Techniques d'enquête, décembre 1995
Vol. 21, n° 2, pp. 161–171
Statistique Canada

Un algorithme de transport à taille réduite pour maximiser le chevauchement des enquêtes

LAWRENCE R. ERNST et MICHAEL M. IKEDA[1]

RÉSUMÉ

Lorsqu'on remanie un échantillon selon un plan stratifié à plusieurs degrés, il est parfois indiqué de maximiser le nombre d'unités primaires d'échantillonnage retenues dans le nouvel échantillon sans modifier les probabilités de sélection inconditionnelle. Il existe à cette fin une solution optimale qui s'appuie sur la théorie du transport pour une classe très générale de plans d'échantillonnage. Toutefois, à la connaissance des auteurs, cette méthode n'a jamais été utilisée pour la refonte d'une enquête. Cela s'explique en partie du fait que même pour une strate de taille modérée, le problème de transport résultant pourrait être trop grand pour se prêter à une solution pratique. Dans le présent article, nous proposons un algorithme de transport modifié à taille réduite permettant de maximiser le chevauchement, ce qui réduit sensiblement les dimensions du problème. Cette méthode a été utilisée lors du remaniement récent de l'enquête sur le revenu et la participation aux programmes (Survey of Income and Program Participation, ou SIPP). Nous décrivons brièvement le rendement de l'algorithme à taille réduite, d'une part pour le chevauchement du SIPP en conditions réelles, et d'autre part pour des simulations artificielles antérieures du chevauchement du SIPP. Même si cette méthode n'est pas optimale et ne risque de produire, en théorie, que des améliorations négligeables du chevauchement attendu comparativement à la sélection indépendante, elle ouvre en pratique la voie à des améliorations importantes du chevauchement par rapport à la sélection indépendante dans le cas du SIPP et produit généralement un chevauchement presque optimal.

MOTS CLÉS: Programmation linéaire; remaniement de l'échantillon; Survey of Income and Program Participation.

1. INTRODUCTION

Le problème de la maximisation du nombre prévu d'unités primaires d'échantillonnage (UPÉ) retenues dans un échantillon lors du remaniement d'une enquête avec un plan stratifié pour lequel les UPÉ sont choisies avec une probabilité proportionnelle à la taille a été abordé pour la première fois dans la documentation scientifique par Keyfitz (1951). Typiquement, la maximisation du chevauchement des UPÉ est motivée par la réduction des coûts supplémentaires tels que ceux requis pour la formation des intervieweurs aux fins d'une enquête auprès des ménages, lesquels s'additionnent à chaque changement d'UPÉ. Les méthodes de maximisation du chevauchement n'influent pas sur la probabilité inconditionnelle de sélection d'un ensemble d'UPÉ dans une nouvelle strate, mais conditionnent sa probabilité de sélection d'une manière telle que la probabilité de sélection d'une UPÉ dans le nouvel échantillon est en règle générale plus grande que la probabilité inconditionnelle lorsque cette UPÉ se trouvait dans l'échantillon initial, et moins grande autrement.

Les méthodes de chevauchement sont applicables lorsque le remaniement entraîne soit une nouvelle stratification des UPÉ, soit un changement de leurs probabilités de sélection. Keyfitz (1951) a proposé une méthode optimale, mais uniquement pour les plans à une UPÉ par strate dans le cas spécial où la strate initiale et la nouvelle strate sont identiques et où seules les probabilités de sélection changent. Causey, Cox et Ernst (1985) ont obtenu une solution optimale du problème du chevauchement sous des conditions très générales en assimilant ce problème à un problème de transport: une forme spéciale de problème de programmation linéaire. Cette méthode n'impose aucune restriction aux changements de définitions des strates ni au nombre d'UPÉ par strate. (Un résultat similaire a été obtenu de façon indépendante par Arthanari et Dodge (1981), même si ces derniers n'ont pas abordé la question des changements dans les définitions des strates. Les deux groupes de chercheurs ont obtenus leurs résultats en généralisant le travail de Raj (1968).) Toutefois, il existe au moins deux autres difficultés avec la méthode de Causey, Cox et Ernst qui peuvent la rendre inutilisable en pratique. La première est examinée par Ernst (1986), et la seconde fait l'objet du présent article.

