Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 147–158
Statistique Canada

Estimation modéliste des taux d'erreur liés au couplage d'enregistrements

J.B. ARMSTRONG et J.E. MAYDA[1]

RÉSUMÉ

On appelle couplage d'enregistrements l'appariement d'enregistrements contenant des données sur des particuliers, des entreprises ou des logements quand on ne dispose pas d'un identificateur unique. Les méthodes utilisées, en pratique, comportent la classification de paires d'enregistrements, comme constituant des liens ou des non-liens, à l'aide d'une procédure automatisée basée sur le modèle théorique présenté par Fellegi et Sunter (1969). L'estimation des taux d'erreur de classification constitue un problème important. Fellegi et Sunter présentent une méthode, afin de calculer des estimations des taux d'erreur de classification, qui découle directement du couplage. Ces estimations faites à l'aide de modèles sont plus faciles à produire que celles obtenues par appariement manuel d'échantillons, méthode généralement utilisée en pratique. Les propriétés des estimations du taux d'erreur de classification fondées sur un modèle, obtenues au moyen de trois estimateurs de paramètre de modèle, sont comparées.

MOTS CLÉS: Modèle mixte; modèle à variable latente; pondération itérative.

1. INTRODUCTION

Dans de nombreuses applications statistiques, on utilise des fichiers informatiques renfermant des renseignements sur des particuliers, sur des entreprises et sur des logements. Il faut souvent effectuer le couplage d'enregistrements qui se rapportent à la même entité. L'opération,qui consiste à coupler des enregistrements portant sur la même entité s'appelle appariement exact. Si l'on a attribué un identificateur unique à tous les enregistrements utilisés dans une application, l'appariement exact ne pose aucun problème. Les méthodes de couplage d'enregistrements s'attaquent au problème de l'appariement exact quand on ne dispose pas d'un identificateur unique. Dans ce cas, chaque enregistrement inclut généralement un certain nombre de champs de données renfermant des renseignements d'identification qui pourraient être utilisés pour effectuer l'appariement. Les problèmes rencontrés au cours de l'appariement sont dus à des erreurs dans ces données ou au fait que la même valeur dans un champ particulier est valable pour plus d'une entité.

Les applications du couplage d'enregistrements comprennent l'élimination des doubles comptes dans des listes de logements ou d'entreprises obtenues de diverses sources afin de créer des bases de sondage. De plus, le couplage d'enregistrements est largement utilisé dans des applications portant sur la santé et l'épidémiologie. Le travail dans ce domaine comporte généralement l'appariement d'enregistrements renfermant des renseignements sur des particuliers dans des cohortes d'industries ou de professions avec des enregistrements renfermant des renseignements sur la maladie ou sur le décès de particuliers. Par exemple, dans Fair, Newcombe et Lalonde (1988), on traite des méthodes de couplage d'enregistrements utilisées dans des études de suivi portant sur des personnes exposées à la radiation.

Le problème du couplage d'enregistrements peut être formulé à l'aide de deux fichiers de données qui correspondent à deux populations. Chaque fichier peut renfermer soit de l'information sur toutes les entités dans la population correspondante, soit de l'information pour un échantillon aléatoire d'entités. Le fichier A contient N_A enregistrements et le fichier B en renferme N_B. L'ensemble de paires d'enregistrements formées comme produit croisé de A et de B, est représenté par $C = \{ (a,b); a \in A, b \in B \}$. C contient $N = N_A \cdot N_B$ paires d'enregistrements. L'objectif du couplage d'enregistrements est de partager l'ensemble C en deux ensembles disjoints, l'ensemble des concordances vraies, représenté par M et l'ensemble des non-concordances vraies, U.

De nombreux travaux d'application sont basés sur le modèle théorique présenté par Fellegi et Sunter (1969). Pour chaque paire d'enregistrements, on prend une décision afin de déterminer si les enregistrements se rapportent ou non à la même entité, après avoir examiné les données enregistrées dans les fichiers A et B. Les décisions possibles sont les suivantes: lien (A_1), non-lien (A_3) et cas indéterminé (A_2). Il y a deux types d'erreurs. Premièrement, la décision A_1 peut être prise pour une paire d'enregistrements qui fait partie de U, l'ensemble des non-concordances vraies. Deuxièmement, la décision A_3 peut être prise pour une paire d'enregistrements qui fait partie de l'ensemble M, l'ensemble des concordances vraies. Des niveaux acceptables d'erreur de classification sont précisés avant le couplage des fichiers. Une paire d'enregistrements est classée comme un cas indéterminé si les données ne

[1] J.B. Armstrong et J.E. Mayda, Statistique Canada, Division des méthodes d'enquêtes-entreprises, 11-Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), K1A 0T6.

fournissent pas une preuve suffisante pour justifier la classification de la paire comme un lien ou un non-lien à des niveaux d'erreur inférieurs ou égaux aux niveaux précisés. Il faut estimer avec précision les taux d'erreur de classification associés à diverses règles de décision afin de déterminer une règle appropriée. Le taux d'erreur de classification pour les non-concordances vraies est $P(A_1 \mid U)$. Le taux d'erreur de classification pour les concordances vraies est $P(A_3 \mid M)$.

On peut obtenir des estimations des taux d'erreur de classification en choisissant un échantillon de paires d'enregistrements de l'ensemble C et en déterminant manuellement l'état véritable vis-à-vis de la concordance des paires échantillonnées. Des applications de cette méthode sont décrites dans Bartlett et coll. (1993). L'échantillonnage peut être à la fois coûteux et peu facile à réaliser, particulièrement quand on doit effectuer le même couplage pour un certain nombre de paires de fichiers, chacune avec des caractéristiques légèrement différentes. Belin et Rubin (1991) décrivent une autre méthode d'estimation des taux d'erreur pour lesquelles il faut disposer de l'état véritable de la concordance pour des paires d'enregistrements dans une étude-pilote. Par opposition à la méthode d'échantillonnage simple, la méthode de Belin et Rubin fournit un modèle pour l'application de renseignements, obtenus à l'aide de l'étude-pilote, à des couplages plus importants dans lesquels des données semblables sont utilisées.

Le modèle de Fellegi-Sunter fournit une méthode pour calculer des estimations des taux d'erreur à l'aide d'estimations des probabilités qu'il y aura concordance entre des paires d'enregistrements pour diverses combinaisons de champs de données. Le calcul de ces estimations du taux d'erreur obtenues à l'aide de modèles est simple et la détermination manuelle de l'état de paires d'enregistrements par rapport à la concordance vraie n'est pas nécessaire. Cependant, ces estimations ont souvent des propriétés qui laissent à désirer dans des applications. Voir, par exemple, Belin (1990). Dans le présent article, on démontre que les propriétés des estimations du taux d'erreur obtenues à l'aide de modèles peuvent être améliorées en faisant une estimation soignée des probabilités d'accord.

Trois méthodes d'estimation qui peuvent être employées sont évaluées. Les façons de procéder décrites n'utilisent que les renseignements dans les fichiers A et B. Elles n'emploient pas de renseignements auxiliaires. Les estimations des taux d'erreur obtenues à l'aide de modèles pour chacune des méthodes évaluées sont comparées aux taux d'erreur réels à l'aide tant de données synthétiques qui incorporent des caractéristiques importantes de données tirées d'applications du couplage d'enregistrements dans le domaine de la santé que de renseignements obtenus à partir d'une application réelle du couplage d'enregistrements.

