Techniques d'enquête, juin 1993
Vol. 19, n° 1, pp. 67–89
Statistique Canada

Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle

A.C. SINGH, H.J. MANTEL, M.D. KINACK et G. ROWE[1]

RÉSUMÉ

Lorsqu'on crée des bases de données de microsimulation, souvent utilisées dans la planification et l'analyse des politiques, on combine plusieurs fichiers de données par des techniques d'appariement statistique afin d'enrichir le fichier receveur. Or, pour effectuer cette opération, il faut poser l'hypothèse de l'indépendance conditionnelle (HIC), ce qui peut fausser sérieusement les relations conjointes entre les variables. On peut éviter de poser cette hypothèse en utilisant des informations supplémentaires appropriées. Dans cet article, nous examinons des méthodes d'appariement statistique qui correspondent à trois méthodes d'imputation – par régression, hot-deck et log-linéaire – appliquées suivant deux scénarios: avec et sans information supplémentaire. La méthode d'imputation log-linéaire consiste essentiellement à introduire des contraintes nominales dans la méthode par régression ou la méthode hot-deck. À partir d'une vaste étude de simulation faite avec des données fictives, nous exécutons des analyses de sensibilité lorsque l'on s'éloigne de l'HIC et nous étudions les gains qui peuvent découler de l'utilisation d'informations supplémentaires. À l'aide de données fictives, nous créons différents scénarios relatifs à la distribution et aux relations des variables pertinentes, par exemple distribution symétrique *vs.* distribution asymétrique et données supplémentaires substitutives *vs.* données supplémentaires non substitutives. Nous faisons aussi quelques recommandations sur l'utilisation des méthodes d'appariement statistique. Notre étude confirme particulièrement que l'HIC peut représenter une contrainte sérieuse, que l'on peut éliminer en utilisant des informations supplémentaires appropriées. L'étude montre aussi que les méthodes hot-deck sont généralement préférables aux méthodes de régression. De plus, lorsqu'on dispose d'informations supplémentaires, les contraintes nominales log-linéaires peuvent accroître l'efficacité des méthodes hot-deck. L'idée de cette étude est née des préoccupations que l'on avait sur l'utilisation de l'HIC dans la construction de la Base de données de simulation des politiques sociales à Statistique Canada.

MOTS CLÉS: Contraintes nominales; corrélation conditionnelle; contaminations log-normales; réduction à la moyenne.

1. INTRODUCTION

On peut envisager l'appariement statistique comme un cas particulier de l'imputation; celle-ci implique deux sources de micro-données qui contiennent des renseignements différents sur des unités différentes. L'une de ces sources sert de fichier cible, ou fichier receveur, dans lequel des données sont imputées pour chaque enregistrement à partir des données de l'autre source, qui est définie comme le fichier donneur. Cependant, ce rapport entre l'appariement statistique et l'imputation disparaît dès lors que le fichier receveur contient des renseignements sur des variables que l'on ne retrouve pas dans le fichier donneur. Le fichier apparié (enrichi) sert le plus souvent de source de paramètres pour les modèles de micro-simulation, qui exigent normalement un fichier complet, comprenant toutes les variables. Les fichiers de micro-données connus peuvent équivaloir à des échantillons tirés de fichiers administratifs ou d'enquêtes. Comme les enregistrements tirés des divers fichiers source se rapportent à des unités différentes, le processus de fusion des données des divers fichiers diffère de l'appariement exact, où l'on recherche plutôt des unités précises dans les diverses sources. En fait, même s'il était possible de faire un appariement exact des fichiers, des règles de protection du secret statistique viendraient imposer certaines contraintes à ce processus.

De façon générale, le problème se formule comme suit. Un fichier receveur A contient de l'information sur les variables (X, Y) et un fichier donneur B, de l'information sur les variables (X,Z). La variable commune aux deux fichiers, X, peut servir à reconnaître les unités identiques dans les deux fichiers. Le problème consiste à compléter les enregistrements du fichier A en imputant des valeurs authentiques (observées) pour Z à l'aide de l'information du fichier B sur la relation (X,Z). Dans la pratique, les variables X, Y et Z sont généralement multidimensionnelles. Un précieux avantage de l'imputation de valeurs authentiques de Z est de préserver les relations entre les éléments de la variable multidimensionnelle Z. Pour des raisons de commodité, nous supposerons dans cet article que X, Y et Z sont des variables unidimensionnelles.

La Base de données de simulation des politiques sociales (BDSPS; voir Wolfson *et al.* 1987), qui est une base de données de microsimulation créée à Statistique Canada,

[1] A.C. Singh, H.J. Mantel et M.D. Kinack, Division des méthodes d'enquêtes sociales; G. Rowe, Division des études sociales et économiques, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

est une application importante de l'appariement statistique en analyse des lignes de conduite économiques (par ex.: calcul des impôts et des transferts pour les familles de la base de données). Dans la construction de la BDSPS, qui se fait en plusieurs étapes, on recourt à l'appariement statistique à certains moments pour enrichir le fichier receveur – l'enquête sur les finances des consommateurs (EFC) – de données d'autres sources. Plus précisément, on ajoute dans les enregistrements de l'EFC des données provenant des demandes de prestations d'assurance-chômage, des déclarations de revenus des particuliers et de l'enquête sur les dépenses des familles. Si le fichier A correspond à l'EFC et le fichier B, au fichier de données fiscales, alors X peut représenter des variables démographiques et économiques, Y peut désigner le revenu sous forme de transferts et Z peut représenter l'impôt à payer, le revenu de placements et les déductions.

Tel qu'il est décrit ci-dessus, l'appariement statistique présente une lacune grave en ceci qu'on ne connaît absolument rien de la variable Y. Cette lacune équivaut à l'hypothèse de l'indépendance conditionnelle de Y et Z étant donné X ($Y \perp Z \mid X$), désignée par HIC (hypothèse d'indépendance conditionnelle). L'importance de cette hypothèse est indéniable puisque l'objectif de l'appariement est d'analyser les relations conjointes entre X, Y et Z. Si les relations réelles entre les variables sont telles que l'indépendance conditionnelle n'existe pas, l'HIC aura pour effet de dissimuler un aspect important de ces relations et de fausser des analyses qui font intervenir toutes les variables. Les risques de l'utilisation de l'HIC ont été soulignés par Sims (1978) et Rubin (1986), et bien que les appariements statistiques fondés sur l'HIC ne soient pas nécessairement entachés d'un biais important, Paass (1986) et Armstrong (1989) montrent, données empiriques à l'appui, que le problème est souvent réel. Si nous avons réalisé cette étude, c'est justement parce que nous cherchions à améliorer le contenu de la BDSPS, où l'appariement statistique repose sur l'HIC; voir aussi les commentaires de Scheuren (1989) sur la méthodologie utilisée dans la BDSPS.

Les premiers ouvrages sur l'appariement statistique remontent à plus de vingt ans. Notons particulièrement ceux de Budd et Radner (1969), de Budd (1971) et d'Okner (1972). Dans ses observations sur l'article d'Okner, Sims (1972) est le premier qui souligne les risques potentiels de l'appariement statistique attribuables à l'existence de l'hypothèse d'indépendance conditionnelle. Fellegi (1977) s'interroge aussi sur la validité des distributions conjointes dans le fichier enrichi et il propose de soumettre les méthodes d'appariement à des tests empiriques complets. Le Département du commerce des États-Unis (U.S. Department of Commerce 1980) fait une bonne analyse des méthodes d'appariement statistique et des méthodes d'appariement exact; voir aussi Kadane (1978) et Rodgers (1984). Barr et Turner (1990) décrivent une étude empirique détaillée qui a pour objet la qualité des opérations dans la fusion de fichiers et présentent une bibliographie intéressante. Pour une analyse plus récente, le lecteur se référera à Cohen (1991).

Dans cet article, nous envisageons l'utilisation d'information supplémentaire en remplacement de l'HIC dans l'appariement statistique. Par conséquent, nous supposons l'existence d'un troisième fichier, C, qui contient de l'information supplémentaire sur l'ensemble complet de variables (X,Y,Z) ou l'ensemble réduit (Y,Z). Il peut s'agir d'informations périmées ou substitutives (c.-à-d. variables différentes mais connexes) ou d'informations disposées en tableaux de fréquences, qui proviennent d'enquêtes spéciales ou de fichiers confidentiels. Nous voulons compléter les enregistrements du fichier A en allant chercher des valeurs de Z dans le fichier B sur la base d'informations contenues dans les fichiers A, B et C au sujet des relations conjointes entre X, Y et Z. On peut mesurer le succès de cette opération en évaluant la proportion des valeurs de Z contenues dans le fichier A qui proviennent vraisemblablement de la distribution réelle subordonnée à X et à Y. Dans une étude de simulation, nous pouvons comparer les valeurs appariées de Z aux valeurs vraies de Z qui ont été supprimées en nous servant de mesures d'évaluation au niveau individuel ou au niveau global. L'écart absolu moyen par rapport à la valeur vraie de Z et l'écart entre la covariance conditionnelle, $\mathrm{Cov}(Y,Z \mid X)$, et la valeur vraie correspondante sont des exemples de mesures d'évaluation au niveau individuel. En ce qui concerne les mesures d'évaluation au niveau global, mentionnons la distance du chi carré et les valeurs P fondées sur les tests du rapport de vraisemblance pour distributions de variables nominales. Il arrive souvent dans la pratique que le fichier A, une fois complété, serve à construire des tableaux de recoupement; dans ce cas, l'attention se portera surtout sur les mesures de niveau global fondées sur des distributions de variables nominales. En outre, pour n'importe quelle distribution arbitraire de (X,Y,Z), qui peut être fort complexe dans la pratique, la transformation en variable nominale est une méthode simple et uniforme pour résumer la distribution conjointe.

Comme nous l'avons mentionné plus haut, le problème de l'appariement statistique a une importance indéniable au point de vue pratique. En effet, pour un problème donné, on devrait choisir la méthode d'appariement en fonction du genre d'information supplémentaire dont on dispose. Le gros des méthodes proposées dans les ouvrages antérieurs est attribuable à Rubin (1986) et à Paass (1986). Le premier a proposé des méthodes de régression paramétrique tandis que le second a proposé des méthodes de régression non paramétrique. Les deux types de méthodes ont trait respectivement aux méthodes d'imputation par régression (REG) et aux méthodes d'imputation hot-deck (HOD), toutes bien connues.

La méthode de Rubin (dont une version est désignée dans cet article par le sigle REG*) consiste essentiellement à déduire tout d'abord une valeur intermédiaire, Z_{int}, du prédicteur par régression de Z par rapport à X et à Y (ce prédicteur étant lui-même obtenu à l'aide d'information sur la corrélation inconditionnelle $\rho_{Y,Z}$ ou la corrélation conditionnelle $\rho_{Y,Z|X}$ dans le fichier C), puis à déterminer une valeur Z authentique à partir du fichier B en utilisant la méthode hot-deck avec une distance euclidienne (X,Z);

se reporter à la section 3 pour plus de détails. Si la forme de la fonction du prédicteur par régression est connue, la méthode d'appariement statistique REG* sera vraisemblablement facile à appliquer. Or, il n'est pas facile en règle générale de trouver un prédicteur adéquat pour Z, surtout si Z est une variable multidimensionnelle. De plus, si l'information du fichier C est sous la forme d'une distribution de variables nominales, ce qui peut être souvent le cas en pratique, la méthode REG* ne sera pas applicable.

Par ailleurs, la méthode de Paass (dont une version est désignée dans cet article par le sigle HOD*) consiste essentiellement à déterminer tout d'abord une valeur intermédiaire, Z_{int}, à l'aide du fichier C en recourant à l'imputation hot-deck (avec une distance Y ou (X,Y) selon le cas), puis à déterminer une valeur Z authentique à partir du fichier B en utilisant de nouveau la méthode hot-deck avec une distance euclidienne (X,Z). Il s'agit là d'une version simplifiée de la méthode originale de Paass, laquelle est itérative et consiste à mettre à jour successivement les valeurs de Z (pour le fichier A), de Y (pour le fichier B) et de X (pour le fichier C – en supposant que celui-ci ne contient que de l'information sur (Y,Z)) en utilisant les fichiers C, A et B respectivement jusqu'à ce qu'un critère de convergence soit satisfait; se reporter à la section 3 pour plus de détails. Pour démarrer l'itération, on impute convenablement les valeurs initiales de Z (pour A), de Y (pour B) et de X (pour C). Dans l'étude exposée ici, nous considérons uniquement la version simplifiée de la méthode de Paass en raison des calculs complexes qu'exige la méthode originale. Comme pour la méthode REG*, la méthode HOD* ne s'applique pas si le fichier C est sous forme de tableau de fréquences. En outre, même si le fichier C renferme des micro-données mais qu'il a une taille réduite (comme dans le cas d'une petite enquête menée à des fins particulières), ou lorsqu'il contient de l'information substitutive ou périmée, il peut être préférable d'extraire de l'information de macro-niveau, comme les distributions de variables nominales fondées sur une partition relativement grossière.

Il convient de souligner qu'en l'absence d'information supplémentaire, c.-à-d. d'un fichier C, les méthodes REG* et HOD* se ramènent simplement aux méthodes d'imputation ordinaires, notamment la méthode par régression (REG) et la méthode hot-deck (HOD). Ces méthodes font l'objet, elles aussi, de l'étude exposée ici.

Nous proposons des versions modifiées des méthodes de Rubin et de Paass désignées respectivement par REG.LOGLIN* et HOD.LOGLIN*; ces versions reposent sur la méthode d'imputation log-linéaire de Singh (1988). Selon ces versions modifiées, on se sert de l'information supplémentaire pour définir des contraintes nominales destinées aux fichiers appariés à l'aide des méthodes REG* et HOD*. De cette manière, les liens catégoriques (estimés par des modèles log-linéaires), qui indiquent dans quelle mesure l'hypothèse d'indépendance conditionnelle n'est pas respectée (au sens qualitatif), se trouvent conservés dans le fichier qui a fait l'objet de l'appariement. Ces contraintes nominales sont censées rendre les distributions conjointes des données du fichier A robustes à l'égard d'informations supplémentaires de qualité inférieure ou imparfaites dans le fichier C. Si l'information supplémentaire est sous forme de distribution de variable nominale et qu'elle n'est pas de l'information de micro-niveau, on peut modifier les méthodes d'appariement fondées sur l'HIC en introduisant des contraintes nominales; dans ce cas, l'HIC est utilisée uniquement pour les classes de valeurs (X et Y). Par exemple, les méthodes d'imputation ordinaires REG et HOD, qui peuvent servir à l'appariement sans que l'on tienne compte de Y, peuvent aboutir à des versions modifiées REG.LOGLIN et HOD.LOGLIN. Ces deux dernières sont aussi considérées dans cet article.