La première difficulté découle du fait que si l'échantillon initial d'UPÉ n'a pas été sélectionné indépendamment d'une strate à l'autre, l'information nécessaire au calcul de l'ensemble des probabilités conjointes requises par cette méthode risque de ne pas être disponible en pratique. Ernst (1986) a mis au point une méthode de programmation linéaire de rechange à utiliser dans de tels cas. Le Bureau of the Census a eu recours à la programmation linéaire pour réaliser le chevauchement de ses enquêtes démographiques à cinq reprises. À quatre de ces occasions (sélection des plans pour les Current Population Surveys (CPS) et pour les National Crime Victimization Surveys (NCVS) des années 1980 et 1990), la méthode établie par Ernst (1986) a été retenue puisque le plan initial

[1] Lawrence R. Ernst, Chief, Research Group, Office of Compensation and Working Conditions, Bureau of Labor Statistics, Washington, DC 20212, U.S.A.; Michael M. Ikeda, Mathematical Statistician, Statistical Research Division, Bureau of the Census, Washington, DC 20233, U.S.A.

n'était pas sélectionné indépendamment d'une strate à l'autre. En particulier, comme l'avait expliqué Ernst (1986), si l'échantillon initial est lui-même sélectionné par chevauchement avec un plan antérieur, l'hypothèse d'indépendance n'est en général pas valide, ce qui expliquait principalement pourquoi elle n'était pas valide dans ces quatre cas particuliers.

Le second problème posé par la méthode optimale est que le problème du transport risque d'être trop grand pour pouvoir être résolu en pratique. Le Bureau of the Census a également utilisé la programmation linéaire pour le chevauchement du plan du SIPP des années 1990 avec celui des années 1980, deux plans à deux UPÉ par strate. L'échantillon initial du SIPP a été sélectionné indépendamment d'une strate à l'autre. Toutefois, le problème du transport pour la méthode optimale aurait été trop grand pour autoriser une solution pratique pour plusieurs des strates. Cette difficulté se pose du fait que pour chaque nouvelle strate constituée de n UPÉ, le nombre de variables du problème de transport pour la méthode optimale peut atteindre $2^n \times \binom{n}{2}$. La valeur de n la plus grande pour laquelle un problème de transport comportant un tel nombre de variables peut être résolu avec les moyens informatiques que nous avions à notre disposition était approximativement $n = 15$.

Nous présentons ci-après une formule à taille réduite de la méthode de chevauchement, assimilée à un problème de transport, qui réduit le nombre de variables du SIPP à $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, une réduction surprenante pour les valeurs de n modérées à grandes. Cette méthode part de l'hypothèse que l'échantillon initial a été sélectionné indépendamment d'une strate à l'autre; elle n'aurait de ce fait pas pu remplacer la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS. Cette méthode à taille réduite a été utilisée avec succès pour des strates comptant jusqu'à 68 UPÉ. À titre de comparaison, pour un $n = 68$, le nombre possible de $2^{68} \times \binom{68}{2}$ variables de la formule non réduite dépasse de loin la taille des problèmes qui peuvent être résolus avec les ordinateurs courants. En outre, même si la réduction de la taille se fait au détriment de l'optimalité, il semble qu'elle donne en pratique des résultats passablement proches des valeurs optimales, comme nous le démontrerons ci-après. La méthode à taille réduite est celle que nous avons utilisée pour le chevauchement du SIPP.

L'examen de la méthode de Causey, Cox et Ernst (1985), présenté dans la section 2, nous sert de toile de fond pour la présentation de la méthode à taille réduite.

La méthode à taille réduite est présentée à la section 3. Même si elle peut se prêter à une application générale, nous nous contenterons, pour simplifier la présentation, de ne décrire en détails que le cas où le plan initial et le nouveau plan comportent deux UPÉ par strate et se font sans remise. Nous présentons également, dans la même section, un petit exemple artificiel de la méthode à taille réduite qui servira à illustrer la méthode et à démontrer que l'ordonnancement des paires d'UPÉ dans la strate d'un nouveau plan, une étape clé de l'algorithme, influe sur le chevauchement attendu. Nous soulignons également dans cette section certains résultats analytiques de la comparaison de la méthode à taille réduite et de la méthode optimale. Nous précisons les limites supérieures de la perte du chevauchement attendu découlant de l'utilisation de la méthode à taille réduite au lieu de la méthode optimale. Nous expliquons également que dans certaines situations, cette perte peut s'approcher de deux UPÉ pour les plans à deux UPÉ par strate – le pire scénario envisageable. On trouvera dans l'article de Ernst et Ikeda (1994) de plus amples détails ainsi que les preuves des résultats présentés dans cette section, en plus de certains résultats d'autres sections du présent article.

Dans la section 4, nous abordons la question du rendement de la méthode à taille réduite, tant pour le chevauchement du SIPP actuel que pour des simulations artificielles du chevauchement des SIPP antérieurs. Le chevauchement attendu avec cette méthode est comparé à celui de la sélection indépendante des UPÉ du nouvel échantillon et à la limite supérieure du chevauchement optimal attendu. Les résultats montrent que pour cette application, et contrairement à certains des résultats théoriques décrits dans la section 3, le chevauchement attendu avec la méthode à taille réduite est beaucoup plus grand que si on avait recouru à une sélection indépendante pour choisir les UPÉ du nouvel échantillon. Ce chevauchement est en fait presque aussi grand que le chevauchement optimal prévu. Nous précisons finalement le temps machine nécessaire pour la méthode à taille réduite en fonction de la taille des strates.