Voici la structure du présent article. La section 2 contient des détails sur la méthode d'estimation de l'erreur de classification à l'aide d'un modèle présentée par Fellegi et Sunter. Le modèle utilisé pour les probabilités d'accord qui forme la base de la discussion ultérieure des méthodes d'estimation est aussi précisé. Deux méthodes d'estimation qui sont fondées sur une hypothèse d'indépendance importante sont décrites dans la section 3. On traite d'une troisième méthode pour laquelle l'indépendance n'est pas exigée dans la section 4. Les résultats des comparaisons des trois méthodes à l'aide de données synthétiques sont présentés dans la section 5. Les résultats du travail d'évaluation effectué avec des renseignements tirés d'une application réelle sont décrits dans la section 6. La section 7 renferme des conclusions.

2. CONCEPTS THÉORIQUES

Nous résumons, dans la présente section, les aspects pertinents de la théorie du couplage d'enregistrements élaborée par Fellegi et Sunter (1969). Dans le modèle de Fellegi-Sunter, les estimations des taux d'erreur de classification sont calculées à l'aide d'estimations des probabilités d'accord pour diverses combinaisons de champs de données. Les applications de la théorie de Fellegi et Sunter supposent habituellement l'hypothèse que la probabilité qu'il y aura accord, pour une paire d'enregistrements, au niveau d'un champ de données particulier est indépendante des résultats des comparaisons pour d'autres champs. La théorie est néanmoins très flexible, pouvant tenir compte de tous les types de dépendance entre les résultats de comparaisons pour différents champs de données. Un paramétrage de la dépendance en fonction d'effets log-linéaires est présenté.

2.1 Estimation des taux d'erreur de classification fondée sur un modèle

Pour obtenir des renseignements portant sur la classification d'une paire d'enregistrements comme un lien (A_1), un non-lien (A_3) ou un cas indéterminé (A_2), on compare des champs de données renfermant des renseignements d'identification. Dans une application comportant des enregistrements relatifs à des personnes, on pourrait faire des comparaisons distinctes des noms de famille, des prénoms et des dates de naissance. Le résultat d'une comparaison est un code numérique représentant un énoncé du genre: "les noms concordent", "les noms ne concordent pas", "un nom manque dans un des fichiers ou dans les deux", "les noms concordent et ils sont tous deux Georges" ou "les noms ne concordent pas mais leurs deux premiers caractères concordent". Les codes de résultat utilisés dans les travaux d'application diffèrent selon les applications et selon les comparaisons dans la même application. Le plus petit nombre de codes de résultat qui peut être utilisé pour toute comparaison est deux, ce qui correspond à l'accord et au désaccord. Dans les applications, il faut habituellement disposer d'un code de résultat correspondant à "manquant dans un des deux fichiers ou dans les deux". Le résultat de l'accord peut être remplacé par un certain nombre de résultats ayant une valeur particulière (telle que "les noms concordent et ils sont tous deux Georges"). Certains désaccords peuvent être codés comme des accords partiels (tels que "les noms ne concordent pas mais leurs deux premiers caractères concordent").

Pour nos fins, nous ne tenons compte que des résultats de l'accord et du désaccord. Dans le cas de K champs d'appariement, nous définissons le vecteur de résultats $\underline{x}^j = (x_1^j, x_2^j, \ldots, x_K^j)$ pour la paire d'enregistrements j. Nous avons $x_k^j = 1$ s'il y a concordance pour la paire d'enregistrements j à propos du champ de données k et $x_k^j = 0$ s'il y a non-concordance pour la paire d'enregistrements j à propos du champ de données k.

Newcombe et coll. (1959) ont proposé l'idée que les décisions relatives au fait qu'une paire d'enregistrements représente ou non la même entité devraient être basées sur le rapport

$$R(\underline{x}) = P(\underline{x} \mid M)/P(\underline{x} \mid U), \tag{1}$$

où $\underline{x} = (x_1, x_2, \ldots, x_K)$ est le vecteur de résultats générique, $P(\underline{x} \mid M)$ est la probabilité que les comparaisons pour une paire d'enregistrements où il y a concordance vraie produiront le vecteur de résultats $\underline{x}$, et $P(\underline{x} \mid U)$ est la probabilité de $\underline{x}$ pour une paire d'enregistrements pour laquelle il y a non-concordance vraie. L'optimalité des méthodes de couplage d'enregistrements qui font appel à ce rapport a été démontrée par Fellegi et Sunter.

Dans le modèle de Fellegi-Sunter, une règle d'appariement attribue une probabilité pour chaque décision de classification (A_1, A_2 et A_3) à chaque vecteur de résultats. La fonction de décision correspondant au vecteur de résultats $\underline{x}$ est $d(\underline{x}) = (P(A_1 \mid \underline{x}), P(A_2 \mid \underline{x}), P(A_3 \mid \underline{x}))$. Des taux d'erreur de classification acceptables pour les non-concordances vraies et les concordances vraies sont précisés avant que le couplage ne soit effectué. Nous désignons ces taux d'erreur précisés à l'avance par μ et λ respectivement. Fellegi et Sunter définissent, parmi la classe de règles d'appariement d'enregistrements qui satisfont aux relations $P(A_1 \mid U) \leq \mu$ et $P(A_3 \mid M) \leq \lambda$ pour des valeurs fixes de μ et de λ, la règle d'appariement optimale comme étant la règle qui minimise $P(A_2)$, la probabilité qu'une paire d'enregistrements sera classée comme un cas indéterminé. La règle optimale a la forme

$$\begin{aligned}
d(\underline{x}^j) &= (1,0,0) \quad \text{si} \quad \omega^j > \tau_1 \\
d(\underline{x}^j) &= (P_\mu, 1 - P_\mu, 0) \quad \text{si} \quad \omega^j = \tau_1 \\
d(\underline{x}^j) &= (0,1,0) \quad \text{si} \quad \tau_2 < \omega^j < \tau_1 \\
d(\underline{x}^j) &= (0, 1 - P_\lambda, P_\lambda) \quad \text{si} \quad \omega^j = \tau_2 \\
d(\underline{x}^j) &= (0,0,1) \quad \text{si} \quad \omega^j < \tau_2
\end{aligned} \tag{2}$$

où $\tau_1 \geq \tau_2$, le "poids" ω^j est défini comme $\omega^j = \log(R(\underline{x}^j))$ et P_μ et P_λ sont des constantes positives sur l'intervalle $[0,1)$. (Pour tous les détails, consulter Fellegi et Sunter (1969)). Pour déterminer τ_1 et τ_2, il faut estimer les taux d'erreur de classification correspondant à divers choix pour ces valeurs seuil, ce qui souligne l'importance d'une estimation précise des taux d'erreur de classification dans le modèle de Fellegi-Sunter.

On peut calculer des estimations des taux d'erreur de classification fondées sur un modèle à l'aide d'estimations des probabilités de résultats pour les concordances vraies et pour les non-concordances vraies. Utilisons $\hat{P}(\underline{x} \mid M)$ et $\hat{P}(\underline{x} \mid U)$ pour représenter les estimations des probabilités du vecteur de résultats $\underline{x}$ pour les concordances vraies et pour les non-concordances vraies et représentons le rapport de ces estimations par $\hat{R}(\underline{x})$. L'estimation du taux d'erreur de classification pour les concordances vraies fondée sur un modèle basé sur la règle de décision (2) est

$$\hat{\lambda} = \sum_{\underline{x} \in L(\tau_2)} \hat{P}(\underline{x} \mid M) + P_\lambda \sum_{\underline{x} \in Q(\tau_2)} \hat{P}(\underline{x} \mid M) \tag{3}$$

où $L(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) < \tau_2\}$ et $Q(\tau_2) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_2\}$.

