un nombre naturel d, supposons que $\mathbf{0}_d$ désigne le d-uplet avec tous les zéros.

Comme c'est le cas précédemment, on peut estimer les paramètres de mélange en maximisant la log-vraisemblance des $\mathbf{n}_i$, c'est-à-dire $M_N(.)$ ou $M_N(;\tau_N)$. Le théorème suivant indique que les estimateurs

qui en résultent sont convergents dans des conditions appropriées, qui comprennent (B.3) ou (B.4). Dans ce dernier cas, on suppose que la fonction ϱ est une injection.

Théorème 2. *Pour $G^* \geq 2$ et $\nu \in (0,1)$, supposons que $\Theta_1, \ldots, \Theta_{G^*}$ désignent des sous-ensembles compacts de $\mathbb{R}^{(2|\Gamma|+1)G^*-1}$ tels que*

$$\begin{aligned}\Theta_G \subset \Big\{ & \Big[\big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big)\Big]_{1\leq g\leq G^*} \in \mathbb{R}^{(2|\Gamma|+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big) \in (\nu,1]\times[0,1]^{|\Gamma|}\times[\nu,\Lambda]^{|\Gamma|} \text{ et } \big|\mathbf{p}_g\big| \leq 1-\nu \text{ et} \\ & \boldsymbol{\lambda}_{g+1} \succ \boldsymbol{\lambda}_g \text{ si } g \geq G^*-G+1, \\ & \big(\alpha_g, \mathbf{p}_g, \boldsymbol{\lambda}_g\big) = (0, \mathbf{0}_{|\Gamma|}, \mathbf{0}_{|\Gamma|}) \text{ si } g \leq G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\},\ G=1,\ldots,G^*.\end{aligned}$$

et supposons que $\Theta = \bigcup_{G=1}^{G^*}\Theta_G$. Pour une application injective $\varrho\colon \mathbb{R}^m \to [0,1]^{|\Gamma|}$, supposons aussi que $\Theta_{\varrho 1}, \ldots, \Theta_{\varrho G^*}$ désigne des sous-ensembles compacts de $\mathbb{R}^{(|\Gamma|+m+1)G^*-1}$ tels que

$$\begin{aligned}\Theta_{\varrho G} \subset \Big\{ & \Big[\big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big)\Big]_{1\leq g\leq G^*} \in \mathbb{R}^{(|\Gamma|+m+1)G^*-1} \text{ s.t.} \\ & \big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big) \in (\nu,1]\times\mathbb{R}^m\times[\nu,\Lambda]^{|\Gamma|} \text{ et } \big|\varrho\big(\boldsymbol{\beta}_g\big)\big| \leq 1-\nu \text{ et} \\ & \boldsymbol{\lambda}_{g+1} \succ \boldsymbol{\lambda}_g \text{ si } g \geq G^*-G+1, \\ & \big(\alpha_g, \boldsymbol{\beta}_g, \boldsymbol{\lambda}_g\big) = (0, \mathbf{0}_m, \mathbf{0}_{|\Gamma|}) \text{ si } g \leq G^*-G, \\ & \alpha_1+\ldots+\alpha_{G^*}=1\Big\}, G=1,\ldots,G^*,\end{aligned}$$

et supposons que $\Theta_\varrho = \bigcup_{G=1}^{G^*}\Theta_{\varrho G}$. Supposons que toutes les règles de couplage sont simples (c'est-à-dire que chaque règle est telle que la décision de coupler deux enregistrements ne dépend d'aucun autre enregistrement), que C.1-C.3 (du théorème 1) s'appliquent et que (B.3) s'applique aussi avec

$$F(\mathbf{p},\boldsymbol{\lambda}) = \sum_{g=1}^{G^*}\alpha_{0g} I\Big((\mathbf{p},\boldsymbol{\lambda}) = \big(\mathbf{p}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big),$$

