Techniques d'enquête, juin 1993
Vol. 19, n° 1, pp. 45–65
Statistique Canada

Analyse de régression de fichiers de données couplés par ordinateur

FRITZ SCHEUREN et WILLIAM E. WINKLER[1]

RÉSUMÉ

Le présent article s'intéresse à la façon de traiter les erreurs de couplage d'enregistrements lorsqu'on effectue une analyse de régression. Des travaux récents de Rubin et Belin (1991) et de Winkler et Thibaudeau (1991) fournissent la théorie, les algorithmes de calcul et le logiciel nécessaires à l'estimation des probabilités de concordance. Ces progrès nous permettent de mettre à jour les travaux de Neter, Maynes et Ramanathan (1965). Des méthodes de redressement sont présentées, et certaines simulations fructueuses sont décrites. Nos résultats sont préliminaires et visent en grande partie à susciter d'autres travaux.

MOTS CLÉS: Couplage d'enregistrements; erreur de couplage; analyse de régression.

1. INTRODUCTION

L'information contenue dans deux bases de données informatiques peut être combinée à des fins d'analyse et de prise de décisions. Par exemple, un épidémiologiste pourrait être intéressé à évaluer l'effet d'un nouveau traitement contre le cancer en couplant l'information provenant d'un ensemble d'études de cas médicales et les données d'un registre de décès, afin de disposer de renseignements sur les causes et les dates des décès (p. ex. Beebe 1985). Un économiste désirant évaluer l'efficacité de décisions en matière de politique énergétique pourrait coupler une base de données contenant de l'information sur les combustibles et les matières premières pour un ensemble d'entreprises, et une base de données contant les quantités et les types de biens produits par ces entreprises (p. ex. Winkler 1985). Si des identificateurs uniques, par exemple des numéros de sécurité sociale ou des numéros d'identification d'employeur ayant été vérifiés, sont disponibles, le couplage des sources de données peut être direct et permettre, sans autres détours, le recours aux méthodes habituelles d'analyse statistique.

Si l'on ne dispose pas d'identificateurs uniques (p. ex. Jabine et Scheuren 1986), le couplage doit être effectué d'après le nom, l'adresse, l'âge ou un autre élément descriptif de l'entreprise ou de la personne. Même en l'absence de variantes d'orthographe ou d'erreurs typographiques, il se peut que le nom, par exemple "Smith" ou "Robert" ne permette pas à lui seul d'identifier une unité. En outre, l'utilisation d'adresses entraîne souvent des erreurs attribuables aux différences de formats, puisque les logiciels d'analyse syntaxique ou de normalisation existants ne permettent pas de s'assurer complètement qu'un numéro de maison ou un nom de rue, par exemple, est bien comparé à un autre numéro de maison ou un autre nom de rue. Les adresses d'une unité qu'on veut coupler peuvent également varier, soit parce que l'une d'elles est erronée, soit en raison d'un déménagement.

Depuis quelques années, beaucoup de nouveaux travaux ont été publiés en Amérique du Nord dans le domaine des techniques de couplage d'enregistrements (p. ex. Jaro 1989; et Newcombe, Fair et Lalonde 1992). Certains de ces résultats sont le produit d'une série de conférences amorcée au milieu des années 1980 (p. ex. Kilss et Alvey 1985; Howe et Spasoff 1986; Coombs et Singh 1987; Carpenter et Fair 1989); les efforts visant à étudier le sous-dénombrement dans le recensement décennal de 1990 ont constitué un autre important facteur de stimulation aux États-Unis (p. ex. Winkler et Thibaudeau 1991). Le nouvel ouvrage de Newcombe (1988) a également joué un important rôle déclencheur. Enfin, les efforts provenant d'autres sources ont aussi été considérables (p. ex. Copas et Hilton 1990).

Ce qui étonne dans l'ensemble de ces travaux récents, c'est le degré de maturité atteint dans l'élaboration des principaux fondements théoriques des méthodes de couplage informatisées. Des applications pratiques bien structurées remontent au moins jusqu'aux années 1950, notamment aux travaux de Newcombe et de ses collaborateurs (p. ex. Newcombe et coll. 1959). Environ dix ans plus tard, un fondement théorique solide pour ces idées de base a été établi dans les articles de Tepping (1968) et, en particulier, de Fellegi et Sunter (1969).

Si l'intérêt à l'égard du couplage d'enregistrements s'est ainsi maintenu, c'est en partie parce que la révolution informatique a permis d'utiliser des techniques de plus en plus efficaces. La prolifération des fichiers accessibles sous forme électronique a également élargi la gamme des applications possibles. Un autre facteur d'intérêt est né du besoin d'établir des ponts entre le domaine relativement étroit (même obscur) du couplage par ordinateur et le reste du domaine des statistiques (p. ex. Scheuren 1985). Le présent article appartient à cette dernière catégorie, et vise à examiner les aspects particuliers des analyses de régression portant sur des ensembles de données couplés.

De façon générale, nous n'examinerons pas ici les techniques de couplage. Nous nous intéresserons plutôt à ce

[1] Fritz Scheuren, U.S. Internal Revenue Service, Washington DC 20224; William E. Winkler, U.S. Bureau of the Census, Washington DC 20233.

qui survient **après** que l'état de concordance a été déterminé. Nous supposerons la situation type où le responsable du couplage effectue son travail séparément de l'analyste. Nous supposerons aussi que l'analyste (ou l'utilisateur) pourrait vouloir appliquer des techniques statistiques classiques – régression, tableaux de contingence, tables de survie, *etc.* – au fichier résultant du couplage. Une question essentielle que l'on souhaite alors explorer est la suivante: "Que peut faire le responsable du couplage pour aider l'analyste?". La question suivante en découle: "Que devrait savoir l'analyste au sujet du couplage et comment cette information devrait-elle être utilisée?".

À notre avis, il est important de définir théoriquement les étapes du couplage et de l'analyse dans le cadre d'un système statistique unique et de concevoir les stratégies appropriées en conséquence. De toute évidence, la qualité du travail de couplage peut avoir un impact direct sur n'importe quelle analyse effectuée. Toutefois, il est rare que des mesures directes de cet impact soient fournies (p. ex. Scheuren et Oh 1975). Rubin (1990) a signalé le besoin de formuler des énoncés inférentiels destinés à résumer l'information que révèlent les données analysées. Les idées de Rubin ont été exprimées dans le contexte du recours à des techniques de préparation des données comme l'édition et l'imputation, dans des situations où, fréquemment, la non-réponse peut rendre inopérantes les méthodes statistiques courantes incluses dans les logiciels existants. Nous croyons que les idées de Rubin s'appliquent, au moins à un degré égal, au couplage d'enregistrements.

La présente analyse est divisée en quatre sections. D'abord, nous faisons un rappel sur certains aspects du couplage, car toute réponse – même partielle – dépendra des fichiers à coupler et de l'utilisation faite des données couplées. Dans la section suivante, nous présentons notre cadre méthodologique en centrant notre attention, comme il a déjà été indiqué, uniquement sur l'analyse de régression. À la section 4, nous présentons quelques résultats de simulations exploratoires. Ces simulations visent à aider le lecteur à soupeser les idées que nous soumettons et à percevoir la nature de certaines difficultés. Dans la dernière section, nous présentons des conclusions préliminaires et nous suggérons des voies de recherche future. Une brève annexe, offrant plus de détails sur des aspects théoriques, est également incluse.

2. FONDEMENTS DU COUPLAGE D'ENREGISTREMENTS

Lorsque deux fichiers ou plus sont couplés, il se peut qu'un enregistrement particulier d'un fichier ne soit pas apparié au bon enregistrement dans l'autre fichier. Si l'on ne dispose pas d'un identificateur unique pour les enregistrements correspondants des deux fichiers – ou encore si un tel identificateur souffre d'inexactitudes – le processus de couplage est sujet aux erreurs. Si la base de données couplée qui en résulte contient une part importante d'information provenant de paires d'enregistrements qui ont été appariées de façon erronée, ou encore une part importante d'enregistrements non appariés qui auraient dû l'être, les analyses statistiques peuvent être suffisamment compromises pour que les techniques statistiques habituelles donnent des résultats trompeurs. Nous traiterons essentiellement, dans le présent article, de l'impact qu'exercent les liens erronés sur les résultats des analyses. L'impact des problèmes causés par la présence de non-liens erronés (un type d'échantillonnage implicite qui peut entraîner des biais de sélection) est brièvement analysé dans la dernière section.

2.1 Modèle de couplage d'enregistrements de Fellegi-Sunter

Le processus de couplage d'enregistrements tente de classer les paires d'un espace produit $\mathbf{A} \times \mathbf{B}$ provenant de deux fichiers A et B en deux ensembles: M, l'ensemble des liens vrais, et U, l'ensemble des non-liens vrais. Énonçant de façon rigoureuse des concepts introduits par Newcombe (p. ex. Newcombe et coll. 1959), Fellegi et Sunter (1969) ont examiné des rapports de probabilités de la forme:

$$R = Pr(\gamma \in \Gamma \mid M)/Pr(\gamma \in \Gamma \mid U), \qquad (2.1)$$

où γ est un profil de concordance arbitraire appartenant à l'espace de comparaison Γ. Par exemple, Γ pourrait être formé de huit profils représentant une concordance simple ou non pour le nom, le prénom et l'âge. Autre possibilité, chaque $\gamma \in \Gamma$ pourrait en outre tenir compte de la fréquence relative à laquelle des noms particuliers, comme Smith ou Zabrinsky, surviennent. Les zones des enregistrements qui sont comparées (nom, prénom, âge) sont appelées *variables de comparaison*.

La règle de décision est donnée par:

Si $R >$ *Limite sup.*, classer la paire comme un lien.

Si *Limite inf.* $\leq R \leq$ *Limite sup.*, classer la paire comme un lien possible et la garder pour la somettre à une vérification manuelle. (2.2)

Si $R <$ *Limite inf.*, classer la paire comme un non-lien.

Fellegi et Sunter (1969) ont montré que la règle de décision est optimale, c'est-à-dire que pour toute paire de bornes fixes de R, la région milieu est minimisée pour l'ensemble des règles de décision relatives au même espace de comparaison Γ. Les seuils *Limite supérieure* et *Limite inférieure* sont déterminés d'après les bornes fixées pour l'erreur. Nous appelons le rapport R ou toute transformation monotone croissante de ce dernier (donnée, par exemple, par un logarithme) un *poids d'appariement* ou un *poids de concordance totale*.