Il convient de souligner que les méthodes d'appariement assujetties à des contraintes nominales diffèrent des méthodes d'appariement sous contrainte ordinaires, dans lesquelles les contraintes correspondent à quelques mesures caractéristiques du fichier B (comme la moyenne ou la variance) auxquelles doivent satisfaire les variables du fichier qui a fait l'objet de l'appariement. De plus, les méthodes d'appariement sous contrainte ordinaires mettent l'accent sur la distribution marginale de Z, tandis que les méthodes étudiées ici mettent l'accent sur la distribution conditionnelle (bien qu'il s'agisse d'une distribution de variable nominale), qui est plus appropriée pour le fichier A; il existe donc une différence fondamentale entre les deux types de méthodes d'appariement sous contrainte.

Conformément à Rubin (1986) et à Paass (1986), nous faisons une analyse empirique de l'efficacité des méthodes d'appariement. Nous avons donc effectué une étude de Monte Carlo pour analyser l'effet des modifications que l'on a proposé d'apporter aux méthodes existantes et ce, suivant deux scénarios: avec et sans information supplémentaire. Nous avons pu ainsi faire des analyses de sensibilité par rapport au non-respect de l'HIC et étudier les gains qui peuvent découler de l'utilisation d'informations supplémentaires. Les données fictives qui ont servi à l'étude de simulation ont été tirées de distributions normales multidimensionnelles accompagnées d'une contamination log-normale visant à induire l'asymétrie. L'utilisation de données fictives présente un avantage notable en ceci qu'on peut modifier les paramètres de contrôle pertinents pour obtenir différents scénarios de distribution pour le problème d'appariement. Nous avons comparé huit méthodes (quatre déjà en usage: REG, REG*, HOD, HOD*, et quatre à l'état de proposition: REG.LOGLIN, REG.LOGLIN*, HOD.LOGLIN, HOD.LOGLIN*) à l'aide de quatre mesures d'évaluation (deux au niveau individuel et deux au niveau global); se reporter à la section 6 pour plus de détails. Les principales conclusions de l'analyse empirique peuvent se résumer comme suit:

i) L'utilisation d'informations supplémentaires dans le but de contourner l'HIC peut améliorer considérablement la qualité du fichier enrichi. Toutefois, parmi les méthodes fondées sur l'HIC (c.-à-d. REG et HOD), la seconde est celle qui, dans l'ensemble, a la plus grande efficacité lorsqu'il n'existe pas d'information supplémentaire. En outre, on observe avec intérêt que

lorsqu'on s'écarte quelque peu de l'hypothèse d'indépendance conditionnelle, l'utilisation d'information supplémentaire n'accroît pas nécessairement l'efficacité de la méthode HOD en ce qui concerne les mesures d'évaluation au niveau global. Cette constatation devrait avoir une incidence notable dans les cas où il est difficile d'obtenir de l'information supplémentaire.

ii) La méthode REG* offre un rendement très acceptable en ce qui concerne les mesures d'évaluation au niveau individuel. On observe tout le contraire en ce qui a trait aux mesures au niveau global; cela est probablement attribuable au phénomène de "réduction à la moyenne", propre aux techniques de régression.

iii) La méthode HOD* est de beaucoup supérieure à la méthode REG* au niveau global mais elle est, en général, un peu moins efficace au niveau individuel.

iv) De façon générale, les contraintes nominales accroissent l'efficacité des méthodes REG* et HOD*. Plus précisément, en ce qui a trait aux mesures d'évaluation au niveau global, la méthode REG.LOGLIN* devient un peu plus efficace grâce à ces contraintes tandis que la méthode HOD.LOGLIN* devient beaucoup plus efficace. En ce qui concerne les mesures au niveau individuel, l'efficacité des deux méthodes demeure essentiellement la même.

v) En ce qui a trait aux mesures au niveau global, la méthode HOD.LOGLIN basée uniquement sur de l'information supplémentaire de type nominal est généralement plus efficace que la méthode HOD.LOGLIN* basée sur de l'information supplémentaire de micro-niveau. En revanche, pour ce qui a trait aux mesures au niveau individuel, la méthode HOD.LOGLIN est relativement moins efficace que la méthode HOD.LOGLIN*. Cette constatation peut être importante du point de vue pratique car la méthode HOD.LOGLIN est beaucoup moins exigeante sur le plan du calcul que la méthode HOD.LOGLIN* et elle ne nécessite pas d'information supplémentaire de micro-niveau. La méthode REG.LOGLIN n'a pas une efficacité comparable, à cause probablement du phénomène de "réduction à la moyenne".

vi) S'il s'agit d'informations supplémentaires substitutives ou périmées, leur utilisation peut néanmoins être profitable. Dans ces circonstances, la méthode HOD.LOGLIN donne des résultats très acceptables et, de fait, est passablement robuste à l'égard d'une information supplémentaire déficiente. Notons que comme cette méthode n'utilise que de l'information touchant les liens catégoriques entre les données supplémentaires, il semble raisonnable de penser qu'elle sera peu influencée par le fait que les données du fichier C sont substitutives ou quelque peu périmées. On ne peut en dire autant, toutefois, de la méthode REG.LOGLIN.

Il convient de souligner que plusieurs études empiriques ont été faites dans le passé afin d'évaluer les méthodes d'appariement statistique. Parmi celles qui ne considèrent pas l'utilisation d'informations supplémentaires, notons principalement Ruggles, Ruggles et Wolff (1977), Paass et Wauschkuhn (1980), Barr, Stewart et Turner (1981) de même que Rodgers et DeVol (1982). Dans Paass (1986), on fait une excellente analyse des tests empiriques portant sur la qualité des méthodes d'appariement.

Toutes les études mentionnées ci-dessus confirment que l'HIC peut représenter une contrainte sérieuse. Elles font ressortir la nécessité d'introduire de nouvelles informations dans le processus d'appariement. Peu d'études empiriques considèrent l'utilisation d'informations supplémentaires et l'incidence de l'HIC; Paass (1986) fait une évaluation en se servant uniquement de données fictives, tandis que Armstrong (1989) effectue des simulations à l'aide de données fictives et de données réelles. Nous pouvons considérer que la présente étude s'inscrit dans la suite de ces études en ce sens que de nouvelles méthodes y sont présentées et que l'éventail des distributions de population correspondantes est assez large.

Le plan de cet article est le suivant. La section 2 présente différents types d'information supplémentaire. La section 3 contient une analyse succincte de diverses méthodes d'appariement qui utilisent de l'information supplémentaire et la section 4 présente les modifications proposées sous forme de contraintes nominales. La section 5 sert à illustrer les diverses méthodes d'appariement par un exemple numérique simple. La section 6 contient la description du plan de l'étude empirique portant sur les méthodes d'appariement proposées et la section 7 contient l'analyse des résultats. Enfin, la section 8 renferme les conclusions de l'étude et propose des voies de recherche.

2. TYPES D'INFORMATION SUPPLÉMENTAIRE

Bien qu'il n'existe pas de fichier de micro-données courant et suffisamment gros qui contienne de l'information sur tout l'ensemble des variables, il peut exister une source auxiliaire qui renferme de l'information sur certaines des relations conjointes entre les éléments de l'ensemble de variables (X,Y,Z) ou les éléments du sous-ensemble (Y,Z). Si tel est le cas, on peut intégrer cette source au processus d'appariement pour éviter de poser l'HIC et pour améliorer la qualité du fichier complété par une réduction des distorsions présentes dans les relations conjointes entre les variables du fichier qui a fait l'objet de l'appariement.

Cette information supplémentaire peut venir de diverses sources et se présenter sous plusieurs formes. Comme cette information a pour seul but de nous permettre de ne pas poser l'HIC, nous en limitons l'usage en ce sens qu'elle ne peut jamais remplacer l'information contenue dans les fichiers receveur ou donneur. Autrement dit, il s'agit de tirer de la source auxiliaire des informations que l'on ne retrouve pas dans les fichiers de base. L'opération se fait en toute conformité avec les règles de protection du secret statistique qui s'appliquent à la source auxiliaire et suppose que cette source peut être le produit d'une petite enquête menée à des fins particulières ou un fichier de données confidentiel.

Deuxièmement, l'information supplémentaire n'a pas besoin d'être parfaite, c'est-à-dire qu'elle peut être déficiente dans une certaine mesure. Par exemple, elle peut venir d'une source périmée (une enquête ou un recensement antérieurs peut-être) – mais les renseignements qui nous intéressent dans cette source peuvent être, eux, tout à fait pertinents – ou elle peut représenter à tout le moins une amélioration par rapport à l'HIC, qui s'imposerait de toute manière. Par ailleurs, l'information supplémentaire peut avoir trait à un ensemble de variables substitutives qui sont supposées évoluer de la même manière que les variables étudiées.

Il peut y avoir de l'information supplémentaire de macro-niveau ou de micro-niveau. L'information de macro-niveau peut prendre la forme de coefficients de corrélation ou de proportions par case (de type nominal) ou encore d'autres paramètres. Si l'information supplémentaire contenue dans le fichier C porte sur la corrélation conditionnelle de Y et Z étant donné X, c'est-à-dire $\rho_{Y,Z|X}$, elle peut être combinée avec l'information des fichiers A et B concernant les corrélations de X et Y et de X et Z pour estimer la corrélation inconditionnelle de Y et Z au moyen de l'équation

$$\rho_{Y,Z} = \rho_{X,Y}\,\rho_{X,Z} + \rho_{Y,Z|X}\,(1 - \rho^2_{X,Y})^{1/2}\,(1 - \rho^2_{X,Z})^{1/2}. \qquad (2.1)$$

Or, on peut se servir des données des fichiers A et B pour faire une régression linéaire de Z par rapport à X et à Y en ce qui concerne la méthode REG* (voir section 3.1). Si l'information supplémentaire dont on dispose ne porte que sur la corrélation inconditionnelle de Y et Z, elle peut néanmoins servir de la même façon.

L'information supplémentaire de macro-niveau du fichier C peut aussi prendre la forme d'une distribution de variable nominale, comme la distribution de (X^*,Y^*,Z^*), où '*' désigne la transformation de la variable initiale en une variable nominale. S'il s'agissait de variables nominales dès le départ, on n'aurait pas à les transformer. On obtient le tableau de fréquences nécessaire pour les méthodes d'appariement sous contrainte nominale en "balayant par itération" le tableau (X^*,Y^*,Z^*) qui correspond au fichier C de manière que les tableaux marginaux (X^*,Y^*) et (X^*,Z^*) concordent respectivement avec le tableau (X^*,Y^*) du fichier A et le tableau (X^*,Z^*) du fichier B. Notons que ce dernier tableau devra avoir été "balayé par itération" au préalable de manière que ses fréquences marginales X^* concordent avec celles du fichier A. Dans la détermination des contraintes nominales, la méthode du "balayage par itération" préserve les liens (Y^*,Z^*) et (X^*,Y^*,Z^*) du tableau (X^*,Y^*,Z^*) du fichier C. L'opération d'ajustement que nous avons effectuée plus haut sur ce tableau est acceptable parce que les données sur la distribution (X^*,Y^*) du fichier A et la distribution (X^*,Z^*) du fichier B sont réputées plus exactes ou plus pertinentes que celles du fichier C. Si l'on ne connaît (ou n'utilise) que la distribution (Y^*,Z^*) du fichier C, il est possible de modifier la procédure de balayage par itération de manière à obtenir des contraintes nominales convenables. Dans ce cas, l'association (Y^*,Z^*) du fichier C serait préservée et le terme d'association à trois facteurs, (X^*,Y^*,Z^*), serait supposé nul. Pour réaliser cela, il faut tout d'abord balayer par itération le tableau (X^*,Z^*) du fichier B de manière que ses fréquences marginales X^* concordent avec celles du fichier A et balayer aussi le tableau (Y^*,Z^*) du fichier C de manière à établir une concordance avec les fréquences marginales Y^* du fichier A et les fréquences marginales Z^* du fichier B. Ensuite, on doit balayer par itération un tableau tridimensionnel de uns de manière à établir une concordance avec le tableau (X^*,Y^*) du fichier A, le tableau (X^*,Z^*) redressé du fichier B et le tableau (Y^*,Z^*) redressé du fichier C. Les valeurs catégoriques obtenues de cette manière n'ont pas besoin d'être des entiers. On les arrondit aléatoirement par une redistribution des fractions que l'on exécute en faisant un échantillonnage aléatoire sans remise de cases avec probabilité proportionnelle à la fraction se rattachant à la case. Cette opération est accomplie pour chaque classe (X^*,Y^*) prise individuellement.

La section suivante traite de l'utilisation de l'information supplémentaire dans l'appariement statistique. Elle décrit aussi l'utilisation de l'information supplémentaire de micro-niveau. Lorsque ce genre d'information existe, on peut, la plupart du temps, ramener cette information au macro-niveau par des totalisations et obtenir des données fiables sur les coefficients de corrélation et les proportions par case (de type nominal). Le succès de cette opération dépendra en partie de la taille du fichier de micro-données.