L'estimation du taux d'erreur de classification pour les non-concordances vraies fondée sur un modèle est

$$\hat{\mu} = \sum_{\underline{x} \in G(\tau_1)} \hat{P}(\underline{x} \mid U) + P_\mu \sum_{\underline{x} \in Q(\tau_1)} \hat{P}(\underline{x} \mid U) \tag{4}$$

où $G(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) > \tau_1\}$ et $Q(\tau_1) = \{\underline{x};\ \log(\hat{R}(\underline{x})) = \tau_1\}$.

2.2 Un modèle pour les probabilités de résultats

Pour calculer les estimations du taux d'erreur de classification fondées sur un modèle, il faut estimer $P(\underline{x} \mid M)$ et $P(\underline{x} \mid U)$, pour chacune des 2^K valeurs possibles de $\underline{x}$. La densité de probabilité pour $\underline{x}$ est une combinaison de deux densités de probabilité données par

$$f(\underline{x}) = pP(\underline{x} \mid M) + (1 - p)\, P(\underline{x} \mid U), \tag{5}$$

où p est la probabilité qu'il y a concordance vraie pour une paire d'enregistrements choisie au hasard. Les probabilités de résultats dépendent de la distribution de fréquences des identificateurs pour des entités représentées dans les fichiers A et B, ainsi que des probabilités qu'il y a introduction d'erreurs quand les identificateurs sont enregistrés dans les fichiers. Fellegi et Sunter (1969, pp. 1192-1194) décrivent, pour estimer les probabilités d'accord, une méthode qui fait appel à leur définition en fonction des distributions de fréquences et des probabilités d'erreur. Ils recommandent d'utiliser la méthode quand on dispose d'informations préalables.

Dans le présent article, nous considérons des situations où les données dans les fichiers A et B, ainsi que les vecteurs de résultats $\underline{x}^j$, $j = 1, 2, \ldots, N$, représentent les seuls renseignements disponibles pour l'estimation des probabilités de résultats. Une structure log-linéaire pour les probabilités de résultats constitue le paramétrage le plus général. Le modèle log-linéaire saturé pour les probabilités de résultats dans le cas des concordances vraies est

$$\log(P(\underline{x} \mid M)) = M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots$$
$$+ M(K)_{x_K} + M(1)\, M(2)_{x_1,x_2} + \ldots$$
$$+ M(K-1)\, M(K)_{x_{K-1},x_K} + \ldots$$
$$+ M(1)\, M(2) \ldots M(K)_{x_1,x_2,\ldots,x_K}, \quad (6)$$

avec les restrictions habituelles

$$\sum_{x_J} M(J)_{x_J} = 0, \quad J = 1, 2, \ldots, K,$$

$$\sum_{x_{J_1}} M(J_1)M(J_2)_{x_{J_1},x_{J_2}} = \sum_{x_{J_2}} M(J_1)M(J_2)_{x_{J_1},x_{J_2}} = 0,$$
$$\forall\, J_1, J_2, \quad etc.,$$

ainsi que la restriction

$$\sum_{\underline{x}} P(\underline{x} \mid M) = 1.$$

Le modèle saturé pour $P(\underline{x} \mid U)$ est analogue.

Si l'on utilise des modèles log-linéaires saturés pour $P(\underline{x} \mid M)$ et pour $P(\underline{x} \mid U)$, la fonction de densité comprend $2^{K+1} - 1$ paramètres inconnus. On ne peut identifier tous ces paramètres quand on ne dispose pas de renseignements auxiliaires. Afin d'obtenir un modèle qui peut être identifié et pour simplifier le problème d'estimation, on fait souvent l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Quand on suppose qu'il y a indépendance, on désigne les probabilités d'accord parmi les paires d'enregistrements qui sont des concordances vraies et des non-concordances vraies, respectivement, par

$$m_k = P(x_k = 1 \mid M), \quad k = 1, 2, \ldots, K,$$

$$u_k = P(x_k = 1 \mid U), \quad k = 1, 2, \ldots, K.$$

Les probabilités de résultats peuvent être écrites sous la forme:

$$P(\underline{x} \mid M) = \prod_{k=1}^{K} m_k^{x_k} (1 - m_k)^{(1-x_k)},$$

$$P(\underline{x} \mid U) = \prod_{k=1}^{K} u_k^{x_k} (1 - u_k)^{1-x_k}.$$

Ce modèle comprend $2 \cdot K + 1$ paramètres inconnus, nommément $(\underline{m}, \underline{u}, p)$, où $\underline{m} = (m_1, m_2, \ldots, m_k)$, $\underline{u} = (u_1, u_2, \ldots, u_k)$. Il y a, bien entendu, un certain nombre de modèles intermédiaires entre le modèle saturé et le modèle où l'on fait appel à l'indépendance. Des méthodes qui peuvent être utilisées pour estimer le modèle qui fait appel à l'indépendance sont décrites dans la section 3. On traite de l'estimation de modèles intermédiaires dans la section 4.

3. ESTIMATION FONDÉE SUR L'HYPOTHÈSE DE L'INDÉPENDANCE

3.1 Méthode des moments

On peut employer un estimateur de $P(\underline{x} \mid M)$ et de $P(\underline{x} \mid U)$ obtenu à l'aide de la méthode des moments quand il y a indépendance. L'estimateur est basé sur un système de $2 \cdot K + 1$ équations qui fournissent des expressions pour des moments fonctionnellement indépendants de $\underline{x}$ en fonction des paramètres. Les équations sont les suivantes:

$$E\left(\prod_{k \neq i}^{K} x_k\right) = pN \prod_{k \neq i}^{K} m_k + (1 - p)\, N \prod_{k \neq i}^{K} u_k,$$
$$i = 1, 2, \ldots, K$$

$$E(x_i) = pNm_i + (1 - p)\, Nu_i, \quad i = 1, 2, \ldots, K, \quad (7)$$

$$E\left(\prod_{k=1}^{K} x_k\right) = pN \prod_{k=1}^{K} m_k + (1 - p)\, N \prod_{k=1}^{K} u_k.$$

Pour obtenir des estimations des paramètres à l'aide de la méthode des moments, il faut résoudre les équations une fois que les valeurs espérées ont été remplacées par des moyennes calculées à l'aide de paires d'enregistrements dans C. Le système d'équations pour $K = 3$ a été présenté par Fellegi et Sunter, qui ont aussi calculé une solution en forme analytique fermée qui existe si certaines conditions modérées sont satisfaites. Leur article comprenait une mise en garde pour ce qui est de l'utilisation de la méthode s'il n'y a pas indépendance. Pour $K > 3$, on ne dispose pas d'une solution en forme analytique fermée, mais on peut employer les méthodes numériques courantes. Les estimations de paramètres obtenues à l'aide de la méthode des moments sont statistiquement convergentes si l'hypothèse d'indépendance est vérifiée.

3.2 Méthode itérative

La méthode itérative a été élaborée par des personnes qui effectuent des couplages d'enregistrements. Bien que la méthode ne soit pas basée sur la distribution de probabilité du vecteur de résultats, elle fait appel à l'hypothèse d'indépendance. L'application de la méthode itérative est décrite par plusieurs auteurs, y compris Newcombe (1988). Le logiciel de couplage d'enregistrements de Statistique Canada, CANLINK, est conçu de façon à faciliter l'utilisation de la méthode itérative.