Dans les applications réelles, l'optimalité de la règle de décision (2.2) est fortement liée à l'exactitude des estimations des probabilités données en (2.1). Ces probabilités sont appelées *paramètres d'appariement*. Les paramètres estimés sont (quasi) *optimaux* s'ils donnent des règles de

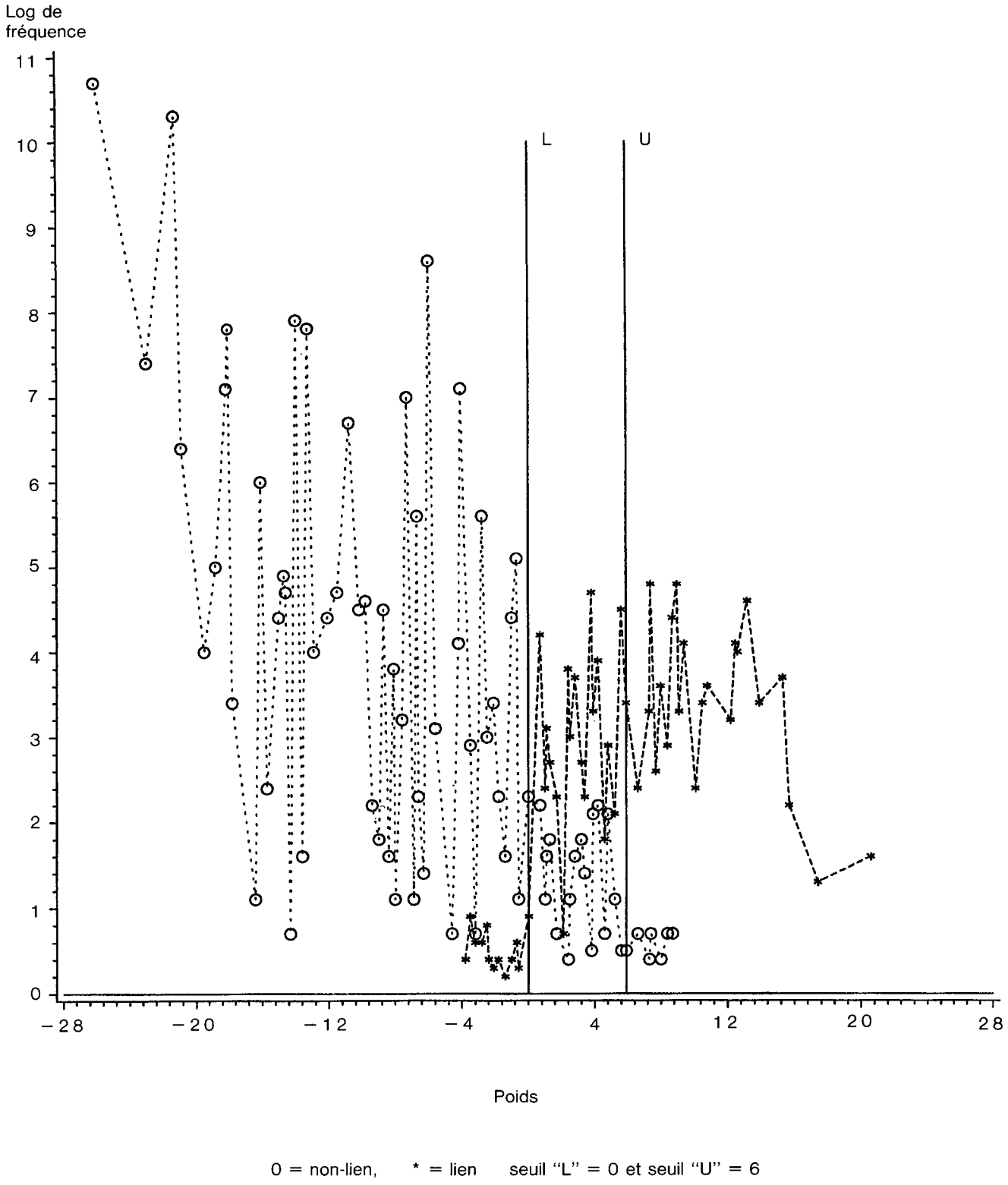

Figure 1. Logarithme de fréquence *vs* poids, Liens et non-liens

décision produisant des résultats (presque) aussi bons que ceux donnés par la règle (2.2) quand les paramètres vrais sont utilisés.

La méthode de Fellegi-Sunter est essentiellement un prolongement direct de la théorie classique des tests d'hypothèses au couplage d'enregistrements. Pour décrire le modèle plus en détail, supposons que nous ayons deux fichiers de tailles n et m avec - sans perte de généralité - $n \leq m$. En vertu du processus de couplage, une comparaison pourrait être faite pour l'ensemble des $n \times m$ paires possibles d'enregistrements (un élément de la paire provenant de chacun des fichiers). Une décision est alors prise quant à savoir si oui ou non chaque paire de comparaison représente la même unité ou si l'on dispose d'éléments suffisants pour déterminer s'il y a un lien ou non.

Schématiquement, il est d'usage d'organiser les $n \times m$ paires du tableau selon une mesure de la probabilité que la paire représente des enregistrements relatifs à la même unité. À la figure 1, par exemple, nous avons tracé deux courbes. La courbe du côté droit est une distribution hypothétique des n liens vrais par rapport au "poids d'appariement" (calculé d'après (2.1), mais en logarithmes naturels). La courbe du côté gauche est la distribution des $n \times (m - 1)$ paires restantes - les non-liens vrais - en fonction des poids d'appariement, également en logarithmes.

En général, comme le montre la figure 1, les distributions des liens et des non-liens se chevauchent. Aux extrêmes, le recoupement n'a pas de conséquence sur la décision de couplage à prendre; toutefois, il existe une région centrale de liens possibles, disons entre "L" et "U", où il serait difficile, d'après la figure 1 seulement, de décider avec un quelconque degré d'exactitude s'il s'agit de liens ou de non-liens.

Le modèle de Fellegi-Sunter est valide pour n'importe quel ensemble de paires examiné. Toutefois, pour faciliter les calculs, nous pouvons faire porter notre analyse non pas sur la totalité des paires possibles dans $\mathbf{A} \times \mathbf{B}$, mais uniquement sur un sous-ensemble de paires, pour lesquelles les enregistrements des deux fichiers concordent au niveau de données clés, ou de "groupage", considérées comme hautement fiables. Comme exemples de *critères de groupage logiques*, mentionnons un identificateur géographique comme le code postal (p. ex. ZIP), un identificateur de nom de famille comme un code Soundex ou NYSIIS (voir, p. ex. Newcombe 1988, p. 182-184). Signalons, en passant, que le modèle de Fellegi-Sunter ne présuppose pas (comme nous le faisons à la figure 1) que parmi les $n \times m$ paires, il y aura n liens, mais plutôt, s'il n'y a pas d'enregistrements en double dans A ou B, qu'il y aura au maximum n liens.

2.2 Traitement des liens possibles

Même lorsqu'un système de couplage par ordinateur utilise la règle de décision de Fellegi-Sunter pour désigner de façon quasi certaine des paires comme des *liens vrais* ou des *non-liens vrais*, il peut rester un imposant sous-ensemble de paires qui ne représentent que des liens possibles. L'une des façons de traiter ces liens possibles consiste à les examiner manuellement, pour tenter de repérer avec exactitude les liens vrais. Quant aux paires désignées par erreur comme des non-liens, on peut aussi effectuer dans leur cas une vérification additionnelle (peut-être elle aussi manuelle). Dans les deux cas, une telle démarche est coûteuse, longue et sujette aux erreurs.

Fait peu surprenant, la recherche dans le domaine du couplage d'enregistrements a surtout porté, depuis les travaux initiaux de Newcombe, sur des moyens de réduire les étapes de vérification manuelle destinées à l'examen des liens possibles. De vastes progrès ont été accomplis en ce qui a trait à l'amélioration des règles de couplage, grâce à une meilleure utilisation de l'information contenue dans les paires d'enregistrements, et à l'estimation des taux d'erreur par l'entremise de modèles probabilistes.

Diverses méthodes ont été mises à contribution pour améliorer les règles de couplage d'enregistrements. Pour traiter les erreurs typographiques légères, comme le fait d'écrire "Smoth" au lieu de "Smith", Winkler et Thibaudeau (1991) ont proposé une extension du comparateur de chaînes de Jaro (1989). Newcombe et coll. (1989), pour leur part, ont élaboré des méthodes pour la création et l'utilisation de tableaux de concordance partielle. Pour certaines catégories de fichiers, Winkler et Thibaudeau (1991) (voir aussi Winkler 1992; Jaro 1989) ont élaboré des algorithmes espérance-maximisation, ainsi que des techniques de modélisation fondées sur une information *a priori* qui donnait automatiquement les paramètres optimaux de (2.1) en vue de leur utilisation dans les règles de décision (2.2).

Rubin et Belin (1991) ont introduit une méthode permettant l'estimation des taux d'erreur lorsque ceux-ci ne peuvent pas être estimés de façon fiable par les méthodes habituelles (Belin 1991, p. 19-20). En utilisant un modèle selon lequel les courbes "poids *vs.* logarithme de fréquence" produites par le processus de couplage pouvaient être exprimées sous forme d'une combinaison de deux courbes (liens et non-liens), Rubin et Belin ont estimé les courbes qui, à leur tour, donnaient des estimations des taux d'erreur. Pour appliquer leur méthode, Rubin et Belin avaient besoin d'un échantillon d'apprentissage donnant une estimation *a priori* de la forme des deux courbes.

Bien que de nombreux problèmes de couplage surviennent rétrospectivement, souvent dans le cadre d'études épidémiologiques, il est parfois arrivé que des responsables du couplage soient en mesure d'indiquer quelle information était nécessaire dans les deux ensembles de données d'après les besoins connus de l'analyse. En exigeant une meilleure information de couplage, comme cela s'est fait dans l'enquête post-dénombrement du recensement de 1990 (voir, p. ex. Winkler et Thibaudeau 1991), on est parvenu à réduire au minimum les ensembles de liens possibles.

Malgré ces progrès, il est possible que le responsable du couplage et l'analyste doivent finalement recourir à une vérification manuelle. Même aujourd'hui, les coûts additionnels en temps, en argent et en erreurs résiduelles cachées peuvent encore être imposants. Y a-t-il des solutions

de rechange fiables à une vérification complète? Nous croyons que oui, et cette conviction motive la perspective que nous adoptons à la section 3, où nous examinons les erreurs de couplage dans un contexte d'analyse de régression. D'autres approches, toutefois, pourraient être nécessaires dans des cadres d'analyse différents.

3. RÉGRESSION PORTANT SUR DES DONNÉES COUPLÉES

Pour les besoins de notre examen de la régression, nous faisons l'hypothèse que le responsable du couplage a aidé l'analyste en lui fournissant un fichier de données combiné comprenant des paires d'enregistrements – un de chacun des fichiers de départ – ainsi que la probabilité de concordance et l'état de concordance de chaque paire. Les liens, les non-liens et les liens possibles sont tous inclus et bien identifiés. Il semble évident qu'il faille garder les liens probables et les liens possibles, mais il est moins évident que les non-liens probables soient utiles. Toutefois, comme l'a signalé Newcombe, l'information sur les non-liens probables est nécessaire au calcul des biais. Nous faisons l'hypothèse qu'il sera suffisant de garder au maximum deux ou trois paires comprenant un enregistrement du fichier B pour chaque enregistrement du fichier A. Les deux ou trois paires ayant les poids d'appariement les plus élevés seraient conservées.

Plus précisément, nous supposerons que le fichier des liens a été élargi de telle façon que chaque enregistrement du plus petit des deux fichiers a été apparié avec, disons, les *deux* enregistrements du fichier le plus gros ayant les poids d'appariement les plus élevés. Puisque $n \leq m$, nous conservons $2n$ des $n \times m$ paires possibles. Pour chaque enregistrement, nous conservons les indicateurs de concordance et les probabilités associées aux enregistrements avec lesquels il forme une paire. Dans certains cas, nous aurons des combinaisons (lien, non-lien) ou des combinaisons (non-lien, non-lien). À des fins de simplicité, nous ne traiterons pas des situations où il peut exister plus d'un lien vrai; par conséquent, les combinaisons (lien, lien) sont par définition écartées.

Comme on peut s'en rendre compte, une telle structure de données se prête à une analyse par différentes méthodes. Par exemple, nous pouvons décomposer le fichier en trois parties – liens désignés, non-liens et liens possibles. L'analyse, quelle qu'elle soit, pourrait être appliquée séparément à chaque groupe ou à des sous-ensembles de ces groupes. Dans la méthode appliquée ici, nous utiliserons les non-liens pour redresser l'ensemble des liens potentiels et, ainsi, obtenir une perspective additionnelle pouvant se traduire par une erreur quadratique moyenne (EQM) moindre que celle de statistiques calculées uniquement à partir des données sur les liens.

Pour les analyses statistiques, si nous devions utiliser seulement les données relatives aux paires d'enregistrements qui constituent presque assurément des liens, nous nous priverions peut-être d'une bonne part d'information additionnelle provenant de l'ensemble des paires formant des liens possibles qui, en tant que sous-ensemble, pourrait contenir autant de liens vrais que l'ensemble des paires désignées comme des liens. En outre, nous pourrions introduire des biais importants dans les résultats, car certains sous-ensembles des liens vrais susceptibles de nous intéresser pourraient résider principalement dans l'ensemble des liens possibles. Par exemple, si nous analysions des aspects liés à l'action positive et aux revenus, certains enregistrements (par exemple ceux ayant trait aux personnes à faible revenu) pourraient être plus difficiles à coupler par l'information sur le nom et l'adresse et donc se trouver en forte concentration dans l'ensemble des liens possibles.