3. ANALYSE DE MÉTHODES D'APPARIEMENT STATISTIQUE

3.1 Méthode de régression

Nous allons décrire tout d'abord une méthode de régression qui utilise de l'information supplémentaire. Il s'agit d'une version de la méthode proposée par Rubin (1986). On suppose une forme paramétrique pour la régression de Z par rapport à X et à Y et on estime ensuite les paramètres correspondants à l'aide des données des fichiers A, B et C. Par exemple, dans le cas d'une régression linéaire, nous avons le modèle

$$E(Z \mid X,Y) = \beta_0 + \beta_1 X + \beta_2 Y, \qquad V(Z \mid X,Y) = \sigma^2, \qquad (3.1)$$

où β_0, β_1, et β_2 sont estimés à l'aide d'équations semblables aux équations des moindres carrés habituelles et par une combinaison adéquate de données des fichiers A, B et C. Nous décrivons ci-dessous une méthode pour estimer ces paramètres, qui diffère quelque peu de celle présentée dans Rubin (1986). Si le fichier C contient de l'information sur $(X, Y$ et $Z)$, on peut obtenir des estimations du coefficient de corrélation conditionnelle $\rho_{Y,Z|X}$ du fichier C, des estimations du coefficient de corrélation $\rho_{X,Z}$, de la

moyenne μ_Z, et de l'écart-type σ_Z du fichier B et des estimations du coefficient de corrélation $\rho_{X,Y}$, des moyennes μ_X, μ_Y et des écarts-types σ_X, σ_Y du fichier A. Par conséquent, on ne se servira du fichier B que si le fichier A ne contient pas tous les renseignements voulus sur la variable étudiée et on ne se servira du fichier C que si les fichiers A et B ne contiennent pas toute l'information voulue. Nous supposons donc l'existence d'une échelle de fiabilité ou de pertinence pour les fichiers A, B et C. Une telle échelle n'existait pas dans l'étude de Rubin. Nous pouvons alors calculer les estimations voulues à l'aide des équations

$$\beta_2 = \rho_{Y,Z|X}\frac{\sigma_{Z|X}}{\sigma_{Y|X}}, \quad \beta_1 = \rho_{X,Z|Y}\frac{\sigma_{Z|Y}}{\sigma_{X|Y}},$$

$$\beta_0 = \mu_Z - \beta_1\mu_X - \beta_2\mu_Y, \qquad (3.2)$$

où

$$\sigma_{Z|X} = (1 - \rho^2_{X,Z})^{1/2}\sigma_Z, \quad \sigma_{Y|X} = (1 - \rho^2_{X,Y})^{1/2}\sigma_Y,$$

$$\sigma_{Z|Y} = (1 - \rho^2_{Y,Z})^{1/2}\sigma_Z, \quad \sigma_{X|Y} = (1 - \rho^2_{X,Y})^{1/2}\sigma_X, \qquad (3.3)$$

et où $\rho_{X,Z|Y}$ est déterminé à l'aide de la formule habituelle après qu'on a calculé $\rho_{Y,Z}$ au moyen de la formule (2.1), c.-à-d.

$$\rho_{X,Z|Y} = (\rho_{X,Z} - \rho_{X,Y}\rho_{Y,Z})\,(1 - \rho^2_{X,Y})^{-1/2}. \qquad (3.4)$$

Notons que suivant l'hypothèse de la normalité, les écarts par rapport à l'indépendance conditionnelle sont représentés par le paramètre $\rho_{Y,Z|X}$. S'il y a indépendance conditionnelle, $\rho_{Y,Z|X} = 0$ et le modèle (3.1) se réduit à la régression linéaire simple de Z par rapport à X, c.-à-d.

$$E(Z \mid X) = \beta_0 + \beta_1 X, \quad V(Z \mid X) = \sigma^2, \qquad (3.5)$$

ce modèle peut être spécifié par la combinaison de données des fichiers A et B ou du fichier B seulement. Les formules (3.2) se ramènent à

$$\beta_2 = 0, \quad \beta_1 = \rho_{X,Z}\frac{\sigma_Z}{\sigma_X}, \quad \beta_0 = \mu_Z - \beta_1\mu_X. \qquad (3.6)$$

Lorsque le fichier C contient de l'information sur $\rho_{Y,Z}$ seulement, on peut estimer facilement les paramètres de (3.1) d'une manière comparable à celle ci-dessus en combinant des données des fichiers A, B et C.

Une fois qu'on a défini le modèle de régression, on peut appliquer la méthode REG* en deux étapes. La seconde étape est importante puisque nous voulons avoir des valeurs authentiques de Z de sorte que les relations entre les éléments de la variable multidimensionnelle Z soient préservées.

REG* (étape 1) Pour chaque couple (X, Y) en A, déterminer une valeur intermédiaire Z_{int} à l'aide du modèle de régression (3.1).

REG* (étape 2) Remplacer chaque triplet (X, Y, Z_{int}) obtenu à l'étape 1 par (X, Y, Z_{match}), où Z_{match} désigne une valeur authentique de Z tirée de B qui est le plus près possible de la valeur intermédiaire selon la distance euclidienne en (X,Z), où les éléments X et Z sont pondérés par leurs écarts-types respectifs. Autrement dit, on se sert de la méthode hot-deck pour déterminer les valeurs authentiques (observées). C'est ce que Rubin appelle l'"appariement par régression avec moyenne prédictive"; voir Little et Rubin (1987).

Ce qui distingue aussi la méthode décrite ci-dessus de celle présentée dans Rubin (1986) est que, selon celle-ci, on prédit une valeur Y pour les enregistrements du fichier B au moyen d'équations analogues à celles de (3.1), puis on prédit les valeurs Z correspondantes; ensuite, on apparie les enregistrements du fichier A à ceux du fichier B en fonction de la différence entre les valeurs Z prédites.

S'il n'y a pas d'information supplémentaire, on peut se servir de la méthode REG, appliquée avec l'HIC. Cette méthode comporte aussi deux étapes:

REG (étape 1) Pour chaque couple (X, Y) en A, déterminer une valeur intermédiaire Z_{int} à l'aide du modèle de régression simple (3.5).

REG (étape 2) Même chose que pour REG*.

Là aussi, on note des différences avec la méthode décrite dans Rubin (1986); selon celle-ci, on prédit une valeur Z pour les enregistrements du fichier B au moyen des équations (3.5), puis on apparie les enregistrements du fichier A à ceux du fichier B en fonction de la différence entre les valeurs Z prédites. Dans le cas qui nous occupe, où X est une variable unidimensionnelle, cela équivaut à un appariement par rapport à X.

3.2 Méthode hot-deck

Nous décrivons tout d'abord une méthode hot-deck qui utilise de l'information supplémentaire. Il s'agit d'une version de la méthode proposée par Paass (1986). Nous parlons ici de régression non paramétrique. Dans la régression paramétrique, la distribution conditionnelle de Z étant donné X et Y est définie au sens large par des fonctions de moyenne et de variance qui font intervenir un petit nombre de paramètres. Dans la régression non paramétrique, on se sert des techniques d'estimation non paramétrique de la densité pour estimer la distribution conditionnelle de Z. Par exemple, en ce qui a trait à la méthode d'estimation de la densité dite du plus proche voisin, on détermine, pour chaque couple (X, Y), K "plus proches voisins" (par rapport à une fonction de distance comme la distance euclidienne en (X, Y)); cet échantillon (probablement pondéré) de K "voisins", où K est un nombre entier spécifié convenablement, sert à représenter la distribution conditionnelle. Par conséquent, $P(Z \in U \mid X, Y)$ peut être spécifiée comme une espérance conditionnelle,

$$E(I_U(Z) \mid X, Y) = \sum_{i=1}^{K} w_i(X, Y)\, I_U(Z_i), \qquad (3.7)$$

où les w_i sont des poids qui diminuent à mesure que s'accroît la distance entre (X_i, Y_i) et (X, Y), et I_U est la fonction indicatrice pour l'ensemble U.

Dans la méthode de Paass, on commence par déterminer la distribution conditionnelle de Z pour chaque couple (X,Y) en A en la représentant au moyen d'un ensemble de K valeurs Z obtenu par une régression non paramétrique. Autrement dit, on ajoute K valeurs Z à chaque couple (X,Y). En deuxième lieu, on détermine, pour chaque couple (X,Y) en A, une valeur Z authentique - Z_{match} - tirée du fichier B, qui est considérée comme la plus proche suivant la distance (X,Z). On obtient ainsi le triplet $(X,Y),Z_{\text{match}})$ pour le fichier qui a fait l'objet de l'appariement (fichier A). Si le fichier C contient de l'information sur (Y,Z), on détermine les distributions conditionnelles pour le fichier A à l'aide du processus itératif suivant: choisir K valeurs initiales comme "plus proches voisins" pour Z, dans le fichier A, Y, dans le fichier B, et X, dans le fichier C; cela peut se faire par la méthode d'imputation hot-deck ordinaire. Or, chaque cycle consiste à établir des distributions conditionnelles pour les éléments (X,Y) en A à l'aide de l'information contenue dans le fichier C, plus précisément en mettant à jour de façon conforme K valeurs Z du fichier A à l'aide de valeurs Z du fichier C en utilisant la distance (X,Y), puis en mettant à jour K valeurs Y du fichier B à l'aide de valeurs Y du fichier A en utilisant la distance (X,Z) et enfin, en mettant à jour K valeurs X du fichier C à l'aide de valeurs X du fichier B en utilisant la distance (Y,Z). Ce cycle est répété jusqu'à ce que la différence maximale entre des statistiques de la distribution tridimensionnelle de (X,Y,Z) obtenue par itérations successives (par ex.: matrice de covariance) soit inférieure à une limite donnée. À la convergence, chaque fichier compte K nouvelles valeurs qui représentent les distributions conditionnelles respectives. Par ailleurs, si le fichier C contient de l'information sur (X,Y,Z), le processus est non itératif. Il suffit de se servir du fichier C pour obtenir K valeurs Z de A en utilisant la distance (X,Y), puis de déterminer une valeur Z_{match}, tirée du fichier B, pour chaque couple (X,Y) en A en utilisant la distance (X,Z). Toutefois, Paass n'envisage pas cette possibilité dans son article.

Dans l'étude empirique que nous exposons ici, nous n'avons pas utilisé la version itérative de la méthode de Paass décrite ci-dessus lorsque le fichier C contenait de l'information sur (Y,Z) parce qu'elle comporte de nombreux calculs. Nous avons plutôt utilisé une version non itérative simplifiée où $K = 1$. Cette méthode, désignée par HOD*, consiste en deux étapes.

HOD* (étape 1) Pour chaque couple (X,Y) en A, déterminer par la méthode hot-deck une valeur intermédiaire Z_{int} à partir du fichier C en utilisant la distance Y si l'information supplémentaire porte sur (Y,Z) et la distance euclidienne (X,Y) si l'information supplémentaire porte sur (X,Y,Z).

HOD* (étape 2) Remplacer chaque triplet (X,Y,Z_{int}) obtenu à l'étape 1 par (X,Y,Z_{match}), où Z_{match} est tiré du fichier B à l'aide de la méthode hot-deck avec la distance euclidienne (X,Z).

S'il n'existe pas de fichier C, on peut se servir de la méthode HOD, appliquée avec l'HIC. Cette méthode comporte deux étapes:

HOD (étape 1) Déterminer des classes X appropriées, comme dans l'imputation hot-deck ordinaire.

HOD (étape 2) Pour chaque couple (X,Y) en A, imputer une valeur Z observée, tirée de la classe X correspondante dans le fichier B, en utilisant la méthode hot-deck avec la distance X.

4. MODIFICATION DES MÉTHODES D'APPARIEMENT PAR L'INTRODUCTION DE CONTRAINTES NOMINALES

Nous proposons de modifier les méthodes d'appariement REG, REG*, HOD et HOD* en définissant des contraintes de type nominal pour les valeurs Z que l'on tire du fichier B pour compléter le fichier A. Ces contraintes ont pour but de conserver les liens catégoriques (définis par des modèles log-linéaires) suivant une partition de (X,Y,Z) acceptable dans le cas du fichier enrichi. On détermine ces liens en combinant des informations des fichiers A, B et C. La notion d'appariement sous contrainte nominale repose sur la méthode d'imputation log-linéaire (voir Singh 1988; Singh *et al.* 1988). Dans ce cas, les contraintes peuvent dépendre de l'information supplémentaire qui peut servir à estimer la distribution conditionnelle de variables nominales, ou certains aspects de cette distribution, mais qui n'a pas le niveau de qualité voulu pour estimer la distribution conditionnelle intégrale.

Nous débutons avec une partition adéquate des variables X, Y et Z. Posons X^*, Y^*, Z^* comme les variables nominales correspondantes (après transformation). On peut alors paramétrer la distribution des proportions par case pour le tableau (X^*, Y^*, Z^*) au moyen d'un modèle log-linéaire

$$\log p_{ijk} = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}, \qquad (4.1)$$

où p_{ijk} est la proportion relative à la case (i,j,k) et les nombres indiciaires 1, 2 et 3 désignent respectivement X^*, Y^*, et Z^*. Il convient de souligner que l'équation de paramétrage (4.1) vaut pour des distributions arbitraires des variables originales (X,Y,Z). Évidemment, les fichiers A et B ne contiennent pas d'information sur les effets à deux facteurs u_{23} et les effets à trois facteurs u_{123}. Si ces effets sont posés égal à zéro, cela revient à poser l'hypothèse de l'indépendance conditionnelle au sens qualitatif, c.-à-d. $Y^* \perp Z^* \mid X^*$. Cependant, si nous disposons d'informations supplémentaires dans le fichier C, nous pouvons nous passer de cette hypothèse parce qu'il est possible d'estimer les paramètres u_{23} et u_{123} à l'aide du fichier C. Par conséquent, peu importe la forme de la distribution conjointe de (X,Y,Z), le modèle log-linéaire ci-dessus offre une façon homogène d'évaluer, à tout le moins au sens qualitatif, la distanciation par rapport à l'HIC. Par ailleurs, dans la régression linéaire, la distanciation par rapport à

l'HIC est paramétrée par $\rho_{Y,Z|X}$ seulement lorsqu'il y a normalité.

Comme nous l'avons expliqué dans la section 2, l'information supplémentaire contenue dans le fichier C (qu'elle porte sur (Y,Z) ou (X,Y,Z)) sert tout d'abord à définir des contraintes nominales sous la forme d'une distribution (X^*,Y^*,Z^*). Cette opération se fait par "balayage" de manière à conserver les effets u_{23} et u_{123} du fichier C. La version de REG* soumise à des contraintes nominales, et désignée par REG.LOGLIN*, peut alors être définie par les étapes suivantes:

REG.LOGLIN* (étape 1) Même chose qu'à l'étape 1 de REG*.