$\boldsymbol{\theta}_0 = \Big[\big(\alpha_{0g},\mathbf{p}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big]_{1\leq g\leq G^*} \in \Theta_{G_0}$ *et* $1 \leq G_0 \leq G^*$, *et supposons que* $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ *et* $\hat{\boldsymbol{\theta}}_{3N}$ *désignent les estimateurs, qui maximisent respectivement $M_N(.)$ sur Θ_{G_0}, $M_N(.)$ sur Θ, et $M_N(.;\tau_N)$ sur Θ, où τ_N est un nombre entier naturel tel que $\tau_N \to \infty$ et $\tau_N = O\big(\log N\big)$. Alors $\hat{\boldsymbol{\theta}}_{1N}$, $\hat{\boldsymbol{\theta}}_{2N}$ et $\hat{\boldsymbol{\theta}}_{3N}$ convergent en probabilité vers $\boldsymbol{\theta}_0$. De plus, supposons que $\mathbf{p}_N(v)$ a aussi la forme $\varrho(\boldsymbol{\beta}_N(v))$ pour $\boldsymbol{\beta}_N\colon \mathcal{V}_N^* \to \mathbb{R}^m$, qui satisfait à (B.4) avec*

$$H(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{g=1}^{G^*}\alpha_{0g} I\Big((\boldsymbol{\beta},\boldsymbol{\lambda}) = \big(\boldsymbol{\beta}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big),$$

$\varrho\big(\boldsymbol{\beta}_{0g}\big) = \mathbf{p}_{0g}$ *pour* $g = G^*-G_0+1,\ldots,G^*$ *et* $\boldsymbol{\theta}_{\varrho 0} = \Big[\big(\alpha_{0g},\boldsymbol{\beta}_{0g},\boldsymbol{\lambda}_{0g}\big)\Big]_{1\leq g\leq G^*} \in \Theta_{\varrho G_0}$. *Alors, on estime aussi $\boldsymbol{\theta}_{\varrho 0}$ de manière convergente en maximisant $M_N(.)$ sur $\Theta_{\varrho G_0}$, $M_N(.)$ sur Θ_ϱ, ou $M_N(.;\tau_N)$ sur Θ_ϱ.*

Dans le théorème ci-dessus, la fonction $\varrho(.)$ doit être une injection. Le lemme suivant montre que cette condition est remplie quand $\varrho(.)$ est basée sur une spécification log-linéaire non saturée des interactions dans les paires appariées.

Lemme 4. *Pour les nombres entiers naturels* $K, H_1, \ldots, H_K$, *supposons que* $\Gamma = \{0, \ldots, H_1\} \times \ldots \times \{0, \ldots, H_K\} - \mathbf{0}_K$ *et* $\mathbf{p} = \left[p^{(\gamma)} \right]_{\gamma \in \Gamma}$ *ont la forme* $p^{(\gamma)} = P(i \in S_A)\, r^{(\gamma)}$, *où* $\sum_{\gamma \in \Gamma \cup \{\mathbf{0}_K\}} r^{(\gamma)} = 1$ *et* $r^{(\gamma)}$ *a la forme log-linéaire suivante sans aucune interaction d'ordre supérieur à* $d < K$.

$$r^{(\gamma)} = \exp\left(u + \sum_{t=1}^{d} \sum_{1 \le k_1 < \ldots < k_t \le K} u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})} \right), \tag{B.8}$$

où le terme $u_{k_1 \ldots k_t (\gamma_{k_1} \ldots \gamma_{k_t})}$ *est mis à zéro si l'un des* $\gamma_{k_1}, \ldots, \gamma_{k_t}$ *est nul, selon la convention de codage « dummy ». Soit* $\boldsymbol{\beta}$ *le vecteur qui comprend* $P(i \in S_A)$ *et les paramètres de* $r^{(\gamma)}$, *qui ne sont pas mis à zéro par cette convention. Alors, la fonction* $\varrho : \boldsymbol{\beta} \mapsto \mathbf{p}$ *est injective.*

C. Procédure d'initialisation

La présente section brosse un tableau de la procédure d'initialisation permettant l'ajustement du modèle de voisinage multivarié dans les simulations. Les paramètres du modèle comprennent les proportions de mélange (les α_g), les paramètres de la distribution des faux positifs (les $\boldsymbol{\lambda}_g$) et ceux de la distribution des vrais positifs (la valeur commune des $\mathbf{p}_g$). Lorsqu'il y a G classes, la proportion de mélange α_g est mise à $1/G$ pour chaque classe. Les autres valeurs initiales sont choisies comme suit.