3.1 Fondement théorique

Neter, Maynes et Ramanathan (1965) ont reconnu que les erreurs introduites au cours du processus de couplage pouvaient avoir un effet négatif sur les analyses ayant comme base les fichiers couplés résultants. Pour montrer comment les idées de Neter et coll. ont inspiré l'approche adoptée dans le présent article, nous fournissons des détails additionnels sur leur modèle. Neter et coll. ont supposé que l'ensemble des enregistrements d'un fichier (1) pouvaient être appariés, (2) avaient toujours la même probabilité p d'être appariés correctement et (3) avaient la même probabilité q d'être appariés incorrectement à n'importe quel enregistrement restant du deuxième fichier (c.-à-d. $p + (N - 1)q = 1$ où N est la taille du fichier). Ils ont généralisé leurs résultats de base en supposant que les ensembles de paires formés à partir des deux fichiers pouvaient être décomposés en classes pour lesquelles (1), (2) et (3) demeuraient valables.

Notre méthode s'inspire de celle de Neter et coll. parce que nous croyons que leur approche est pertinente. Nous souscrivons à leur conclusion selon laquelle un niveau modéré d'erreur de couplage peut introduire un biais élevé dans les coefficients de régression. Nous ne croyons pas, toutefois, que la condition (3) – qui a constitué leur principal moyen de simplification des formules de calcul – sera jamais respectée en pratique. Si le couplage se fonde sur des identificateurs uniques comme les numéros de sécurité sociale, qui sont sujets aux erreurs typographiques, il est improbable qu'un enregistrement comportant une erreur typographique aura la même probabilité d'être couplé incorrectement à tous les enregistrements restants du deuxième fichier. Si les variables de comparaison sont le nom et l'adresse (qui souffrent souvent d'un nombre beaucoup plus grand d'erreurs typographiques), le respect de la condition (3) est encore plus improbable.

Pour bien montrer comment nos travaux sont un prolongement et une généralisation des résultats de Neter et coll. nous allons examiner un cas spécial. Supposons que nous utilisions les moindres carrés ordinaires dans le cadre d'une régression simple de la forme

$$y = a_0 + a_1 x + \epsilon. \qquad (3.1)$$

Supposons, en outre, que des erreurs de couplage se soient produites, c'est à dire que les variables y (d'un

fichier) et les variables x (d'un autre fichier) ne représentent *pas* toujours la *même unité*.

Dans une telle situation, l'estimateur non redressé de a_1 serait biaisé; toutefois, en vertu d'hypothèses comme celle qui consiste à supposer x et y indépendants en cas d'erreur de couplage, il peut être démontré que, si nous connaissons le taux d'erreurs de couplage h, nous pouvons obtenir un estimateur redressé non biaisé par une simple correction de l'estimateur ordinaire, en le multipliant par $(1/(1 - h))$. Intuitivement, on peut croire que les paires faisant l'objet d'une erreur de couplage entraînent une sous-estimation de la véritable corrélation (positive ou négative) entre x et y. Le coefficient redressé compense pour cette sous-évaluation. Une fois connu le coefficient de pente redressé $\hat{a}_1$, l'ordonnée à l'origine appropriée peut être obtenue de l'expression habituelle $\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$, où $\hat{a}_1$ est le coefficient redressé.

Des méthodes d'estimation des erreurs-types de régression peuvent aussi être élaborées pour les cas où il existe des erreurs de couplage. Plutôt que de continuer d'examiner ce cas spécial, toutefois, nous allons examiner comment l'idée de faire un redressement multiplicatif peut être généralisée. Considérons

$$Y = X\beta + \epsilon, \tag{3.2}$$

le modèle de régression unidimensionnel ordinaire, pour lequel tous les termes d'erreur ont comme moyenne zéro et sont indépendants avec variance constante σ^2. Si nous travaillions avec une base de données de taille n, nous ferions une régression de Y en X de la façon normale. Mais, puisque chaque enregistrement est apparié à deux autres, nous disposons globalement de $2n$ paires. Nous souhaitons utiliser (X_i, Y_i), mais nous utilisons plutôt (X_i, Z_i). Z_i peut Y_i, mais peut prendre aussi une autre valeur, Y_j, en raison de l'erreur de couplage.

Pour $i = 1, \ldots, n$,

$$Z_i = \begin{cases} Y_i \text{ avec probabilité } p_i \\ Y_j \text{ avec probabilité } q_{ij} \text{ pour } j \neq i, \end{cases} \tag{3.3}$$

$p_i + \sum_j q_{ij} = 1.$

La probabilité p_i peut être zéro ou un. Nous définissons $h_i = 1 - p_i$ et divisons l'ensemble de paires en n classes mutuellement exclusives. Les classes sont établies d'après les enregistrements d'un des fichiers. Chaque classe comprend la variable x indépendante X_i, la valeur vraie de la variable y dépendante, les valeurs des variables y des enregistrements du deuxième fichier avec lesquels l'enregistrement du premier fichier contenant X_i forme une paire, et les probabilités (ou poids) du couplage par ordinateur. Les liens, les non-liens et les liens possible sont inclus. Dans une hypothèse de couplage biunivoque, pour chaque $i = 1, \ldots, n$, il existe au plus un j tel que $q_{ij} > 0$. Nous posons aussi ϕ, défini par $\phi(i) = j$.

L'idée intuitive de notre approche (et de celle de Neter et coll.) est que nous pouvons, en vertu des hypothèses du modèle, exprimer chaque paire de points de données observés (X,Z) en termes des valeurs vraies (X,Y) et d'un terme de biais (X,b). Toutes les équations nécessaires aux techniques de régression habituelles peuvent alors être obtenues. Nos formules de calcul sont beaucoup plus complexes que celles de Neter et coll. car leur hypothèse contraignante (3) a permis une importante simplification de leurs formules de calcul. En particulier, en vertu de leurs hypothèses, Neter et coll. ont prouvé que la moyenne et la variance des valeurs Z observées étaient toutes deux nécessairement égales à la moyenne et à la variance des valeurs Y vraies.

En vertu du modèle considéré ici, nous observons (voir l'annexe) que

$$\begin{aligned} E(Z) &= (1/n)\sum_i E(Z|i) = (1/n)\sum_i (Y_i p_i + \sum_j Y_j q_{ij}) \\ &= (1/n)\sum_i Y_i + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &= \bar{Y} + B. \end{aligned} \tag{3.4}$$

Puisque chaque X_i, $i = 1, \ldots, n$, peut être apparié avec soit Y_i, soit $Y_{\phi(i)}$, la deuxième égalité dans (3.4) représente $2n$ points. De la même façon, nous pouvons représenter σ_{zy} en fonction de σ_{xy} et d'un terme de biais B_{xy}, et σ_z^2 en fonction de σ_y^2 et d'un terme de biais B_{yy}. Nous ne supposons ni que les termes de biais ont une espérance zéro ni qu'ils ne sont pas corrélés avec les données observées.

Avec les différentes représentations, nous pouvons redresser les coefficients de régression β_{zx} et leurs erreurs-types associées et les ramener aux valeurs vraies β_{yx} et à leurs erreurs-types associées. Notre hypothèse d'un couplage biunivoque (qui n'est pas nécessaire à la théorie générale) est posée en vue de faciliter les calculs et de réduire le nombre d'enregistrements ainsi que la quantité d'information devant être prise en considération au cours du processus de couplage.

Pour les besoins des redressements, nous faisons deux hypothèses cruciales. La première est que, pour $i = 1, \ldots, n$, nous pouvons estimer avec exactitude les probabilités vraies d'une concordance p_i. Voir l'annexe pour une description de la méthode de Rubin et Belin (1991). La deuxième est que, pour chaque $i = 1, \ldots, n$, la valeur vraie Y_i associée à la variable indépendante X_i est celle de la paire ayant le poids d'appariement le plus élevé, tandis que la valeur fausse $Y_{\phi(i)}$ est celle de la paire dont le poids d'appariement vient au deuxième rang. (D'après les simulations effectuées, il semble qu'au moins la première de ces deux hypothèses ait un grand rôle lorsqu'une portion importante des paires sont des liens possibles.)

3.2 Application simulée

À l'aide des méthodes qui viennent d'être décrites, nous avons tenté une simulation avec des données réelles. Notre

méthode de base a consisté à prendre deux fichiers dont les concordances réelles étaient connues et à refaire le couplage selon différentes variables - ou en réalité des versions des mêmes variables avec introduction de différents degrés de distorsion, rendant de plus en plus difficile la distinction entre un lien et un non-lien. Nous avons ainsi créé un cadre comportant assez de pouvoir discriminant pour que l'algorithme de Rubin-Belin visant l'estimation des probabilités fonctionne, mais pas assez de pouvoir discriminant pour que la zone de chevauchement des liens possibles devienne non significative.

Nous avons obtenu les résultats de base de la simulation en commençant avec une paire de fichiers de taille 10,000 comportant une information utile pour les besoins du couplage et dont les concordances réelles étaient connues. Pour les besoins des simulations, divers degrés d'erreur ont été introduits dans les variables de comparaison, différentes quantités de données ont été utilisées pour le couplage et des écarts plus grands par rapport aux probabilités de concordance optimales ont été permis.

Trois scénarios de couplage ont été examinés: (1) *efficace*, (2) *médiocre*, (3) *pauvre*. Le scénario de couplage efficace a consisté à utiliser la plupart des méthodes disponibles qui avaient servi au couplage durant le recensement américain de 1990 (p. ex. Winkler et Thibaudeau 1991). Les variables de comparaison étaient le nom, le prénom, l'initiale, le numéro de maison, le nom de rue, l'identificateur d'appartement ou d'unité, le numéro de téléphone, l'âge, l'état matrimonial, le lien avec le chef de ménage, le sexe et la race. Les probabilités de concordance utilisées dans les ratios de vraisemblance cruciaux nécessaires aux règles de décision ont été fixées à des niveaux voisins du niveau optimal.

Le scénario de couplage médiocre a consisté à utiliser le nom, le prénom, l'initiale, deux variantes de l'adresse, l'identificateur d'appartement ou d'unité, et l'âge. De légères erreurs typographiques ont été introduites de façon indépendante dans un septième des noms de famille et un cinquième des prénoms. Les probabilités de concordance ont été fixées à des niveaux s'écartant du niveau optimal, mais demeurant comparables à celles que pourrait sélectionner un spécialiste expérimenté du couplage par ordinateur.

Dans le scénario de couplage pauvre, les variables de comparaison utilisées étaient le nom, le prénom, une variante de l'adresse et l'âge. De légères erreurs typographiques ont été introduites de façon indépendante dans un cinquième des noms de famille et un tiers des prénoms. Des erreurs typographiques assez graves ont été introduites dans un quart des adresses. Les probabilités de concordance ont été fixées à des niveaux s'écartant substantiellement du niveau optimal. Nous visions ainsi à simuler le choix que pourrait faire un responsable du couplage n'ayant que peu d'expérience dans ce domaine.

Notre capacité de faire la distinction entre les liens vrais et les non-liens vrais varie sensiblement d'un scénario à l'autre. Dans le cas du scénario efficace, nous voyons que le diagramme de dispersion des liens et des non-liens vrais se compose de deux parties presque entièrement séparées (figure 2). Dans le cas du scénario médiocre, les ensembles de points correspondants se recoupent quelque peu (figure 3). Dans le cas du scénario pauvre, le chevauchement est considérable (figure 4).

Ainsi, nous avons transformé le scénario de couplage efficace en un scénario de couplage pauvre (figures 2 à 4) principalement en réduisant l'information de comparaison et en introduisant des erreurs typographiques dans les variables de comparaison. Même si nous avions gardé les mêmes variables de comparaison que dans le scénario de couplage efficace (figure 2), nous aurions pu obtenir un chevauchement des courbes (comme à la figure 4) simplement en faisant varier les paramètres d'appariement donnés par l'équation (2.1). Le scénario de couplage pauvre peut se produire lorsqu'on ne dispose pas d'un logiciel convenable d'analyse syntaxique des noms permettant la comparaison des noms de famille et des prénoms correspondants, ou encore d'un logiciel convenable d'analyse syntaxique des adresses permettant la comparaison des numéros de maison et des noms de rue correspondants. En l'absence d'une analyse syntaxique appropriée, des variables de comparaison déterminantes pour la désignation de nombreux liens vrais ne seront pas adéquatement utilisées.

Notre capacité d'estimer la probabilité d'une concordance varie beaucoup. Les valeurs vraies et estimées de ces probabilités, par classes de poids, sont présentées au tableau 1. Pour les scénarios de couplage efficace et médiocre, les probabilités estimées étaient assez voisines des valeurs vraies. Dans le cas du scénario pauvre, dans lequel la plupart des paires sont des liens possibles, les écarts sont très prononcés.