REG.LOGLIN* (étape 2) Même chose qu'à l'étape 2 de REG*, sauf que des contraintes nominales sont appliquées, ce qui suppose la nécessité d'un ordre d'appariement lorsqu'on tire des valeurs authentiques de Z du fichier B. On détermine tout d'abord le cas d'appariement qui a la distance minimum en (X,Z). On note la classe (X^*,Y^*,Z^*) de l'enregistrement complété et si le nombre d'appariements réalisés dans cette classe ne dépasse pas le chiffre prescrit par les contraintes nominales, la concordance est reconnue. Autrement, on rejette la concordance et on examine le cas d'appariement qui vient au deuxième rang en ce qui concerne la distance minimum. Le processus se poursuit jusqu'à ce que le fichier A soit complété; la distribution de (X^*,Y^*,Z^*) dans le fichier complété doit alors satisfaire aux contraintes nominales.

De la même manière, la version de HOD* soumise à des contraintes nominales, et désignée par HOD.LOGLIN*, comporte les deux étapes suivantes:

HOD.LOGLIN* (étape 1) Pour chaque couple (X,Y) en A, déterminer par la méthode hot-deck une valeur intermédiaire Z_{int} à partir du fichier C en utilisant la distance Y ou (X,Y) selon le cas, de sorte que les contraintes nominales soient respectées. Cette étape est semblable à l'étape 2 de REG.LOGLIN*.

HOD.LOGLIN* (étape 2) Pour chaque triplet (X,Y,Z_{int}), déterminer à l'aide de la méthode hot-deck une valeur observée – Z_{match} – du fichier B avec la distance (X,Z), en ayant soin de respecter la classe de la valeur Z_{int}.

Une variante de cette méthode consisterait à imputer une valeur intermédiaire Z_{int} sans contrainte, puis à recourir à l'appariement sous contrainte nominale fondé sur la distance pour tirer une valeur observée du fichier B, comme à l'étape 2 de REG.LOGLIN*. Nous avons essayé cette variante, mais celle-ci n'a pas donné de bons résultats; nous l'avons donc exclue de l'étude à cause des nombreux calculs qu'elle comporte. On peut attribuer le piètre rendement de cette variante au phénomène de "réduction à la moyenne" qui s'applique aux valeurs Z_{int} du fichier C du fait que celui-ci est trop petit. Autrement dit, les valeurs Z_{int} sont concentrées au centre de la distribution et lorsqu'on introduit les contraintes nominales, les valeurs Z finales se trouvent concentrées vers les limites intérieures des classes extérieures de Z.

Supposons que le fichier C ne renferme que de l'information de macro-niveau sous forme de distribution de variable nominale, ou bien que l'information de micro-niveau du fichier C soit jugée incertaine alors que l'information sous forme de distribution de variable nominale (suivant une partition relativement grossière), elle, soit jugée fiable. Nous pouvons alors définir des versions des méthodes REG et HOD assujetties à des contraintes nominales, et désignées respectivement par REG.LOGLIN et HOD.LOGLIN. La première comporte deux étapes:

REG.LOGLIN (étape 1) Même chose qu'à l'étape 1 de REG.

REG.LOGLIN (étape 2) Même chose qu'à l'étape 2 de REG.LOGLIN*.

De la même manière, HOD.LOGLIN comprend les deux étapes suivantes:

HOD.LOGLIN (étape 1) Même chose qu'à l'étape 1 de HOD.

HOD.LOGLIN (étape 2) Même chose qu'à l'étape 2 de REG.LOGLIN*, sauf qu'il n'existe pas de valeur intermédiaire Z_{int}, de sorte que l'appariement est fondé sur la distance X au lieu de la distance (X,Z).

Pour REG.LOGLIN comme pour HOD.LOGLIN, qui ne nécessitent pas d'information de micro-niveau dans le fichier C, l'HIC n'est utilisée que dans les classes X,Y. On suppose donc une forme réduite d'indépendance conditionnelle et les conséquences de cette hypothèse ne devraient pas être aussi grandes que si l'on posait l'HIC dans son intégralité.

5. EXEMPLE TYPIQUE

Avant d'étudier les propriétés empiriques des modifications proposées pour les méthodes d'appariement présentées plus haut, il serait utile de considérer un exemple numérique simple qui illustre les genres de calcul qu'impliquent les huit méthodes. Supposons que les fichiers A, B et C sont définis comme dans le tableau 1; ces fichiers reposent sur des échantillons aléatoires tirés d'une distribution normale multidimensionnelle caractérisée par une moyenne nulle et une matrice de covariance définie par $\sigma_X = \sigma_Y = \sigma_Z = 1$, $\rho_{X,Y} = \rho_{X,Z} = .5$ et $\rho_{Y,Z} = .7$ (ce qui implique que $\rho_{Y,Z|X} = .6$). On suppose ici que le fichier C contient de l'information uniquement sur (Y,Z). En ce qui concerne le fichier A, les valeurs Z ont été supprimées dans le tableau 1 mais elles figurent dans le tableau 3 pour le calcul des mesures d'évaluation. Supposons que, pour des raisons de simplicité et à cause de la taille relativement modeste des fichiers, nous utilisions une transformation en variable nominale plutôt grossière pour X,Y,Z en ne considérant que deux classes, $(-\infty,0)$ et $[0,\infty)$. On obtient ainsi les trois tableaux bidimensionnels qui correspondent aux fichiers A, B et C et qui sont représentés dans le tableau 2(a). Dans le tableau 2(b) se trouvent

les tableaux redressés des fichiers B et C; ce redressement est effectué de sorte qu'il y ait concordance entre ces tableaux et les fréquences marginales pertinentes, comme nous l'expliquons dans la section 2. Le tableau 2(c) correspond au tableau tridimensionnel obtenu à la suite d'un balayage par itération et le tableau 2(d) montre les contraintes nominales obtenues à la suite de l'arrondissement aléatoire des éléments du tableau 2(c), comme nous l'avons expliqué dans la section 2.

Les huit méthodes ont été appliquées aux données du tableau 1 et les résultats de l'appariement figurent dans le tableau 3 avec les valeurs vraies de Z, qui avaient été supprimées dans le tableau 1.

Les mesures d'évaluation indiquées dans le tableau 3 ont été mentionnées brièvement dans l'introduction et sont expliquées en détail dans la section 6. La partition utilisée pour la mesure χ^2 est la même que celle qui a servi à la détermination des contraintes nominales. Comme la partition n'a pas été modifiée pour l'évaluation, il convient de noter que les valeurs de χ^2 pour M3, M4, M7 et M8 seront identiques. Notons aussi que les mesures d'évaluation ne sont présentées ici que dans le but d'illustrer le calcul et ne doivent pas être considérées comme des indicateurs du rendement relatif des méthodes puisqu'elles ne reposent que sur un petit échantillon.

La méthode M8 (HOD.LOGLIN*) est celle qui comporte le plus grand nombre de calculs; la tableau 4 en décrit les étapes. À partir de ce tableau, on peut se représenter assez facilement les étapes requises pour les autres méthodes.

6. ANALYSE EMPIRIQUE DES MÉTHODES D'APPARIEMENT PROPOSÉES

Dans cette section, nous décrivons une analyse empirique qui prend la forme d'une étude de simulation approfondie qui utilise des données fictives tirées de distributions multidimensionnelles symétriques et asymétriques. La symétrie provient de distributions normales tandis que la dissymétrie est introduite par des contaminations engendrées par des distributions log-normales multidimensionnelles. Nous utilisons des données fictives afin d'exercer un contrôle sur tous les paramètres pertinents, y compris ceux qui définissent les relations conjointes des différentes variables. On peut alors évaluer les diverses méthodes d'appariement lorsque les relations conjointes s'écartent d'une manière systématique de l'indépendance conditionnelle. On peut aussi comparer ces méthodes lorsque la distribution d'où sont tirées les données s'écarte de la symétrie. Nous avons obtenu de l'information supplémentaire substitutive en modifiant les paramètres de la distribution normale d'où était tiré le fichier C ou en induisant des contaminations log-normales. Nous avons donc quatre types de problème d'appariement: d'une part, deux distributions, l'une symétrique et l'autre asymétrique, avec des données du fichier C qui sont non substitutives et d'autre part, deux distributions symétriques avec, chacune, une catégorie de données substitutives du fichier C différente. La programmation s'est faite sur micro-ordinateur à l'aide du logiciel GAUSS.

Tableau 1

Données des fichiers A, B, C

Identificateur d'enregistrement	Fichier A		Identificateur d'enregistrement	Fichier B		Identificateur d'enregistrement	Fichier C	
	X	Y		X	Y		Y	Z
A1	−0.86	−0.32	B1	−0.95	−0.69	C1	−0.40	−0.60
A2	−0.77	−0.33	B2	−0.64	−0.83	C2	−2.33	−2.81
A3	−0.09	−0.26	B3	−1.58	−0.11	C3	−0.79	−0.47
A4	−0.42	0.62	B4	−0.42	0.36	C4	0.67	−0.29
A5	−0.81	0.56	B5	0.97	−0.42	C5	−0.65	1.19
A6	−0.56	0.00	B6	1.09	−1.16	C6	−1.32	0.05
A7	0.37	−0.04	B7	0.44	−0.49	C7	−0.55	0.70
A8	0.06	−1.29	B8	0.14	−0.38	C8	0.55	0.66
A9	0.95	−2.15	B9	1.33	1.24	C9	1.31	1.12
A10	1.90	−1.07	B10	0.80	0.85	C10	1.46	2.58
A11	1.32	0.61	B11	1.60	0.31			
A12	1.38	0.79	B12	1.42	0.99			
A13	1.63	1.03						
A14	0.50	1.24						
A15	0.90	1.19						

Tableau 2

Distributions de variables nominales pour les fichiers A, B et C suivant la partition donnée 2 x 2 x 2

(a)

	Fichier A	
	$Y < 0$	$Y \geq 0$
$X < 0$	3	3
$X \geq 0$	4	5

	Fichier B	
	$Z < 0$	$Z \geq 0$
$X < 0$	3	1
$X \geq 0$	4	4

	Fichier C	
	$Z < 0$	$Z \geq 0$
$Y < 0$	3	3
$Y \geq 0$	1	3

(b)

	Tableau non redressé du fichier A	
	$Y < 0$	$Y \geq 0$
$X < 0$	3	3
$X \geq 0$	4	5

	Tableau redressé du fichier B	
	$Z < 0$	$Z \geq 0$
$X < 0$	4.5	1.5
$X \geq 0$	4.5	4.5

	Tableau redressé du fichier C	
	$Z < 0$	$Z \geq 0$
$Y < 0$	5.15	1.85
$Y \geq 0$	3.85	4.15

(c) Tableau obtenu à partir d'un tableau $2 \times 2 \times 2$ de uns "balayé par itération" en vue d'une concordance avec les fréquences marginales du tableau 2(b)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2.55	1.95	0.45	1.05
$X \geq 0$	2.60	1.90	1.40	3.10

(d) Contraintes nominales établies par l'arrondissement aléatoire des éléments du tableau 2(c)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2	2	1	1
$X \geq 0$	2	2	2	3

Tableau 3

Comparaison de huit méthodes d'appariement permettant de compléter le fichier A

Fichier A			Valeurs Z appariées							
			Versions de la méthode REG				Versions de la méthode HOD			
X	Y	Z	M1	M2	M3	M4	M5	M6	M7	M8
−0.86	−0.32	−0.97	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69
−0.77	−0.33	0.16	−0.69	−0.69	−0.69	−0.69	−0.83	−0.69	−0.83	−0.69
−0.09	−0.26	0.19	−0.38	−0.38	0.36	0.36	0.36	−0.38	0.36	0.36
−0.42	0.62	−0.44	−0.38	0.36	0.36	−0.38	0.36	−0.38	0.36	−0.38
−0.81	0.56	−0.76	−0.69	0.36	−0.69	0.36	−0.69	0.36	−0.69	0.36
−0.56	0.00	1.06	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83
0.37	−0.04	−1.18	−0.38	−0.38	−0.38	0.36	−0.49	−0.49	−0.49	−0.49
0.06	−1.29	0.33	−0.38	−0.38	−0.36	−0.38	−0.38	−0.38	0.85	0.36
0.95	−2.15	−1.26	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16
1.90	−1.07	0.01	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
1.32	0.61	2.08	0.31	0.99	0.31	0.99	1.24	−0.42	1.24	−0.42
1.38	0.79	0.32	0.31	0.99	0.31	0.99	0.99	−0.42	0.99	−0.42
1.63	1.03	1.53	0.31	0.99	0.31	0.99	0.31	0.99	0.31	0.99
0.50	1.24	1.34	−0.49	0.85	−0.49	−0.38	−0.49	0.85	−0.49	0.85
0.90	1.19	−1.01	−0.42	0.85	−0.42	−0.42	−0.42	0.85	−0.42	0.85
Mesures d'évaluation		MAD-Z	0.79	0.81	0.76	0.78	0.79	0.85	0.78	0.78
		x^2	13.07	13.34	1.75	1.75	2.70	10.78	1.75	1.75

Nota: M1: REG M2: REG* M3: REG.LOGLIN M4: REG.LOGLIN* M5: HOD M6: HOD* M7: HOD.LOGLIN M8: HOD.LOGLIN*.

