Techniques d'enquête, juin 1993
Vol. 19, n° 1, pp. 15–33
Statistique Canada

Évaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle

THOMAS R. BELIN[1]

RÉSUMÉ

Le couplage d'enregistrements désigne une technique algorithmique qui sert à identifier des paires d'enregistrements ayant trait au même individu dans des fichiers distincts. Dans cet article, nous étudions un modèle destiné à évaluer les sources de variation dans le couplage d'enregistrements en comparant ce procédé à une "boîte noire" qui reçoit des données d'entrée et restitue un produit (un ensemble de concordances désignées, c.-à-d. de paires dont les éléments représentent la même entité) qui a certaines caractéristiques. Nous illustrons nos propos au moyen d'une expérience factorielle dans laquelle nous nous servons de données du recensement et d'enquêtes postcensitaires afin d'évaluer l'influence de divers facteurs qui sont réputés pour réduire la fiabilité du procédé. Avec ce cadre expérimental, l'évaluation du couplage d'enregistrements devient un problème statistique comme les autres. L'étude permet de répondre à plusieurs questions de recherche et nous prétendons qu'il est essentiel de recourir à des méthodes expérimentales comme celle proposée ici si l'on veut mieux comprendre les sources d'erreur qui interviennent dans les techniques de couplage d'enregistrements.

MOTS CLÉS: Poids limite; taux d'erreur d'appariement; algorithme de Fellegi-Sunter; variables d'appariement; enquête postcensitaire; comparaison de chaînes de caractères; méthode de pondération.

1. ÉVALUATION DES MÉTHODES DE COUPLAGE D'ENREGISTREMENTS

Le couplage d'enregistrements désigne une technique algorithmique qui sert à distinguer les paires d'enregistrements dont les éléments se rapportent au même individu mais proviennent de fichiers distincts. Le but de l'opération est de définir par une méthode informatique les paires d'enregistrements de fichiers différents qui devraient être désignées comme des "concordances" et celles qui devraient être désignées comme des "non-concordances" et ce, avec un taux d'erreur raisonnable. Ainsi, on évite les coûts d'un traitement manuel.

Pour définir une méthode de couplage d'enregistrements, il faut avoir une méthode qui permet de mesurer le degré de concordance d'enregistrements ainsi qu'une règle qui permet d'identifier des paires d'enregistrements comme des concordances ou des non-concordances. Dans les ouvrages portant sur le couplage d'enregistrements, on a beaucoup parlé du problème qui peut se poser lorsqu'on attribue des "poids" à des zones d'information dans un enregistrement à variables multiples pour obtenir un "poids composite" qui résume le degré de concordance de deux entités (par exemple, Newcombe *et al.* 1959; Fellegi et Sunter 1969; Newcombe 1988; Copas et Hilton 1990). On a moins parlé d'autres aspects du couplage d'enregistrements, comme la manière de traiter les cas de concordance imparfaite de zones d'information, et des conséquences de l'utilisation combinée de diverses approches (traitements).

Dans certaines circonstances, un identificateur personnel comme le numéro de sécurité sociale pourra servir de critère pour le couplage. Or, on ne connaît pas toujours ces identificateurs, et même si on les connaissait, il faudrait probablement s'appuyer sur d'autres données repères pour un grand sous-ensemble de cas (voir, par exemple, Rogot, Sorlie et Johnson 1986).

Cet article décrit une vaste expérience factorielle dans laquelle nous avons comparé diverses méthodes permettant d'apparier des enregistrements du recensement à des enregistrements d'une enquête postcensitaire (EP). Comme le numéro de sécurité sociale n'est pas demandé dans un recensement, nous sommes devant une situation où l'évaluation du degré de concordance repose sur plusieurs variables. Nous cherchons à répondre à deux questions en particulier:

(1) Quels sont les principaux facteurs qui influent sur la précision du couplage d'enregistrements?

(2) Quelle combinaison de facteurs donne les meilleurs résultats dans la pratique?

Au delà du fait que cette étude tente de répondre à ces questions par une comparaison du recensement et de l'enquête postcensitaire, son plus grand mérite est probablement de faire valoir que l'analyse des méthodes de couplage d'enregistrements doit se faire au moyen d'expériences menées soigneusement. Si de nombreux facteurs sont laissés sous le contrôle de l'opérateur du programme, il y a peu de chances que l'on comprenne toute la complexité d'un algorithme d'appariement en faisant varier un seul facteur à la fois (ou pire encore, en ne faisant aucune

[1] Thomas R. Belin, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA, 90024-1766, U.S.A.

évaluation systématique). L'idée de réaliser une expérience paraîtrait tout à fait naturelle pour un scientifique agricole ou un ingénieur du contrôle de la qualité, mais il semble qu'elle ne soit pas retenue en ce qui concerne le couplage d'enregistrements, si l'on fait abstraction de la présente étude et de certains de nos ouvrages antérieurs (Belin 1989a, 1989b).

2. DOMAINES D'APPLICATION DU COUPLAGE D'ENREGISTREMENTS

2.1 Applications pour le couplage d'enregistrements

Les méthodes de couplage d'enregistrements sont utilisées dans diverses circonstances. On peut diviser les applications en deux grandes catégories: 1) celles où on cherche à faire des inférences sur les relations entre des variables provenant de grands fichiers différents et 2) celles où on s'intéresse spécialement au nombre d'individus qui sont représentés dans un fichier ou les deux fichiers (ou à une fonction de ces quantités).

Les exemples d'applications de la première catégorie abondent. En effet, citons les études où l'on rapproche les données d'enquêtes de santé et de nutrition avec les données des registres des décès afin d'étudier le rapport entre les facteurs de risque alimentaires et les décès attribuables à diverses causes (Johansen 1986), les études où l'on rapproche des données de l'enquête sur la population active avec des données sur la mortalité afin d'évaluer les effets de l'exploitation des gîtes d'uranium sur la santé (Newcombe, Smith, Howe, Mingay, Strugnell et Abbatt 1983; Abbatt 1986), les études où l'on rapproche des données sur la formation scolaire d'individus avec des données sur les traitements que reçoivent ces individus afin d'évaluer les avantages d'une formation collégiale (Fagerlind 1975), les études où l'on compare le revenu indiqué dans les dossiers de l'aide sociale avec celui indiqué dans les dossiers d'impôt (Kershaw et Fair 1979), et les études où l'on rapproche les dossiers d'individus exposés à des radiations lors de tests nucléaires et les dossiers d'une cohorte d'individus-témoins avec les registres nationaux des décès afin d'évaluer les différences de régime de mortalité entre les individus exposés et les individus-témoins (Dulberg Spasoff et Raman 1986). Dans toutes ces études, l'utilisation de méthodes de couplage d'enregistrements présente de l'intérêt surtout à cause du coût relativement peu élevé de ces méthodes et de leur rapidité d'exécution, puisqu'il est beaucoup plus long et beaucoup plus coûteux de réaliser les mêmes études avec une ou plusieurs étapes de suivi que de se servir des données existantes.

Le principal exemple que nous présentons dans cet article est représentatif des applications de la seconde catégorie, qui visent à déterminer le nombre de cas communs à deux fichiers de données. Dans l'exemple en question, on utilise une méthode de couplage d'enregistrements dans la première étape d'une opération d'appariement de grande envergure qui permet de comparer les enregistrements d'un recensement aux enregistrements d'une enquête postcensitaire (EP) générale exécutée à la suite du recensement afin d'évaluer le taux de couverture de ce recensement. Comme autres exemples d'application de la seconde catégorie, mentionnons l'étude de Nicholl (1986) sur les erreurs de classification concernant les genres de blessures subies par les victimes d'accidents de la circulation (étude fondée sur le rapprochement de dossiers d'hôpitaux et de rapports de police), l'étude de Johnson (1991) sur le volume de travail des procureurs du ministère de la justice des É.-U. dans différents districts (étude fondée sur le couplage d'une liste de cas dressée par le ministère de la justice et d'une liste de cas dressée par les tribunaux fédéraux de première instance), ainsi que diverses études portant sur l'exactitude et le champ d'observation des fichiers de données sur la mortalité (Wentworth *et al.* 1983; Curb *et al.* 1985; Boyle et Decouflé 1990; Williams *et al.* 1992).

L'estimation du sous-dénombrement dans le recensement a été un sujet notable et, par moments, controversé en recherche statistique, surtout durant la dernière décennie. La controverse porte essentiellement sur une proposition voulant que l'on redresse les chiffres du recensement en se fondant sur des estimations du sous-dénombrement tirées d'une EP (enquête postcensitaire). Pour se documenter sur la question, le lecteur se référera aux ouvrages suivants: Ericksen et Kadane (1985), Citro et Cohen (1985), Freedman et Navidi (1986), Wolter (1986), Schirm et Preston (1987), Ericksen, Kadane et Tukey (1989), Cohen (1990) ainsi que les sections spéciales sur l'erreur de couverture dans le recensement des numéros de juin et décembre 1988 de cette revue. Le couplage d'enregistrements est la première étape dans l'appariement de données du recensement et de données de l'EP; cette étape est suivie d'un appariement d'enregistrements effectué par des commis, puis d'une interview de rappel menée auprès de ménages lorsqu'il semble y avoir des divergences entre les données du recensement et celles de l'EP et, enfin, d'une autre séance d'appariement effectué par des commis. En s'appuyant sur des rapports de l'opération d'appariement et certaines hypothèses relatives à la probabilité qu'un individu figure uniquement dans les enregistrements du recensement, uniquement dans ceux de l'EP, dans les deux à la fois ou ni dans l'un ni dans l'autre, on peut estimer les taux de sous-dénombrement (ou de surdénombrement) dans le recensement.

2.2 Généralités sur la théorie du couplage d'enregistrements

Le développement de l'approche probabiliste dans la théorie du couplage d'enregistrements remonte à Newcombe, Kennedy, Axford et James (1959), qui ont élaboré une méthode de pondération dans le but de représenter la probabilité que deux enregistrements concordent l'un avec l'autre. Fellegi et Sunter (1969) approfondissent les fondements théoriques des règles de pondération courantes et font remarquer que la méthode proposée par Newcombe *et al.* équivaut à calculer un rapport de vraisemblance suivant un modèle simple pour couplage

d'enregistrements qui présuppose une relation de concordance indépendante pour chacune des zones de données d'un enregistrement. Fellegi et Sunter montrent que si on utilise une méthode de pondération semblable à celle de Newcombe *et al.* avec des poids limites qui dépendent d'un certain taux d'erreur dans les concordances et d'un certain taux d'erreur dans les non-concordances, on se trouve à définir une méthode de couplage qui est optimale en ce sens qu'elle réduit au maximum la proportion d'enregistrements qui ne peuvent être classés ni dans les concordances désignées ni dans les non-concordances désignées, cela en supposant que le modèle utilisé est valide.