Pour chaque scénario de couplage, des données empiriques ont été créées. Chaque base de données comprenait un poids d'appariement par ordinateur, des probabilités de concordance vraies et estimées, la variable x indépendante pour la régression, la variable y dépendante vraie, la variable y observée dans l'enregistrement ayant le poids d'appariement le plus élevé, et la variable y observée dans l'enregistrement ayant le poids d'appariement venant au deuxième rang.

Les variables x indépendantes pour la régression ont été construites à l'aide de la procédure RANUNI du SAS, de façon à être uniformément distribuées entre 1 et 101. Pour les fins du présent article, elles ont été choisies indépendamment de toute variable de comparaison. (Bien que nous ayons examiné la situation dans laquelle les variables de régression dépendent d'une ou plusieurs variables de comparaison (Winkler et Scheuren 1991), nous ne présentons pas ces résultats ici.)

Trois scénarios de régression, correspondant à des valeurs de R^2 de plus en plus faibles, ont ensuite été examinés, c.-à-d. R^2 (1) entre 0.75 et 0.80; (2) entre 0.40 et 0.45; et (3) entre 0.20 et 0.22. Les variables dépendantes ont été créées avec des valeurs de départ indépendantes au moyen de la procédure RANNOR du SAS. Dans chaque scénario de couplage (efficace, médiocre ou pauvre), tous les appariements d'enregistrements produits par le processus de couplage et, donc, l'erreur de couplage, étaient fixes.

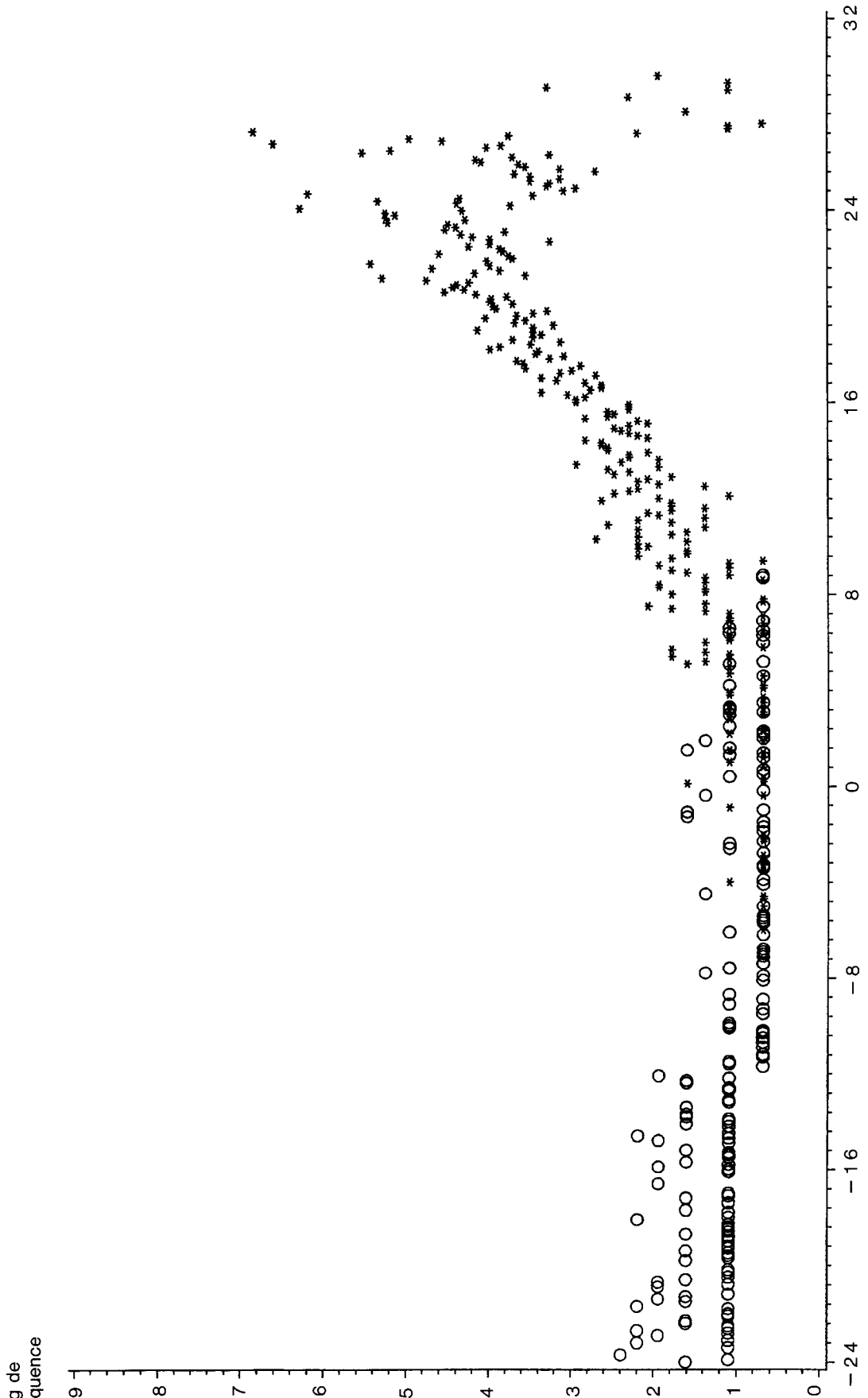

Figure 2. Logarithme de fréquence *vs* poids, Scénario de couplage efficace, liens et non-liens

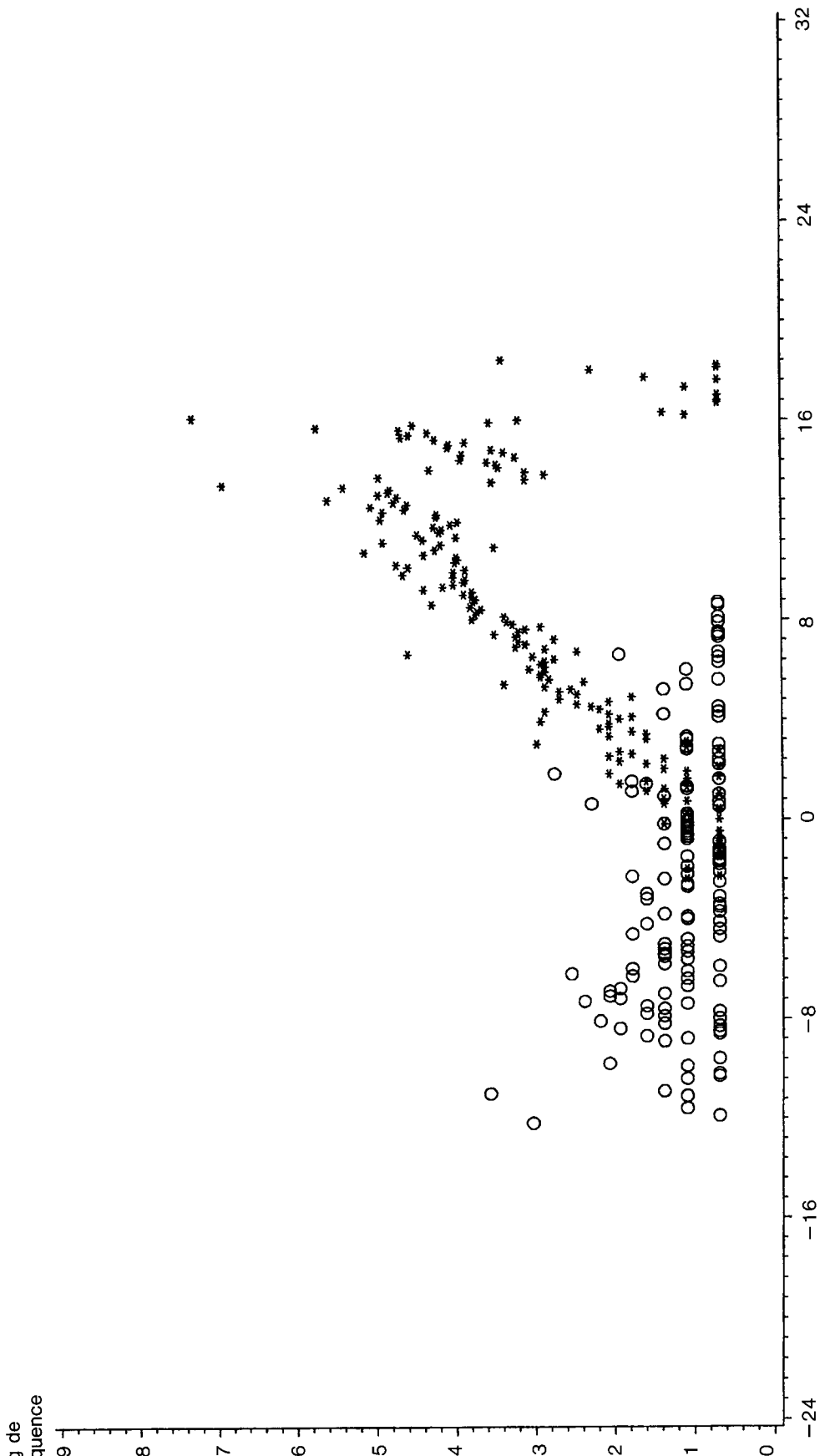

Figure 3. Logarithme de fréquence *vs* poids, Scénario de couplage médiocre, liens et non-liens

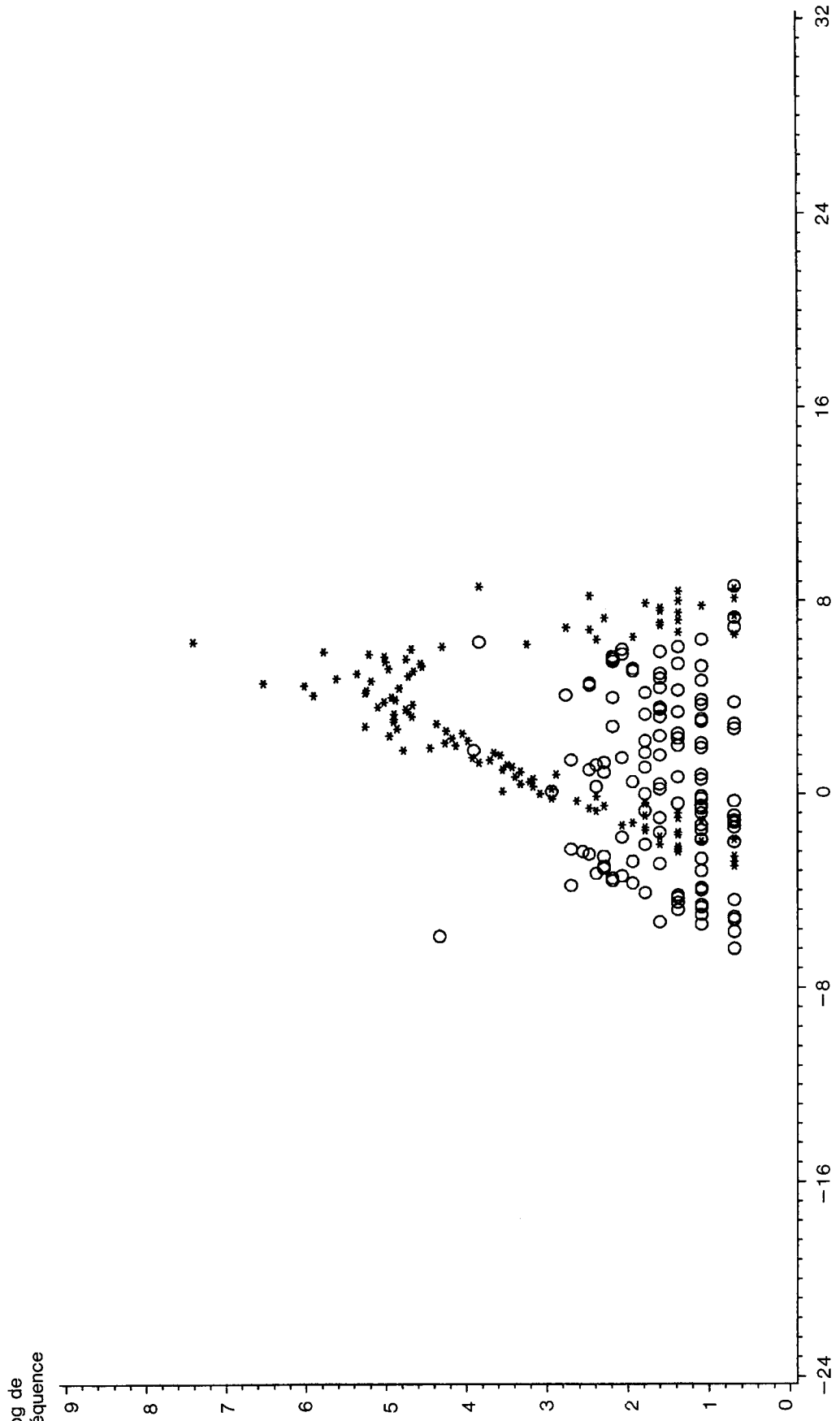

Figure 4. Logarithme de fréquence *vs* poids, Scénario de couplage pauvre, liens et non-liens