Tableau 4

Étapes de la méthode M8 (HOD.LOGLIN*)

Case (X,Y)	X	Y		Ordre d'appariement	Dist. Y	Z_{int}		Dist. (X,Z)	Z_{match}	
$X < 0$	−0.86	−0.32	(A1)	2	.08	−0.60	(C1)	.12	−0.69	(B1)
$Y < 0$	−0.77	−0.33	(A2)	1	.07	−0.60	(C1)	.19	−0.69	(B1)
	−0.09	−0.26	(A3)	3	.29	0.70	(C7)	.47	0.36	(B4)
$X < 0$	−0.42	0.62	(A4)	2	.05	−0.29	(C4)	.56	−0.38	(B8)
$Y \geq 0$	−0.81	0.56	(A5)	1	.01	0.66	(C8)	.50	0.36	(B4)
	−0.56	0.00	(A6)	3	.40	−0.60	(C1)	.25	−0.83	(B2)
$X \geq 0$	0.37	−0.04	(A7)	4	.36	−0.60	(C1)	.14	−0.49	(B7)
$Y < 0$	0.06	−1.29	(A8)	1	.03	0.05	(C6)	.57	0.36	(B4)
	0.95	−2.15	(A9)	2	.18	−2.81	(C2)	1.66	−1.16	(B6)
	1.90	−1.07	(A10)	3	.25	0.05	(C6)	.40	0.31	(B11)
$X \geq 0$	1.32	0.61	(A11)	1	.06	−0.29	(C4)	.37	−0.42	(B5)
$Y \geq 0$	1.38	0.79	(A12)	3	.12	−0.29	(C4)	.43	−0.42	(B5)
	1.63	1.03	(A13)	5	.28	1.12	(C9)	.25	0.99	(B12)
	0.50	1.24	(A14)	2	.07	1.12	(C9)	.41	0.85	(B10)
	0.90	1.19	(A15)	4	.12	1.12	(C9)	.29	0.85	(B10)

Nota: L'ordre d'appariement des enregistrements des fichiers A et C est renouvelé pour chaque case (X,Y) selon les contraintes nominales établies dans le tableau 2(d).

6.1 Plan de l'étude de Monte Carlo

Il faut trois fichiers pour simuler l'appariement statistique: un fichier receveur A, un fichier donneur B et un fichier auxiliaire C. Ces fichiers sont créés synthétiquement à l'aide de distributions spécifiées, chacun d'eux contenant les trois variables, X, Y et Z. On supprime la variable Z du fichier A et la variable Y du fichier B. Les valeurs de la variable supprimée du fichier A servent à évaluer le rendement des diverses méthodes d'appariement statistique. Le fichier C peut contenir de l'information sur (Y,Z) seulement (si X est supprimée) ou sur les trois variables. Les résultats empiriques présentés dans cet article sont ceux qui correspondent à la situation où C contient de l'information sur (Y,Z) seulement, bien que l'autre situation - C contenant de l'information sur (X,Y,Z) - ait été envisagée aussi dans l'étude (voir Singh *et al.* 1990).

Des passages de 100 simulations chacun ont été exécutés pour chaque combinaison de paramètres de plan étudiés. On calculait quatre mesures d'évaluation pour chaque simulation, puis on faisait la moyenne de chacune de ces mesures pour les 100 simulations.

Les fichiers A et B étaient toujours tirés de la même distribution, chacun d'eux contenant 500 observations indépendantes et identiquement distribuées. Le fichier C contenait 250 observations qui ne provenaient pas nécessairement de la même distribution que les observations des fichiers A et B; autrement dit, le fichier C pouvait contenir de l'information supplémentaire substitutive ou non substitutive.

Les observations (X,Y,Z) suivaient une distribution normale multidimensionnelle mais dans certains cas, on induisait une contamination log-normale en calculant l'exponentielle de X, de Y et de Z. On déterminait si une observation était contaminée ou non au moyen d'un processus de Bernoulli avec une probabilité déterminée pour tout passage de 100 simulations. Avant la contamination, X, Y et Z étaient des variables normales centrées réduites; de plus, les covariances de (X,Y) et de (X,Z) étaient toujours de .5 tandis que la covariance de (Y,Z) variait d'un passage à l'autre. Par conséquent, la corrélation conditionnelle de Y et Z étant donné X, $\rho_{Y,Z|X}$, variait aussi d'un passage à l'autre.

Pour la plupart des passages, les observations du fichier C suivaient la même distribution que les observations des fichiers A et B. Toutefois, si, dans une application, l'information supplémentaire vient d'une source historique ou se rapporte à des variables de substitution, cette hypothèse peut être déraisonnable. Deux séries de passages ont été exécutées avec de l'information supplémentaire substitutive. Dans la première série, les données supplémentaires avaient un coefficient $\rho_{Y,Z|X}$, différent. Dans la seconde série, les données étaient soumises dans une certaine proportion à une contamination log-normale.

Pour les méthodes proposées qui utilisent des contraintes nominales et pour la définition de classes d'appariement pour la méthode HOD, il fallait choisir une partition. Nous en avons utilisé deux. La première, appelée partition à intervalles standard, répartissait les valeurs des variables X, Y et Z dans les classes suivantes: < -1, $[-1,0)$, $[0,1)$, ≥ 1; autrement dit, la partition était centrée sur la moyenne de la distribution marginale avant contamination, avec des valeurs seuil au centre et à 1 écart-type. La seconde partition, appelée partition à probabilités égales, était semblable à la première sauf que les valeurs seuil coïncidaient avec les quartiles des distributions marginales pré-contamination; autrement dit, la partition consistait dans les classes suivantes: $< -.6745$, $[-.6745,0)$, $[0,.6745)$, $\geq .6745$. Les partitions étaient définies en fonction des distributions pré-contamination; pour des raisons de simplicité, on utilisait les mêmes partitions lorsque des contaminations log-normales étaient induites. Cependant, il aurait été plus réaliste de faire dépendre les partitions des données.

6.2 Les méthodes d'appariement

Nous avons étudié les huit méthodes décrites plus haut. Toutes, sauf REG et HOD, utilisent de l'information supplémentaire. Il existe donc deux versions de chacune de ces méthodes selon que le fichier C renferme de l'information sur (Y,Z) ou (X,Y,Z). En ce qui concerne les méthodes HOD et HOD.LOGLIN, nous avons considéré trois versions de la hot-deck (notamment, celle fondée sur le rang, la méthode aléatoire et celle fondée sur la distance X) pour tirer des valeurs authentiques de Z du fichier B, mais seuls les résultats de la troisième sont reproduits ici. Pour ce qui a trait aux six autres méthodes, nous avons considéré aussi trois versions de hot-deck (à savoir, distance Z, distance (X,Z) et distance (X,Y,Z)) pour tirer des valeurs authentiques de Z du fichier B, mais nous ne présentons ici, pour des raisons de simplicité, que les résultats relatifs à la distance (X,Z). La section 7.3 expose sommairement les résultats obtenus avec différentes mesures de distance. On trouvera dans le document de travail de Singh *et al.* (1990) d'autres détails qui ne figurent pas dans le présent article. Notons que si l'on veut utiliser la méthode hod-deck fondée sur la distance (X,Y,Z) pour tirer une valeur authentique de Z du fichier B, il faut tout d'abord déterminer à partir du fichier C des valeurs intermédiaires Y pour B, à l'instar de la valeur Z_{int} pour le fichier A. Notons aussi que nous nous sommes servis de la distance euclidienne à chaque fois qu'une version de la hot-deck fondée sur une mesure de distance était utilisée. Toutefois, nous n'avons pas redressé au préalable les variables par leurs écarts-types respectifs pour des raisons de commodité et parce que toutes les variables de la population fictive avaient des variances communes.

6.3 Les mesures d'évaluation

Quatre mesures ont servi à évaluer le degré d'efficacité des diverses méthodes d'appariement. Toutes les évaluations reposent sur des comparaisons du fichier enrichi avec le fichier duquel ont été supprimées les valeurs vraies de Z. Deux des mesures reposent sur des comparaisons de type nominal, mais il n'est pas nécessaire que les classes qui servent aux évaluations soient celles qu'utilisent les méthodes LOGLIN pour les contraintes nominales. Les résultats présentés ici sont ceux que nous avons obtenus en utilisant la partition à probabilités égales (voir section 6.1).

pour l'appariement et la partition à intervalles standard pour les évaluations. La première des quatre mesures d'évaluation repose sur une comparaison par unité des valeurs Z appariées et des valeurs Z supprimées. Cependant, une méthode d'appariement statistique n'a pas pour but de reproduire fidèlement les valeurs Z supprimées mais de produire des valeurs Z qui originent de la même distribution étant donné ce que l'on connaît, en l'occurrence X et Y. Les trois autres mesures d'évaluation reposent plus sur la comparaison des propriétés des distributions conditionnelles de Z.

i) Moyenne des écarts absolus moyens de Z ($\overline{\text{MAD-}Z}$)

La mesure de rendement la plus simple est l'écart absolu moyen entre les valeurs appariées et les valeurs supprimées de Z pour les enregistrements du fichier A. Nous avons calculé des moyennes de Monte Carlo de ces écarts et les erreurs types correspondantes.

La formule de la statistique MAD-Z pour n'importe quelle simulation donnée est

$$\text{MAD-}Z = \sum_i | Z_{s,i} - Z_{m,i} | /500, \qquad (6.1)$$

où $Z_{s,i}$ est la valeur Z supprimée pour le $i^{\text{ème}}$ enregistrement du fichier A, $Z_{m,i}$ est la valeur Z appariée et la sommation est étendue aux 500 enregistrements du fichier A. $\overline{\text{MAD-}Z}$ désigne la moyenne des statistiques MAD-Z pour toutes les simulations.

ii) Moyenne des écarts des covariances ($\overline{\text{AD-Cov}}$)

La deuxième mesure de rendement est l'écart entre les covariances conditionnelles de Y et Z étant donné X pour le fichier enrichi et le fichier duquel ont été supprimées les valeurs vraies de Z. Nous avons calculé des moyennes de Monte Carlo de ces écarts et les erreurs types correspondantes.

Pour un fichier contenant les variables X, Y et Z, nous pouvons définir la relation

$$\text{Cov}(Y,Z \mid X) = \text{Cov}(Y,Z) - \text{Cov}(X,Y)\text{Cov}(X,Z)/\text{Var}(X), \qquad (6.2)$$

où Cov et Var sont les opérateurs de covariance et de variance d'échantillon respectivement. Dans le cas d'une distribution normale multidimensionnelle, cela correspond à la covariance de Y et Z étant donné X; autrement, on peut le voir comme la covariance des résidus d'une régression linéaire de Y par rapport à X et des résidus d'une régression linéaire de Z par rapport à X. Pour n'importe quelle simulation donnée, la statistique AD-Cov serait l'écart entre ces quantités pour le fichier enrichi et le fichier duquel ont été supprimées les valeurs vraies de Z. Comme pour la première mesure, $\overline{\text{AD-Cov}}$ désigne la moyenne de ces statistiques pour toutes les simulations.

iii) Moyenne des statistiques chi carré ($\overline{\chi^2}$)

La troisième mesure de rendement, basée sur des comparaisons de type nominal, est une mesure de distance fondée sur le critère chi carré de Pearson. Ce qu'on présente, en fait, est la statistique chi carré moyenne calculée pour les 100 simulations, transformée de manière à être incluse dans l'intervalle (0,1).

La formule de la statistique chi carré est

$$\chi^2 = \sum_{i,j,k} (m_{ijk} - n_{ijk})^2/(m_{ijk} + .5), \qquad (6.3)$$

où m_{ijk} est le nombre d'enregistrements qui appartiennent à la classe X^* i, à la classe Y^* j, et à la classe Z^* k dans le fichier enrichi, n_{ijk} est l'équivalent pour le fichier duquel ont été supprimées les valeurs vraies de Z, et où la sommation est étendue à toutes les classes (X^*, Y^*, Z^*). Une valeur constante de .5 est ajoutée à tous les dénominateurs de cette somme pour éviter le problème des zéros.

Une fois calculée la moyenne des statistiques chi carré pour 100 simulations, disons $\overline{\chi^2}$, cette valeur est transformée de manière à être incluse dans l'intervalle (0,1). La formule de transformation est la suivante (voir Bishop, Fienberg et Holland, 1975, p. 383; le chiffre 500 représente la taille du fichier A)

$$\overline{\chi^2} \text{ transformé} = \{\overline{\chi^2}/(\overline{\chi^2} + 500)\}^{1/2}. \qquad (6.4)$$

iv) Test du rapport des vraisemblances (TRV)

La dernière mesure de rendement est aussi basée sur des comparaisons de type nominal. Pour chaque classe (X^*, Y^*) qui contient un nombre minimum d'observations (dans la présente étude, ce nombre est fixé à 20), on exécute un test du rapport des vraisemblances pour tester l'hypothèse que les valeurs Z de type nominal du fichier enrichi et du fichier amputé proviennent de la même distribution multinomiale. On combine ensuite les tests faits pour les différentes classes (X^*, Y^*) afin d'obtenir une valeur P globale. Ce qui nous intéresse est le nombre de fois, sur 100 simulations, où la valeur P globale était inférieure à .05. Plus cette proportion est élevée, plus il y a de différence entre la distribution des valeurs vraies de $(Z^*$ et celle des valeurs appariées étant donné les classes (X^*, Y^*).

Le critère minimum de 20 unités pour les classes (X^*, Y^*) du fichier A était nécessaire pour que soit acceptable l'approximation chi carré de la distribution de la variable à tester. Si le nombre de classes de $(Z^*$ augmentait, il faudrait probablement hausser le critère minimum.

Utilisant la même notation que pour la mesure précédente, nous pouvons écrire la formule de la statistique TRV pour la classe (X^*, Y^*) (i,j)

$$\mathrm{TRV} = 2 \sum_k \{ (n_{ijk} + .5) ln((n_{ijk} + .5)/ (n_{ijk} + m_{ijk} + 1)) + (m_{ijk} + .5) ln((m_{ijk} + .5)/(n_{ijk} + m_{ijk} + 1)) \} + (4n_{ij} + 2K) ln2, \quad (6.5)$$

où

$$n_{ij} = \sum_k n_{ijk} = \sum_k m_{ijk}, \quad i = 1, \ldots, \mathrm{I}, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K. \quad (6.6)$$

Lorsque les m_{ijk} et les n_{ijk} proviennent de la même distribution multinomiale, la distribution asymptotique de cette statistique est de type chi carré avec $(K - 1)$ degrés de liberté. On obtient une valeur P globale en faisant la somme des statistiques et des degrés de liberté correspondants pour les classes (X^*, Y^*) qui répondent au critère minimum et en calculant la probabilité que la valeur d'une variable chi carré, étant donné le nombre de degrés de liberté approprié, soit plus grande que la valeur observée.