Pour la distribution des faux positifs, chaque $\boldsymbol{\lambda}_g$ est mis à une valeur commune désignée par $\hat{\boldsymbol{\lambda}} = \left[\hat{\lambda}^{(\gamma)} \right]_{\gamma \in \Gamma}$, où $\Gamma = \{0,1\}^3 \setminus \{(0,0,0)\}$. Pour $\gamma \in \Gamma$, $\hat{\lambda}^{(\gamma)}$ est mis à l'estimation de $\bar{\lambda}$, qui est obtenue par l'ajustement du modèle de voisinage univarié, quand les enregistrements sont liés selon γ. Par exemple, quand $\gamma = (1,0,1)$, cela signifie qu'il faut lier une paire si elle est liée par la première règle de couplage et s'il y a une concordance exacte concernant le nom de famille et le mois de naissance, mais pas de concordance sur le jour de naissance.

À partir de $\hat{\boldsymbol{\lambda}}$, les valeurs initiales sont choisies pour la probabilité de couverture $P(i \in S_A)$ et les paramètres log-linéaires (c'est-à-dire $u_{1(1)}$, $u_{2(1)}$, $u_{3(1)}$, $u_{12(11)}$, $u_{13(11)}$ et $u_{23(11)}$). Ces valeurs s'obtiennent en trois étapes, expliquées ci-dessous. Dans la première étape, on calcule une estimation $\hat{\mathbf{p}} = \left[\hat{p}^{(\gamma)} \right]_{\gamma \in \Gamma}$ du vecteur des probabilités de vrais positifs en maximisant la log-vraisemblance du modèle multivarié avec une seule classe, où $\hat{\boldsymbol{\lambda}}$ remplace $\boldsymbol{\lambda}$, et les probabilités de vrais positifs n'ont pas nécessairement une forme log-linéaire. Cette étape correspond à une optimisation convexe, car la log-vraisemblance du modèle multivarié est concave par rapport aux probabilités de vrais positifs, quand les autres paramètres sont donnés. Dans la deuxième étape, les valeurs initiales pour les paramètres log-linéaires de $r^{(\gamma)}$ (dans (5.2)) sont trouvées par une méthode des moments, basée sur $\hat{\mathbf{p}}$ comme suit. En cas d'ajustement du modèle sans interactions, supposons que $\hat{q}_k = \sum_{\gamma \in \Gamma : \gamma_k = 1} \hat{p}^{(\gamma)}$ (pour $k = 1, 2, 3$) et $\hat{q}_{k_1 k_2} = \sum_{\gamma \in \Gamma : \gamma_{k_1} = \gamma_{k_2} = 1} \hat{p}^{(\gamma)}$ (pour $1 \le k_1 < k_2 \le 3$) et choisissons les valeurs initiales comme étant

$$\hat{u}_{1(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_2} + \frac{\hat{q}_{13}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{2(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{12}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_3}\right)\right),$$

$$\hat{u}_{3(1)} = \text{logit}\left(\frac{1}{2}\left(\frac{\hat{q}_{13}}{\hat{q}_1} + \frac{\hat{q}_{23}}{\hat{q}_2}\right)\right),$$

avec $\hat{u}_{12(11)} = \hat{u}_{13(11)} = \hat{u}_{23(11)} = 0$. Lorsque les interactions sont incluses, choisissons les valeurs initiales comme suit :

$$\begin{aligned}
\hat{u}_{1(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{2(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{3(1)} &= \log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right) - \left(\log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right)\right), \\
\hat{u}_{12(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(1,0,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{13(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,1,0)}}{\hat{p}^{(1,1,1)}}\right), \\
\hat{u}_{23(11)} &= -\left(\log\left(\frac{\hat{p}^{(1,1,0)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(1,0,1)}}{\hat{p}^{(1,1,1)}}\right) + \log\left(\frac{\hat{p}^{(0,1,1)}}{\hat{p}^{(1,1,1)}}\right)\right) + \log\left(\frac{\hat{p}^{(0,0,1)}}{\hat{p}^{(1,1,1)}}\right).
\end{aligned}$$

Enfin, supposons que $\hat{\mathbf{r}} = \left[\hat{r}^{(\gamma)}\right]_{\gamma \in \{0,1\}^3}$ est le vecteur de probabilités qui correspond aux valeurs initiales ci-dessus des paramètres log-linéaires (c'est-à-dire $\hat{r}^{(\gamma)}$ est égal au deuxième membre de (5.2), où les valeurs initiales sont obtenues en remplaçant les vraies valeurs des paramètres par les valeurs estimées et choisissons la couverture initiale comme suit :

$$\hat{P}(i \in S_A) = \frac{\sum_{\gamma \in \Gamma} \hat{p}^{(\gamma)}}{\sum_{\gamma \in \Gamma} \hat{r}^{(\gamma)}}.$$

Bibliographie

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

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.