Par la suite, le développement des méthodes de couplage d'enregistrements s'est surtout fait à la faveur d'applications, les chercheurs mettant en pratique les théories exposées dans les ouvrages antérieurs. Parmi les applications les plus notables, mentionnons la Oxford Record Linkage Study (Acheson 1967; Goldacre 1986), l'opération d'appariement effectuée avec des enregistrements de trois sources: la Current Population Survey, la Social Security Administration et l'Internal Revenue Service (Kilss et Scheuren 1978), et la National Longitudinal Mortality Study (Rogot, Sorlie, Johnson, Glover et Treasure 1988). Les actes de colloques portant sur le couplage d'enregistrements (Kilss et Alvey 1985; Howe et Spasoff 1986; Carpenter et Fair 1990), les recueils de symposiums annuels (Kilss et Alvey 1984a; Kilss et Alvey 1984b; Kilss et Alvey 1984c; Kilss et Alvey 1987; Kilss et Jamerson 1990), de même que les actes de colloques portant plus spécialement sur les utilisations des données administratives (Coombs et Singh 1988) décrivent de nombreuses autres applications où l'on utilise des méthodes de couplage d'enregistrements.

L'élaboration de logiciels a fourni de nouveaux outils pour poursuivre les recherches dans le domaine du couplage d'enregistrements. À Statistique Canada comme au Bureau of the Census des É.-U., on a élaboré pour des applications variées des logiciels qui apportent des améliorations aux méthodes de pondération et de groupage. Howe et Lindsay (1981) font état du "Système itératif général de chaînage d'articles" (SIGCA) de Statistique Canada; les ouvrages de Hill (1981) et de Hill et Pring-Mill (1986) portent aussi sur ce système. Du côté américain, les ouvrages de Jaro (1989), Winkler (1989) et Winkler et Thibaudeau (1992) font état du système d'appariement mis au point par la division du couplage d'enregistrements du Bureau of the Census, tandis que les deux ouvrages de Laplant (1988 et 1989) et celui de Winkler (1991) renferment de la documentation sur ce système.

Copas et Hilton (1990) présentent de nouveaux modèles qui reflètent les particularités des fichiers de données qui peuvent servir à l'élaboration d'une méthode de pondération probabiliste. Newcombe, Fair et Lalonde (1992) décrivent d'autres méthodes de couplage conçues pour tirer profit de l'information contenue dans les noms de personnes. Comme ouvrages de référence sur le couplage d'enregistrements, mentionnons l'article de synthèse de Jabine et Scheuren (1986), le manuel de Newcombe (1988) ainsi que l'ouvrage édité par Baldwin, Acheson et Graham (1987).

2.3 Étapes d'une méthode type de couplage d'enregistrements

Les étapes habituelles d'une méthode de couplage d'enregistrements sont les suivantes: 1) collecte de données, 2) prétraitement des données, 3) définition des règles permettant d'évaluer le degré de concordance de deux éléments susceptibles de former la même paire, 4) rapprochement d'éléments susceptibles de former la même paire et 5) confirmation de la concordance. Nous employons l'expression "éléments susceptibles de former la même paire" pour décrire des couples d'enregistrements qui ont des chances de constituer la meilleure concordance possible parmi tous les éléments de leurs fichiers respectifs (voir les termes "hits" dans Rogot, Sorlie et Johnson (1986), "pairs" dans Winkler (1989) et "assigned pairs" dans Jaro (1989)). Il se peut que des éléments susceptibles de former la même paire soient confirmés dans leur concordance à la suite de l'application d'une règle de décision à l'étape (5), mais ils ne le seront pas nécessairement par la règle proprement dite.

Comme nous l'avons mentionné plus tôt, dans beaucoup de méthodes de couplage, on évalue le degré de concordance de deux éléments susceptibles de former la même paire au moyen d'une statistique globale unidimensionnelle souvent appelée "poids composite". Dans de telles méthodes, l'étape (3) mentionnée ci-dessus consisterait dans la définition de règles de pondération et l'étape (5) consisterait à fixer un poids limite au-delà duquel des paires d'enregistrements seraient reconnues comme des concordances.

Le couplage d'enregistrements peut être envisagé comme un problème décisionnel où l'ordinateur a deux possibilités d'action ou plus. En règle générale, on parle de trois possibilités d'action (par ex., reconnaître une concordance, reconnaître une non-concordance ou confier les enregistrements à un observateur pour que celui en fasse un examen plus profond, comme dans Fellegi et Sunter 1969), bien que parfois seulement deux possibilités soient envisagées (reconnaître une concordance, reconnaître une non-concordance). Il y a même des cas où le nombre de possibilités d'action atteint cinq (Tepping 1968).

Le fait de postuler que la distance entre des enregistrements multidimensionnels peut être résumée par un poids composite unidimensionnel réduit l'éventail de méthodes qui peuvent servir au couplage d'enregistrements. Nous connaissons très peu d'études où l'on examine d'autres voies que celle évoquée ci-dessus, si nous faisons abstraction de la simple définition d'un ensemble déterministe de règles qui permettent de dire si une paire d'enregistrements constitue une concordance. L'étude de Smith et Newcombe (1975) est l'une des rares où, effectivement, denouvelles voies sont envisagées. Cependant, toute cette question déborde le cadre de notre article.

2.4 Description détaillée de la méthode servant à apparier les enregistrements du recensement à ceux de l'EP

Diverses techniques peuvent être incluses dans chacune des cinq étapes mentionnées plus haut. La figure 1 donne

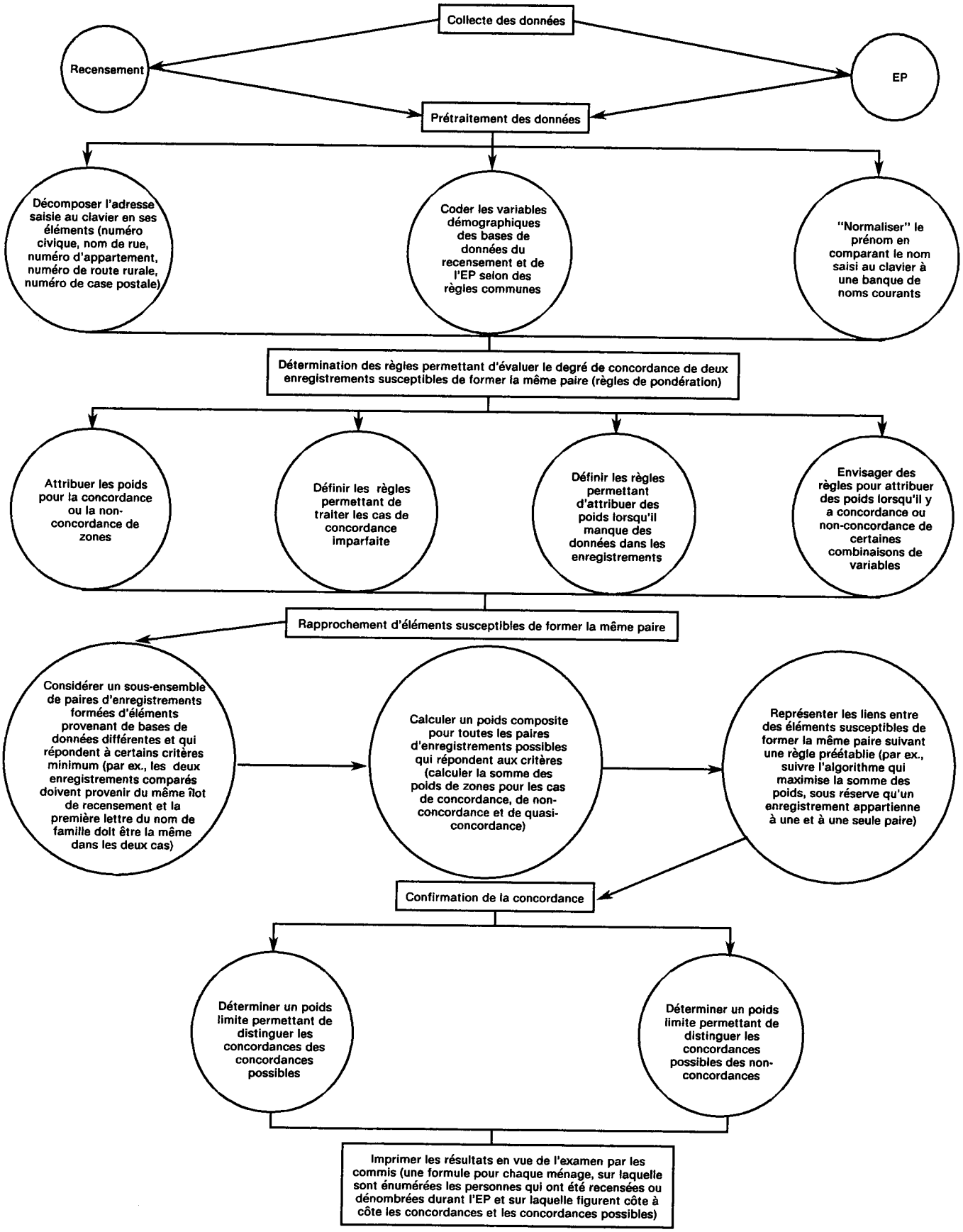

Figure 1. Étapes de la méthode de couplage des enregistrements du recensement avec ceux de l'EP

une représentation graphique des principales étapes du couplage des enregistrements du recensement et de l'EP.

La base du recensement est constituée de listes d'adresses d'unités de logement. Ces listes sont formées à l'aide de diverses méthodes, celles-ci variant généralement selon qu'il s'agit d'une région urbaine ou d'une région rurale. En ce qui concerne les régions urbaines et suburbaines, on envoie les questionnaires du recensement par la poste en espérant que les ménages les retourneront dûment remplis; pour ce qui est des autres régions, des recenseurs se rendent à domicile. Lorsqu'un ménage ne retourne pas le questionnaire qu'il a reçu par la poste, un recenseur se rend au domicile de ce ménage. Les données sont versées dans les fichiers informatiques du Census Bureau par une combinaison de techniques d'exploration informatisée et d'opérations de saisie clavier. Pour un aperçu de la méthodologie du recensement, voir Citro et Cohen (1985); on trouvera une description détaillée de diverses opérations du recensement dans le recueil du Bureau of the Census intitulé "1990 Decennial Census Information Memorandum Series" (Bureau of the Census 1988-1991).