7. RÉSULTATS DE L'ÉTUDE DE MONTE CARLO

Dans cette section, nous présentons les résultats de l'étude de simulation. On trouvera une description plus complète de ces résultats dans Singh *et al.* (1990). Les personnes qui voudraient obtenir les tableaux de chiffres qui ont servi à l'élaboration des figures 1 à 5 peuvent en faire la demande.

Notre présentation n'insiste pas sur les erreurs types de Monte Carlo des mesures d'évaluation parce qu'elles sont en général relativement faibles; par exemple, les coefficients de variation pour $\overline{\text{AD-Cov}}$ étaient généralement inférieurs à deux pour cent. En outre, les évaluations de différentes méthodes d'appariement devraient normalement être corrélées positivement, de sorte que les erreurs types refléteront beaucoup plus précisément que l'on croit les différences relatives entres les méthodes. Un autre indice de la qualité de l'évaluation des diverses méthodes d'appariement par la méthode de Monte Carlo est la régularité générale des tendances observées (voir, par exemple, les figures 2 à 5). Bref, le moindre écart est susceptible de signaler une différence réelle.

7.1 Méthodes n'utilisant pas d'information supplémentaire (REG et HOD)

Les figures 2 à 5 montrent comment l'efficacité des méthodes d'appariement qui appliquent l'HIC est influencée lorsqu'on s'écarte de l'indépendance conditionnelle. Il semblerait que l'utilisation de méthodes qui appliquent l'HIC crée un biais notable pour la relation conjointe de (X, Y, Z) dans le fichier enrichi. On peut constater par exemple dans la figure 2 une détérioration progressive de la situation (sauf pour $\overline{\text{MAD-}Z}$ lorsque le coefficient vrai de corrélation conditionnelle, $\rho_{Y,Z|X}$, s'éloigne de zéro. Le cas d'exception que représente $\overline{\text{MAD-}Z}$ s'explique probablement par le caractère inconditionnel de cette mesure, qui repose sur une comparaison par unité des valeurs Z appariées et des valeurs Z supprimées, alors que les autres mesures reposent sur des comparaisons des distributions conditionnelles de Z. Il est intéressant de noter dans la figure 2 que lorsque la valeur vraie de $\rho_{Y,Z|X}$, est faible, la méthode HOD*, qui utilise de l'information supplémentaire, peut, en ce qui concerne les mesures d'évaluation de type nominal ou les mesures au niveau global, être moins efficace que la méthode HOD, qui, elle, n'utilise pas cette information. Les conditions dans lesquelles l'utilisation d'informations supplémentaires devient avantageuse dépendraient du degré de précision de ces informations.

7.2 Méthodes utilisant de l'information supplémentaire

Comme il fallait s'y attendre, les résultats de notre étude empirique confirment que l'utilisation d'information supplémentaire est un moyen de compenser la non-vérification de l'HIC. Le degré de compensation dépendrait de la méthode et du genre d'information supplémentaire utilisées. Dans l'introduction, nous avons brossé un tableau du rendement des diverses méthodes étudiées. Maintenant, nous allons examiner plus en détail le rendement de ces méthodes à l'aide des figures 2 à 5.

Parmi les méthodes fondées sur la régression, celles qui utilisent de l'information supplémentaire portant sur la corrélation conditionnelle – notamment REG* et REG. LOGLIN* – ont, pour des populations symétriques, un rendement très acceptable en ce qui concerne les mesures au niveau individuel (c.-à-d. $\overline{\text{MAD-}Z}$ et $\overline{\text{AD-Cov}}$) (voir figure 2). Elles surclassent aussi les méthodes hot-deck pour des populations asymétriques (figure 3), bien que le biais tende à s'accroître lorsque le degré d'asymétrie augmente. Toutefois, lorsque le coefficient de corrélation conditionnelle pour l'information supplémentaire substitutive varie (figure 4), les méthodes de régression ont un rendement inégal, c.-à-d. qu'elles peuvent être plus efficaces ou moins efficaces que les méthodes hot-deck au niveau individuel. De fait, pour ce qui a trait à la deuxième catégorie d'information supplémentaire substitutive (c'est-à-dire celle avec contamination log-normale; voir figure 5), les méthodes de régression sont généralement un peu moins efficaces que la méthode HOD.LOGLIN en ce qui concerne la mesure $\overline{\text{AD-Cov}}$). Si nous nous limitons aux méthodes fondées sur la régression, nous pouvons recommander l'usage de la méthode REG* en ce qui regarde les mesures d'évaluation au niveau individuel. Toutefois, pour ce qui a trait aux mesures au niveau global, toutes les méthodes de régression ont un très mauvais rendement. Cela s'explique probablement par l'effet de réduction à la moyenne dont il est question à la sous-section 7.3.

Méthodes d'appariement étudiées dans les figures 1 à 5

REG	valeur Z_{int} tirée de la régression de Z par rapport à X, valeur Z_{match} fondée sur la distance (X,Z)
REG*	valeur Z_{int} tirée de la régression de Z par rapport à X et à Y, valeur Z_{match} fondée sur la distance (X,Z)
REG.LOGLIN	valeur Z_{int} tirée de la régression de Z par rapport à X, valeur Z_{match} fondée sur la distance (X,Z) avec des contraintes nominales
REG.LOGLIN*	valeur Z_{int} tirée de la régression de Z par rapport à X et à Y, valeur Z_{match} fondée sur la distance (X,Z) avec des contraintes nominales
HOD	méthode hot-deck utilisant la distance X dans les classes X
HOD*	valeur Z_{int} tirée du fichier C à l'aide de la méthode hot-deck utilisant la distance Y, valeur Z_{match} tirée du fichier B à l'aide de la méthode hot-deck utilisant la distance (X,Z)
HOD.LOGLIN	méthode hot-deck utilisant la distance X dans les classes X avec contraintes nominales
HOD.LOGLIN*	valeur Z_{int} obtenue à l'aide de la méthode hot-deck utilisant la distance Y avec contraintes nominales, valeur Z_{match} obtenue à l'aide de la méthode hot-deck utilisant la distance (X,Z) dans les classes (X,Y,Z)

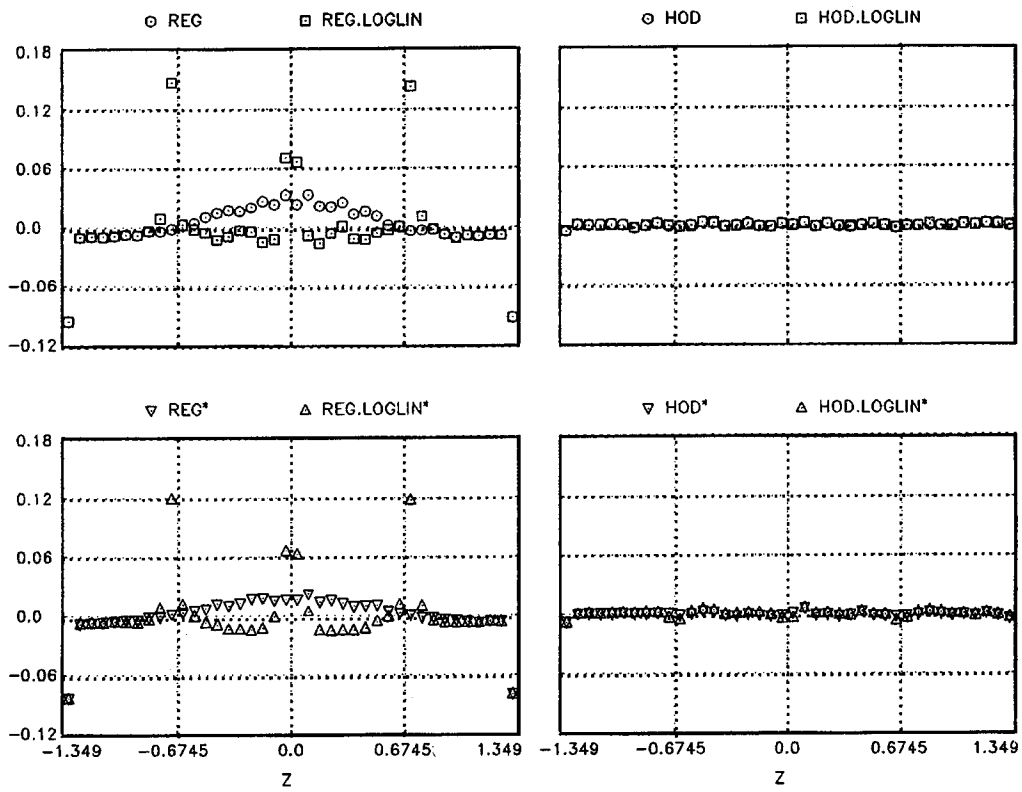

Figure 1. Différence entre les histogrammes marginaux des valeurs Z appariées et des valeurs Z supprimées (données symétriques, $\rho_{Y,Z|X} = .4$)

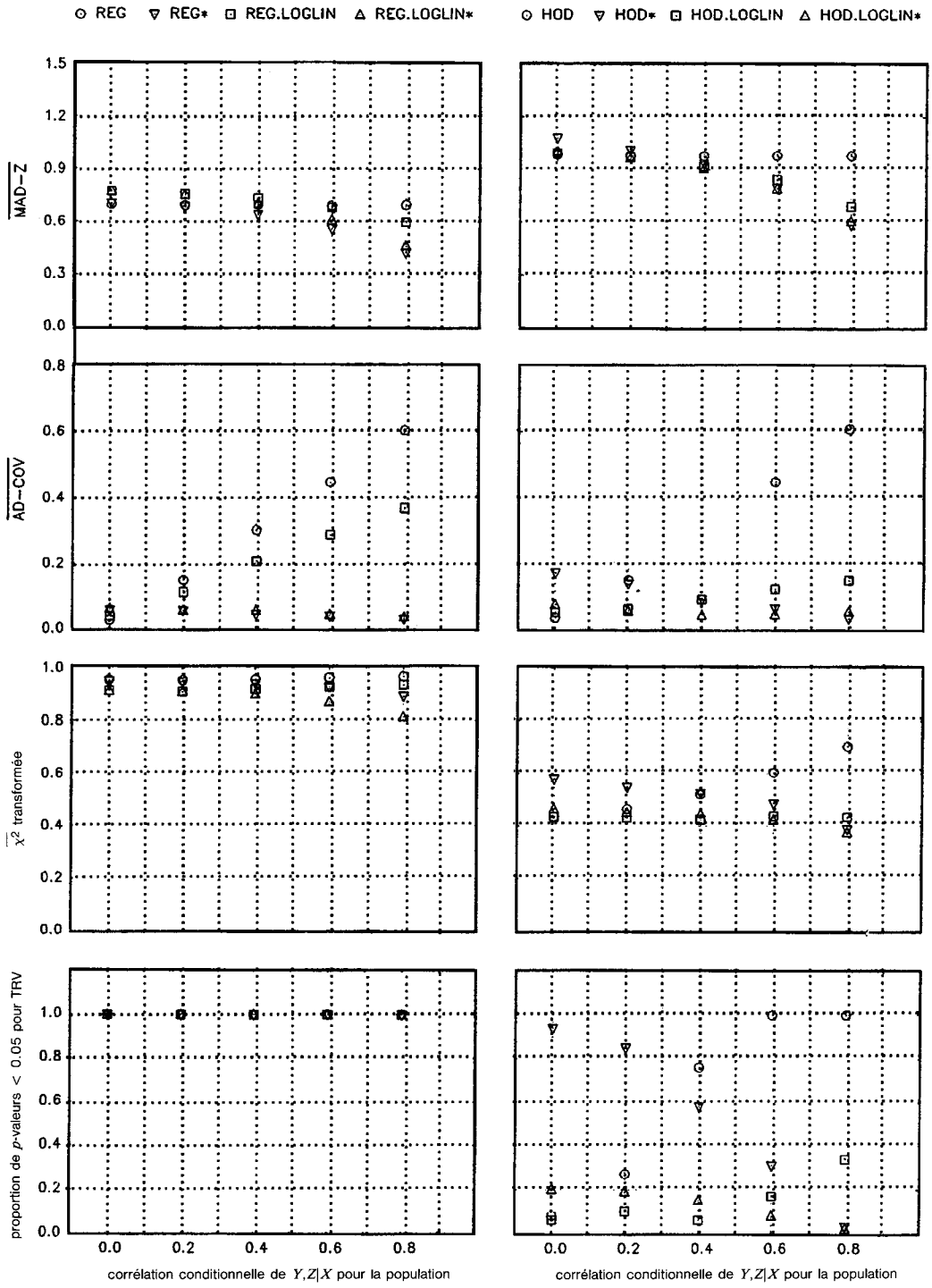

Figure 2. Comparaison de méthodes d'appariement statistique lorsque $\rho_{Y,Z|X}$ varie pour une population symétrique; information supplémentaire non substitutive