Finalement, les conclusions de notre étude sont présentées à la section 5.

2. EXAMEN DE LA MÉTHODE DE CHEVAUCHEMENT DE CAUSEY, COX ET ERNST (1985)

La méthode de chevauchement de Causey, Cox et Ernst (1985), comme toutes les méthodes de chevauchement, conditionne d'une certaine manière la sélection des UPÉ des échantillons de chaque nouvelle strate au choix des UPÉ de la strate qui faisaient partie de l'échantillon initial. Cette méthode de chevauchement particulière réalise une optimalité réelle en faisant plein usage de cette information et en assimilant la procédure à un problème de transport. Nous en présentons ci-après la description.

Précisons tout d'abord certaines des notations qui nous serviront tout au long de notre exposé. Désignons par S une strate du nouveau plan d'échantillonnage. Chacune de ces strates correspond à un problème de chevauchement séparé. Désignons par n le nombre d'UPÉ comprises dans S, et par $A_1, \ldots, A_n$ les UPÉ en question comprises dans S. Désignons par I le sous-ensemble aléatoire de $\{1, \ldots, n\}$ tel que $k \in I$ si et seulement si A_k était présent dans l'échantillon initial, et désignons par N l'ensemble correspondant du nouvel échantillon. Par exemple, si A_2 et A_3 étaient les UPÉ de S qui faisaient partie de l'échantillon initial et A_1 et A_3 les UPÉ comprises dans le nouvel échantillon, alors, $I = \{2,3\}$ et $N = \{1,3\}$. Désignons par m^*, n^* le nombre de valeurs possibles de I et N respectivement.

Désignons par ailleurs par J_i, $i = 1, \ldots, m^*$ les valeurs possibles de I et par S_j, $j = 1, \ldots, n^*$ les valeurs possibles de N. L'objet de toutes les méthodes de chevauchement est de maximiser le nombre prévu d'UPÉ dans $N \cap I$, tout en préservant les valeurs des $P(S_j)$.

Pour illustrer plus avant certaines de ces notions, supposons un exemple où $n = 3$. Alors, $n^* = 3$ si le nouveau plan compte une ou deux UPÉ par strate, les valeurs de N, c'est-à-dire les S_j, étant $\{1\},\{2\},\{3\}$ dans le cas à une UPÉ par strate et $\{1,2\},\{1,3\},\{2,3\}$ dans le cas à deux UPÉ par strate. Supposons maintenant que les UPÉ A_1 et A_2 étaient dans une strate initiale et que l'UPÉ A_3 était dans une autre, et qu'il y avait trois UPÉ dans chacune de ces strates initiales. Si le plan initial comptait une UPÉ par strate, alors $m^* = 6$, et les valeurs de I, c'est-à-dire les J_i, étaient $\varnothing$, $\{1\},\{2\},\{3\},\{1,3\},\{2,3\}$. Si le plan initial comptait deux UPÉ par strate, alors $m^* = 6$, et les valeurs des J_i étaient $\{1\},\{2\},\{1,2\},\{1,3\},\{2,3\}$, $\{1,2,3\}$.

Examinons maintenant le problème de transport pour la méthode de chevauchement de Causey, Cox et Ernst (1985). Soit $P(J_i)$ la probabilité que $I = J_i$ et $P(S_j)$ la probabilité que $N = S_j$. En outre, désignons par x_{ij} la variable dénotant la probabilité conjointe de ces deux événements, et désignons par c_{ij} le nombre d'éléments dans $J_i \cap S_j$. Les valeurs des $P(J_i)$, des $P(S_j)$ et des c_{ij} sont connues alors que celles des x_{ij} sont des variables dont il s'agit de déterminer les valeurs optimales. Ainsi, le problème de transport à résoudre consiste à déterminer la valeur de $x_{ij} \geq 0$ qui maximise

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \tag{2.1}$$

sous réserve que

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \quad i = 1, \ldots, m^*, \tag{2.2}$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \quad j = 1, \ldots, n^*. \tag{2.3}$$

Noter que dans ce problème de transport, la fonction objective (2.1) est le nombre prévu d'UPÉ de S qui sont dans $N \cap I$. Noter également que les contraintes (2.2) et (2.3) sont requises par les définitions des $P(J_i)$, des $P(S_j)$ et des x_{ij}.

Une fois obtenues les valeurs optimales x_{ij}, la probabilité conditionnelle que $N = S_j$ étant donné $I = J_i$ est donnée par $x_{ij}/P(J_i)$ pour tous les i,j.