La méthode exige des estimations initiales des probabilités d'accord pour les concordances vraies et pour les non-concordances vraies. Pour les concordances vraies, on doit utiliser des estimations au jugé basées sur l'expérience acquise. Pour obtenir des estimations initiales des probabilités d'accord parmi les paires d'enregistrements qui correspondent à des non-concordances vraies, on suppose habituellement que ces probabilités sont égales aux probabilités d'accord parmi les paires d'enregistrements choisies de façon aléatoire, nommément que:

$$u_k = P(x_k = 1), \quad k = 1, 2, \ldots, K.$$

Supposons que $J(k)$ valeurs différentes apparaissent, pour le champ de données k, dans le fichier A et/ou dans le fichier B. Désignons les fréquences de ces valeurs dans le fichier A par $f_{k1}, f_{k2}, \ldots, f_{kJ(k)}$ et désignons les fréquences pour le fichier B par $g_{k1}, g_{k2}, \ldots, g_{kJ(k)}$. Pour une valeur particulière, un des chiffres, mais non les deux, peut être nul. L'estimation initiale de u_k est

$$\hat{u}_k^0 = \sum_{j=1}^{J(k)} (f_{kj}\, g_{kj})/N. \tag{8}$$

Compte tenu de ces estimations de probabilités, des ensembles initiaux de concordances et de non-concordances, désignés par M^0 et par U^0, respectivement, sont obtenus à l'aide d'une règle de décision

$$j \in M^0 \quad \text{si} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{si} \quad \omega^j < \tau_2^0.$$

Puis, on utilise les chiffres des fréquences parmi les paires d'enregistrements dans les ensembles M^0 et U^0 comme nouvelles estimations des probabilités d'accord. Ces estimations sont employées pour obtenir de nouveaux ensembles de concordances et de non-concordances et le processus itératif est repris jusqu'à ce que des estimations consécutives des probabilités d'accord soient suffisamment rapprochées.

Dans la majorité des applications, l'hypothèse que la probabilité d'accord parmi les paires d'enregistrements qui correspondent à des concordances vraies est égale à la probabilité d'accord parmi toutes les paires d'enregistrements est justifiée et l'itération ne mène pas à des modifications importantes dans les estimations des probabilités d'accord pour les non-concordances vraies. Toutefois, il arrive souvent que la première itération produise des variations considérables dans les estimations de la probabilité d'accord pour les concordances vraies. Généralement, il n'y a pas de variations importantes lors de la deuxième itération.

Il faudrait remarquer que les propriétés statistiques de la méthode itérative ne sont pas connues avec précision. En pratique, le rendement de la méthode dépendra du choix des seuils initiaux τ_1^0, τ_2^0. Ces seuils sont généralement choisis de façon subjective. Les simulations mentionnées dans la section 5 fournissent des renseignements à propos des effets des divers seuils initiaux.

4. ASSOUPLISSEMENT DE L'HYPOTHÈSE DE L'INDÉPENDANCE - ESTIMATION À L'AIDE DE LA PONDÉRATION ITÉRATIVE

On peut utiliser des méthodes d'estimation pour les modèles à variable latente afin d'estimer les probabilités d'accord quand on effectue le paramétrage, en fonction d'effets log-linéaires, de la dépendance entre les résultats de comparaisons pour différents champs d'appariement. Winkler (1989) et Thibaudeau (1989) ont estimé les probabilités d'accord avec des modèles log-linéaires qui comprennent tous les termes d'interaction jusqu'au troisième ou quatrième ordre afin de paramétrer les dépendances. La formulation présentée ici facilite l'utilisation de modèles log-linéaires, y compris certaines interactions. On peut considérer l'état par rapport à la concordance comme une variable latente avec deux niveaux (concordance vraie et non-concordance vraie). Représentons par $c_{0,\underline{x}}$ et $c_{1,\underline{x}}$, le nombre de non-concordances vraies et de concordances vraies, respectivement, avec vecteur de résultats $\underline{x}$ dans une application de couplage d'enregistrements pour laquelle on utilise K variables d'appariement. Bien entendu, ces chiffres ne peuvent être observés puisque la valeur de la variable latente pour chaque paire d'enregistrements est inconnue. On observe plutôt $c_{\underline{x}} = c_{0,\underline{x}} + c_{1,\underline{x}}$.

À l'aide du paramétrage de la dépendance réalisé au moyen d'effets log-linéaires et d'un modèle saturé pour les concordances vraies, nous pouvons poser

$$\begin{aligned}\log(c_{1,\underline{x}}/(Np)) = {} & M(0) + M(1)_{x_1} + M(2)_{x_2} + \ldots \\ & + M(K)_{x_K} + M(1)M(2)_{x_1,x_2} + \ldots \\ & + M(K-1)M(K)_{x_{K-1},x_K} + \ldots \\ & + M(1)M(2) \ldots M(K)_{x_1,x_2,\ldots,x_K},\end{aligned}$$

avec les restrictions habituelles. Nous disposons d'une expression semblable pour les non-concordances vraies. Le modèle à variable latente correspondant à ces modèles log-linéaires saturés est:

$$\begin{aligned}\log(c_{s,\underline{x}}/w_s) = {} & G(0) + Z_s + G(1)_{x_1} + \ldots \\ & + G(K)_{x_K} + ZG(1)_{s,x_1} + \ldots + ZG(K)_{s,x_K} \\ & + \ldots + G(1)G(2)\ldots G(K)_{x_1,x_2,\ldots,x_K} \\ & + ZG(1)G(2) \ldots G(K)_{s,x_1,x_2,\ldots x_K},\end{aligned}$$

où la valeur de l'indice s est zéro pour les non-concordances vraies et un pour les concordances vraies,

$w_0 = (1 - p)N$ et $w_1 = pN$. Les paramètres sont analogues aux paramètres d'un modèle log-linéaire saturé pour un tableau de contingence de dimensions 2^{K+1}. Les restrictions habituelles s'appliquent. Par exemple, le terme $ZG(1)_{s,x_1}$ représente l'interaction de la variable latente et de la première variable d'appariement et

$$\sum_s ZG(1)_{s,x_1} = \sum_{x_1} ZG(1)_{s,x_1} = 0.$$

Ce modèle est conforme au modèle à variable latente général de Haberman (1979, p. 561). On doit imposer des restrictions additionnelles pour identifier et estimer les paramètres. Pour plus de simplicité, nous ne considérerons que les modèles hiérarchiques. De plus, nous n'étudierons que des modèles qui permettent à tous les effets non nuls d'interagir avec la variable latente.

Dans la discussion qui suit, nous désignerons les modèles à variable latente par les symboles $G(1)$, $G(2)$, ... les modèles log-linéaires pour les concordances vraies par $M(1), M(2)$, ... et les modèles log-linéaires pour les non-concordances vraies par $U(1)$, $U(2)$, ... Dans le cas de quatre variables d'appariement, par exemple, le modèle $G(1)G(2)$, $G(3)$, $G(4)$ est un modèle à variable latente qui comprend un terme de niveau général, des effets principaux pour les quatre variables d'appariement et un terme pour l'interaction des variables d'appariement un et deux, ainsi qu'un terme pour les effets principaux pour la variable latente (l'interaction du terme de niveau général et de la variable latente), des termes pour l'interaction de chaque variable d'appariement et de la variable latente ainsi qu'un terme pour l'interaction des variables d'appariement un et deux et de la variable latente. Le modèle comprend douze paramètres qui doivent être estimés. Le nombre de paramètres qui doivent être estimés dans un des modèles à variable latente examinés dans le présent document est deux fois plus élevé que le nombre de paramètres dans le modèle log-linéaire correspondant.