Tableau 1

Nombres de liens vrais et de non-liens vrais et probabilités de lien erroné par intervalle de poids pour divers scénarios de couplage; probabilités estimées selon la méthode de Rubin-Belin

Poids	Taux de concordance fausses											
	Scénario efficace				Scénario médiocre				Scénario pauvre			
	Vrais		Prob		Vrais		Prob		Vrais		Prob	
	Liens	NL	Vraies	Est	Liens	NL	Vraies	Est	Liens	NL	Vraies	Est
15+	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
−1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
−2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
−3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
−4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
−5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
−6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
−7	1	6	.86	.98	0	40	.99	.99			.99	.99
−8	0	8	.99	.99	0	41		.99			.99	.99
−9	0	4	.99	.99	0	4		.99			.99	.99
−10−	0	22			0	22		.99			.99	.99

Notes: Dans la première colonne, le poids 10 signifie l'intervalle entre 10 et 11. Les classes de poids 15 et supérieures, ainsi que les classes de poids −9 et inférieures, ont été combinées. Les poids sont des logarithmes de rapports basés sur les probabilités de concordance estimées. **NL** désigne les non-liens et **Prob.** désigne la probabilité.

Signalons qu'il y a deux raisons pour lesquelles nous avons créé les données (x,y) utilisées dans les analyses. Premièrement, nous voulions exercer sur les données de régression un contrôle suffisant pour être en mesure de déterminer quel était l'effet de l'erreur de couplage. Il s'agissait d'un aspect important dans le cas des très vastes simulations de Monte Carlo dont il est fait état dans Winkler et Scheuren (1991). Deuxièmement, nous n'avions pas à notre disposition une paire de bases de données pour laquelle une information de couplage de grande précision était disponible et qui contenait des données quantitatives convenables.

Dans le cadre de nos simulations, dont certaines sont présentées ici, nous avons créé plus de 900 bases de données, correspondant à un grand nombre de variantes des trois scénarios de couplage de base. Chaque base de données contenait trois paires de variables (x,y) correspondant aux trois scénarios de régression de base. Nous avons effectué une analyse de ces bases de données pour examiner la sensibilité des régressions au couplage et les redressements connexes de la méthode d'échantillonnage. Les différentes bases de données correspondant aux diverses valeurs de départ sont appelées *échantillons différents*.

Les redressements des données de régression ont été effectués séparément pour chaque classe de poids présentée au tableau 1, en utilisant aussi bien les probabilités de couplage estimatives que les probabilités vraies. Dans le tableau 1, la classe de poids 10 vise les paires ayant un poids entre 10 et 11, tandis que la classe de poids −1 vise les paires ayant un poids entre −1 et −2. Toutes les paires

ayant un poids de 15 ou plus sont combinées dans la classe 15+, et toutes les paires ayant un poids de −9 ou moins sont combinées dans la classe −10−. Bien qu'il ait été possible, dans le cas des résultats de Rubin-Belin, d'effectuer des redressements individuels pour tenir compte des probabilités de couplage, nous avons choisi d'effectuer des redressements moyens, pour chaque classe de poids du tableau 1. (Voir Czajka et coll. 1992, pour des observations sur une décision semblable.) Notre méthode se rapproche un peu des travaux relatifs aux scores de propension (p. ex. Rosenbaum et Rubin 1983, 1985). Les techniques de mesure de la propension, bien qu'elles aient été proposées pour d'autres catégories de problèmes, pourraient aussi s'appliquer dans le présent cas.

4. POINTS SAILLANTS ET LIMITES DES RÉSULTATS DES SIMULATIONS

En raison de contraintes d'espace, nous ne présenterons que quelques résultats représentatifs des simulations effectuées. Pour plus d'information, y compris un vaste ensemble de tableaux, voir Winkler et Scheuren (1991).

Les deux mesures résultant de nos simulations auxquelles nous nous attarderons sont le biais relatif et l'erreur-type relative. Nous n'examinerons de façon détaillée que le scénario de couplage médiocre, en nous limitant au cas où R^2 se situe entre 0.40 et 0.45. Les figures 5 à 7 présentent le biais relatif pour un échantillon représentatif unique. Toutefois, un sommaire global visant les autres scénarios est présenté au tableau 2. Certaines limites des simulations sont également signalées à la fin de la présente section.

4.1 Exemple de résultats - couplage médiocre

Plutôt que d'utiliser toutes les paires, nous ne tenons compte que de celles ayant un poids de 10 ou moins. L'utilisation de ce sous-ensemble nous permet d'examiner des méthodes de redressement des données de régression pour des classes de poids comportant une proportion de non-liens vrais allant de faible à élevée. Notons que les paires éliminées (ayant un poids de 10 ou plus) correspondent seulement à des liens vrais. Les figures 5 et 6 présentent nos résultats pour les données de régression redressées et non redressées, respectivement. Les résultats obtenus avec les données non redressées sont basés sur des formules de régression classiques (p. ex. Draper et Smith 1981). Les classes de poids présentées sont cumulatives à partir des paires ayant le poids le plus élevé. La classe de poids w correspond à toutes les paires ayant un poids entre w et 10.

Nous observons ce qui suit:

- Le *cumul* se fait par poids d'appariement décroissants (c.-à-d. à partir des classes les plus susceptibles de contenir presque seulement des liens vrais, en allant vers les classes contenant des proportions croissantes de non-liens vrais). En particulier, pour la classe de poids $w = 8$, premier point de données indiqué aux figures 5 à 7, il y avait 3 non-liens et 439 liens. Une fois les données cumulées, disons, jusqu'à la classe $w = 5$, il y avait 24 non-liens; le nombre de liens, toutefois, avait grimpé à 1,121 - ce qui nous donnait une taille d'échantillon globale beaucoup plus grande, avec une réduction correspondante de l'erreur-type de régression.
- Les *biais* relatifs sont fournis pour le coefficient de pente original et redressé $\hat{a}_1$ sous forme du rapport entre le coefficient vrai (environ 2) et le coefficient calculé pour chaque classe de poids cumulative.
- Les résultats de régression redressés se fondent aussi bien sur les probabilités de concordance estimées que sur les probabilités vraies. En particulier, la figure 5 correspond aux résultats obtenus d'après les probabilités estimées (qui seraient normalement les seules disponibles en pratique). La figure 7 correspond à la situation irréaliste dans laquelle nous connaissons les probabilités vraies.
- Les *erreurs quadratiques moyennes* relatives (non présentées) sont obtenues par le calcul de l'EQM de chaque classe de poids cumulative. Pour chaque classe, le biais est élevé au carré, ajouté au carré des erreurs-types, puis la racine carrée est prise.

Les observations faites se déduisent assez directement de nos résultats et sont à peu près celles que nous avions prévues. Par exemple, à mesure que la taille d'échantillon augmentait, nous avons constaté que les erreurs quadratiques moyennes relatives diminuaient passablement pour les coefficients redressés. Si les coefficients de régression n'étaient pas redressés, les erreurs-types diminuaient quand même avec l'augmentation de la taille d'échantillon, mais au prix d'une augmentation inacceptable du biais.

Un sujet d'inquiétude a trait à notre capacité d'estimer avec précision les probabilités de concordance, car elles ont un effet crucial sur l'exactitude des estimations des coefficients. Si nous pouvons estimer précisément les probabilités (comme c'est le cas ici), la méthode de redressement fonctionne raisonnablement bien; si nous ne pouvons pas le faire (voir plus loin), la méthode de redressement pourrait donner de piètres résultats.

4.2 Sommaire général des résultats

Nos résultats ont varié quelque peu pour les trois différentes valeurs de R^2, se révélant meilleurs pour les R^2 les plus élevés. Ces différences liées à R^2, toutefois, ne modifient pas nos principales conclusions; le tableau 2, par conséquent, n'en fait pas état. Notons que dans le cas du scénario de couplage efficace, l'effort de redressement est peu bénéfique et peut même être légèrement préjudiciable. Il est en tout cas inutile, et nous ne l'avons inclus dans nos simulations que par souci d'exhaustivité. À l'autre extrême, même pour un couplage pauvre, nous avons obtenu des résultats satisfaisants, mais seulement lorsque les probabilités vraies étaient utilisées - ce qui est impossible en pratique.