Nous présentons ci-après un exemple de l'utilisation des formules (2.1) à (2.3) dans un cas où le plan initial et le plan nouveau sont tous les deux à deux UPÉ par strate et sans remise. Dans cet exemple comme partout ailleurs dans le présent article, nous désignerons par p_i, π_i la probabilité prédéterminée que $i \in I$ et $i \in N$, respectivement.

Supposons une strate finale S avec $n = 3$. Toutes les UPÉ étaient dans des strates différentes. Soit $p_1 = .6$, $p_2 = .75$, $p_3 = .7$, $\pi_1 = .5$, $\pi_2 = .8$, $\pi_3 = .7$. Puisque les UPÉ étaient toutes dans des strates initiales différentes, il existe 8 possibilités différentes pour I, dont les probabilités sont présentées dans le tableau 1.

Tableau 1

Probabilités des groupes possibles d'UPÉ de l'échantillon initial

i	1	2	3	4	5	6	7	8
J_i	$\{1,2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\varnothing$
$P(J_i)$	.315	.135	.105	.21	.045	.09	.07	.03

Puisque le nouveau plan d'échantillonnage est à deux UPÉ par strate et sans remise, il existe trois possibilités pour N: les paires $S_1 = \{1,2\}$, $S_2 = \{1,3\}$, $S_3 = \{2,3\}$. Ainsi, $P(S_1) = .30$, $P(S_2) = .20$ et $P(S_3) = .50$.

Les valeurs c_{ij} sont présentées dans le tableau 2. En maximisant ensuite (2.1) sous réserve de (2.2) et (2.3) avec les valeurs données des $P(J_i)$, $P(S_j)$ et c_{ij}, on obtient le groupe de valeurs optimales de x_{ij} présentées dans le tableau 2. Finalement, en divisant chacune des entrées x_{ij} d'un rang i du tableau 2 par $P(J_i)$, on obtient un groupe optimal de probabilités conditionnelles $P(S_j \mid J_i)$. Par exemple, puisque $x_{12} = .025$ et $P(J_1) = .315$, il s'ensuit que $P(S_2 \mid J_1) = 5/63$.

Tableau 2

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode optimale

	c_{ij}			x_{ij}		
	j			j		
i	1	2	3	1	2	3
1	2	2	2	.000	.025	.290
2	2	1	1	.135	.000	.000
3	1	2	1	.000	.105	.000
4	1	1	2	.000	.000	.210
5	1	1	0	.045	.000	.000
6	1	0	1	.090	.000	.000
7	0	1	1	.000	.070	.000
8	0	0	0	.030	.000	.000

Pour cet exemple, comme on peut le calculer à partir de (2.1) et du tableau 2, le chevauchement attendu en vertu de la procédure optimale est de 1.735 UPÉ. À titre comparatif, le chevauchement attendu si les plans initial et final sont sélectionnés indépendamment est de $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ UPÉ.

Pour les plans d'échantillonnage sans remise à deux UPÉ par strate, les valeurs possibles de N sont toujours données par les sous-ensembles $\binom{n}{2}$ de $\{1, \ldots, n\}$ de

taille 2, c'est-à-dire $n^* = \binom{n}{2}$. Toutefois, m^* peut varier largement. $m^* = \binom{n}{2}$ lorsque les UPÉ de S comprennent une seule strate initiale. La limite supérieure de 2^n sur m^* est atteinte lorsque toutes les UPÉ de S étaient dans des strates initiales différentes, comme le montre l'exemple précédent, et comme dans certaines autres situations. Ernst et Ikeda (1994) présentent une expression générale exacte de m^*.

En ce qui concerne le chevauchement des plans d'échantillonnage sans remise à deux UPÉ par strate, le nombre de variables du problème de transport pour la méthode optimale est m^*n^*, et peut atteindre jusqu'à $2^n\binom{n}{2}$. Si $n = 15$, $2^n\binom{n}{2} = 3{,}440{,}640$, ce qui correspond à peu près à la capacité de résolution des ordinateurs que nous avions à notre disposition. Toutefois, $n > 15$ pour près de la moitié des strates qui ne sont pas auto-représentatives (c'est-à-dire celles constituées d'UPÉ sélectionnées avec probabilité) dans l'enquête SIPP, et il s'est donc avéré nécessaire d'élaborer une méthode, décrite dans la prochaine section, qui réduit le problème du transport tout en produisant en pratique un chevauchement attendu presque maximal.