Dans des enquêtes postcensitaires comme celle qui a été réalisée en 1990 (et dans des recensements d'essai comme ceux qui ont mené à l'EP de 1990), la collecte de données débute par la confection de listes d'adresses; cette opération est confiée à des recenseurs qui parcourent les quartiers. La collecte proprement dite se fait uniquement par interview sur place, contrairement à la méthode "envoi et retour par la poste" utilisée dans le recensement pour les régions urbaines et suburbaines. Les données pertinentes sont versées dans les fichiers informatiques uniquement par saisie clavier. Hogan (1992) donne un aperçu de l'EP; on trouvera une description détaillée des opérations de l'EP dans le recueil du Bureau of the Census intitulé "STSD Decennial Census Memorandum Series" (Bureau of the Census 1987-1991).

Le prétraitement des données est un sujet qui est peu souvent abordé dans les ouvrages sur le couplage d'enregistrements, même si cette étape est celle où existe la possibilité de tirer toutes les informations voulues des données d'enquête comme la possibilité de rejeter inconsidérément des informations utiles. Winkler (1985a, 1985b) présente des méthodes particulières qui ont pour propriété de faciliter la distinction entre de vraies concordances et de fausses concordances, et Jabine et Scheuren (1986) de même que Newcombe (1988) proposent de grandes lignes directrices à ce sujet. Dans l'opération d'appariement des enregistrements du recensement à ceux de l'EP, le prétraitement des données consiste à coder les variables démographiques selon des normes, à décomposer la zone "adresse" en ses éléments préalablement définis (numéro civique, rue, numéro d'appartement, numéro de route rurale, numéro de case postale) (Laplant 1989), et à "normaliser" les prénoms d'individus en confrontant le prénom introduit au clavier à une liste de diminutifs puis en convertissant les diminutifs observés dans leur forme au long (Paletz 1989).

Le couplage des enregistrements du recensement à ceux de l'EP est un procédé fondé sur la pondération. Pour déterminer une méthode de pondération, il faut envisager des règles de modèle et des règles *ad hoc* pour l'attribution de poids pour la concordance ou la non-concordance de zones d'information, pour la concordance imparfaite de zones d'information, pour des enregistrements incomplets et pour la concordance ou la non-concordance de certaines combinaisons de variables.

La définition d'éléments susceptibles de former la même paire dans l'opération d'appariement des enregistrements du recensement à ceux de l'EP reflète certaines contraintes auxquelles est assujetti le processus d'appariement. Premièrement, à cause de contraintes de temps et de limites financières, il est peu réaliste de comparer chacun des enregistrements d'un fichier à chaque enregistrement d'un autre fichier. Par conséquent, on limite la comparaison à des paires d'enregistrements qui répondent à certains critères minimum (par ex., les deux enregistrements comparés doivent provenir du même îlot de recensement et la première lettre du nom de famille doit être la même dans les deux cas). Le sous-ensemble d'enregistrements ainsi constitué est appelé "groupe" et les variables par rapport auxquelles on doit observer une concordance de deux enregistrements sont appelées "variables de groupage" (Jaro 1989).

Une autre contrainte à laquelle est assujetti le processus d'appariement des enregistrements du recensement à ceux de l'EP est celle voulant qu'un enregistrement donné d'un fichier ne peut être couplé à plus d'un enregistrement d'un autre fichier. La méthode utilisée pour rapprocher des éléments susceptibles de former la même paire s'inspire des techniques de recherche opérationnelle appliquées au problème de transport (Jaro 1989). L'algorithme rapproche des éléments susceptibles de former une concordance afin de maximiser la somme des poids composites qui se rattachent à chacune des paires d'enregistrements d'un groupe défini par les variables de groupage, sous réserve qu'un enregistrement d'un fichier ne peut être couplé à plus d'un enregistrement d'un autre fichier. Supposons par exemple que dans un groupe particulier, l'enregistrement A du fichier 1 a un poids de concordance plus élevé pour l'enregistrement B du fichier 2 que pour tout autre enregistrement du même fichier. L'algorithme de rapprochement peut néanmoins coupler l'enregistrement A à un autre enregistrement que B, disons C, et coupler l'enregistrement B à un autre enregistrement que A, disons D, si la somme des poids de concordance pour (A,C) et (D,B) est plus élevée que celle pour toute autre combinaison de couples.

Selon la méthode utilisée actuellement pour apparier les enregistrements du recensement à ceux de l'EP, l'ordinateur a trois possibilités d'action: reconnaître une paire d'enregistrements comme une concordance, reconnaître une paire d'enregistrements comme une "concordance possible" ou reconnaître une paire d'enregistrements comme une non-concordance. Tous les cas de non-concordance et de concordance possible sont soumis à l'attention de commis pour être examinés, et on tente d'organiser une interview de suivi avec les ménages pour

lesquels on observe un écart entre les données du recensement et les données de l'EP. La distinction entre les concordances possibles et les non-concordances a trait uniquement aux méthodes qu'appliquent les commis lorsqu'ils examinent ces cas (Childers 1989; Donoghue 1990). Dans le traitement des données du recensement et de l'EP de 1990, l'opérateur qui voit à l'exécution du programme d'appariement détermine manuellement des poids limites qui permettent d'identifier les concordances, les concordances possibles et les non-concordances après avoir exploré des ensembles d'éléments susceptibles de former la même paire et affectés de poids dont la valeur se situe dans un intervalle donné. Une nouvelle méthode proposée par Belin et Rubin (1991) permet de déterminer automatiquement les poids limites.

3. EXPÉRIENCE FACTORIELLE

3.1 Facteurs qui influent sur les résultats des méthodes de couplage d'enregistrements

L'efficacité d'une méthode de couplage d'enregistrements tient à un certain nombre de facteurs, dont voici la liste:

(1) le choix des variables d'appariement;

(2) le choix des variables de groupage;

(3) l'attribution de poids de concordance ou de non-concordance pour diverses variables d'appariement;

(4) le traitement des cas de concordance imparfaite entre des variables d'appariement;

(5) le traitement de données manquantes dans l'un ou l'autre des enregistrements formant une paire ou dans les deux enregistrements;

(6) l'algorithme servant à rapprocher des éléments susceptibles de former la même paire;

(7) le choix d'un poids limite au-delà duquel des paires d'enregistrements seront reconnues comme des concordances;

(8) le lieu ou le milieu d'où proviennent les données.

Parmi tous ces facteurs, seul le huitième représente une source de variation sur laquelle l'opérateur qui voit à l'exécution du programme d'appariement n'a aucun contrôle.

Comme nous l'avons mentionné plus tôt, deux questions en particulier méritent d'être approfondies dans l'expérience factorielle. Primo, si on parvient à déterminer les principales sources de variation dans le couplage d'enregistrements, on sera en mesure d'orienter les recherches en matière de couplage pour l'avenir et de mieux comprendre le processus par lequel des erreurs sont introduites dans les méthodes de couplage. Secundo, il serait intéressant de déterminer la combinaison de facteurs qui permet d'obtenir le plus grand nombre de concordances avec le taux d'erreur le plus faible possible puisque dans la pratique, l'utilisateur a généralement le choix entre de nombreuses possibilités pour chaque facteur que nous venons de décrire.

3.2 Expérience factorielle réalisée avec des données du recensement et de l'enquête postcensitaire

On a réalisé une étude à l'aide de données provenant de chacune des trois régions d'essai (St-Louis, Missouri; centre est du Missouri, y compris la région de Columbia (Mo); et une région rurale dans la partie Est de l'État du Washington) de la répétition générale et de l'EP de 1988. Ces ensembles de données ont été appariées par ordinateur, puis examinées par des commis. Pour les besoins des analyses qui vont suivre, nous tenons pour acquis que le classement opéré par les commis en ce qui a trait aux concordances (vraie concordance/fausse concordance) est juste. Par conséquent, malgré que les analyses qui vont suivre ne pourront jamais être plus précises que celles des commis, ces fichiers de données nous offrent une excellente occasion d'étudier le couplage d'enregistrements.

On trouvera une description des méthodes particulières utilisées pour coupler les enregistrements du recensement à ceux de l'EP dans Jaro (1989), Winkler (1991) et Winkler et Thibaudeau (1992). Selon la méthode actuellement en usage, l'utilisateur jouit de plusieurs possibilités par rapport à tous les facteurs qui ont été énumérés dans la section 3.1, sauf en ce qui concerne le choix de l'algorithme servant à rapprocher des éléments susceptibles de former la même paire (on utilise un algorithme de "rapprochement à somme linéaire"; voir Jaro 1989).

Les variables qui peuvent servir à l'appariement des enregistrements du recensement avec ceux de l'EP sont le nom, l'adresse, l'âge, l'origine raciale, le sexe, le numéro de téléphone, l'état matrimonial et le lien avec le chef de ménage. En pratique, le nom est habituellement décomposé en trois éléments: prénom, nom de famille et second prénom, chacun d'eux servant de variable d'appariement. On utilise habituellement un programme de prétraitement pour décomposer l'adresse en ses éléments: numéro civique, rue, numéro d'appartement, numéro de route rurale et numéro de case postale (Laplant 1989). Il arrive qu'en raison d'"irrégularités" dans la zone "adresse", causées probablement par une erreur de saisie ou une erreur d'inscription de la part d'un interviewer, on ne soit pas capable de décomposer une adresse en ses éléments; dans ces circonstances, la zone "adresse" intégrale (appelée "adresse globale") sert de variable d'appariement. On peut aussi se servir d'un programme de prétraitement pour convertir les diminutifs dans leur forme "normalisée" en utilisant une banque de noms, y compris les variantes les plus courantes (Paletz 1989). Il existe diverses méthodes permettant d'attribuer des poids en fonction d'une quasi-concordance de variables, et il existe aussi une méthode qui permet d'accroître ou de réduire (par une simple opération d'addition ou de soustraction) le poids composite associé à une paire d'enregistrements lorsque certaines combinaisons de zones concordent ou ne concordent pas entre elles (Winkler 1991).