On peut utiliser la méthode de pondération itérative de Haberman (1976) pour estimer des modèles à variable latente. La méthode d'estimation de Haberman consiste à soumettre à un balayage des tableaux qui renferment des chiffres estimés pour chaque résultat parmi les concordance vraies et les non-concordances vraies. Représentons les chiffres estimés pour le vecteur de résultats $\underline{x}$ après i itérations de l'algorithme de Haberman par $\hat{C}^i_{1,\underline{x}}$ et par $\hat{C}^i_{0,\underline{x}}$ pour les concordances vraies et les non-concordances vraies, respectivement. On peut construire les valeurs initiales $\hat{C}^0_{1,\underline{x}}$ et $\hat{C}^0_{0,\underline{x}}$ à l'aide d'estimations des probabilités d'accord et de la proportion de concordances vraies obtenues quand on applique l'hypothèse de l'indépendance. Chaque itération de l'algorithme comporte une série d'opérations de balayage appliquées au tableau courant des concordances vraies ainsi que les opérations analogues sur le tableau courant des non-concordances vraies. À l'aide de la notation pour les modèles hiérarchiques présentée plus haut, on effectue un ensemble d'opérations de balayage pour chacun des termes d'interaction qui définissent le modèle. Pour quatre variables d'appariement ainsi que le modèle $G(1)G(2)$, $G(3)G(4)$, on effectue deux ensembles d'opérations de balayage, une pour l'interaction $G(1)G(2)$ et la seconde pour l'interaction $G(3)G(4)$. Pour chaque interaction, une opération de balayage est effectuée pour chaque niveau de la variable de classification correspondante. Représentons par S_{gl} l'ensemble de vecteurs de résultats au niveau l du terme g. L'opération de balayage appliquée au tableau des concordances vraies lors de l'itération i pour le niveau l du terme g comporte le calcul de

$$\gamma_{1,\underline{x}} = c_{\underline{x}}\hat{c}^{i-1}_{1,\underline{x}} / (\hat{c}^{i-1}_{1,\underline{x}} + \hat{c}^{i-1}_{0,\underline{x}}),$$

$$\hat{c}^i_{1,\underline{x}} = \hat{c}^{i-1}_{1,\underline{x}} \sum_{\underline{x}\in s_{gl}} \gamma_{1,\underline{x}} \Big/ \sum_{\underline{x}\in s_{gl}} \hat{c}^{i-1}_{1,\underline{x}}, \quad \forall\ \underline{x}\in S_{gl}.$$

Le traitement de l'algorithme prend fin quand les variations entre les chiffres estimés pour des itérations consécutives sont inférieures à une tolérance donnée.

Haberman (1976) fait remarquer qu'il se peut que l'algorithme de pondération itérative converge vers un maximum local de la fonction de vraisemblance plutôt que vers l'estimation obtenue à l'aide de la méthode du maximum de vraisemblance. Des expériences réalisées avec des valeurs initiales différentes qui utilisent des ensembles de données employés dans l'évaluation dont on traite à la section 5 n'ont permis d'obtenir aucun exemple de ce problème.

5. COMPARAISON DES MÉTHODES D'ESTIMATION - DONNÉES SYNTHÉTIQUES

On présente, dans cette section, les résultats de comparaisons des méthodes d'estimation décrites dans les sections 3 et 4. Les comparaisons comprenaient l'application de chaque méthode à une série d'ensembles de données synthétiques produits à l'aide de méthodes de Monte Carlo.

On a employé des enregistrements de données synthétiques renfermant quatre identificateurs de personne (nom de famille, initiale d'un deuxième prénom, prénom, date de naissance). Les renseignements sur les valeurs possibles de chaque identificateur ainsi que leurs fréquences relatives, ont été tirés de la Base canadienne de données sur la mortalité pour 1988. Cette base de données, qui est souvent utilisée dans des applications de couplage d'enregistrements dans le domaine de la santé, renferme un enregistrement distinct pour chaque mortalité.

L'hypothèse d'indépendance n'était pas respectée parmi les concordances vraies dans chaque ensemble de données synthétiques. Pendant la production des données, on a utilisé des renseignements, sur la fréquence des vecteurs de résultats pour les concordances vraies, obtenus à partir de divers projets de couplage d'enregistrements réalisés par le Centre canadien d'information sur la santé de Statistique Canada. La majorité des projets comportaient l'appariement d'un fichier de cohorte à la Base canadienne de données

sur la mortalité. La fréquence de chaque vecteur de résultats parmi les concordances vraies est présentée au tableau 1. L'absence d'indépendance dans ces données est évidente. Bien que pour environ 88.3% des concordances vraies, il y ait accord pour le prénom, la probabilité d'un accord pour le prénom quand il y a désaccord pour l'initiale d'un deuxième prénom et accord pour le nom de famille et l'année de naissance n'est que de 381/1366, soit environ 27.9%. La valeur de la statistique du test du rapport des vraisemblances pour l'hypothèse d'indépendance est 3604. Cette valeur est extrêmement élevée par rapport à la distribution de référence chi carré avec 10 degrés de liberté. (Il faut remarquer qu'un degré de liberté est perdu parce que la fréquence pour la case (1,0,0,0) est zéro.)

Tableau 1

Fréquences des résultats, ensemble de concordances vraies, données synthétiques

Résultat selon l'identificateur: 0 = Désaccord, 1 = Accord				Fréquence	
Prénom	Initiale d'un deuxième prénom	Nom de famille	Année de naissance	Chiffre	Pourcentage
0	0	0	0	7	0.03
0	0	0	1	33	0.12
0	0	1	0	125	0.45
0	0	1	1	985	3.54
0	1	0	0	5	0.02
0	1	0	1	39	0.14
0	1	1	0	202	0.73
0	1	1	1	1,848	6.65
1	0	0	0	0	0.0
1	0	0	1	13	0.05
1	0	1	0	50	0.18
1	0	1	1	381	1.37
1	1	0	0	44	0.16
1	1	0	1	451	1.62
1	1	1	0	1,751	6.30
1	1	1	1	21,860	78.65
			Total	27,794	100

Pour chaque ensemble de données synthétiques, on a produit les enregistrements du fichier A en choisissant des identificateurs selon les fréquences relatives dans la Base canadienne de données sur la mortalité pour 1988. Afin de simplifier le processus de production des données, le choix des noms de famille était limité aux 100 noms de famille non francophones les plus courants et aux 100 noms de famille francophones les plus courants qui figuraient dans le fichier de 1988. Le choix des prénoms était limité aux 50 prénoms francophones les plus courants et aux 50 prénoms non francophones les plus courants. On n'a pas tenu compte des variantes orthographiques lors du choix des noms. On a toutefois pris en considération toutes les initiales d'un deuxième prénom et de toutes les années de naissance qui figuraient dans le fichier pour 1988. La probabilité que des enregistrements avec des prénoms anglophones se voient attribuer un nom de famille anglophone était plus élevée que dans le cas des prénoms francophones (ce qui reflète la répartition des noms dans la population canadienne). À part ces restrictions, les identificateurs étaient choisis indépendamment.

Le point de départ pour le fichier B était une copie exacte du fichier A. Pour chaque enregistrement du fichier B, il y avait une concordance vraie avec exactement un enregistrement du fichier A. Pour introduire une absence d'indépendance parmi les concordances vraies, on a tiré un vecteur de résultats de la distribution de fréquences du tableau 1 pour chaque enregistrement du fichier B. Les identificateurs correspondant à des zéros dans le vecteur de résultats ont été choisis à nouveau. Par conséquent, l'ensemble de vecteurs de résultats pour les concordances vraies était un échantillon de la distribution du tableau 1. Les ensembles de données synthétiques comprenaient aussi de légers écarts par rapport à l'hypothèse d'indépendance pour les non-concordances vraies puisque la sélection des prénoms et des noms de famille n'était pas complètement indépendante.