3. L'ALGORITHME DE LA MÉTHODE À TAILLE RÉDUITE

Dans les travaux antérieurs portant sur la réduction de la taille du problème de transport défini par les formules (2.1) à (2.3), on s'est surtout attaché à réaliser la réduction de la taille tout en maintenant l'optimalité. Par exemple, Aragon et Pathak (1990) conservent l'optimalité et réduisent la taille du problème de 75% lorsque $m^* = n^*$. Malheureusement, lorsque m^* est beaucoup plus grand que n^*, c'est-à-dire lorsque la réduction de la taille serait la plus utile, leur méthode produit une réduction de taille négligeable en termes relatifs. Pathak et Fahimi (1992) proposent une généralisation de cette méthode, mais rien n'indique qu'elle puisse aboutir à coup sûr à une réduction de taille sensible en termes relatifs.

Dans la présente section, nous abordons le problème de la réduction de la taille sous un angle différent. Nous sacrifions l'optimalité, à tout le moins en théorie, pour obtenir en retour une réduction du problème de transport qui en autorise une résolution pratique. Nous y arrivons, dans les cas où les plans initial et nouveau comportent tous les deux deux UPÉ par strate par exemple, en ordonnant toutes les paires d'UPÉ dans une nouvelle strate puis en conditionnant les nouvelles probabilités de sélection pour tout ensemble initial d'UPÉ de l'échantillon de taille supérieure à 2 sur la première paire d'UPÉ de l'ensemble initial plutôt que sur la totalité de cet ensemble initial. Autrement dit, chaque ensemble initial possible d'UPÉ de l'échantillon comportant plus de deux UPÉ est combiné à un ensemble de taille 2. Comme nous l'expliquons à la section 4, cette méthode peut en pratique donner un chevauchement presque optimal, en particulier avec un ordonnancement approprié des paires d'UPÉ tel qu'il est décris à la section 3.1.2.

La méthode de réduction de la taille s'applique dans tous les cas où les UPÉ des plans initial et nouveau sont tirées sans remise. Toutefois, nous limiterons notre description détaillée, dans la section 3.1, aux cas où le plan initial et le nouveau plan comportent tous les deux deux UPÉ par strate. Nous expliquerons ensuite brièvement, dans la section 3.2, les changements nécessaires pour appliquer cette méthode à d'autres types de plans initiaux et nouveaux. Finalement, dans la section 3.3, nous présenterons certains résultats analytiques portant sur les rapports entre le chevauchement attendu de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante. Tout au long de la présente section, nous présumons au départ que les UPÉ de l'échantillon initial ont toutes été sélectionnées indépendamment d'une strate à l'autre.

3.1 Méthode de réduction de la taille pour deux plans à deux UPÉ par strate

La méthode que nous décrivons ci-après présente les aspects clés suivants: l'ordonnancement particulier des paires d'UPÉ; la refonte du problème de transport (2.1) à (2.3) aux fins de la méthode à taille réduite; le calcul des probabilités correspondant aux résultats initiaux de la nouvelle formulation; le calcul des coefficients de coût (c_{ij}) de la fonction objective. Dans la section 3.1.1, nous présentons une description détaillée de la méthode à taille réduite, et une nouvelle définition du problème de transport. L'ordonnancement des paires est décrit dans la section 3.1.2. Finalement, nous présentons dans la section 3.1.3 le calcul des probabilités des résultats initiaux et des coefficients de coût.

3.1.1 Description générale de la méthode

Dans une première étape, les sous-ensembles $\binom{n}{2}$ de $\{1, \ldots, n\}$ de taille 2 sont ordonnés selon une méthode que nous décrirons ultérieurement. (Qu'il nous suffise pour l'instant de mentionner que n'importe quel ordonnancement peut servir à réduire la taille du problème de transport. La méthode précise que nous retenons permet de parvenir à ce résultat en conservant la plus grande part possible des gains de la méthode optimale en matière de chevauchement.) Désignons par I_i, $i = 1, \ldots, \binom{n}{2}$, le i-ième élément de l'ordonnancement et par $I_{\binom{n}{2}+1}, \ldots, I_{\binom{n}{2}+n}$ les n sous-ensembles à un seul élément. Posons finalement que $I_{\binom{n}{2}+n+1} = \varnothing$. Ainsi, les I_i constituent chacun des sous-ensembles de $\{1, \ldots, n\}$ de deux éléments ou moins. À chaque possibilité de I correspond un ensemble I^* unique parmi ces $\binom{n}{2} + n + 1$ sous-ensembles, et les nouvelles probabilités de sélection sont conditionnées sur ce I^* plutôt que sur I. Ainsi, les nouvelles probabilités de sélection sont conditionnés sur $\binom{n}{2} + n + 1$ événements plutôt que sur un nombre possible de 2^n événements, ce qui explique la réduction de la taille. Le paramètre I^* est le premier I_i pour lequel $I_i \subset I$. C'est-à-dire que si I consiste en au moins deux nombres entiers, I^* correspondra à la première paire de l'ordonnancement contenue dans I, tandis que si I est un ensemble à un seul élément ou un ensemble vide, on obtiendra alors $I^* = I$.