L'expérience a consisté en huit facteurs de "traitement" et en un facteur de "groupage" (ce terme signifie ici, par rapport à la planification d'expériences, que l'on forme un groupe d'unités qui sont censées produire des résultats

le plus semblables possible en l'absence d'effets déterminés par le traitement) avec répétition dans trois régions d'essai par un plan factoriel $2^5 \times 3^3 \times 5 \times 13$. La variable de résultat de l'expérience, décrite plus loin dans la section 3.5, consistait en une transformation du taux d'erreur d'appariement, cette transformation étant destinée à stabiliser la variance des résultats. Les facteurs utilisés dans l'expérience peuvent être décrits comme suit:

Symbole	Description du facteur	Nombre de niveaux du facteur	Description des niveaux
A	Attribution de poids pour les zones de nom.	5	1. Attribution de poids d'une valeur de ± 2 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille. 2. Attribution de poids d'une valeur de ± 4 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille. 3. Attribution de poids d'une valeur de ± 6 selon qu'il y a concordance ou non concordance pour le prénom, le nom de famille. 4. Attribution de poids en fonction des estimations de la probabilité de concordance du prénom, du nom de famille, calculées selon l'algorithme de Fellegi-Sunter (voir Winkler et Thibaudeau 1992). 5. Utilisation de la pondération basée sur la fréquence pour le prénom, le nom de famille (voir Winkler et Thibaudeau 1992).
B	Attribution de poids pour les cas de concordance imparfaite dans les zones de nom.	3	1. Attribution d'un poids de non-concordance pour toute différence entre prénoms, entre noms de famille. 2. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de prénoms, de noms de famille, en utilisant le comparateur de chaînes de caractères de Jaro (Jaro 1989; Winkler 1991). 3. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de prénoms, de noms de famille, en utilisant la métrique linéaire progressive décrite dans Winkler (1991).
C	Attribution de poids pour les zones autres que les zones de nom.	2	1. Attribution de poids d'une valeur de ± 2 selon qu'il y a concordance ou non-concordance pour l'âge, le numéro de téléphone ou l'adresse, et attribution de poids d'une valeur de ± 1 selon qu'il y a concordance ou non-concordance pour le sexe, l'origine raciale, l'état matrimonial, le lien avec le chef de ménage ou le second prénom. 2. Attribution de poids en fonction des estimations de la probabilité de concordance calculées selon l'algorithme de Fellegi-Sunter.

Symbole	Description du facteur	Nombre de niveaux du facteur	Description des niveaux
D	Attribution de poids pour les cas de concordance imparfaite dans les zones autres que les zones de nom.	3	1. Attribution d'un poids de non-concordance pour toute différence entre des zones autres que les zones de nom. 2. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de numéros civiques, de noms de rue, de numéros de téléphone, d'âges, en utilisant le comparateur de chaînes de caractères de Jaro. 3. Attribution d'une fraction de poids de concordance s'il y a quasi-concordance de noms de rue, en utilisant le comparateur de chaînes de caractères de Jaro; s'il y a quasi-concordance d'âges, en utilisant la métrique "au prorata de l'écart absolu" de Jaro; s'il y a quasi-concordance de numéros civiques ou de numéros de téléphone, en utilisant le comparateur linéaire progressif de chaînes de caractères de Winkler.
E	Utilisation du prénom introduit au clavier ou de son équivalent normalisé.	2	1. Utilisation du prénom introduit au clavier dans chaque fichier pour la comparaison de prénoms. 2. Utilisation du prénom produit par le logiciel de normalisation des noms (Paletz 1989).
F	Correction de poids pour les cas de concordance corrélée.	2	1. Ne pas corriger les poids composites pour les cas possibles de concordance corrélée. 2. Corriger les poids composites pour les cas possibles de concordance corrélée des prénoms et des seconds prénoms et de concordance corrélée des prénoms, des sexes et des âges.
G	Inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement.	2	1. Ne pas compter l'état matrimonial et le lien avec le chef de ménage parmi les variables d'appariement. 2. Compter l'état matrimonial et le lien avec le chef de ménage parmi les variables d'appariement.
H	Utilisation de quatre ou sept chiffres pour le numéro de téléphone.	2	1. Utiliser seulement les quatre derniers chiffres du numéro de téléphone comme variable d'appariement. 2. Utiliser les sept chiffres du numéro de téléphone.
I	Région d'essai du recensement ou de l'enquête postcensitaire.	3	1. Est de l'État du Washington. 2. Columbia, Missouri. 3. St-Louis, Missouri.
J	Proportion du fichier de l'EP ayant pu être couplée.	13	1.-13. Le nombre d'enregistrements ayant pu être couplés équivaut à 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, 90% du nombre total d'enregistrements de l'EP dans la région d'essai donnée.

Si l'on se reporte aux sources de variation énumérées dans la section 3.1, on peut dire que les facteurs E, G et H ont trait au choix des variables d'appariement; les facteurs A, C et F ont trait au choix de la méthode de pondération; les facteurs B et D concernent le traitement des cas de concordance imparfaite; enfin, le facteur J évoque le choix d'un poids limite et le facteur I, l'influence de la région d'essai sur l'efficacité de la méthode d'appariement.

En raison d'un manque de ressources, on a décidé de ne pas étudier dans cette expérience l'effet de l'utilisation de différentes méthodes de traitement des données manquantes ou de l'utilisation de différentes combinaisons de variables de groupage; en outre, l'inexistence de certains logiciels a rendu impossible l'étude de nouveaux algorithmes permettant de rapprocher des éléments susceptibles de former la même paire. Belin (1989a, 1989b) a étudié l'effet de l'utilisation de différentes méthodes de traitement des données manquantes et de différentes combinaisons de variables de groupage dans une expérience semblable à l'expérience factorielle décrite ici. Les résultats de son étude permettent de croire que la méthode de traitement des données manquantes change peu de chose aux taux d'erreur d'appariement jumelés à des poids limites différents dans l'appariement des données du recensement avec celles de l'EP, mais que le choix des variables de groupage a un effet appréciable sur ces taux.

Dans la présente étude, comme dans celle de Belin (1989a, 1989b), nous ne considérons que des méthodes d'appariement "à une seule phase", c'est-à-dire que l'opération d'appariement informatisé consiste à exécuter en une seule phase les opérations suivantes: choix des variables de groupage, calcul des poids et fixation d'un poids limite. Cela contraste avec les méthodes "à phases multiples", qui utilisent d'abord des variables de groupage très restrictives afin d'isoler les concordances quasi-parfaites, puis assouplissent graduellement les critères de groupage à chaque phase suivante. Nous connaissons très peu d'études qui aient été faites sur les méthodes d'appariement à phases multiples. Belin (1989b) fait observer que lorsqu'on utilise des méthodes à une seule phase, les méthodes qui appliquent des critères de groupage relativement moins restrictifs sont plus avantageuses que les méthodes qui appliquent des critères relativement plus restrictifs, ce qui confirme l'idée intuitive que le processus de groupage peut avoir pour inconvénient de camoufler des concordances vraies.

3.3 Aspects particuliers des traitements expérimentaux

3.3.1 Traitements relatifs à l'attribution de poids pour les cas de concordance ou de non-concordance dans des zones d'information

Afin de préciser l'expérience, nous décrivons plus en détail chacun des facteurs expérimentaux. Les facteurs A et C ont trait à l'attribution de poids de concordance ou de non-concordance pour diverses variables d'appariement. Les méthodes de pondération utilisées dans les circonstances comprennent des méthodes tout à fait ponctuelles de même que des méthodes fondées sur des estimations de paramètres de modèles probabilistes explicites. L'étude de méthodes de pondération ponctuelles nous donne l'occasion de jauger l'importance d'introduire des méthodes plus complexes.

Selon les méthodes de pondération ponctuelles, on doit ajouter un poids U, par exemple, au poids composite s'il y a concordance des zones comparées, ou soustraire ce même poids du poids composite s'il y a non-concordance des zones comparées. Trois valeurs de U sont étudiées pour le facteur A; des poids identiques sont attribués pour la concordance de prénoms et la concordance de noms de famille. En ce qui concerne le facteur C, on étudie une méthode ponctuelle qui permet d'attribuer des poids plus élevés à certaines variables, le choix de ces variables étant fondé sur des jugements *a priori*. Belin (1989b) estime que cette méthode présente des avantages par rapport à la méthode "à pondération uniforme", où toutes les variables d'appariement reçoivent le même poids pour la concordance comme pour la non-concordance.

L'expression "algorithme de Fellegi-Sunter" désigne la méthode exposée dans l'article de Fellegi et Sunter (1969), qui repose sur un modèle probabiliste qui intègre de l'information sur des schémas de concordance et de non-concordance pour des paires d'enregistrements. Le modèle suppose que les probabilités de concordance de zones d'information, étant donné qu'une paire représente une vraie concordance ou bien une fausse concordance, sont indépendantes d'une zone à l'autre. Dans leur article, Fellegi et Sunter montrent que ce modèle implique certaines propriétés d'optimalité pour le genre de méthode de pondération qu'utilisent Newcombe *et al.* (1959), selon laquelle on calcule les poids relatifs aux zones d'information par le logarithme du rapport de la probabilité de concordance étant donné une concordance vraie à la probabilité de concordance étant donné une fausse concordance, et selon laquelle on calcule les poids composites en faisant la somme des poids de chaque zone.

Dans les applications, la probabilité de concordance étant donné une concordance vraie ainsi que la probabilité de concordance étant donné une fausse concordance doivent être estimées. En ce qui concerne les traitements de l'expérience qui reposent sur la méthode de pondération de Fellegi-Sunter, on estime les probabilités de concordance étant donné une concordance vraie à l'aide d'une version d'un algorithme EM (Dempster, Laird et Rubin 1977) qui produit des estimations les plus vraisemblables de ces probabilités en fonction de données relatives aux schémas de concordance observés dans les fichiers disponibles (Winkler 1989; Jaro 1989). Quant aux probabilités de concordance étant donné une fausse concordance, on les estime en se fondant sur des données relatives à la concordance, pour une zone, de toutes les paires d'enregistrements possibles à l'égard des variables de groupage, en tenant compte du fait que la plupart des couples d'enregistrements susceptibles de constituer une concordance ne sont pas des concordances vraies (Winkler et Thibaudeau 1992).

Le programme de couplage d'enregistrements du Census Bureau prévoit l'application d'une autre méthode

de pondération, où la fréquence relative des noms dans les fichiers disponibles est prise en considération. Cette méthode consiste à attribuer un poids de concordance plus grand pour des noms comme Abramowicz, qui peuvent être relativement rares, que pour des noms comme Smith, qui peuvent être très courants. Il se pourrait, évidemment, que Abramowicz soit plus courant que Smith dans une région particulière; dans ce cas, on attribuera un poids de concordance plus grand pour le nom Smith. L'idée de tenir compte de fréquences marginales tirées des fichiers courants a été lancée par Newcombe *et al.* (1959) et est évoquée depuis par de nombreux auteurs, dont Fellegi et Sunter (1969). (Par conséquent, la distinction qui est faite entre l'"algorithme de Fellegi-Sunter" et la "pondération fondée sur la fréquence" sert en réalité à distinguer deux méthodes de calcul de poids analysées l'une et l'autre par Fellegi et Sunter.) On trouvera dans Winkler et Thibaudeau (1992) des détails sur la mise en application de la méthode de pondération fondée sur la fréquence dans le programme du Census Bureau.