Chaque ensemble de résultats de simulations mentionné plus loin est basé sur 50 essais de Monte Carlo. Chaque essai comportait la production de fichiers A et B comprenant 500 personnes, l'estimation de $\underline{m}$ et de $\underline{u}$, la détermination de seuils correspondant à diverses estimations du taux d'erreur de classification fondées sur un modèle et le calcul de taux d'erreur réels correspondant aux seuils. La même série de 50 ensembles de données synthétiques était utilisée pour chaque ensemble de simulations. Il faut remarquer que l'ensemble C renferme 250,000 paires d'enregistrements, y compris 249,500 non-concordances vraies pour chaque essai de Monte Carlo. Afin de réduire le temps de calcul nécessaire pour effectuer les simulations, on n'a utilisé que 49,500 non-concordances vraies pour chaque essai. (On a effectué un essai à petite échelle afin de s'assurer que la réduction du nombre de non-concordances vraies avait un effet négligeable sur les probabilités d'accord estimées.) On a supprimé les non-concordances vraies contenues dans C en divisant les fichiers A et B en cinq blocs correspondants de taille 100 et en excluant les paires d'enregistrements dans lesquelles on trouvait des enregistrements provenant de blocs qui ne correspondaient pas.

Le système d'équations utilisé pour la méthode des moments a été résolu à l'aide d'une variation de la méthode de Newton, décrite en détail dans Moré et coll. (1980). Un logiciel fourni par IMSL (1987) a été utilisé. On a employé des probabilités d'accord de 0.9 pour les concordances vraies et de 0.1 pour les non-concordances vraies, pour tous les champs d'appariement, comme valeurs initiales pour la solution du système d'équations. La méthode ne semblait pas sensible aux valeurs initiales.

Tableau 2

Taux d'erreur de classification, non-concordances vraies, données synthétiques
(erreurs-types Monte Carlo entre parenthèses)

Taux estimé (× 99)	Taux réel (× 99)				
	Méthode des moments	Méthode itérative $\mu^0 = 0.0000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$	Pondération itérative
0.02	0.0188	0.0208	0.0208	0.0207	0.0195
	(0.0008)	(0.0008)	(0.001)	(0.001)	(0.001)
0.04	0.0381	0.0408	0.0407	0.0405	0.0397
	(0.001)	(0.0013)	(0.0016)	(0.0016)	(0.0016)
0.06	0.057	0.0626	0.0615	0.0602	0.059
	(0.0012)	(0.0015)	(0.0018)	(0.0019)	(0.0018)
0.08	0.076	0.0855	0.0838	0.0804	0.0785
	(0.0015)	(0.0017)	(0.0019)	(0.0022)	(0.0019)
0.10	0.095	0.1086	0.1061	0.1007	0.0978
	(0.0019)	(0.0021)	(0.0022)	(0.0026)	(0.0021)

Les propriétés de la méthode itérative dépendent des définitions des ensembles initiaux de concordances et de non-concordances, M^0 et U^0. Il faut se rappeler que, compte tenu de probabilités initiales, les paires d'enregistrements sont classées selon la règle suivante:

$$j \in M^0 \quad \text{si} \quad \omega^j > \tau_1^0,$$

$$j \in U^0 \quad \text{si} \quad \omega^j < \tau_2^0.$$

Quand la méthode itérative a été appliquée pour les simulations mentionnées ici, on a donné à τ_2^0 la valeur de τ_1^0. Pour chaque essai de Monte Carlo, τ_1^0 a été fixé de façon à ce que

$$\hat{P}(j \in U \mid \omega^j > \tau_1^0) + \gamma \cdot \hat{P}(j \in U \mid \omega^j = \tau_1^0) = \mu^0,$$

pour un $\gamma \in [0,1)$, où les probabilités estimées sont basées sur les estimations itératives initiales de $\underline{u}$. Les paires d'enregistrements avec poids τ_1^0 ont été classifiés dans M^0 avec probabilité γ. C'est-à-dire que l'ensemble initial de concordances utilisé par la méthode itérative était défini de façon à ce que le taux de fausses concordances estimé correspondant soit μ^0. Les valeurs initiales pour m_k, $k = 1, 2, \ldots, 4$, ont été fixées à 0.9.

Le chiffre zéro dans le tableau 1 (accord pour le prénom, désaccord pour tous les autres identificateurs) était traité comme un zéro structurel pendant la production des données. Parmi les modèles log-linéaires pour lesquels on n'utilisait pas plus de six paramètres, c'est le modèle $M(1)M(2), M(3), M(4)$ qui donne le meilleur ajustement avec les données du tableau 1. Ce modèle, dans lequel on utilise la dépendance pour les résultats de comparaisons pour le prénom et l'initiale d'un deuxième prénom, ne donne pas un très bon ajustement. La statistique du test du rapport des vraisemblances pour le manque d'ajustement est 57.95, une valeur extrême par rapport à la distribution de référence chi carré avec 9 degrés de liberté. Le modèle à variable latente $G(1)G(2), G(3), G(4)$ était estimé, à l'aide de la pondération itérative, pour chaque ensemble de données synthétiques. L'ajustement de ce modèle avec les ensembles de données synthétiques était légèrement meilleur que l'ajustement du modèle $M(1)M(2), M(3), M(4)$ avec les données sur les concordances vraies. La plus élevée des statistiques du test du manque d'ajustement parmi les cinquante ensembles de données synthétiques était 25.03 et le modèle n'était rejeté que dix fois au niveau de signification 5%.

Les moyennes des estimations du taux d'erreur de classification obtenues à l'aide des ensembles de données synthétiques ainsi que les erreurs-types de Monte Carlo correspondantes sont présentées au tableau 2 pour les non-concordances vraies et au tableau 3 pour les concordances vraies. Après multiplication par 99, les taux d'erreur pour les non-concordances vraies représentent le nombre de non-concordances vraies mal classées divisé par le nombre de concordances vraies. Les résultats sont présentés pour la méthode des moments et pour la pondération itérative, ainsi que pour la méthode itérative avec $\mu^0 = 0.0000625$, 0.00025 et 0.001. Les biais dans les taux d'erreur estimés pour les non-concordances vraies sont généralement faibles. La méthode itérative avec $\mu^0 = 0.001$ fournit les meilleures estimations, vient ensuite la pondération itérative. Pour les concordances vraies, le rendement de la méthode itérative dépend beaucoup du choix de μ^0. Bien que la méthode itérative donne de bons résultats pour $\mu^0 = 0.001$, les biais pour $\mu^0 = 0.0000625$ et $\mu^0 = 0.00025$ sont considérables. Les estimations du taux d'erreur de classification pour les concordances vraies obtenues à l'aide de la méthode des moments comprennent aussi des biais importants. Les biais dans les estimations obtenues par pondération itérative sont relativement faibles.