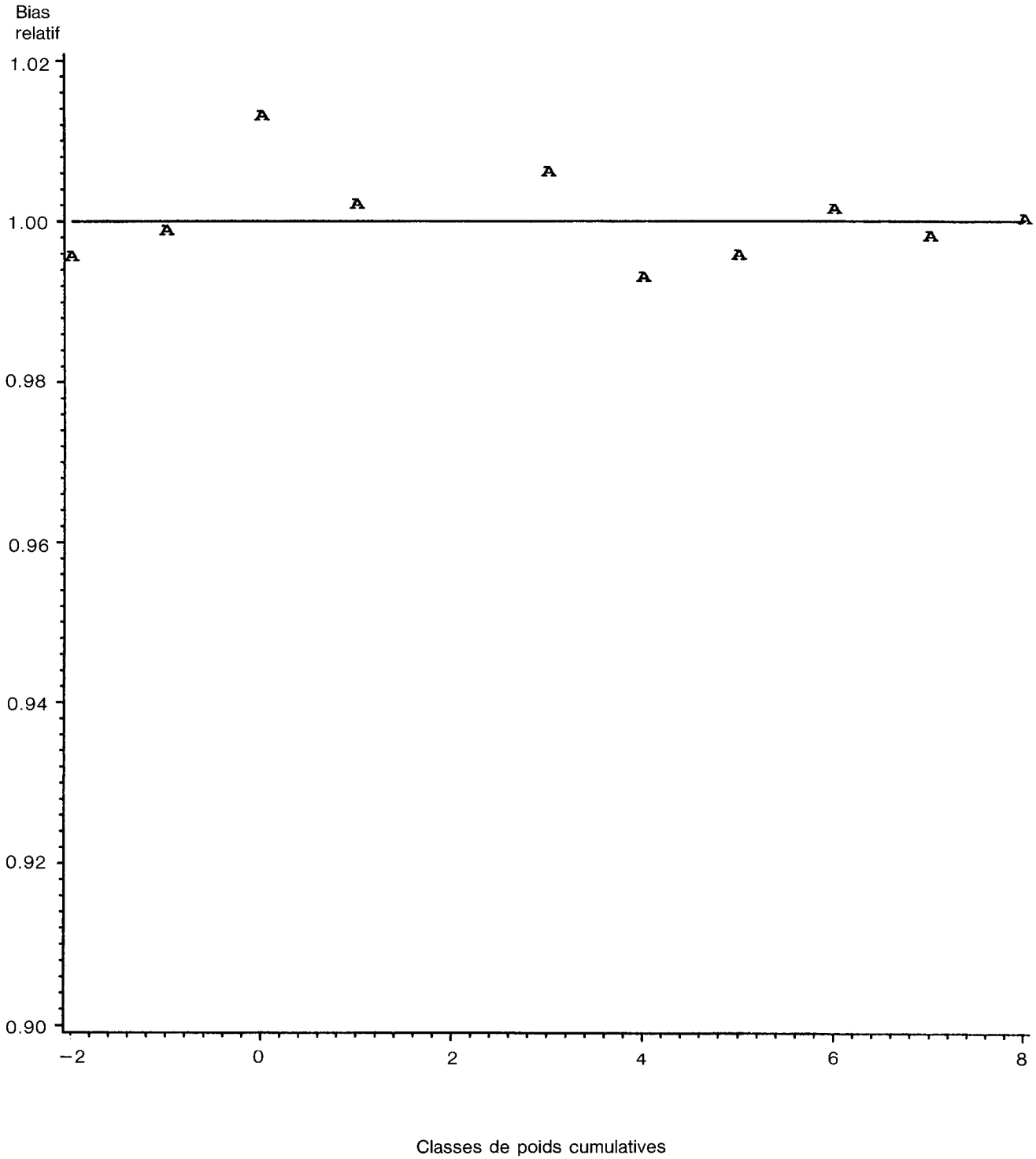

Figure 5. Biais relatif pour les estimateurs redressés probabilités estimées

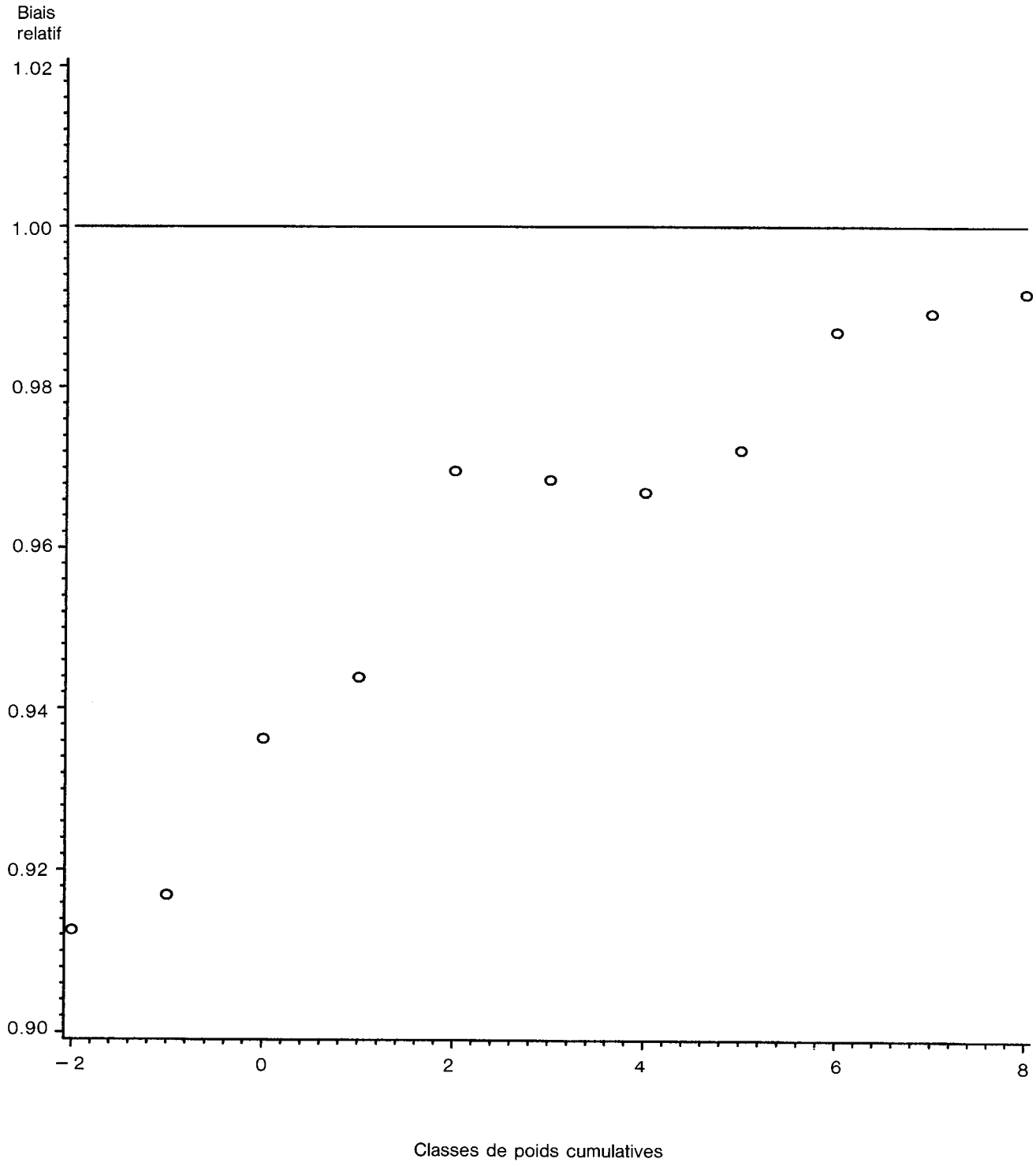

Figure 6. Biais relatif pour les estimateurs non redressés

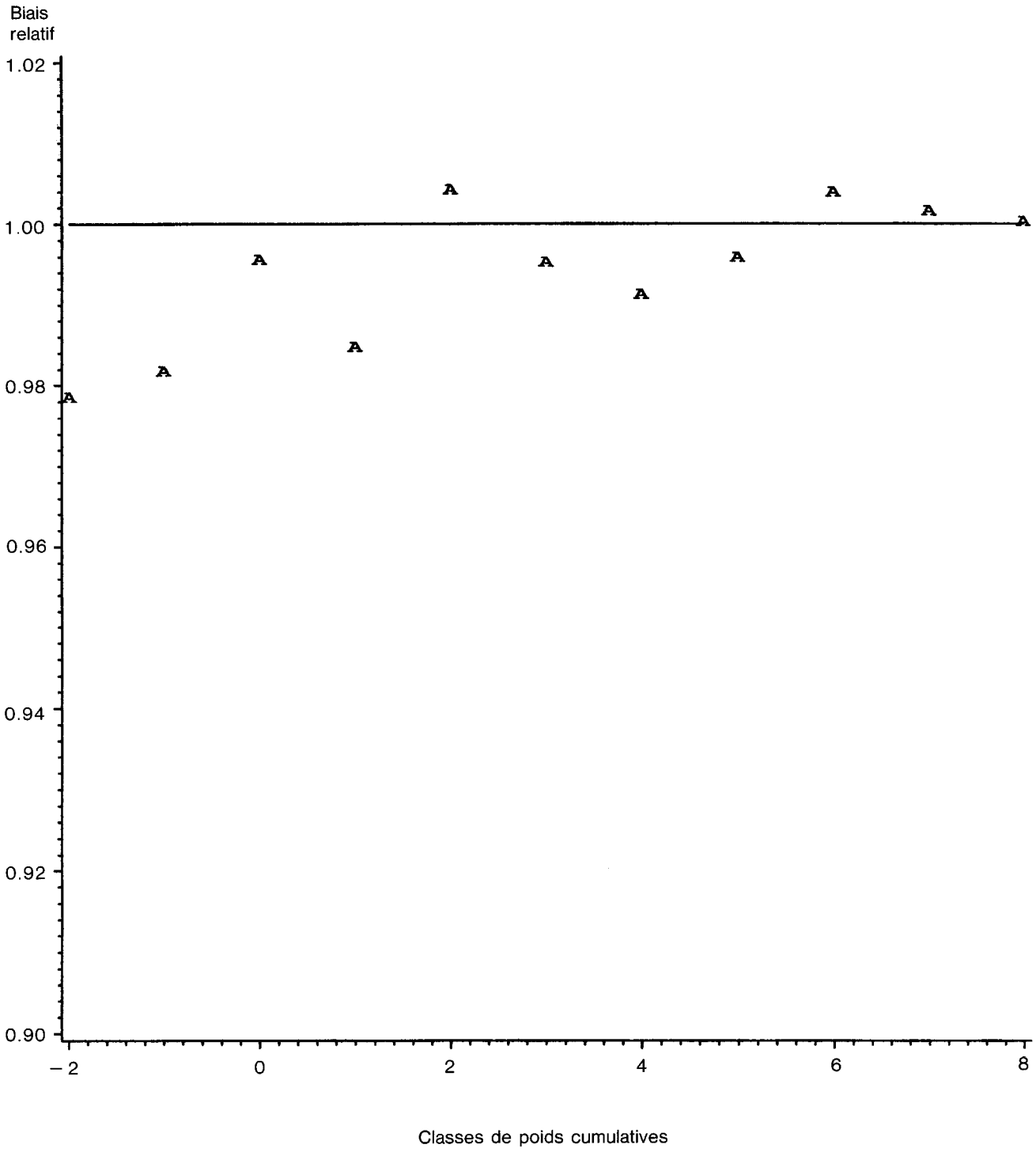

Figure 7. Biais relatif pour les estimateurs redressés probabilités vraies

Tableau 2

Sommaire des résultats des redressements pour des simulations représentatives

Base des redressements	Scénario de couplage		
	Efficace	Médiocre	Pauvre
Probabilités vraies	Le redressement n'a pas été utile parce que non nécessaire	Bons résultats comme ceux de la section 4.1	Bons résultats comme ceux de la section 4.1
Probabilités estimées	Même commentaire que ci-dessus	Même commentaire que ci-dessus	Piètres résultats, car la méthode de Rubin-Belin ne permettait pas l'estimation des probabilités

Il sera difficile, quelle que soit la méthode d'estimation statistique, d'obtenir de bons résultats avec le scénario de couplage pauvre, en raison de l'énorme chevauchement des courbes. Voir la figure 4. Nous croyons que le scénario médiocre couvre une vaste gamme de cadres généralement rencontrés. Le scénario de couplage pauvre, toutefois, pourrait également survenir assez souvent, en particulier si les responsables du couplage sont moins expérimentés. Il faudra ou bien que de nouvelles méthodes d'estimation soient élaborées pour le scénario pauvre, ou bien que la méthode d'estimation des probabilités de Rubin-Belin – qui n'a pas été conçue pour cette situation – soit améliorée.

4.3 Limites des simulations

Les résultats des simulations souffrent d'un certain nombre de limites. Quelques-unes pourront avoir des répercussions pratiques importantes, d'autres moins. En voici une liste partielle:

- Dans l'exécution des simulations relatives au présent article, nous avons supposé que la paire ayant le poids le plus élevé était un lien vrai et que la paire dont le poids venait au deuxième rang était un non-lien vrai. Cette hypothèse n'est pas valable, car c'est parfois le poids venant au deuxième rang qui correspond à un lien vrai, et le poids le plus élevé qui correspond à un non-lien vrai. (Nous ne savons pas trop quelle pourrait être l'importance de ce fait en pratique. Il s'agirait sûrement d'un facteur important dans le cas des scénarios de couplage pauvres.)
- Une deuxième limite des ensembles de données ayant servi aux simulations vient de la possibilité que l'enregistrement réellement lié ne figure pas du tout dans le fichier auquel le premier fichier est couplé. (Ce pourrait être un facteur important. Dans beaucoup de situations pratiques, il serait normal de s'attendre aussi à ce que les "critères de groupage logique" fassent en sorte que les deux paires servant aux redressements soient de faux liens.)
- Une troisième limite de notre approche est qu'aucune utilisation n'a été faite des outils diagnostiques classiques de la régression. (Selon l'environnement examiné, les valeurs aberrantes créées par la présence de non-liens pourraient bouleverser les relations sous-jacentes. Cela n'a pas représenté un gros problème dans nos simulations, en grande partie, peut-être, parce que les valeurs X et Y créées étaient confinées à un intervalle modérément étroit.)

5. CONCLUSIONS ET TRAVAUX FUTURS

Les résultats théoriques et les données de simulation connexes présentés ici sont évidemment quelque peu fabriqués et artificiels. Il reste donc beaucoup à faire pour confirmer et généraliser nos efforts initiaux. On peut toutefois en tirer certaines recommandations valables pour les applications actuelles, ainsi que des suggestions de recherche future. Nous allons d'abord examiner quelques aspects qui ont attiré notre attention et qui mériteraient qu'on les étudie plus à fond afin d'améliorer les redressements touchant les liens possibles. Ensuite, nous formulerons certaines remarques sur le problème connexe du traitement à réserver aux non-liens (restants). Enfin, nous énoncerons certaines idées sommaires et nous rappellerons notre perspective quant à l'unité des tâches accomplies par les responsables du couplage et les analystes.