3.3.2 Traitement des cas de concordance imparfaite

Les facteurs B et D ont rapport au traitement des cas de concordance imparfaite (c.-à-d. des cas où deux zones d'information se rapprochent sensiblement l'une de l'autre sans être parfaitement identiques). Plusieurs techniques ont été proposées pour le traitement des cas de concordance imparfaite; ces techniques sont souvent aussi nombreuses que les théories qui cherchent à expliquer l'absence de concordance parfaite.

Le comparateur de chaînes de caractères de Jaro sert à mesurer le degré de concordance de deux zones à caractères multiples; la métrique qui définit le degré de similitude est fonction de la longueur des zones de caractères des deux fichiers, du nombre de caractères communs aux deux zones et du nombre de caractères qui peuvent être transposés d'une zone à l'autre. Le poids attribué pour une concordance partielle se situe entre celui attribué pour une concordance de zones et celui attribué pour une non-concordance de zones et est une fonction linéaire du comparateur de chaînes de caractères entre le poids de concordance et le poids de non-concordance.

La méthode linéaire progressive de Winkler utilise la même métrique que le comparateur de chaînes de caractères de Jaro pour définir le degré de similitude, sauf que le taux de diminution des poids de concordance partielle entre le poids de concordance et le poids de non-concordance est une fonction linéaire progressive du comparateur de chaînes de caractères, ce qui nécessite deux paramètres de taux et deux seuils fournis par l'utilisateur au point de changement de la pente.

La méthode proportionnelle de Jaro permet d'attribuer un poids dont la valeur se situe entre le poids de concordance et le poids de non-concordance, en se fondant sur la valeur absolue de la différence entre deux zones numériques. Comme pour les techniques précédentes, le poids de concordance partielle se trouve être une fonction linéaire de la valeur absolue de la différence.

Dans le cas de certaines zones numériques (numéro de téléphone, par exemple), il serait plus raisonnable d'avoir une méthode de comparaison conçue en fonction des erreurs typographiques mineures qu'une méthode fondée sur l'écart. En revanche, pour des variables comme l'année de naissance ou l'âge, il est difficile de dire s'il faut se concentrer sur les erreurs typographiques (auquel cas une méthode de comparaison de chaînes de caractères serait la solution la plus appropriée), les erreurs de déclaration (auquel cas une méthode fondée sur l'écart serait la solution la mieux indiquée), ou d'autres types d'erreur, comme l'arrondissement de l'âge déclaré au multiple de cinq le plus près, (auquel cas ni l'une ni l'autre des méthodes de comparaison mentionnées ci-dessus ne conviendrait). C'est pourquoi nous poursuivons nos analyses empiriques dans le but d'approfondir ces questions.

3.3.3 Traitements relatifs au choix des variables d'appariement

Nous avons mentionné plus tôt que le Census Bureau avait élaboré une méthode pour convertir les diminutifs dans une forme "normalisée". Le logiciel élaboré par Paletz (1989) applique le programme de normalisation des noms.

Le traitement qui n'inclut pas l'état matrimonial et le lien avec le chef de ménage dans les variables d'appariement permet d'évaluer l'importance de deux variables démographiques de base pour la qualité de l'appariement. Chernoff (1980) élabore une théorie relativement à l'information que "renferme" une variable d'appariement et il montre qu'une variable qui est enregistrée incorrectement, ne serait-ce que très peu de fois, peut "perdre" une partie appréciable de l'information qu'elle contient pour l'appariement (par exemple, l'information de Kullback-Leibler rattachée à une variable binaire qui a été enregistrée incorrectement dans trois pour-cent des cas n'équivaut qu'à environ la moitié de l'information rattachée à une variable binaire qui a enregistrée correctement à tout coup). Compte tenu de ce que le lien avec le chef de ménage pourra avoir changé entre le recensement et l'EP si la personne identifiée à l'origine comme le chef de ménage n'est plus la même, et de ce que l'état matrimonial de certaines personnes changera dans l'intervalle, on ne peut dire d'avance quelle quantité d'information ces variables renferment-elles pour l'appariement. En revanche, il est difficile de penser que l'utilisation de variables d'appariement additionnelles serait nuisible; par conséquent, ce traitement peut servir de norme pour l'évaluation de l'importance pratique de certains des autres traitements.

L'utilisation de quatre ou sept chiffres du numéro de téléphone comme variable d'appariement est explicite. Ce traitement est justifié par le fait que l'élaboration de l'une des méthodes de comparaison de chaînes de caractères de Winkler (métriques linéaires progressives) reposait sur l'analyse des quatre derniers chiffres du numéro de téléphone comme variable d'appariement.

3.3.4 Traitement relatif à la correction de poids composites pour les cas de concordance corrélée

On doit aussi à Winkler la méthode qui sert à corriger les poids composites de manière à tenir compte des cas possibles de concordance corrélée; cette méthode est décrite dans Winkler et Thibaudeau (1992). Des études de Kelley (1986) et de Thibaudeau (1989) révèlent que la concordance qui peut être observée pour des zones qui font l'objet de l'opération d'appariement entre les fichiers du recensement et ceux de l'EP n'est pas du tout indépendante d'une zone à l'autre. Les études donnent à croire, en particulier, que la concordance de prénoms est corrélée avec la concordance de seconds prénoms et que les concordances de prénoms, d'âges et de sexes sont corrélées l'une avec l'autre. Par suite de ces observations, on en est venu à appliquer des règles de modification du poids composite dans certaines circonstances (par exemple, s'il y a non-concordance des prénoms, des âges et des sexes simultanément, on soustrait du poids composite une valeur élevée). La méthode utilisée actuellement pour corriger les poids composites est une méthode entièrement *adaptée aux circonstances*; l'étude des méthodes qui tiennent compte de la concordance corrélée semble encore en être au premier stade.

3.4 Fichiers de données utilisés dans l'expérience

Comme nous l'avons mentionné plus haut, les trois régions d'essai de la répétition générale et de l'enquête postcensitaire de 1988 ont permis de constituer des fichiers de données distincts qui ont pu faire l'objet de notre analyse. Le fichier de l'EP de St-Louis contenait 12,072 enregistrements, celui du centre est du Missouri en contenait 6,581 et celui de l'est de l'État du Washington, 2,782. Nous avons aussi souligné que, pour les besoins de l'analyse, nous considérions comme juste le classement opéré par les commis qui avaient examiné ces fichiers. D'autres recensements d'essai ont eu lieu durant les années 1980; si nous n'avons pas tenu compte des données de ces recensements dans notre expérience, c'est surtout parce que la préparation d'un ensemble de données en vue des analyses effectuées ici exige beaucoup de temps système.

3.5 Variable de résultat

La principale variable de résultat envisagée dans cette expérience était une transformation du taux d'erreur d'appariement. Le taux d'erreur d'appariement est défini comme le quotient du nombre de fausses concordances par le nombre de concordances désignées et est une mesure de rendement couramment utilisée dans les ouvrages sur le couplage d'enregistrements (par ex., Fellegi et Sunter (1969) tentent d'obtenir des résultats qui respectent un critère fixé par l'opérateur; ce critère correspond à un taux d'erreur d'appariement). Afin de stabiliser la variance des résultats, nous utilisons l'arc sinus de la racine carrée du taux d'erreur d'appariement comme variable de résultat dans notre analyse.

3.6 Choix du poids limite comme facteur de groupage

Il est évident que dans le couplage d'enregistrements, le taux d'erreur d'appariement est susceptible de dépendre largement du choix d'un point de démarcation entre les concordances désignées et les non-concordances désignées. C'est pourquoi on introduit un facteur de groupage (facteur J) qui permet de déterminer des seuils de manière à faciliter la comparaison d'autres traitements relatifs au couplage d'enregistrements. Afin de pouvoir établir des comparaisons entre des régions d'essai qui n'ont pas le même nombre d'enregistrements, le seuil d'exclusion est exprimé en fonction de la proportion du fichier de l'EP ayant pu être couplée.

Comme les poids de couplage d'enregistrements sont des valeurs discrètes, il peut exister des liens entre les poids attribués à des paires d'enregistrements qui se trouvent dans le voisinage du seuil d'exclusion. Dans un fichier de 10,000 enregistrements par exemple, on peut en compter 40 qui ont un poids W (dont 10 peuvent appartenir à de fausses concordances), 7,980 qui ont un poids supérieur à W (dont 3 peuvent appartenir à de fausses concordances) et 1,980 qui ont un poids inférieur à W. Si, selon le facteur J, la proportion des enregistrements du fichier de l'EP pouvant être couplés doit être de 80%, il se peut que la manière de calculer le taux d'erreur d'appariement ne s'impose pas à l'évidence puisque 40 enregistrements qui ont le même poids se trouvent dans le voisinage du point où devrait être fixé le seuil d'exclusion. Dans de telles circonstances, on calcule le taux d'erreur d'appariement au moyen de la relation suivante:

$$\text{fmr} = \frac{f_{\text{abv}} + \dfrac{f_{\text{bdy}}}{n_{\text{bdy}}} \times (n_{\text{cut}} - n_{\text{abv}})}{n_{\text{cut}}},$$

où fmr désigne le taux d'erreur d'appariement, f_{abv} est le nombre de fausses concordances et n_{abv}, le nombre de concordances désignées dont le poids est supérieur au poids limite, f_{bdy} est le nombre de fausses concordances et n_{bdy}, le nombre de concordances désignées dont le poids est égal au poids limite; enfin, n_{cut} est le nombre de concordances désignées nécessaires pour respecter la condition selon laquelle une certaine proportion des enregistrements du fichier de l'EP doivent pouvoir être couplés. S'il fallait calculer le taux d'erreur d'appariement en déterminant de façon aléatoire le nombre d'enregistrements limites voulu pour respecter le seuil d'exclusion, la formule ci-dessus permettrait de connaître le taux d'erreur d'appariement prévu pour plusieurs itérations de cette procédure; par conséquent, la logique sous-jacente à cette définition est claire.