Tableau 3

Taux d'erreur de classification, concordances vraies, données synthétiques
(erreur-types Monte Carlo entre parenthèses)

Taux estimé	Taux réel				
	Méthode des moments	Méthode itérative $\mu^0 = 0.0000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$	Pondération itérative
0.02	0.0580 (0.0013)	0.1179 (0.0041)	0.0507 (0.0014)	0.0149 (0.0008)	0.025 (0.0012)
0.04	0.0773 (0.0014)	0.1362 (0.004)	0.0735 (0.0012)	0.0359 (0.0018)	0.0455 (0.0016)
0.06	0.0966 (0.0014)	0.1542 (0.0038)	0.0954 (0.0012)	0.0660 (0.0014)	0.0646 (0.0018)
0.08	0.1159 (0.0014)	0.1722 (0.0036)	0.1165 (0.0012)	0.0866 (0.0017)	0.0841 (0.0019)
0.10	0.1348 (0.0014)	0.1904 (0.0035)	0.1319 (0.0014)	0.1025 (0.002)	0.1043 (0.002)

Tableau 4

Taux d'erreur de classification, non-concordances vraies, données synthétiques modifiées
(erreur-types Monte Carlo entre parenthèses)

Taux estimé (× 99)	Taux réel (× 99)	
	Méthode des moments	Pondération itérative
0.02	0.0189 (0.0008)	0.0194 (0.001)
0.04	0.0385 (0.0011)	0.0396 (0.0016)
0.06	0.0577 (0.0013)	0.0589 (0.0019)
0.08	0.0767 (0.0016)	0.0785 (0.002)
0.10	0.0957 (0.002)	0.0978 (0.0021)

Tableau 5

Taux d'erreur de classification, concordances vraies, données synthétiques modifiées
(erreurs-types Monte Carlo entre parenthèses)

Taux estimé	Taux réel	
	Méthode des moments	Pondération itérative
0.02	0.0553 (0.0014)	0.0208 (0.0011)
0.04	0.0747 (0.0014)	0.0415 (0.0016)
0.06	0.094 (0.0014)	0.0608 (0.0018)
0.08	0.1134 (0.0014)	0.0805 (0.002)
0.10	0.1325 (0.0015)	0.1007 (0.002)

Les renseignements présentés dans les tableaux 4 et 5 sont basés sur une série d'ensembles de données synthétiques produits à l'aide d'une version modifiée du tableau 1. Les valeurs probables des chiffres dans les cases du tableau 1 selon le modèle $M(1)M(2), M(3), M(4)$ ont été utilisées pour produire les données. Les biais dans les estimations du taux d'erreur de classification fondées sur un modèle obtenues à l'aide de la méthode des moments sont fortement réduits quand on utilise le modèle à variable latente $G(1)G(2), G(3), G(4)$ estimé à l'aide de la pondération itérative, particulièrement dans le cas des concordances vraies.

6. COMPARAISON DES MÉTHODES D'ESTIMATION - DONNÉES RÉELLES

Les résultats des comparaisons des trois méthodes d'estimation effectuées à l'aide de données provenant d'une application de couplage d'enregistrements sont présentées dans cette section. Deux fichiers de données utilisés dans un travail empirique présenté par Fair et Lalonde (1987) ont été employés. Le premier fichier renfermait des renseignements sur les mineurs ontariens obtenus de la Commission des accidents du travail. Le deuxième fichier comprenait des renseignements tirés de la Base canadienne de données sur la mortalité (BCDM) pour les décès de

particuliers pendant la période allant de 1964 à 1977 inclusivement. Le fichier des mineurs n'incluait que les enregistrements avec un numéro d'assurance sociale valable. Le deuxième fichier renfermait des enregistrements qui avaient été retenus après une comparaison initiale visant à éliminer les enregistrements qui n'avaient aucune similitude avec un quelconque des enregistrements dans les fichiers des mineurs. Le statut vital de chaque mineur, à la fin de 1977, avait été classé comme "décès confirmé", "survie confirmée" ou "non retrouvé lors du suivi", basé sur un couplage antérieur, combiné avec des procédures de suivi complètes, y compris un examen manuel. Les enregistrements dans le fichier des mineurs, pour les personnes dont le statut vital est "décès confirmé", incluaient le numéro d'enregistrement du décès dans la BCDM. On peut trouver plus de renseignements sur la construction des fichiers et sur les procédures utilisées pour déterminer l'état véritable du couplage dans Fair et Lalonde.

Quatre identificateurs, le premier prénom, le code "NYSIIS" du nom de jeune fille de la mère, le jour de naissance et le mois de naissance, ont été choisis comme champs d'appariement pour la comparaison. Les enregistrements dans le fichier des mineurs pour lesquels le statut vital était "non retrouvé lors du suivi" ont été éliminés. Après que les enregistrements pour lesquels des valeurs manquaient soit dans au moins un champ d'appariement, soit dans le champ de l'année de naissance, aient aussi été supprimés, le fichier A (basé sur le fichier des mineurs) renfermait 45,638 enregistrements et le fichier B (basé sur la BCDM) comprenait 24,597 enregistrements. En limitant les comparaisons des deux fichiers aux enregistrements pour lesquels il y avait accord pour le code "NYSIIS" du nom de jeune fille de la mère et la différence entre les années de naissance est d'au plus un, il y avait 26,500 non-concordances vraies et 2,063 concordances vraies.

Les fréquences des résultats pour les concordances vraies et pour les non-concordances vraies sont présentées au tableau 6. Tous les modèles log-linéaires correspondant à un modèle à variable latente non saturé (c'est-à-dire, tous les modèles comprenant moins de huit paramètres) sont rejetés par les données sur les fréquences pour les non-concordances vraies à un niveau de signification très faible. Parmi les modèles comprenant moins de huit paramètres, le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ correspond à la statistique la plus faible du test du rapport des vraisemblances pour le manque d'ajustement, soit 35.29. Le modèle $M(1)$, $M(2)M(4)$, $M(3)M(4)$ fournit un ajustement adéquat avec les données pour la concordance vraie (la statistique du test du rapport des vraisemblances est 10.29).

Les estimations des probabilités d'accord ont été calculées à l'aide de la méthode des moments, de la méthode itérative et de la pondération itérative, en utilisant le modèle à variable latente $G(1)$, $G(2)G(4)$, $G(3)G(4)$. La statistique du test du rapport des vraisemblances pour le modèle fondé sur l'hypothèse d'indépendance correspondant à l'estimateur de la méthode des moments est 108 (six degrés de liberté). Le modèle fondé sur l'hypothèse d'indépendance est rejeté par les données à un niveau de signification très faible. Par contre, la statistique du test du rapport des vraisemblances pour le modèle à variable latente $G(1)$, $G(2)G(4)$, $G(3)G(4)$ est 1.44 (deux degrés de liberté), ce qui laisse supposer un ajustement adéquat. Les estimations, fondées sur un modèle, du taux d'erreur de classification correspondant à chaque ensemble d'estimations de probabilité ont été calculées pour divers seuils. Les taux d'erreur de classification réels sont comparés aux estimations fondées sur un modèle pour les non-concordances vraies dans le tableau 7 et pour les concordances vraies dans le tableau 8. On a modifié l'échelle des taux d'erreur pour les non-concordances vraies afin que le nombre de concordances vraies se trouve au dénominateur.

Tableau 6

Fréquences des résultats, données réelles

Résultat selon l'indenticateur: 0 = Désaccord, 1 = Accord				Fréquence	
Premier prénom	Code NYSIIS du nom de jeune fille de la mère	Jour de naissance	Mois de naissance	Concordances vraies	Non-concordances vraies
0	0	0	0	4	22,100
0	0	0	1	3	888
0	0	1	0	11	2,322
0	0	1	1	128	211
0	1	0	0	3	199
0	1	0	1	7	19
0	1	1	0	27	27
0	1	1	1	242	13
1	0	0	0	9	576
1	0	0	1	10	32
1	0	1	0	52	94
1	0	1	1	392	4
1	1	0	0	27	13
1	1	0	1	32	1
1	1	1	0	115	0
1	1	1	1	1,001	1
			Total	2,063	26,500

Les estimations, fondées sur un modèle, du taux d'erreur de classification obtenues à l'aide de la méthode itérative sont très imprécises, particulièrement pour les non-concordances vraies, quelle que soit la valeur de μ^0. Les estimations du taux d'erreur obtenues à l'aide de la

Tableau 7

Taux d'erreur de classification, non-concordances vraies, données réelles

Taux estimé (× 12.84)	Taux réel (× 12.84)				
	Méthode des moments	Méthode itérative $\mu^0 = 0.0000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$	Pondération itérative
0.02	0.0368	1.311	0.1859	0.186	0.0339
0.04	0.0796	1.314	0.1888	0.193	0.0649
0.06	0.1224	1.317	0.1917	0.1967	0.0684
0.08	0.1573	1.323	0.1990	0.1994	0.1106
0.10	0.1863	1.333	0.60	0.4066	0.1282

Tableau 8

Taux d'erreur de classification, concordances vraies, données réelles

Taux estimé	Taux réel				
	Méthode des moments	Méthode itérative $\mu^0 = 0.0000625$	Méthode itérative $\mu^0 = 0.00025$	Méthode itérative $\mu^0 = 0.001$	Pondération itérative
0.02	0.0166	0.0141	0.0193	0.0225	0.0105
0.04	0.0318	0.0264	0.029	0.0278	0.0263
0.06	0.0598	0.0383	0.0472	0.0326	0.0529
0.08	0.0782	0.0416	0.1372	0.0488	0.0784
0.10	0.0966	0.045	0.1393	0.1371	0.0958

pondération itérative sont un peu moins précises que celles basées sur la méthode des moments pour les concordances vraies. Toutefois, elles sont beaucoup plus précises que les estimations obtenues avec la méthode des moments pour les non-concordances vraies.