Le problème de transport à taille réduite cherche à conserver les UPÉ correspondant aux éléments de l'ensemble I^* dans le nouvel échantillon, mais n'utilise pas l'information sur les éléments dans $I \sim I^*$. Ce problème de transport à taille réduite fondé sur l'ensemble de I_i prend la forme décrite ci-après. Désignons par p_i^* la probabilité que $I^* = I_i$, $i = 1, \ldots, \binom{n}{2} + n + 1$, et abrégeons $\pi_j^* = P(S_j)$, $j = 1, \ldots, \binom{n}{2}$. Pour chaque i,j, la variable x_{ij} correspond à la probabilité conjointe que $I^* = I_i$ et que $N = S_j$, alors que c_{ij} désigne le nombre attendu d'éléments dans $I \cap S_j$ étant donné $I^* = I_i$. Le problème consiste à déterminer le $x_{ij} \geq 0$ qui maximise

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij}\, x_{ij}, \tag{3.1}$$

sous réserve que

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \quad i = 1, \ldots, \binom{n}{2} + n + 1, \tag{3.2}$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \quad j = 1, \ldots, \binom{n}{2}. \tag{3.3}$$

Une fois obtenues les valeurs optimales de x_{ij}, les probabilités conditionnelles de nouvelle sélection pour S_j, $j = 1, \ldots, \binom{n}{2}$, étant donné $I^* = I_i$, sont x_{ij}/p_i^*. À noter que le nombre de variables x_{ij} dans les formules (3.1) à (3.3) est $\left(\binom{n}{2} + n + 1\right) \times \binom{n}{2}$, comparativement à un maximum de $2^n \times \binom{n}{2}$ dans les formules (2.1) à (2.3).

Il nous reste à expliquer la méthode générale utilisée pour ordonner les $\binom{n}{2}$ paires ainsi que les méthodes de calcul des valeurs de p_i^* et de c_{ij}. Avant cela, nous procéderons à une démonstration de la méthode à taille réduite avec l'exemple à deux UPÉ par strate utilisé dans la section 2 afin d'illustrer la formulation du problème de transport pour la méthode optimale.

L'ordonnancement des paires de cet exemple, comme il sera démontré plus tard, est $\{2,3\}, \{1,2\}, \{1,3\}$. En conséquence, les valeurs I_i sont présentées dans le tableau 3. Noter que si $I = \{1,2,3\}$ ou $I = \{2,3\}$, l'ensemble associé devient alors $I_1 = \{2,3\}$. Pour les six autres possibilités pour I l'ensemble associé est I lui-même.

Par conséquent, en utilisant les valeurs du tableau 1, nous obtenons

$$p_1^* = P(I = \{1,2,3\}) + P(I = \{2,3\}) = .525, \tag{3.4}$$

$p_i^* = P(J_i)$, $i = 2,3$, et $p_i^* = P(J_{i+1})$, $i = 4, \ldots, 7$, donnant les valeurs présentées au tableau 3. Puisque $\pi_j^* = P(S_j)$, nous obtenons $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Tableau 3

Probabilités des ensembles associés: méthode à taille réduite

	i						
	1	2	3	4	5	6	7
I_i	{2,3}	{1,2}	{1,3}	{1}	{2}	{3}	$\varnothing$
p_i^*	.525	.135	.105	.045	.09	.07	.03

Les valeurs de c_{ij} pour cet exemple sont présentées au tableau 4. Pour les obtenir, nous avons simplifié les calculs en posant que

$$b_{it} = P(t \in I \mid I^* = I_i),$$

$$i = 1, \ldots, \binom{n}{2} + n + 1, \quad t = 1, \ldots, n, \tag{3.5}$$

et en notant que si $S_j = \{s,t\}$, alors

$$c_{ij} = b_{is} + b_{it}. \tag{3.6}$$

Ainsi, le nombre attendu d'éléments dans $I \cap S_j$, sous réserve que $I^* = I_i$, est simplement la somme des probabilités que chacun des deux éléments de S_j soit dans I, étant donné $I^* = I_i$. Observer en outre que si le problème de transport pour la méthode optimale connaît la valeur exacte de I et connaît donc avec certitude si chaque élément de S_j est également dans I, tel n'est pas le cas en ce qui concerne la méthode à taille réduite puisque dans ce cas, seul l'ensemble associé I_i est connu. À titre d'illustration, examinons le premier rang du tableau 4. Puisque $I_1 = \{2,3\}$, nous savons que $2 \in I$ et $3 \in I$, et donc que $b_{12} = b_{13} = 1$. Toutefois, nous ne savons pas avec certitude si $1 \in I$ puisque I_1 est l'ensemble associé à la fois pour $I = \{1,2,3\}$ et $I = \{2,3\}$. En fait, en se reportant au tableau 1,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Tableau 4