5.1 Améliorations de la méthode de redressement

Une question évidente consiste à déterminer si nos méthodes de redressement pourraient s'inspirer des méthodes générales concernant les erreurs sur les variables (p. ex. Johnston 1972). Nous n'avons pas exploré cet aspect, mais l'effort pourrait en valoir la peine.

Les techniques qui émanent des outils diagnostiques classiques de la régression sont plus intéressantes pour nous. Une combinaison de ces dernières avec notre approche pourrait se révéler très intéressante. Rappelons que nous effectuons des redressements s'appliquant tour à tour à chacune des classes de poids. Supposons que nous devancions l'examen du diagramme de dispersion résiduel d'une classe de poids particulière, dans laquelle les résidus ont été calculés autour de la régression obtenue d'après les classes de poids cumulatives situées au-dessus de la classe en question. Des valeurs aberrantes pourraient alors être identifiées et traitées comme des non-liens plutôt que comme des liens possibles.

Nous nous proposons d'explorer cette possibilité à l'aide de données plus concentrées aux extrémités que celles que nous avons utilisées ici. En outre, nous examinerons les résultats obtenus en faisant varier la longueur des classes de poids et le nombre minimum de cas dans chaque classe. Nous craignons en effet que le nombre de cas de chaque classe ait été trop faible dans certains cas. (Voir le tableau 1.) Par contre, nous n'avons pas exploité

le fait que les classes de poids étaient de longueur égale et nous n'avons pas non plus étudié ce qui serait survenu si elles avaient été de longueurs différentes.

Dernière observation: comme nous l'avons déjà indiqué, nous croyons que notre approche a beaucoup en commun avec la méthode des scores de propension; toutefois, nous n'avons pas fait appel explicitement à cette théorie plus générale, mais nous croyons qu'il vaudrait la peine de le faire. Par exemple, les notions propres à cette théorie pourraient être particulièrement utiles dans le cas où les variables de régression et les variables de couplage sont dépendantes. (Voir Winkler et Scheuren (1991) pour un rapport sur les simulations limitées qui ont été effectuées et les difficultés additionnelles rencontrées.)

5.2 Traitement des non-liens erronés

Dans l'application des méthodes de couplage d'enregistrements, le problème général du biais de sélection survient en raison de la présence de non-liens erronés. Il existe plusieurs façons de traiter ce problème. Par exemple, les liens pourraient être redressés par l'analyste pour tenir compte du manque de représentativité, au moyen de méthodes familières aux responsables des redressements effectués au titre de la non- réponse d'unités ou, peut-être même, de la non-réponse à des questions particulières (p. ex. Scheuren et coll. 1981).

La présente méthode de traitement des liens possibles pourrait aider à réduire l'ampleur du problème des non-liens erronés, mais, en général, ne l'éliminerait pas. Plus précisément, supposons que nous ayons un cadre de couplage dans lequel, en raison d'un manque de ressources, il est impossible de faire le suivi des liens possibles. De nombreux praticiens pourraient simplement laisser de côté les liens possibles, accroissant du même coup le nombre de non-liens erronés. (Par exemple, dans la détermination des membres d'une cohorte qui sont vivants ou décédés, une troisième possibilité - indéterminé - est souvent utilisée.)

Notre approche à l'égard des liens possibles aurait *implicitement* produit un redressement pour la partie des non-liens erronés qui pouvaient éventuellement donner un lien (avec, disons, une étape de suivi). Les autres non-liens erronés demeureraient généralement irrésolus et la possibilité d'effectuer un autre redressement à leur sujet pourrait être une question à envisager.

Il se produit fréquemment des situations où les fichiers couplés comportent des sous-groupes qui recèlent une information de couplage de qualité variable, et qui présentent par conséquent des taux de liens et de non-liens erronés qui diffèrent. En principe, nous pourrions appliquer les techniques décrites ici à chacun de ces sous-groupes séparément. La façon de procéder face à des sous-groupes de très petite taille est un problème toujours sans réponse, et l'effet sur les différences estimées d'un sous-groupe à l'autre, même quand les deux sont de taille modeste, bien qu'il puisse sembler évident, mérite d'être étudié.

5.3 Derniers commentaires

Au début du présent article, nous avons posé deux questions "clés". Maintenant que nous arrivons à la fin, il est utile de revoir ces questions et de tenter, de façon sommaire, d'y répondre.

- *"Que peut faire le responsable du couplage pour aider l'analyste?"* Si possible, le responsable du couplage devrait jouer un rôle dans la conception des ensembles de données à coupler, de façon que l'information d'identification des deux ensembles soit de haute qualité. Plusieurs organismes disposent déjà d'algorithmes puissants, capables d'effectuer un excellent couplage (c'est le cas à Statistique Canada et au U.S. Bureau of the Census, pour n'en nommer que deux). Les responsables du couplage devraient résister à la tentation de concevoir et d'élaborer leurs propres logiciels. Il est hautement recommandé, dans la plupart des cas, de modifier ou simplement d'utiliser les logiciels existants (Scheuren 1985). De toute évidence, dans l'intérêt de l'analyste, le responsable du couplage doit fournir autant d'information que possible au sujet des fichiers couplés, afin que l'analyste puisse faire des choix éclairés dans son travail. Dans le présent article, nous avons proposé que les liens, les non-liens *et* les liens possibles soient fournis à l'analyste - et non pas uniquement les liens. Nous recommandons fortement qu'il en soit ainsi, même si une étape de révision manuelle a été réalisée. Nous ne recommandons *pas* nécessairement les choix particuliers que nous avons faits au sujet de la structure des fichiers, du moins tant qu'ils n'auront pas été étudiés davantage. Nous faisons valoir, toutefois, que nos choix présentent une certaine utilité.
- *"Que devrait savoir l'analyste au sujet du couplage et comment cette information devrait-elle être utilisée?"* L'analyste doit disposer de l'information sur les liens, les non-liens *et* les liens possibles, ainsi que les probabilités de concordance, si ces données sont disponibles. Il se peut que dans de nombreuses situations, la simple exécution des étapes d'analyse des données séparément pour chaque état de concordance permette d'en apprendre beaucoup sur la sensibilité des résultats. Le présent article énonce des idées initiales sur la façon dont cela peut se faire dans un contexte de régression. Il semble également que des améliorations soient possibles grâce aux redressements effectués ici, notamment dans le scénario de couplage médiocre. La mesure dans laquelle ces améliorations sont générales reste à vérifier. Malgré tout, nous sommes assez satisfaits de nos résultats et nous sommes impatients de pousser le sujet plus à fond. En effet, il y a des liens directs à établir entre notre approche du problème de la régression et d'autres techniques courantes, comme les modèles log-linéaires à tableaux de contingence.

De toute évidence, nous n'avons pas proposé de réponses complètes et générales aux questions soulevées. Nous espérons, toutefois, que cet article pourra au moins stimuler l'intérêt et susciter d'autres contributions qui nous aideront à améliorer nos méthodes.

REMERCIEMENTS ET RESPONSABILITÉS

Les auteurs voudraient remercier Yahia Ahmed et Mary Batcher pour leur contribution à la préparation de cet article, ainsi que deux arbitres pour leurs commentaires détaillés et pertinents. Des échanges fructueux ont eu lieu avec Tom Belin. Wendy Alvey nous a également fourni une aide importante sur le plan de la rédaction.

Selon la coutume, il importe de préciser que cet article reflète les opinions des auteurs et pas nécessairement celles des organismes dont ils font partie. Tout problème, par exemple un manque de clarté dans les idées ou les sujets présentés, est entièrement attribuable aux auteurs.

ANNEXE

L'annexe est divisée en quatre sections. La première explique de façon détaillée comment l'erreur de couplage se répercute sur les modèles de régression dans le cas unidimensionnel simple. L'approche est très voisine de celle présentée par Neter et coll. (1965) et sert d'inspiration aux généralisations présentées à la deuxième et à la troisième section de l'annexe. Les formules de calcul sont beaucoup plus complexes que celles présentées par Neter et coll. puisque nous utilisons un modèle plus réaliste du processus de couplage. Dans la deuxième section, nous étendons le modèle unidimensionnel au cas où toutes les variables indépendantes proviennent d'un fichier, tandis que la variable dépendante vient de l'autre, et dans la troisième section, nous étendons le deuxième cas à celui où certaines variables indépendantes viennent d'un fichier et certaines viennent d'un autre. La quatrième section résume les méthodes de Rubin et Belin (1991) (voir aussi Belin 1991) permettant l'estimation de la probabilité d'un lien.

A.1. Modèle de régression unidimensionnel

Dans la présente section, nous examinons la situation de régression la plus simple, dans laquelle nous couplons deux fichiers et examinons un ensemble de paires numériques dans lequel la variable indépendante est tirée d'un enregistrement d'un fichier et la variable dépendante, de l'enregistrement correspondant de l'autre fichier.

Soit $Y = X\beta + \epsilon$ le modèle de régression unidimensionnel ordinaire pour lequel les termes d'erreur sont indépendants avec espérance zéro et variance constante σ^2. Si nous examinions une seule base de données, la régression serait faite de Y en X de la façon habituelle. Pour $i = 1, \ldots, n$, nous souhaitons utiliser (X_i, Y_i), mais nous utilisons plutôt (X_i, Z_i), où Z_i est habituellement Y_i, mais pourrait prendre une autre valeur, Y_j, en raison de l'erreur de couplage.

Autrement dit, pour $i = 1, \ldots, n$,

$$z_i = \begin{cases} Y_i & \text{avec probabilité } p_i \\ Y_j & \text{avec probabilité } q_{ij} \text{ pour } j \neq i, \end{cases}$$

où $p_i + \sum_{j \neq i} q_{ij} = 1$.

La probabilité p_i peut être zéro ou un. Nous définissons $h_i = 1 - p_i$. Comme dans Neter et coll. (1965), nous divisons l'ensemble de paires en n classes mutuellement exclusives. Chaque classe est formée d'exactement un (X_i, Z_i), de sorte qu'il y a n classes. L'idée intuitive de notre méthode est que nous redressons Z_i dans chaque (X_i, Z_i) pour tenir compte du biais introduit par le processus de couplage. L'exactitude du redressement est fortement liée à l'exactitude des estimations des probabilités de concordance dans notre modèle.

Pour simplifier les formules de calcul, nous supposons un couplage biunivoque, c.-à-d. que pour chaque $i = 1, \ldots, n$, il existe au plus un j tel que $q_{ij} > 0$. Nous définissons ϕ par $\phi(i) = j$. Notre modèle demeure valable même si nous ne faisons pas l'hypothèse d'un couplage biunivoque.

À titre d'étapes intermédiaires dans l'estimation des coefficients de régression et de leurs erreurs-types, nous devons trouver $\mu_z \equiv E(Z)$, σ_z^2 et σ_{zx}. Comme dans Neter et coll. (1965),

$$\begin{aligned} E(Z) &\equiv (1/n)\sum_i E(Z|i) \equiv (1/n)\sum_i (Y_i p_i + \sum_{j \equiv i} Y_j q_{ij}) \\ &= (1/n)\sum_i Y_i \\ &\quad + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &\equiv \bar{Y} + B. \end{aligned} \tag{A.1.1}$$

La première et la deuxième égalité découlent des définitions et la troisième s'obtient par addition et soustraction. C'est à la troisième inégalité que nous appliquons pour la première fois l'hypothèse d'un couplage biunivoque. Le dernier terme du côté droit de l'égalité est le biais, que nous désignons par B. Notons que le biais global B est la moyenne statistique (espérance) des biais individuels $[Y_i(-h_i) + Y_{\phi(i)} h_i]$ pour $i = 1, \ldots, n$. De même, nous avons

$$\begin{aligned} \sigma_z^2 &\equiv E(Z - EZ)^2 = E(Z - (\bar{Y} + B))^2 \\ &= (1/n)\sum_i (Y_i - \bar{Y})^2 p_i + (1/n)\sum_{j \neq i} \\ &\quad (Y_j - \bar{Y})^2 q_{ij} - 2B\,E(Z - \bar{Y}) + B^2 \\ &= (1/n)S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned} \tag{A.1.2}$$

où $B_{yy} = (1/n)\sum_i [(Y_i - \bar{Y})^2(-h_i) + (Y_{\phi(i)} - \bar{Y})^2 h_i]$, $S_{yy} = \sum_i (Y_i - \bar{Y})^2$ et $\sigma_y^2 = (1/n)S_{yy}$.

$$\sigma_{zx} \equiv E[(Z - EZ)(X - EX)]$$

$$= (1/n) \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) p_i$$

$$+ (1/n) \sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X}) q_{ij}$$

$$= (1/n) S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \quad \text{(A.1.3)}$$

où $B_{yx} = (1/n) \sum_i [(Y_i - \bar{Y})(X_i - \bar{X})(-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X}) h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$ et $\sigma_{yx} = (1/n) S_{yx}$. Le terme B_{yy} est le biais pour les deuxièmes moments et le terme B_{yx} est le biais pour le produit croisé de Y et de X. Les formules (A.1.1), (A.1.2) et (A.1.3) correspondent respectivement aux formules (A.1), (A.2) et (A.3) de Neter et coll. Les formules diffèrent nécessairement dans leurs détails, parce que nous utilisons un modèle plus général du processus de couplage.