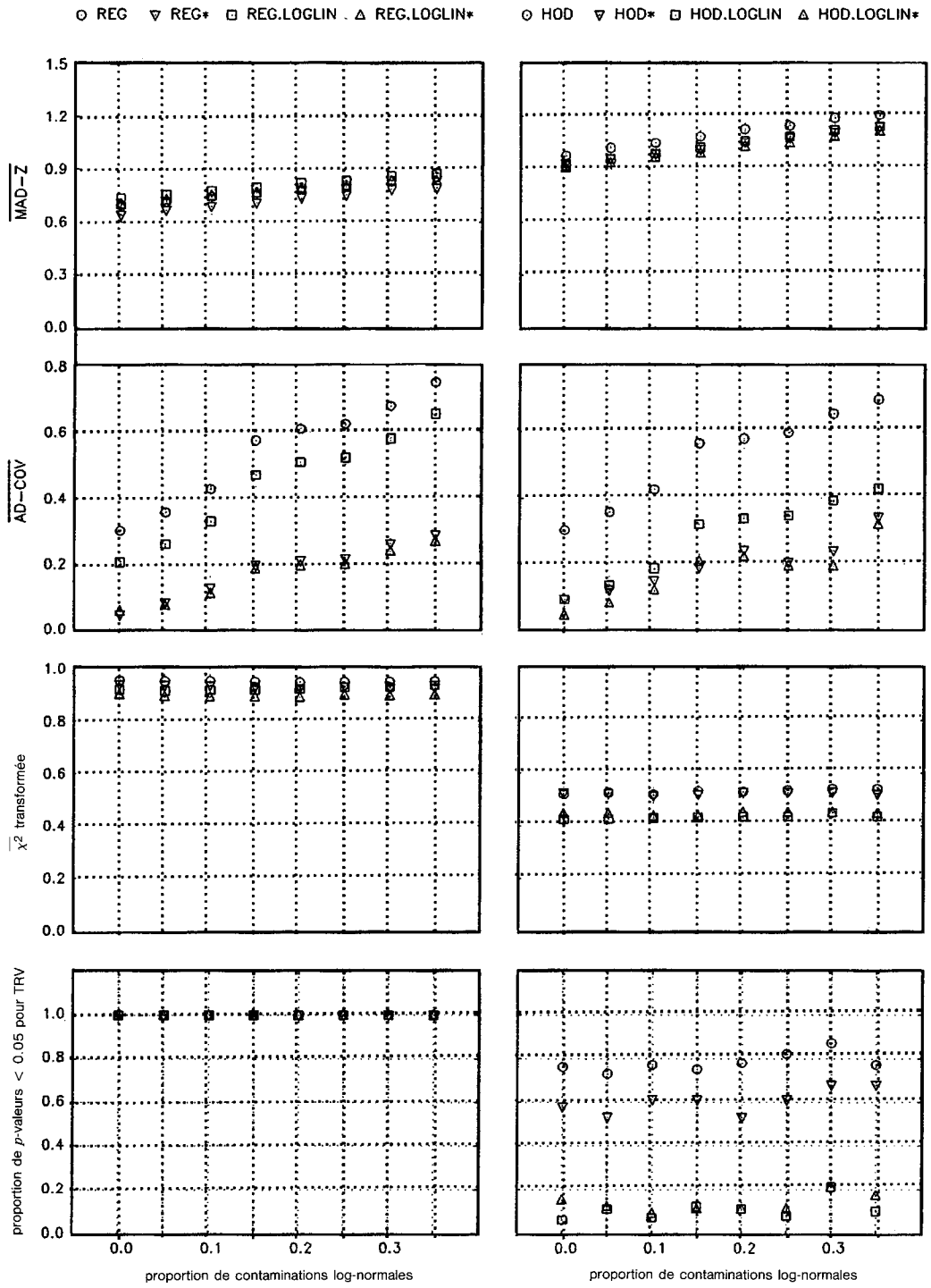

Figure 3. Comparaison de méthodes d'appariement statistique lorsque la proportion de contaminations log-normales varie ($\rho_{Y,Z|X}$ avant contamination); information supplémentaire non substitutive

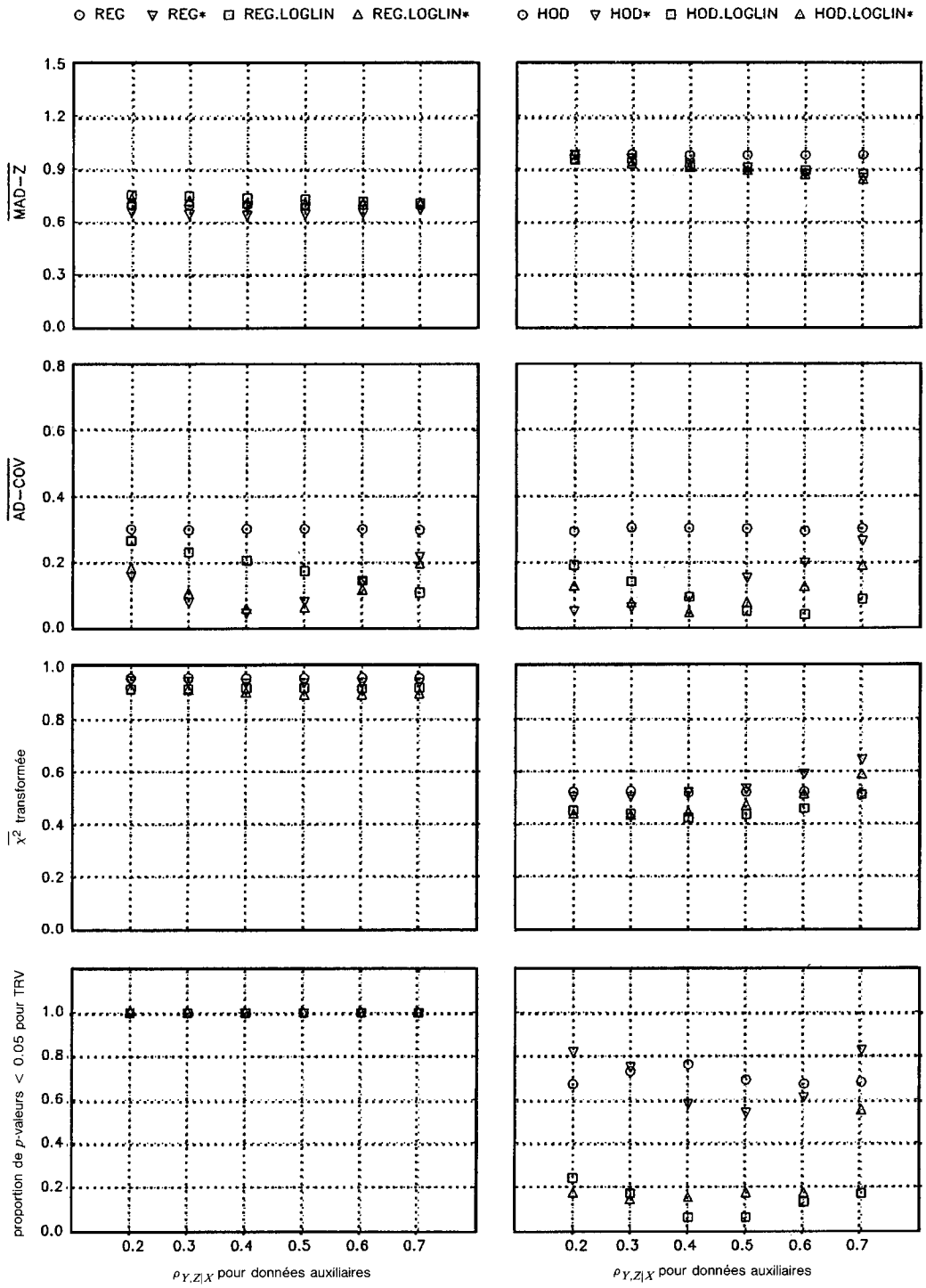

Figure 4. Comparaison de méthodes d'appariement statistique lorsque $\rho_{Y,Z|X}$ pour le fichier d'information supplémentaire C varie ($\rho_{Y,Z|X} = .4$ pour les fichiers A et B)

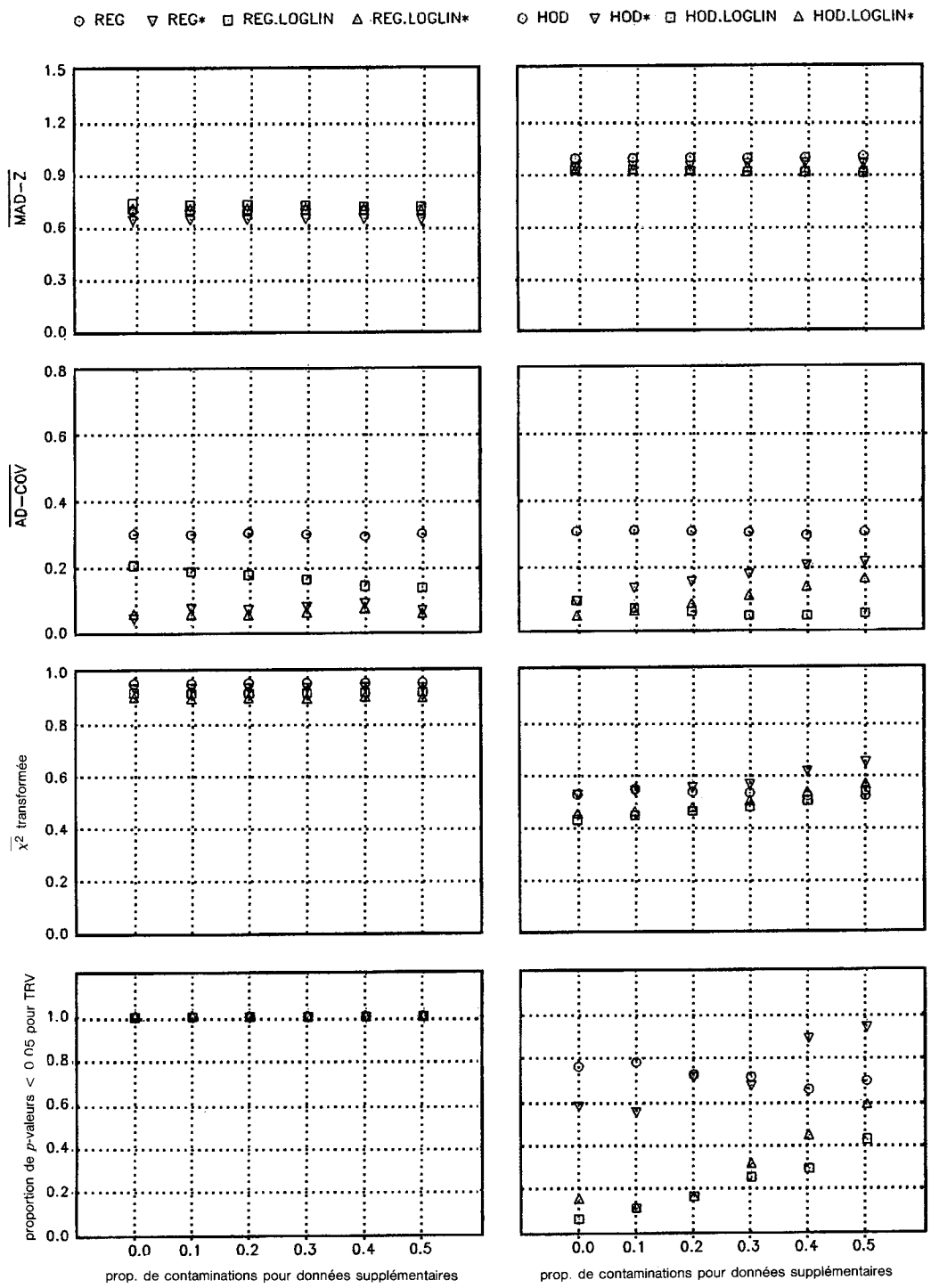

Figure 5. Comparaison de méthodes d'appariement statistique lorsque la proportion de contaminations log-normales pour le fichier C varie ($\rho_{Y,Z|X}$ = .4 avant contamination et absence de contamination log-normale pour les fichiers A et B)

En ce qui concerne les méthodes hot-deck avec information supplémentaire, les deux méthodes assujetties à des contraintes nominales - notamment HOD.LOGLIN et HOD.LOGLIN* - ont un rendement très acceptable au niveau global (c.-à-d. par rapport à χ^2 transformé et au TRV) pour tous les types de population; voir figures 2 à 5. De façon générale, la méthode HOD.LOGLIN est supérieure à la méthode HOD.LOGLIN*. Considérons maintenant les mesures d'évaluation au niveau individuel. Pour des populations symétriques et dissymétriques (figures 2 et 3), les méthodes HOD* et HOD.LOGLIN* sont, en règle générale, aussi efficaces l'une que l'autre et sont un peu plus efficaces que HOD.LOGLIN mais un peu moins efficaces que REG*. Cependant, lorsqu'il s'agit d'information supplémentaire substitutive (figures 4 et 5), la méthode HOD.LOGLIN peut être soit plus efficace ou moins efficace que les méthodes HOD.LOGLIN* et REG* en ce qui concerne la mesure $\overline{\text{AD-Cov}}$, et elle est souvent plus efficace lorsqu'il s'agit de données substitutives avec des contaminations log-normales. En outre, avec des données substitutives, HOD.LOGLIN tend à offrir assez de robustesse à l'égard des quatre mesures d'évaluation. Par conséquent, du point de vue du rendement global, nous pouvons recommander l'usage de la méthode HOD.LOGLIN* parmi toutes les méthodes hot-deck. Cependant, au point de vue pratique, il peut être préférable d'opter pour la méthode HOD.LOGLIN comme solution de compromis parce qu'elle est assez efficace au niveau individuel et extrêmement efficace au niveau global, qu'elle est beaucoup moins exigeante sur le plan du calcul et qu'elle montre de la robustesse à l'égard des données supplémentaires substitutives. En outre, HOD.LOGLIN ne nécessite pas d'information supplémentaire de micro-niveau.

7.3 Observations diverses

Dans cette sous-section, nous exposons quelques conclusions intéressantes qui sont ressorties de l'étude de Monte Carlo; les résultats empiriques relatifs à certaines de ces conclusions ne figurent pas ici mais sont présentés dans Singh *et al.* (1990).

i) Réduction à la moyenne

Une conclusion importante qui ressort de façon constante est que les méthodes d'appariement fondées sur la régression sont peu efficaces en ce qui concerne les mesures de type nominal. Ce manque d'efficacité peut s'expliquer par le phénomène de réduction à la moyenne, c'est-à-dire le phénomène selon lequel les valeurs Z appariées sont plus concentrées autour de leur moyenne que ne le sont les valeurs Z supprimées. C'est ce que l'on peut voir à la figure 1, qui montre la différence entre les histogrammes marginaux des valeurs Z appariées et des valeurs Z supprimées pour diverses méthodes d'appariement.

Les écarts positifs que l'on observe près du centre pour REG et REG* indiquent que le fichier enrichi compte un plus grand nombre de valeurs Z dans cette région que le fichier duquel ont été supprimées les valeurs vraies de Z. Les très fortes valeurs négatives observées aux extrémités des graphiques sont associées à des intervalles ouverts, et il semble fort probable que si ces intervalles avaient été divisés en de plus petits intervalles, on aurait observé plusieurs valeurs négatives moindres aux extrémités des graphiques, de sorte qu'on aurait pu déduire que les méthodes REG et REG* placent un trop grand nombre de valeurs Z au centre de la distribution par rapport aux extrémités.