Une mise en garde s'impose. Bien que le modèle $U(1)$, $U(2)U(4)$, $U(3)U(4)$ ne décrive pas de façon adéquate les dépendances parmi les non-concordances vraies, l'algorithme de pondération itérative a permis d'obtenir un bon ajustement à l'aide d'une estimation de la proportion d'enregistrements appariés (0.0747) qui diffère un peu de la valeur vraie (0.0722). On peut aussi obtenir un ajustement semblable à l'aide du modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ainsi qu'une estimation de 0.077 pour la proportion des appariements. Les estimations du taux d'erreur basées sur le modèle $G(1)G(2)$, $G(1)G(3)$, $G(4)$ ne sont pas meilleures que les estimations obtenues à l'aide de la méthode des moments.

7. CONCLUSIONS

Dans cet article, on a traité de la question de l'estimation des taux d'erreur de classification pour le couplage d'enregistrements. Le modèle de Fellegi-Sunter permet de calculer des estimations des taux d'erreur de classification à l'aide d'estimations des probabilités d'accord. Ces estimations fondées sur un modèle ont généralement de mauvaises propriétés en pratique. Il a été démontré que leurs propriétés peuvent être améliorées en estimant avec soin les probabilités d'accord. Trois méthodes d'estimation ont été évaluées à l'aide de données synthétiques ainsi que de renseignements provenant d'une application réelle.

Pour deux des trois méthodes, on a utilisé l'hypothèse que les résultats des comparaisons pour différents champs de données sont indépendants. Cette hypothèse n'était pas valable soit pour les données synthétiques, soit pour les données réelles. Les données synthétiques incluaient de fortes dépendances pour les concordances vraies et des dépendances mineures pour les non-concordances vraies. Les dépendances pour les données réelles étaient particulièrement importantes dans le cas des non-concordances vraies. Les estimations du taux d'erreur de classification obtenues à l'aide de la méthode des moments, qui est fondée sur l'hypothèse d'indépendance, comportaient des biais considérables pour les données synthétiques et étaient relativement imprécises pour les données réelles. L'importance du biais dans les estimations des taux d'erreur de classification obtenues à l'aide de la méthode itérative dépendait de la définition d'un ensemble initial de concordances. Bien que certaines définitions de l'ensemble initial de concordances aient mené à des biais relativement faibles, d'autres ont produit des estimations avec des biais

beaucoup plus considérables que ceux obtenus à l'aide des autres méthodes. Pour les données réelles, toutes les définitions de l'ensemble initial d'appariements considérées ont mené à des estimations très imprécises du taux d'erreur. Il n'existe pas de règles mathématiques qui permettent de choisir un ensemble initial de concordances pour la méthode itérative. Rien dans les résultats présentés dans cet article ne permet de recommander l'utilisation de cette dernière méthode.

La troisième méthode est basée sur un paramétrage de dépendances entre les résultats de comparaisons, pour différents champs de données, à l'aide d'effets log-linéaires. Avec ce paramétrage, on peut obtenir des estimations des probabilités d'accord qui ne sont pas fondées sur l'hypothèse d'indépendance en utilisant la méthode de pondération itérative pour estimer les paramètres d'un modèle à variable latente. Pour les ensembles de données synthétiques, avec absence d'indépendance, les estimations des taux d'erreur de classification fondées sur un modèle calculé par pondération itérative comprenaient des biais de beaucoup inférieurs à ceux qui s'appliquaient aux estimations basées sur l'hypothèse d'indépendance. Bien que l'ajustement obtenu à l'aide du modèle à variable latente pour la plupart des ensembles de données synthétiques ait été meilleur que celui obtenu pour un modèle fondé sur l'hypothèse d'indépendance, on relevait parfois un manque d'ajustement important avec le premier modèle. Quand les données synthétiques ont été modifiées afin d'améliorer l'ajustement obtenu avec le modèle à variable latente, rien ne montrait qu'il y avait un biais dans les estimations du taux d'erreur de classification fondées sur un modèle. Pour les données réelles, il y avait des écarts importants par rapport à l'hypothèse d'indépendance tant pour les concordances vraies que pour les non-concordances vraies. Les estimations du taux d'erreur fondées sur un modèle obtenues à l'aide de la pondération itérative étaient un peu moins précises que les estimations fondées sur la méthode des moments pour les concordances vraies et beaucoup plus précises pour les non-concordances vraies.

Les résultats présentés ici montrent que l'on peut améliorer les propriétés des estimations du taux d'erreur de classification fondées sur un modèle quand on utilise un estimateur approprié des probabilités d'accord. Les modèles à variable latente ainsi que la pondération itérative fournissent une méthode pour incorporer des dépendances entre des résultats de comparaisons pour différents champs de données pendant l'estimation des probabilités d'accord.

REMERCIEMENTS

Les auteurs désirent remercier William Winkler qui a fourni le code machine sur lequel est basé le programme d'estimation par pondération itérative que nous avons utilisé pour obtenir nos résultats, ainsi que Fritz Scheuren et trois arbitres anonymes pour leurs commentaires, sur une version antérieure de cet article, qui ont mené à des améliorations considérables tant pour ce qui est du contenu que de la présentation. Nous désirons aussi remercier Martha Fair et Pierre Lalonde qui nous ont permis d'obtenir les données sur les mineurs ontariens ainsi que les données sur la fréquence des résultats pour les concordances vraies.

BIBLIOGRAPHIE

BARTLETT, S., KREWSKI, D., WANG, Y., et ZIELINSKI, J.M. (1992). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé. *Techniques d'enquête*, 19, 3-13.

BELIN, T.R. (1990). A proposed improvement in computer matching techniques. Dans *Statistics of Income and Related Administrative Record Research*: 1988-1989, U.S. Internal Revenue Service, 167-172.

BELIN, T.R., et RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 657-668.

FAIR, M.E., et LALONDE, P. (1987). Identificateurs manquants et justesse de l'observation suivie. *Recueil: Les utilisations statistiques des données administratives, Statistique Canada*, 111-125.

FAIR, M.E., NEWCOMBE, H.B., et LALONDE, P. (1988). Improved mortality searches for Ontario miners using social insurance index identifiers. Rapport de recherche, Commission de contrôle de l'énergie atomique.

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

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*. London: Academic Press.

IMSL (1987). Math/Library FORTRAN subroutines for mathematical applications. Houston: IMSL Inc.

MORÉ, J., GARBOW, B., et HILLSTROM, K. (1980). User guide for MINPACK-1. Argonne National Labs Report ANL-80-74.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, 145-155.