Pour les coefficients de régression, on a

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \quad \text{(A.1.4)}$$

Pour obtenir une estimation de la variance de β_{yx}, nous estimons d'abord s^2 pour obtenir la variance σ^2 de la façon habituelle.

$$(n - 2) s^2 = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx} S_{xy}$$

$$= n \sigma_y^2 - n \beta_{yx} \sigma_x^2. \quad \text{(A.1.5)}$$

L'utilisation de (A.1.2) et de (A.1.3) nous permet d'exprimer s^2 en termes des quantités observables σ_z^2 et σ_{zx} et des termes de biais B_{yy}, B_{yx} et B qui peuvent être calculés en vertu de nos hypothèses. La variance estimée de β_{yx} est alors calculée au moyen de la formule habituelle (p. ex. Draper et Smith 1981, p. 18-20)

$$\text{Var}(\beta_{yx}) = s^2/(n \sigma_x^2).$$

Nous observons que la première égalité dans (A.1.5) fait intervenir l'hypothèse habituelle des régressions selon laquelle les termes d'erreur sont indépendants et ont une variance identique.

Dans les exemples numériques du présent article, nous avons supposé que la valeur indépendante vraie X_i associée à chaque Y_i provenait de l'enregistrement ayant le poids d'appariement le plus élevé et que la valeur indépendante fausse provenait de l'enregistrement ayant le poids d'appariement venant au deuxième rang. Cette hypothèse est plausible, car nous n'avons examiné que la régression simple dans le présent article, et parce que le poids d'appariement venant au deuxième rang était généralement beaucoup plus faible que le poids le plus élevé. Par conséquent, il est beaucoup plus naturel de supposer que l'enregistrement dont le poids vient au deuxième rang donne la valeur fausse. Dans nos exemples empiriques, nous effectuons des redressements directs et posons des hypothèses simples qui donnent de bons résultats, car ils sont conformes à la nature des données et au processus de couplage. Dans des situations de régression plus complexes, ou avec d'autres modèles comme les modèles log-linéaires, il nous faudra vraisemblablement poser d'autres hypothèses. Ce besoin d'hypothèses additionnelles s'apparente au cas des modèles simples de non-réponse, qui exigent des hypothèses additionnelles lorsqu'on passe de la non-réponse dont on n'a pas à tenir compte à la non-réponse dont il faut tenir compte (voir Rubin 1987).

Dans la présente section, nous avons choisi de redresser les valeurs x indépendantes et de garder fixes les valeurs y dépendantes, afin de préserver l'uniformité avec le raisonnement de Neter et coll. Nous aurions pu tout aussi facilement redresser les valeurs y dépendantes et garder fixes les valeurs x.

A.2. Régression multiple avec variables indépendante venant d'un fichier et variables dépendantes venant de l'autre fichier

À ce stade, nous passons à la notation matricielle courante (p. ex. Graybill 1976). Notre modèle de base est

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \epsilon,$$

où $\boldsymbol{Y}$ est une matrice $n \times 1$, $\boldsymbol{X}$ est une matrice $n \times p$, $\boldsymbol{\beta}$ est une matrice $p \times 1$ et ϵ est une matrice $n \times 1$. En suivant le même raisonnement qu'en (A.1.1), nous pouvons écrire

$$\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{B}, \quad \text{(A.2.1)}$$

où $\boldsymbol{Z}$, $\boldsymbol{Y}$ et $\boldsymbol{B}$ sont des matrices $n \times 1$ ayant des termes liés entre eux, pour $i = 1, \ldots, n$, par

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Parce que nous observons $\boldsymbol{Z}$ et $\boldsymbol{X}$ seulement, nous considérons l'équation

$$\boldsymbol{Z} = \boldsymbol{X C} + \epsilon. \quad \text{(A.2.2)}$$

Nous obtenons une estimation $\hat{\boldsymbol{C}}$ en effectuant la régression par rapport aux données observées de la façon habituelle. Nous voulons redresser l'estimation $\hat{\boldsymbol{C}}$ pour obtenir une estimation $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ d'une manière analogue à (A.1.1).

En utilisant (A.2.1) et (A.2.2), nous obtenons

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{XB} = \hat{\boldsymbol{C}}. \quad \text{(A.2.3)}$$

Le premier terme du côté gauche de (A.2.3) est l'estimation habituelle $\hat{\boldsymbol{\beta}}$. Le deuxième terme du côté gauche de (A.2.3) est notre redressement tenant compte du biais. $\boldsymbol{X}^T$ est la matrice transposée de $\boldsymbol{X}$.

La formule habituelle (Graybill 1976, p. 176) permet l'estimation de la variance σ^2 associée aux composantes d'erreur indépendantes et identiquement distribuées de ϵ

$$(n - p)\,\hat{\sigma}^2 = (Y - X\beta)^T(Y - X\beta)$$

$$= Y^TY - \hat{\beta}\,X^TY, \qquad \text{(A.2.4)}$$

où $\hat{\beta} = (X^TX)^{-1}\,X^TY$.

En utilisant (A.2.1), on peut représenter $\hat{\beta}\,X^TY$ en termes des variables observables Z et X, de la même façon qu'en (A.1.2) et en (A.1.3). Puisque

$$Y^TY = Z^TZ - B^TZ - Z^TB + B^TB, \qquad \text{(A.2.5)}$$

nous pouvons obtenir la portion restante du côté droit de (A.2.4) qui permet l'estimation de σ^2.

Selon la formule habituelle (p. ex. Graybill 1976, p. 276), la covariance de $\hat{\beta}$ est

$$\text{cov}[\hat{\beta}] = \sigma^2\,(X^TX)^{-1}, \qquad \text{(A.2.6)}$$

ce que nous pouvons estimer.

A.3. Régression multiple avec variables indépendantes venant des deux fichiers

Lorsque certaines des variables indépendantes viennent du même fichier que Y, nous devons les redresser d'une manière analogue à celle employée pour redresser Y dans les équations (A.1.1) et (A.2.1). La matrice X peut alors s'écrire sous la forme

$$X_d = X + D, \qquad \text{(A.3.1)}$$

où D est la matrice des redressements tenant compte du biais qui redonne aux termes de X venant du même fichier que Y leurs valeurs vraies qui sont représentées dans X_d. En utilisant (A.2.1) et (A.2.2), nous obtenons

$$Y + B = (X_d - D)C. \qquad \text{(A.3.2)}$$

Après quelques opérations algébriques, (A.3.2) devient

$$(X_d^TX_d)^{-1}X_d^TY = (X_d^TX_d)^{-1}X_d^T(-B)$$

$$+ (X_d^TX_d)^{-1}X_d^T(X_d + D)C$$

$$= (X_d^TX_d)^{-1}X_d^T(-B)$$

$$+ (X_d^TX_d)^{-1}X_d^TDC + C. \qquad \text{(A.3.3)}$$

Si D est zéro (c.-à-d. que toutes les valeurs x indépendantes proviennent d'un même fichier), (A.3.3) correspond à (A.2.3). Le premier terme du côté gauche de (A.2.3) est l'estimation de $\hat{\beta}$. L'estimation $\hat{\sigma}^2$ est obtenue d'une manière analogue à la façon dont (A.2.3), (A.2.4) et (A.2.5) ont été utilisés. La covariance de $\hat{\beta}$ découle de (A.2.6).

A.4. Modèle de Rubin-Belin

Pour estimer la probabilité d'un lien vrai dans n'importe quel intervalle de poids, Rubin et Belin (1991) considèrent l'ensemble de paires produites par le programme de couplage par ordinateur, rangées par ordre décroissant de poids. Ils supposent que la probabilité d'un lien vrai est une fonction monotone du poids; autrement dit, plus le poids est élevé, plus grande est la probabilité d'un lien vrai. Ils supposent que la distribution des poids observés est une combinaison des distributions des liens vrais et des non-liens vrais.

Leur méthode d'estimation est la suivante:

1. Modéliser chacune des deux composantes de la combinaison comme une distribution normale avec moyenne et variance inconnues après des transformations puissance distinctes.
2. Estimer la puissance des deux transformations à l'aide d'un échantillon d'apprentissage.
3. À l'aide des deux transformations ainsi déterminées, ajuster un modèle de combinaison normale aux données courantes sur les poids, pour obtenir des estimations du maximum de vraisemblance (et des erreurs-types).
4. Utiliser les paramètres du modèle ajusté pour obtenir des estimations ponctuelles du taux de liens faux en fonction du niveau seuil et obtenir les erreurs-types visant le taux de liens faux au moyen d'une approximation par la méthode delta.

Bien que la méthode de Rubin-Belin exige un échantillon d'apprentissage, celui-ci sert surtout à obtenir la forme des courbes. Autrement dit, si la transformation puissance est donnée par

$$\psi(w_i; \delta, \omega) = \begin{cases} (w_i^\delta - 1)/(\delta\,\omega^{\delta-1}) & \text{si} \quad \delta \neq 0 \\ \omega \log(w_i) & \text{si} \quad \delta = 0, \end{cases}$$

où ω est la moyenne géométrique des poids w_i, $i = 1, \ldots, n$, alors ω et δ peuvent être estimés pour les deux courbes. Dans le cas des exemples du présent article et d'une vaste gamme d'autres situations de couplage (Winkler et Thibaudeau 1991), la méthode d'estimation de Rubin-Belin donne de bons résultats. Dans certaines autres situations, une méthode différente (Winkler 1992) utilisant plus d'information que la méthode de Rubin-Belin et n'exigeant pas d'échantillon d'apprentissage donne des estimations exactes, tandis que le logiciel (voir p. ex. Belin 1991) basé sur la méthode de Rubin-Belin ne donne pas d'estimation convergente, même si de nouvelles données d'étalonnage sont obtenues. Puisque les données d'étalonnage visant les scénarios efficace et médiocre du présent article sont appropriées, la méthode de Rubin-Belin donne de meilleures estimations que la méthode de Winkler.

BIBLIOGRAPHIE

BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? Dans *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service.

BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.

CARPENTER, M., et FAIR, M.E. (Éditeurs) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, Statistique Canada.

COOMBS, J.W., et SINGH, M.P. (Éditeurs) (1987). *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada.

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

CZJAKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregrates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.

DRAPER, N.R., et SMITH, H. (1981). *Applied Regression Analysis*, 2ième édition. New York: J. Wiley.

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

GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

HOWE, G., et SPASOFF, R.A. (Éditeurs) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.

JABINE, T.B., et SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHNSTON, J. (1972). *Econometric Methods*, 2ième Édition. New York: McGraw-Hill.

KILSS, B., et ALVEY, W. (Éditeurs) (1985). *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service, Publication 1299, 2-86.

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

NEWCOMBE, H.B., FAIR, M.E., et LALONDE, P. (1992). The use of names for linking personal records. *Journal of the Americal Statistical Association*, 87, 1193-1208.

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

NETER, J., MAYNES, E.S., et RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

ROSENBAUM, P., et RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P., et RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.

RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the Bureau of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.

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

SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques* - 1985. U.S. Internal Revenue Service.

SCHEUREN, F., et OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.

SCHEUREN, F., OH, H.L., VOGEL, L., et YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfield identification, and information theory. Dans *Record Linkage Techniques* - 1985, (Éds. B. Kilss et W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.

WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

WINKLER, W.E., et SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.

WINKLER, W.E., et THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical report.