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

Chambers, R. (2009). Regression analysis of probability-linked data. Dans *Research Series in Official Statistics*. Gouvernement de Nouvelle-Zélande.

Chipperfield, J.O., et Chambers, R.L. (2015). Using the bootstrap to analyse binary data obtained via probabilistic linkage. *Journal of Official Statistics*, 31, 397-414.

Chipperfield, J., Hansen, N. et Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *Revue Internationale de Statistique*, 86, 219-236.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*. New York: Springer.

Daggy, J.K., Xu, H. Hui, S.J., Gamache, R.E. et Grannis, S.J. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Medical Informatics and Decision Making*, 13, 1-8.

Dasylva, A., Abeysundera, M., Akpoué, B., Haddou, M. et Saïdi, A. (2016). Measuring the quality of a probabilistic linkage through clerical reviews. Recueil : *Symposium 2016, Croissance de l'information statistique : défis et bénéfices*, Statistique Canada.

Dasylva, A., et Goussanou, A. (2020). Estimating linkage errors under regularity conditions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 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, 2, 325-338. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2021002/article/00002-fra.pdf.

Dasylva, A., et Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*, 5, 181-216. DOI: https ://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., et Goussanou, A. (2024). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*, 7, 17-56. DOI: https ://doi.org/10.1007/s42081-023-00228-9.

Dasylva, A., Goussanou, A. et Nambeu, C.O. (2024). *Models of Linkage Error for Capture-Recapture Estimation without Clerical Reviews*. https://arxiv.org/pdf/2403.11438.pdf.

De Wolf, P.-P., van der Laan, J. et Zult, D. (2019). Connection correction methods for linkage error in capture-recapture. *Journal of Official Statitsics*, 35, 577-597.

Di Consiglio, L., et Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31, 415-429.

Ding, Y., et Fienberg, S. (1994). Dual system estimation of census undercount in the presence of matching error. *Journal of the American Statistical Association*, 20, 149-158.

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

Fortini, M., Liseo, B., Nuccitelli, A. et Scanu, M. (2001). On bayesian record linkage. *Research in Official Statistics*, 4, 185-198.

Haque, S., et Mengersen, K. (2022). Improved assessment of the accuracy of record linkage via an extended macsim approach. *Journal of Official Statistics*, 38, 429-451.

Haque, S., Mengersen, K. et Stern, S. (2021). Assessing the accuracy of record linkages with markov chain based monte carlo simulation approach. *Journal of Big Data*, 8, 1-26.

Herzog, T.N., Scheuren, F.J. et Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Lahiri, P., et Larsen, D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-227.

Larsen, M., et Rubin, D. (2001). Iterated automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Lincoln, F. (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular*, 118, 1-4.

Little, R., et Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

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

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station*, 6, 5-84.

Račinskij, V., Smith, P.A. et van der Heijden, P. (2019). *Linkage Free Dual System Estimation*. https://arxiv.org/abs/1903.10894, 2019.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112, 600-612.

Sariyar, M., Borg, A. et Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

Steorts, R., Hall, R. et Fienberg, S.E. (2016). A bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660-1672.

Tancredi, A., et Liseo, B. (2011). A hierarchical bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5, 1553-1585.

Thibaudeau, Y. (1993). Le pouvoir discriminant des structures de dépendance dans le couplage d'enregistrements. *Techniques d'enquête*, 19, 1, 35-43. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1993001/article/14477-fra.pdf.

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).

Winglee, M., Valliant, R. et Scheuren, F. (2005). Une étude de cas en couplage d'enregistrements. *Techniques d'enquête*, 31, 1, 3-13. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2005001/article/8085-fra.pdf.

Winkler, W.E. (1993). Improved decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 274-279.

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.