Dans l'exemple ci-dessus, le quart des cas limites sont de fausses concordances, et il faut vingt enregistrements de plus pour respecter le critère de 80%. De fait, cinq fausses concordances sont ajoutées aux trois qui existent déjà parmi les enregistrements qui forment les paires dont le poids est supérieur au poids limite, ce qui donne un taux d'erreur d'appariement de $(3 + 0.25(40-20))/8{,}000 = 8/8{,}000 = 0.001$.

3.7 Autres considérations touchant l'analyse des résultats expérimentaux

L'analyse des résultats expérimentaux partait de l'idée que des indices généraux de signification sont plus importants que des valeurs p précises, à plus forte raison s'il s'agit d'une expérience exploratoire. Belin (1991) précise que les méthodes destinées spécialement à évaluer le degré de signification tiré de ces données sont assez complexes parce que la région d'essai devrait, en réalité, être considérée comme un facteur aléatoire (si nous voulons pouvoir inférer sur les "effets déterminés par le traitement" à partir de l'échantillon de trois régions d'essai pour une population de régions possibles); cependant, les méthodes courantes qui utilisent l'interaction "région d'essai-traitement" comme terme d'erreur pour un traitement particulier ont peu d'efficacité à cause du faible nombre de régions d'essai. Belin (1991) se sert du graphique de Johnson-Tukey pour les ratios (Johnson et Tukey 1987), qui se rapproche sensiblement du graphique semi-normal de Daniel (1959), pour estimer les niveaux de bruit liés à l'évaluation de la signification des effets. Il n'est pas dans nos intentions d'exposer ici des conclusions formelles sur cette question.

4. RÉSULTATS

4.1 Analyse de variance des résultats expérimentaux

Nous commençons par faire une analyse de variance des résultats de l'expérience factorielle en distinguant les effets déterminés par le traitement, les effets déterminés par la région d'essai, les effets déterminés par le seuil et les interactions des uns avec les autres, ces effets étant groupés selon le nombre de critères. Le tableau 4.1 contient un extrait de l'analyse de variance, qui montre les interactions de traitements (de deux à cinq critères) et les termes d'erreur correspondants.

On calcule la statistique F en divisant la moyenne des carrés pour l'effet donné par la moyenne des carrés pour le terme d'interaction de l'effet et de la région d'essai. Par exemple, la statistique F pour les interactions à trois critères des traitements est calculée à l'aide du rapport $0.0120/0.00470 = 2.551$, le dénominateur étant le chiffre qui figure sur la ligne des interactions à quatre critères du traitement et de la région d'essai.

Tableau 4.1

Extrait de l'analyse de variance des résultats de l'expérience factorielle (effets groupés selon le nombre de critères)

Source	*dl*	Sommes des carrés	Moyennes des carrés	*F*
Effets majeurs - région d'essai	2	35.195	17.598	
Effets majeurs - traitement	13	30.917	2.378	10.570
Effets majeurs - seuil d'exclusion	12	147.515	12.293	7.548
Interactions à deux critères du traitement et de la région d'essai	26	5.850	0.225	
Interactions à deux critères du seuil d'exclusion et de la région d'essai	24	39.089	1.629	
Interactions à deux critères des traitements	70	6.992	0.100	4.041
Interactions à deux critères du traitement et du seuil d'exclusion	156	1.410	0.009	3.553
Interactions à trois critères du traitement et de la région d'essai	140	3.461	0.0247	
Interactions à trois critères du seuil d'exclusion, du traitement et de la région d'essai	312	0.794	0.0025	
Interactions à trois critères des traitements	206	2.472	0.0120	2.551
Interactions à trois critères du traitement et du seuil d'exclusion	840	0.530	0.0006	1.866
Interactions à quatre critères du traitement et de la région d'essai	412	1.938	0.00470	
Interactions à quatre critères du seuil d'exclusion, du traitement et de la région d'essai	1,680	0.568	0.00034	
Interactions à quatre critères des traitements	365	0.747	0.00205	2.365
Interactions à quatre critères du traitement et du seuil d'exclusion	2,472	0.267	0.00011	0.236
Interactions à cinq critères du traitement et de la région d'essai	730	0.632	0.00087	
Interactions à cinq critères du seuil d'exclusion, du traitement et de la région d'essai	4,944	0.226	0.00046	
Total	**56,159**	**279.169**		

Si l'on interprète les statistiques F de la manière habituelle, on constate que toutes les valeurs F du tableau 4.1, sauf celle relative aux interactions à quatre critères du traitement et du seuil d'exclusion, permettent de conclure à la signification statistique à un niveau de 0.0001. Néanmoins, ces résultats devraient être interprétés avec prudence, d'abord parce que les valeurs des carrés moyens laissent supposer que les effets d'ordre supérieur n'ont pas une grande importance pratique; deuxièmement, la comparaison des statistiques F calculées ci-dessus avec une distribution F de référence repose sur certaines hypothèses d'interchangeabilité (*ex.*: que la variabilité des effets majeurs des régions d'essai est la même pour tous les autres effets majeurs) qui ne sont pas nécessairement solides. Par exemple, il peut être illogique de combiner la variabilité d'une région à l'autre de l'effet de l'utilisation de quatre ou sept chiffres du numéro de téléphone avec la variabilité d'une région à l'autre de l'effet des diverses méthodes de pondération pour estimer un terme d'erreur pour des effets majeurs.

4.2 Importance du choix du seuil d'exclusion par rapport à d'autres facteurs contrôlables

Il est évident (d'après les carrés moyens pour les effets majeurs) que la variabilité due à la région d'essai et celle due au choix d'un seuil d'exclusion sont beaucoup plus grandes que la variabilité due aux différences de traitement. Même si on peut expliquer en partie cette différence par la similitude de certains traitements (par ex., utiliser quatre chiffres du numéro de téléphone plutôt que sept), il n'en reste pas moins que certaines des différences qualitatives entre les traitements sont très appréciables (par ex., éliminer ou conserver deux variables d'appariement). L'analyse de variance met aussi en évidence le fait qu'il faille s'attendre à une forte variabilité des taux d'erreur d'appariement entre les régions d'essai. Dans leur méthode visant à calibrer les méthodes de couplage d'enregistrements, Belin (1991) et Belin et Rubin (1991) tiennent compte explicitement de la variabilité dué à la région d'essai en présentant des estimations du taux d'erreur d'appariement qui correspondent à différents seuils d'exclusion.

4.3 Effets majeurs des traitements

Dans le tableau 4.2, nous donnons la valeur moyenne de la variable de résultat observée pour chaque niveau des facteurs de traitement. Comme arc sinus (x) est une fonction croissante monotone de x, des valeurs faibles pour cette variable signifient des taux d'erreur d'appariement peu élevés et, donc, un meilleur rendement.

Tableau 4.2

Valeurs marginales moyennes de arc sinus ($\sqrt{\text{fmr}}$) pour chaque niveau des facteurs expérimentaux (moyenne calculée par rapport à toutes les autres conditions expérimentales)

Facteur	A	(Poids, zones de nom)	Facteur	B	(Concordance imparfaite, poids pour zones de nom)	Facteur	C	(Poids, autres zones)
Niveau	1	0.106	Niveau	1	0.113	Niveau	1	0.101
	2	0.096		2	0.094		2	0.101
	3	0.093		3	0.095			
	4	0.130						
	5	0.079						
Facteur	D	(Concordance imparfaite, poids pour autres zones)	Facteur	E	(Normalisation des noms)	Facteur	F	(Corr. de poids pour concordance corrélée)
Niveau	1	0.111	Niveau	1	0.102	Niveau	1	0.106
	2	0.108		2	0.100		2	0.095
	3	0.084						
Facteur	G	(Incl. de l'état matrimonial et du lien avec le chef de ménage)	Facteur	H	(Quatre ou sept chiffres du numéro de téléphone)			
Niveau	1	0.103	Niveau	1	0.102			
	2	0.098		2	0.100			

Belin (1991) construit un ensemble de contrastes orthogonaux complémentaires à partir des résultas expérimentaux. Les plus importants contrastes d'effets majeurs parmi ceux définis préalablement par Belin (1991) sont ceux entre les poids basés sur la fréquence pour les noms (A = 5) et les poids pour zones de nom de Fellegi-Sunter (A = 4), entre les comparateurs de chaînes de caractères de Winkler pour les zones autres que les zones de nom (D = 3) et les comparateurs de chaînes de caractères correspondants de Jaro (D = 2), entre un comparateur de chaînes de caractères quelconque pour les zones de nom (B = 2 ou 3) et aucun comparateur de chaînes de caractères pour les zones de nom (B = 1), entre un comparateur de chaînes de caractères quelconque pour les zones autres que les zones de nom (D = 2 ou 3) et aucun comparateur de chaînes de caractères pour ces mêmes zones (D = 1), et entre corriger des poids pour les cas de concordance corrélée (F = 2) et ne pas en corriger (F = 1).

4.4 Interactions à deux critères des traitements

Le plus important contraste d'interactions à deux critères des traitements parmi ceux analysés par Belin (1991) est l'effet F × G, c'est-à-dire l'interaction de la correction de poids pour les cas de concordance corrélée (des prénoms et des seconds prénoms ou des prénoms, des sexes et des âges) et de l'inclusion (ou de la non-inclusion) de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement. Ce contraste est statistiquement significatif selon n'importe laquelle des méthodes utilisées dans Belin (1991) pour estimer un niveau de bruit sous-jacent. Dans le tableau 4.3 ci-dessous, nous donnons la valeur moyenne des variables de résultat pour les quatre combinaisons des traitements.

Tableau 4.3

Rendement moyen de combinaisons des traitements F et G

F	G	Taux d'erreur d'appariement	Arc sinus ($\sqrt{fmr}$)
1	1	0.0182	0.116
1	2	0.0143	0.097
2	1	0.0128	0.091
2	2	0.0151	0.100

Les données de ce tableau laissent supposer que la correction de poids pour les cas de concordance corrélée (niveau 2 du facteur F) est très utile lorsque l'état matrimonial et le lien avec le chef de ménage ne sont pas inclus dans les variables d'appariement (niveau 1 du facteur G) mais qu'elle ne change rien, en moyenne, lorsque ces deux mêmes variables sont comprises dans les variables d'appariement. Le fait que nous soyons capables de constater ces effets souligne l'importance de poursuivre les études empiriques dans un cadre expérimental.