Valeurs de c_{ij} et de x_{ij} qui maximisent le chevauchement pour la méthode à taille réduite

		c_{ij}			x_{ij}		
		j			j		
i	I_i	1	2	3	1	2	3
1	{2,3}	1.6	1.6	2.0	0.000	0.025	0.500
2	{1,2}	2.0	1.0	1.0	0.135	0.000	0.000
3	{1,3}	1.0	2.0	1.0	0.000	0.105	0.000
4	{1}	1.0	1.0	0.0	0.045	0.000	0.000
5	{2}	1.0	0.0	1.0	0.090	0.000	0.000
6	{3}	0.0	1.0	1.0	0.000	0.070	0.000
7	$\varnothing$	0.0	0.0	0.0	0.030	0.000	0.000

Ainsi, $c_{11} = b_{11} + b_{12} = 1.6$, et c_{12}, et c_{13} se calculent de la même façon. Pour les six autres rangs du tableau 4, $I_i = I$ et nous savons donc avec certitude quels sont les nombres entiers de I. En conséquence, les c_{ij} de ces six rangs se calculent facilement.

Finalement, nous maximisons le chevauchement attendu (3.1) sous réserve de (3.2) et (3.3), obtenant ainsi les valeurs x_{ij} du tableau 4. Les probabilités conditionnelles $P(N = S_j \mid I^* = I_i)$ du tableau 5 sont ensuite calculées en divisant chacune des entrées x_{ij} dans le i-ième rang du tableau 4 par p_i^*.

Tableau 5

Probabilités conditionnelles pour la méthode à taille réduite

		j		
i	I_i	1	2	3
1	{2,3}	0	1/21	20/21
2	{1,2}	1	0	0
3	{1,3}	0	1	0
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	1	0
7	∅	1	0	0

Le chevauchement attendu pour la méthode à taille réduite est inférieur de .01 à la valeur optimale, c'est-à-dire 1.725 UPÉ. Cet écart par rapport à l'optimalité est dû uniquement au fait que le chevauchement attendu est de 1.6 pour l'événement conjoint $I^* = \{2,3\}$ et $N = \{1,3\}$. Puisque la probabilité de cet événement conjoint est .025 et que la méthode optimale pour cet exemple donne toujours un chevauchement de 2 lorsqu'il existait au moins 2 UPÉ dans l'échantillon initial, l'écart par rapport à l'optimalité est de $.025(2 - 1.6) = .01$.

La méthode à taille réduite ne permet pas de réaliser l'optimalité parce que la paire {2,3} est assortie d'une probabilité de sélection moindre dans le nouvel échantillon que dans l'échantillon initial. Ainsi, tant pour la méthode optimale que pour la méthode à taille réduite, il convient parfois de sélectionner une autre paire (toujours {1,3} pour les deux méthodes de cet exemple) lorsque l'échantillon initial était {2,3}. Les deux méthodes se distinguent en ce que la méthode optimale choisit uniquement {1,3} lorsque $1 \in I$. Avec la méthode à taille réduite, il n'est pas possible d'utiliser l'information concernant $1 \in I$. Ainsi, lorsque $\{2,3\} \subset I$, $1 \in N$ indépendamment de $1 \in I$. C'est là la cause de l'écart par rapport au chevauchement optimal.

3.1.2 Ordonnancement des paires

Nous examinons ci-après comment on parvient généralement à ordonner les paires. Nous désignons à cette fin par p_{st}, π_{st}, s, $t = 1, \ldots, n$, $s \neq t$, la probabilité conjointe que s, $t \in I$ et s, $t \in N$, respectivement.

Nous procédons à l'ordonnancement des paires pour les motifs suivants. Si la i-ième paire de l'ordonnancement est $\{s,t\}$ le problème de transport pourra alors conserver cette paire dans le nouvel échantillon lorsque $I^* = I_i$ avec la probabilité conditionnelle $\min\{1, \pi_{st}/p_i^*\}$. (La probabilité de conservation conditionnelle ne peut dépasser ce seuil puisqu'une valeur plus élevée donnerait une probabilité de sélection inconditionnelle supérieure à π_{st} pour la paire du nouveau plan). Ainsi, l'objectif général de l'ordonnancement est de rendre ces probabilités conditionnelles aussi grandes que possible en moyenne pour toutes les paires.