La figure 1 permet de constater qu'il y a aussi une réduction à la moyenne dans le cas des méthodes REG. LOGLIN et REG.LOGLIN*. Toutefois, dans ce cas-ci la réduction est limitée par les contraintes nominales, de sorte que même si nous constatons que les extrémités de la distribution de Z pour le fichier enrichi ne sont pas assez étendues, les valeurs Z ne sont pas concentrées au centre de la distribution mais s'agglomèrent autour des limites d'intervalle, qui ont l'effet d'un mur. On peut expliquer de la même manière les fortes valeurs positives observées de part et d'autre de la limite centrale si l'on tient compte du fait que ces graphiques représentent en réalité une moyenne des différences entre histogrammes pour 100 simulations indépendantes. Il semble raisonnable de penser que si nous devions étudier séparément chacune des 100 différences d'histogrammes, nous observerions une forte valeur positive tantôt juste à gauche de la limite centrale, tantôt juste à droite, mais jamais des deux côtés à la fois.

La figure 1 montre aussi que l'effet de réduction à la moyenne et l'effet des limites sont négligeables pour les méthodes de type hot-deck.

ii) Information supplémentaire *(Y,Z)* *vs* information supplémentaire *(X,Y,Z)*

Bien que cet article ne présente que les résultats fondés sur l'information supplémentaire (Y,Z), l'étude de simulation considérait aussi, comme nous l'avons dit dans la section 6, l'information supplémentaire (X,Y,Z). Nous avons constaté avec intérêt que dans le cas des méthodes HOD.LOGLIN et HOD.LOGLIN*, l'utilisation d'informations supplémentaires (Y,Z) produit, en règle générale, de meilleurs résultats au niveau global que l'utilisation d'informations (X,Y,Z). Cette constatation ne semble pas s'appliquer à la méthode HOD*. Ce phénomène s'explique probablement par la variabilité de l'estimation des effets des facteurs (X^*,Y^*,Z^*) qui entrent dans les contraintes nominales, cette variabilité étant le fait d'un trop petit volume de données supplémentaires. Nous pouvons en conclure que les valeurs vraies étaient probablement voisines de zéro et que les assimiler à la valeur nulle donne de meilleurs résultats. C'est ce qui nous porte à croire qu'il faudrait tenir compte dans les analyses futures de l'effet de la taille de l'échantillon sur le rendement des méthodes d'appariement. Les considérations ci-dessus suggèrent aussi une nouvelle catégorie intéressante de méthodes qui combineraient l'information supplémentaire de micro-niveau (X,Y,Z) qui sert à déterminer les valeurs Z_{int} avec la distribution de variables nominales (Y^*,Z^*) tirée du fichier C qui sert à définir des contraintes. Cependant, ces méthodes n'ont pas été considérées dans la présente étude.

iii) Comparaison de différentes versions de la méthode hot-deck

Dans toutes les méthodes d'appariement étudiées, sauf HOD et HOD.LOGLIN, la deuxième étape, qui consiste à déterminer la valeur Z_{match}, fait intervenir l'imputation hot-deck avec la distance (X,Z). Pour les deux cas d'exception, on utilise la distance X. On peut aussi utiliser (pour d'autres méthodes que HOD et HOD.LOGLIN) la distance Z ou la distance (X,Y,Z). Dans ce dernier cas, il faut voir tout d'abord à introduire la valeur Y_{int} dans le fichier B. C'est ce qui a été fait dans l'étude de simulation originale, sauf que les résultats empiriques pertinents ne sont pas présentés ici. Les résultats ont montré que le choix de la distance change peu de choses; pour les méthodes REG et REG*, la distance (X,Z) donnait parfois des résultats supérieurs en ce qui a trait à la mesure $\overline{\text{AD-Cov}}$. C'est ce qui nous a fait opter pour la distance (X,Z) pour les méthodes étudiées ici. Cependant, dans la pratique, il serait préférable d'utiliser la distance Z pour les méthodes hot-deck à cause de la simplicité des calculs.

Par ailleurs, notons qu'en ce qui concerne les méthodes HOD et HOD.LOGLIN, on a la possibilité d'utiliser à l'étape 2 la version aléatoire de la méthode hot-deck ou la version fondée sur le rang, plutôt que celle fondée sur la distance X. Selon la version fondée sur le rang, les enregistrements des fichiers A et B sont ordonnés séparément en fonction de la valeur de X, puis ils sont appariés selon le rang. C'est la procédure qu'a proposée G. Rowe pour l'application relative à la BDSPS évoquée dans l'introduction. De toute évidence, cette méthode ne convient qu'aux variables X unidimensionnelles. Un avantage de la méthode fondée sur le rang est qu'un enregistrement du fichier B ne peut servir de donneur pour plus d'un enregistrement du fichier A. Les trois versions de hot-deck mentionnées ci-dessus étaient incluses dans l'étude de Monte Carlo mais seuls les résultats relatifs à la version fondée sur la distance X sont présentés ici. Notre étude a permis de constater que le choix de la version n'a pas beaucoup d'incidences en règle générale. Nous avons choisi la version fondée sur la distance X pour les méthodes HOD et HOD.LOGLIN parce qu'elle était conforme à la version de la hot-deck utilisée pour les autres méthodes (c.-à-d. celle fondée sur la distance). En pratique, la version léatoire serait la moins exigeante sur le plan du calcul; cependant, dans une application réelle, nous ne pourrions pas mesurer les inconvénients rattachés à l'utilisation de la méthode aléatoire par opposition à l'utilisation de la méthode fondée sur le rang ou la distance et pourtant, nous voudrions probablement disposer du plus de renseignements possible.

8. CONCLUSIONS

Cet article nous a permis de traiter le problème de l'utilisation d'informations supplémentaires pour l'appariement statistique. Nous avons vu que les deux principales méthodes d'appariement proposées antérieurement venaient de Rubin (1986) et de Paass (1986); des versions de ces méthodes sont désignées par les sigles REG* et HOD* respectivement. Nous avons proposé des versions modifiées de ces méthodes, désignées par REG.LOGLIN* et HOD.LOGLIN*, en établissant des contraintes nominales à partir d'informations supplémentaires. Ces deux dernières versions se ramènent à REG.LOGLIN et à HOD.LOGLIN si l'on ne dispose que d'informations supplémentaires de type nominal. En l'absence d'information supplémentaire, on utilise les méthodes d'imputation habituelles, REG et HOD. Nous avons fait une étude empirique pour évaluer le rendement de huit méthodes d'appariement statistique par rapport à quatre mesures d'évaluation (deux au niveau individuel et deux au niveau global). L'étude a permis de constater que lorsqu'il n'y a pas d'information supplémentaire, la méthode HOD est préférable. Cependant, lorsqu'on dispose d'information supplémentaire, la situation est plus complexe. Si l'on juge importantes seulement les mesures d'évaluation au niveau individuel, alors la méthode REG* est recommandée. Si les mesures au niveau global sont elles aussi jugées importantes, la méthode HOD.LOGLIN* est recommandée si l'information supplémentaire est non substitutive. Par ailleurs, la méthode HOD.LOGLIN est une bonne solution de compromis si le nombre de calculs à effectuer est une source de préoccupation majeure ou si l'on croit que l'information supplémentaire peut être substitutive. Si les mesures au niveau individuel sont moins importantes ou qu'elles ne présentent aucun intérêt (cela peut être fréquent parce que les données appariées sont souvent présentées en tableaux dans la pratique), alors HOD.LOGLIN sera la méthode recommandée. Compte tenu du rendement équivalent des trois versions de la méthode hot-deck (celles fondées sur la distance et le rang et la version aléatoire) pour les méthodes HOD et HOD.LOGLIN, on pourrait envisager d'utiliser la version aléatoire dans la pratique en raison de la simplicité des calculs.

Il convient de souligner que nous n'avons pas considéré la version entièrement itérative de la méthode de Paass. Il serait intéressant d'examiner dans des analyses futures le rendement de cette version. Une autre question qui mérite d'être étudiée est l'application de contraintes nominales avec de nombreuses variables. L'application de l'algorithme de "balayage" pourrait comporter un nombre exorbitant de calculs dans ce cas. À cet égard, les résultats de l'étude de Paass (1989) devraient être utiles.

Dans l'étude que nous venons d'exposer, nous n'avons pas fait varier systématiquement le degré d'exactitude de la source de données supplémentaires à cause de limites imposées par le calcul; autrement dit, nous n'avons pas fait varier la taille du fichier C, ni d'ailleurs celle des fichiers A et B. Une question intéressante aurait été de savoir comment la taille de ces fichiers influe sur le rendement des diverses méthodes.

Enfin, soulignons que même si les résultats de cette étude reposent sur des données fictives (ce qui était nécessaire pour créer des scénarios qui simulent des données réelles), nous croyons qu'ils seraient pertinents pour des applications concrètes. De toute évidence, il serait intéressant et utile de faire une étude de simulation avec

des données réelles pour vérifier la solidité des résultats obtenus avec des données fictives et pour voir quel genre d'incidence peuvent avoir les biais contenus dans la distribution conjointe du fichier enrichi. On peut aussi chercher à savoir comment prendre en compte ces biais dans des inférences qui reposent sur le fichier enrichi, c'est-à-dire comment produire des mesures d'incertitude pour les estimations de paramètres tirées du fichier enrichi qui reflètent non seulement la variabilité dans ce fichier, mais aussi la variabilité inhérente à la procédure d'appariement. Bien que nous ne puissions répondre à cette question, il ne fait aucun doute que les méthodes d'appariement qui utilisent de l'information supplémentaire accroîtront l'utilité générale du fichier enrichi. Nous nous pencherons sur ces questions et sur d'autres questions connexes dans des études ultérieures.

REMERCIEMENTS

Les auteurs tiennent à remercier J. Armstrong, G. Gray, G. Hole, D. Royce et M. Wolfson pour leurs précieux commentaires. La recherche faite par le premier auteur a été rendue possible en partie grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada accordée à l'Université Carleton.

BIBLIOGRAPHIE

ARMSTRONG, J. (1989). An evaluation of statistical matching methods. Direction de la méthodologie, document de travail, BSMD, 90-003E. Statistique Canada.

BARR, R.S., et TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Surveys. Rapport technique, U.S. Department of the Treasury, Office of Tax Analysis.

BARR, R.S., et TURNER, J.S. (1990). Quality issues and evidence in statistical file merging. Dans *Data Quality Control: Theory and Pragmatics* (Éds. G.E. Liepins et V.R.R. Uppuluri). New York: Marcel Dekker, 245-313.

BARR, R.S., STEWART, W.H., et TURNER, J.S. (1981). An empirical evaluation of statistical matching methodologies. Rapport technique, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

BISHOP, Y.M.M., FIENBERG, S.E., et HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass: MIT Press.

BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.

BUDD, E.C., et RADNER, D.B. (1969). The OBE size distribution series: methods and tentative results for 1964. *American Economic Review, Papers and Proceedings*, LIX, 435-449.

COHEN, M.L. (1991). Statistical matching and microsimulation models. Dans *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Volume II, Technical Papers, (Éds. C.F. Citro et E.A. Hanushek). Washington, D.C.: National Academy Press, 62-85.

FELLEGI, I.P. (1977). Discussion paper. *Proceedings of the Section on Social Statistics, American Statistical Association*, 762-764.

FORD, B.L. (1983). An overview of hot-deck procedures. Dans *Incomplete Data in Sample Surveys*, (Vol. 2), (Éds. W.G. Madow, I. Olkin et D.B. Rubin). New York: Academic Press, 185-207.

KADANE, J.B. (1978). Some statistical problems in merging data files. Dans *1978 Compendium of Tax Research*. Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.

PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. Dans *Microanalytic Simulation Models to Support Social and Financial Policy* (Éds. G.H. Orcutt, J. Merz et H. Quinke). Amsterdam: Elsevier Science.

PAASS, G. (1989). Stochastic generation of a synthetic sample from marginal information. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 431-445.

PAASS, G., et WAUSCHKUHN, U. (1980). Experimentelle erprobung und vergleichende Bewertung statistischer Matchverfahren. Rapport interne, IPES.80.201, St. Augustin, *Gesellschaft für Mathematik und Datenverarbeitung*.

PURCELL, N.J., et KISH, L. (1980). Postcensal estimates for local areas (or Domains). *Revue Internationale de Statistique*, 48,3-18.

RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.

RODGERS, W.L., et DeVOL, E. (1982). An evaluation of statistical matching. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132.

RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.

RUGGLES, N., RUGGLES, R., et WOLFF, E. (1977). Merging microdata: Rationale, practice and testing. *Annals of Economic and Social Measurement*, 6, 407-428.

SCHEUREN, F.J. (1989). Comment on Wolfson *et al.* (1989). *Survey of Current Business*, 69, 40-41.

SIMS, C.A. (1972). Comment on Okner (1972). *Annals of Economic and Social Measurement*, 1, 343-345.

SIMS, C.A. (1978). Comment on Kadane (1978). Dans *1978 Compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.

SINGH, A.C. (1988). Log-linear imputation. Direction de la méthodologie, document de travail, SSMD, 90-016E. Statistique Canada; publié aussi dans *Proceedings of the Fifth Annual Research Conference* , U.S. Bureau of the Census, 118-132.

SINGH, A.C., ARMSTRONG, J.B., et LEMAÎTRE, G.E. (1988). Statistical matching using log linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.

SINGH, A.C., MANTEL, H., KINACK, M., et ROWE, G. (1990). On methods of statistical matching with and without auxiliary information: Some modifications and an empirical evaluation. Direction de la méthodologie, document de travail, SSMD, 90-016E. Statistique Canada.

U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Washington, D.C.: Federal Committee on Statistical Methodology.

WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B., et ROWE, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Receuil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada, Ottawa, (Éds. J.W. Coombs et M.P. Singh), 201-229; une autre version publiée dans *Survey of Current Business* (1989), 69, 36-40.