Les deux plus importants contrastes d'interactions à deux critères des traitements relevés par Belin (1991) après l'interaction F × G forment une partie de l'interaction A × B (qui consiste dans le choix de poids et de comparateurs de chaînes de caractères pour les zones de nom). Nous donnons ci-dessous dans le tableau 4.4 les résultats moyens pour toutes les combinaisons des traitements faisant intervenir les facteurs A et B.

Tableau 4.4

Rendement moyen des combinaisons des traitements A et B

A	B	Taux d'erreur d'appariement	Arc sinus ($\sqrt{fmr}$)
1	1	0.0192	0.120
1	2	0.0140	0.099
1	3	0.0143	0.100
2	1	0.0170	0.110
2	2	0.0120	0.087
2	3	0.0123	0.089
3	1	0.0177	0.113
3	2	0.0118	0.084
3	3	0.0119	0.083
4	1	0.0254	0.145
4	2	0.0193	0.123
4	3	0.0189	0.122
5	1	0.0109	0.079
5	2	0.0109	0.079
5	3	0.0109	0.078

Par ce tableau nous voyons que lorsqu'on utilise des poids basés sur la fréquence pour le nom (A = 5), la méthode de comparaison de chaînes de caractères ne fait pas de différence, mais lorsqu'on utilise des poids quelconques ou des poids de "Fellegi-Sunter", l'emploi de méthodes de comparaison de chaînes de caractères accroît sensiblement l'efficacité moyenne de l'opération d'appariement informatisé.

Voici quelques-unes des autres observations intéressantes de Belin (1991) tirées de l'analyse des effets d'interactions à deux critères les plus importants:

1) La méthode de Winkler pour le traitement des cas de concordance imparfaite dans les zones autres que les zones de nom (c.-à-d., D = 3), qui est en moyenne le meilleur traitement en ce qui concerne le facteur D, est plus qu'utile en règle générale lorsque l'état matrimonial et le lien avec le chef de ménage sont inclus dans les variables d'appariement (c.-à-d., G = 2), même si ces variables ne sont associées d'aucune manière au traitement des cas de concordance imparfaite.
2) Contrairement aux autres méthodes de pondération des zones de nom, qui s'accommodent très bien de l'inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement, la méthode basée sur la fréquence semble être dérangée par une telle inclusion.
3) L'attribution de poids d'une valeur de ±6 pour les zones de nom donne de meilleurs résultats en moyenne lorsqu'elle est combinée à la méthode de pondération ponctuelle pour les zones autres que les zones de nom; l'attribution de poids d'une valeur de ±4 et de ±2 pour les zones de nom donne de meilleurs résultats lorsque combinée à la méthode de pondération fondée sur l'algorithme de Fellegi-Sunter pour les zones autres que les zones de nom.

4) Les poids de "Fellegi-Sunter" pour les zones autres que les zones de nom donnent de meilleurs résultats que les poids déterminés de façon ponctuelle lorsqu'il n'y a pas de correction de poids pour les cas de concordance corrélée, alors que c'est l'inverse lorsqu'on effectue une telle correction. (Cependant, d'après la méthode décrite dans Belin (1991) pour estimer le niveau de bruit sous-jacent, ce phénomène ne devrait pas nécessairement être observé dans d'autres régions d'essai.)

4.5 Quelle combinaison de traitements donne les meilleurs résultats?

Pour conclure l'analyse des résultats expérimentaux, nous allons chercher à déterminer les combinaisons de traitements qui sont le plus efficaces. Pour mesurer le rendement d'une combinaison de traitements donnée, nous considérons la valeur moyenne de la variable de résultat pour les trois régions d'essai données. La variable de résultat étudiée est le taux d'erreur d'appariement pour diverses proportions des enregistrements du fichier de l'EP qui peuvent être couplés (60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5% et 90%). Les résultats pertinents figurent dans le tableau 4.5.

Tableau 4.5

Les meilleurs combinaisons de traitements pour treize seuils d'exclusion d'après l'expérience factorielle

Seuil d'exclusion	Niveaux des facteurs formant les meilleures combinaisons (A B C D E F G H)								Taux d'erreur d'appariement moyenne pour les trois régions d'essai
60%	3	3	2	3	2	1	1	1	0.00042
62.5%	3	3	1	3	1	1	2	1	0.00047
65%	3	3	1	3	2	2	2	2	0.00052
67.5%	3	3	2	3	2	2	1	1	0.00071
70%	2	3	2	3	1	2	1	2	0.00079
72.5%	5	2	2	3	1	1	1	2	0.00081
75%	5	1	1	3	2	1	2	1	0.00112
77.5%	3	3	1	3	2	1	2	1	0.00133
80%	2	3	2	3	1	2	1	2	0.00188
82.5%	3	3	1	3	2	2	1	1	0.00571
85%	5	1	2	3	2	2	1	2	0.01556
87.5%	2	3	2	3	1	2	1	2	0.03023
90%	2	3	2	3	1	2	1	2	0.05174

Ces résultats contrastent avec ce que nous avons observé plus tôt; en effet, ils donnent à penser que la méthode de pondération basée sur la fréquence pour les zones de nom (niveau 5 du facteur A) est supérieure, en moyenne, à la méthode de pondération ponctuelle qui attribue des poids d'une valeur de ± 6 aux zones de nom (niveau 3 du facteur A). Cet état de fait serait imputable apparemment à certains effets d'interaction. Lorsque la méthode ponctuelle est utilisée conjointement avec les niveaux appropriés des autres facteurs, elle semble être au moins aussi efficace que la méthode basée sur la fréquence. Nous remarquons aussi que la combinaison de traitements 2-1 pour les facteurs F et G respectivement n'est pas toujours la meilleure combinaison possible dans leur cas, malgré ce que nous avons constaté plus haut. Seul le traitement 3 du facteur D (le volet qui correspond à l'utilisation de la méthode de Winkler pour les cas de concordance imparfaite dans les zones autres que les zones de nom) s'impose comme le meilleur traitement peu importe la façon dont on mesure les résultats de l'expérience. En ce qui concerne le meilleur traitement pour l'attribution de poids pour les zones de nom, le choix doit se faire entre l'une ou l'autre des méthodes déterministes qui attribuent des poids de ± 6 ou ± 4 et la méthode basée sur la fréquence. Lorsque l'une ou l'autre des méthodes déterministes est utilisée, l'usage de la méthode de comparaison de chaînes de caractères de Winkler est recommandé; lorsque c'est la méthode basée sur la fréquence qui est utilisée, il n'est pas sûr que l'on doive recourir à un comparateur de chaînes de caractères pour les noms.

Il est difficile de trancher entre les poids de "Fellegi-Sunter" pour les zones autres que les zones de nom et les poids déterminés de façon ponctuelle, mais les analyses antérieures laissent supposer que l'effet est négligeable d'un côté comme de l'autre. On peut faire les mêmes commentaires quant à l'utilisation de prénoms normalisés ou non et l'utilisation de quatre ou sept chiffres du numéro de téléphone.

Si l'on tient compte du fait qu'il n'existe pas de combinaison de traitements qui soit supérieure à toutes les autres sur tous les plans, on peut examiner le rendement de différentes combinaisons dans une région particulière (par exemple, une région où le taux d'erreur d'appariement se situe autour de 0.001). Or, si nous regardons les meilleures combinaisons qui existent pour une région où 70 à 80% des enregistrements du fichier de l'EP appartiennent à des concordances désignées (c'est-à-dire que l'on réduit l'analyse à cinq seuils d'exclusion), aucune combinaison ne se détache vraiment des autres. Si on fait la moyenne pour les cinq seuils d'exclusion, la meilleure combinaison de traitements possible est (2,3,2,3,1,2,1,2), c'est-à-dire, attribution de poids d'une valeur de ± 4 pour les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones de nom, estimation de poids au moyen de l'algorithme de Fellegi-Sunter pour les zones autres que les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones autres que les zones de nom, utilisation de la forme non normalisée du prénom, correction des poids composites pour les cas de concordance corrélée, non-inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement, et utilisation des sept chiffres du numéro de téléphone.

Tableau 4.6

Taux d'erreur d'appariement moyen pour diverses combinaisons de traitements (moyenne calculée pour trois régions d'essai et cinq seuils d'exclusion: (70%, 72.5%, 75%, 77.5% et 80% des enregistrements du fichier de l'EP ayant pu être couplés)

Niveaux des facteurs formant la combinaison (A B C D E F G H)	Taux d'erreur d'appariement 70%, 72.5%, 75%, 77.5% et 80% moyenne pour trois régions d'essai et cinq seuils d'exclusion
3 3 2 3 2 1 1 1	0.00493
3 3 1 3 1 1 2 1	0.00154
3 3 1 3 2 2 2 2	0.00137
3 3 2 3 2 2 1 1	0.00161
2 3 2 3 1 2 1 2	0.00124
5 2 2 3 1 1 1 2	0.00191
5 1 1 3 2 1 2 1	0.00153
3 3 1 3 2 1 2 1	0.00138
3 3 1 3 2 2 1 1	0.00156
5 1 2 3 2 2 1 2	0.00155

À des fins de comparaison, nous présentons dans le tableau 4.6 le rendement moyen de quelques-unes des combinaisons de traitements susceptibles de figurer parmi les plus efficaces. Ainsi donc, il semble que les meilleures combinaisons après (2,3,2,3,1,2,1,2) soient (3,3,1,3,2,2,2,2) et (3,3,1,3,2,1,2,1). Celles-ci se distinguent de la première par les points suivants: attribution de poids d'une valeur de ± 6 pour les zones de nom, attribution de poids d'une valeur préétablie pour les zones autres que les zones de nom, application de la méthode de Winkler pour l'attribution de poids pour les cas de concordance imparfaite dans les zones de nom comme dans les autres zones, utilisation de la forme normalisée du prénom, et inclusion de l'état matrimonial et du lien avec le chef de ménage dans les variables d'appariement. Ces deux mêmes combinaisons diffèrent entre elles sur deux points: la première prévoit la correction des poids composites pour les cas de concordance corrélée ainsi que l'utilisation des sept chiffres du numéro de téléphone alors que la seconde ne prévoit pas de correction pour la concordance corrélée et exige l'utilisation des quatre derniers chiffres du numéro de téléphone. Les combinaisons qui prévoient l'utilisation de la pondération basée sur la fréquence pour les zones de nom ne sont pas aussi efficaces que les combinaisons qui prévoient l'attribution de poids déterminés de façon ponctuelle.

La combinaison de traitements qui a été utilisée dans les opérations d'appariement informatisé de l'EP de 1990 ressemblait beaucoup à la combinaison (5,3,2,3,2,2,2,1). En ce qui concerne les ensembles de données de prérecensement analysés ici, cette combinaison a produit un taux d'erreur d'appariement moyen (pour les cinq seuils d'exclusion) de 0.00179.

4.6 Conclusions

Tandis que les résultats de cette étude tentent de mettre en balance le nombre d'enregistrements qui appartiennent à des concordances désignées et le taux d'erreur d'appariement, un arbitre anonyme a souligné que "(...) chaque gain réalisé grâce à une meilleure méthode de couplage d'enregistrements doit être mis en parallèle avec le coût de mise en oeuvre de cette méthode." Cela est un autre arbitrage qui mérite d'être considéré par le praticien. Nous espérons que les résultats qui ont été présentés dans cette étude sur l'importance relative de divers facteurs dans le couplage d'enregistrements orienteront le travail des personnes qui élaborent et appliquent des logiciels de couplage. Comme certains résultats peuvent dépendre de caractéristiques particulières des données du recensement et de l'EP qui sont appariées, on peut se demander quel genre de rapport existe entre ces résultats et d'autres cadres de couplage d'enregistrements. Mais comme nous l'avons souligné au départ, l'étude du couplage d'enregistrements doit se faire idéalement à l'aide de méthodes expérimentales pour les besoins de la généralisation. Les études empiriques fondées sur des plans d'expérience constituent une source d'orientation authentique et éprouvée qui permet, par un cadre bien défini, d'élargir les connaissances des spécialistes du couplage d'enregistrements.

REMERCIEMENTS

Cette étude a été réalisée en majeure partie pendant que l'auteur travaillait à la division du couplage des enregistrements du Bureau of the Census des É.-U. à Washington, DC. L'auteur tient à exprimer toute sa gratitude à Don Rubin, Bill Winkler et Alan Zaslavsky ainsi qu'à l'arbitre anonyme pour les échanges fructueux auxquels ils ont participé et les commentaires utiles qu'ils ont formulés. L'auteur tient à souligner aussi l'aide qu'il a reçue grâce aux conventions sur la statistique n° 88-02 et 89-07 pendant qu'il poursuivait des études de doctorat à l'Université Harvard.

BIBLIOGRAPHIE

ABBATT, J.D. (1986). A cohort study of eldorado uranium workers. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Éds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 51-57.

ACHESON, E.D. (1967). *Medical Record Linkage*. Oxford: Oxford University Press.

ACHESON, E.D. (Éd.) (1968). *Record Linkage in Medicine*, Edinburgh: E. & S. Livingstone.

BALDWIN, J.A., ACHESON, E.D., et GRAHAM, W.J. (Éds.) (1987). *A Textbook of Medical Record Linkage*. Oxford: Oxford University Press.

BELIN, T.R. (1989a). Outline of procedure for evaluating computer matching in a factorial experiment. Note de service non-publiée, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1989b). Results from evaluation of computer matching. Note de service non publiée, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1991). Using mixture models to calibrate error rates in record-linkage procedures, with application to computer-matching for census undercount estimation. Thèse de doctorat, Department of Statistics, Harvard University. (Publié par University Microfilms, Inc.)

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

BOYLE, C.A., et DECOUFLÉ, P. (1990). National sources of vital status information: Extent of coverage and possible selectivity in reporting. *American Journal of Epidemiology*, 131, 160-168.

BROWN, P., LAPLANT, W., LYNCH, M., ODELL, S., THIBAUDEAU, Y., et WINKLER, W. (1988). Collective Documentation for the 1988 PES Computer Match Processing and Printing. Vols. I-III, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BUREAU OF THE CENSUS (1988-1991). 1990 Decennial Census Information Memorandum Series, Decennial Planning Division, Bureau of the Census, Washington, D.C.

[Nota: Tous les rapports de la série mentionnée ci-dessus sont accompagnés de l'avertissement suivant:

"Ces documents ont été rédigés à l'intention des divisions de la planification et de l'exploitation du Census Bureau; ces divisions connaissent bien les données de base, les expériences antérieures, la terminologie et les procédés de même que le cadre général du recensement décennal, ses objectifs et la relation mutuelle des opérations et des systèmes. Ces documents ne sont donc PAS [soulignement tiré de l'original] destinés à la diffusion et ne doivent pas sortir du Census Bureau sans l'autorisation préalable de Jim Dinwiddie ([301]-763-5270) de la division de la planification du recensement décennal."]

BUREAU OF THE CENSUS (1987-1991). STSD Decennial Census Memorandum Series, Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

CARPENTER, M., et FAIR, M.E. (Éds.) (1990). Canadian Epidemiology Research Conference - 1989: *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Centre for Health Information, Statistique Canada, Ottawa, Ontario.

CHERNOFF, H. (1980). The identification of an element of a large population in the presence of noise. *Annals of Statistics*, 8, 1179-1197.

CHILDERS, D. (1989). 1990 PES Within Block Matching - Clerical Matching Group. STSD Decennial Census Memorandum Series #V-69, U.S. Bureau of the Census, Washington, D.C.

CITRO, C.F., et COHEN, M.L. (Éds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*. Washington, D.C.: National Academy Press.

COHEN, M.L. (1990). Adjustment and reapportionment - Analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.

COOMBS, J.W., et SINGH, M.P. (Éds.) (1988). *Recueil: Symposium sur les utilisations statistiques des données Administratives*, Statistique Canada, Ottawa, Ontario.

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

CURB, J.D., FORD, C.E., PRESSEL, S., PALMER, M., BABCOCK, C., et HAWKINS, C.M. (1985). Ascertainment of vital status through the National Death Index and the Social Security Administration. *American Journal of Epidemiology*, 121, 754-766.

DANIEL, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.

DEMPSTER, A.P., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Séries B, 39, 1-38.

DONOGHUE, G. (1990). Clerical Specifications for the 1990 Post Enumeration Survey Before Followup Matching - Special Matching Group. STSD Decennial Census Memorandum Series #V-92, U.S. Bureau of the Census, Washington, D.C.

DULBERG, C.S., SPASOFF, R.A., et RAMAN, S. (1986). Reactor clean-up and bomb test exposure study. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Éds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 59-62.

ERICKSEN, E.P., et KADANE, J.B. (1985). Estimating the population in a census year: 1980 and Beyond (avec discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSON, E.P., KADANE, J.B., et TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.

FAGERLIND, I. (1975). *Formal Education and Adult Earnings: A Longitudinal Study on the Economic Benefits of Education*, Stockholm: Almquist and Wiksell.

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

FREEDMAN, D.A., et NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (avec discussion). *Statistical Science*, 1, 1-39.

GOLDACRE, M.J. (1986). The Oxford record linkage study: Current position and future prospects. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press, 106-129.

HILL, T. (1981). Generalized Iterative Record Linkage System: GIRLS. (Glossary, Concepts, Strategy Guide, User Guide). Division du développement de système, Statistique Canada, Ottawa, Ontario.

HILL, T., et PRING-MILL, F. (1986). Generalized iterative record linkage system: GIRLS, (édition revisé). Division du développement de système, Statistique Canada, Ottawa, Ontario.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269.

HOWE, G.R., et LINDSAY, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research*, 14, 327-340.

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

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

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

JOHANSEN, H.L. (1986). Record linkage of national surveys: The Nutrition Canada example. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Éds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 153-163.

JOHNSON, E.G., et TUKEY, J.W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson et Tsao Data. Dans *Design, Data and Analysis*, (Éd. C.L. Mallows) New York: John Wiley and Sons.

JOHNSON, R.A. (1991). Methodology for Evaluating Errors in U.S. Department of Justice Attorney Workload Data. Rapport techniques non-publié, General Accounting Office, Washington, D.C.

KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.

KERSHAW, D., et FAIR, J. (1979). *The New Jersey Income and Maintenance Experiment: Operations, Surveys, and Administration*, Volume I. New York: Academic Press.

KILSS, B., et ALVEY, W. (Éds.) (1984a). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. I, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et ALVEY, W. (Éds.) (1984b). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. II, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et ALVEY, W. (Éds.) (1984c). *Statistics of Income and Related Administrative Record Research:* 1984, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et ALVEY, W. (Éds.) (1985). *Record Linkage Techniques - 1985*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et ALVEY, W. (Éds.) (1987). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et JAMERSON, B. (Éds.) (1990). *Statistics of Income and Related Administrative Record Research 1988-1989*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., et SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin*, Vol. 41, 10, 14-22.

LAPLANT, W. (1988). User's Guide for the Generalized Record Linkage Program Generator (GENLINK). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

LAPLANT, W. (1989). User's Guide for the Generalized Address Standardizer (GENSTAN). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

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

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

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

NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., et ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. *Computers in Biology and Medicine*, 13, 157-169.

NICHOLL, J.P. (1986). The use of hospital in-patient data in the analysis of the injuries sustained by road accident casualties. Dans *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Éds. G.R. Howe et R.A. Spasoff). Toronto: University of Toronto Press, 243-244.

PALETZ, D. (1989). Name standardization software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

ROGOT, E., SORLIE, P.D., et JOHNSON, N.J. (1986). Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*, 39, 719-734.

ROGOT, E., SORLIE, P.D., JOHNSON, N.J., GLOVER, C.S., et TREASURE, D.W. (1988). A Mortality Study of One Million Persons. Public Health Service, National Institutes of Health, Washington, D.C.

SCHIRM, A.L., et PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (avec discussion). *Journal of the American Statistical Association*, 82, 965-990.

SMITH, M.E., et NEWCOMBE, H.B. (1975). Methods for computer linkage of hospital admission-separation records for cumulative health histories. *Methods of Information in Medicine*, 14, 118-125.

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

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

WENTWORTH, D.N., NEATON, J.D., et RASMUSSEN, W.L. (1983). An evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the ascertainment of vital status. *American Journal of Public Health*, 73, 1270-1274.

WILLIAMS, B.C., DEMITRACK, L.B., et FRIES, B.E. (1992). The accuracy of the National Death Index when personal identifiers other than Social Security Number are used. *American Journal of Public Health*, 82, 1145-1147.

WINKLER, W.E. (1985a). Preprocessing of lists and string comparison. Dans *Record Linkage Techniques - 1985*, (Éds. W. Alvey et B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 181-187.

WINKLER, W.E. (1985b). Exact matching lists of businesses: blocking, subfield identification, and Information Theory. Dans *Record Linkage Techniques - 1985*, (Éds. W. Alvey et B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 227-241.

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

WINKLER, W.E. (1991). Documentation of record-linkage software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WINKLER, W.E., et THIBAUDEAU, Y. (1992). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.