Pour montrer comment l'ordonnancement des paires influe sur le chevauchement attendu, nous examinons l'exemple du tableau 3. Notre méthode d'ordonnancement, comme nous le montrerons plus tard, donne les résultats indiqués et un chevauchement attendu de 1.725 UPÉ. Examinons maintenant une autre solution d'ordonnancement pour cet exemple. Soit {1,3} la première paire de l'ordonnancement, {1,2} la seconde et {2,3} la dernière, on obtient $I^* = \{1,3\}$ dans tous les cas où $I = \{1,2,3\}$ ou $I = \{1,3\}$. Ainsi, pour cet ordonnancement, p_1^* est la probabilité que $I^* = \{1,3\}$, et sa valeur est .42. En outre, dans ce cas, $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21$, alors que les 5 autres colonnes du tableau 3 restent inchangées. L'ordonnancement de rechange donne un tableau de probabilités conditionnelles semblable au tableau 5 sauf dans le rang 1 où les colonnes I_i, $j = 2$ et $j = 3$ deviennent maintenant {1,3}, 10/21 et 11/21 respectivement, et dans le rang 3 où les colonnes correspondantes deviennent {2,3}, 0 et 1 respectivement.

En utilisant la méthode de l'ordonnancement original, il est possible de calculer que le chevauchement attendu pour l'ordonnancement de rechange est inférieur de .055 à la valeur optimale, c'est-à-dire qu'il est de 1.68 UPÉ. La raison pour laquelle cet ordonnancement de rechange donne un chevauchement attendu plus faible est donnée ci-après. En général, l'entrée plus tardive d'une paire dans l'ordonnancement entraîne une valeur p_i^* correspondante plus faible et donc une probabilité de rétention conditionnelle plus élevée lorsque $I^* = I_i$. Ainsi, avec {1,3} au premier rang de l'ordonnancement, $\pi_{13}/p_1^* = 10/21$, ce qui correspond à la probabilité conditionnelle de rétention pour cette paire lorsque $I^* = \{1,3\}$. Par contre, lorsque {1,3} est au troisième rang de l'ordonnancement, $\pi_{13}/p_3^* > 1$ et cette paire est retenue avec certitude. Par ailleurs, la probabilité de rétention conditionnelle de la paire {2,3} lorsque $I^* = \{2,3\}$ augmente aussi pour atteindre 1 lorsque {2,3} passe du premier au troisième rang de l'ordonnancement, mais l'augmentation n'est que de 20/21 et l'ordonnancement original du tableau 3 produit donc une augmentation plus grande que prévue du chevauchement attendu que l'ordonnancement de rechange.

Ainsi, comme le montre cet exemple, l'objectif de l'ordonnancement consiste à placer plus tôt dans l'ordonnancement les paires assorties d'une probabilité de rétention conditionnelle relativement élevée même avec un placement précoce. Pour obtenir l'ordonnancement voulu des paires de nombres entiers, on obtiendra d'abord par récursion

un ordonnancement $f(1), \ldots, f(n)$ de $\{1, \ldots, n\}$. Ensuite, pour chaque $k = 1, \ldots, n-1$, on établira par récursion un ordonnancement $g_k(1), \ldots, g_k(n-k)$ de $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\}$. Un ordonnancement linéaire des paires distinctes de $\{1, \ldots, n\}$ pourra alors être déterminé comme suit. Chacune de ces paires peut être représentée uniquement sous la forme d'une paire ordonnée $(f(k), g_k(\ell))$ pour un certain $k \in \{1, \ldots, n-1\}$, $\ell \in \{1, \ldots, n-k\}$. Une seconde paire pouvant être représentée sous la forme $(f(k'), g_{k'}(\ell'))$ précède $(f(k), g_k(\ell))$ si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$. À titre d'explication pour l'exemple que nous venons d'examiner, il sera montré plus tard que $f(1) = 2$, $f(2) = 3, f(3) = 1, g_1(1) = 3, g_1(2) = 1$ et $g_2(1) = 1$, et que l'ordonnancement des paires est donc le suivant: $\{2,3\}, \{2,1\}, \{3,1\}$. L'ordonnancement de f et l'ordonnancement de g_k seront tous les deux effectués pour répondre à l'objectif énoncé au début du présent paragraphe.

Pour obtenir l'ordonnancement $f(1), \ldots, f(n)$, on définit de façon récursive $f(k), k = 1, \ldots, n$, en choisissant un $f(k) \in T_k$ qui satisfait à

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)} : i \in T_k\},$$

où

$$T_1 = \{1, \ldots, n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\},$$
$$k = 2, \ldots, n, \quad p_i^{(k)} = P(i \in I \text{ et } I \subset T_k),$$
$$k = 1, \ldots, n, \quad i \in T_k. \qquad (3.7)$$

Puisque $p_i^{(1)} = p_i$, l'ordonnancement que nous venons de définir équivaut à placer d'abord une UPÉ dont la valeur de π_i/p_i^*, est la plus élevée. Pour tous les k, $p_{f(k)}^{(k)}$ est la probabilité que $f(k)$ faisait partie de I et que aucun des éléments $k-1$ précédant $f(k)$ dans l'ordonnancement f ne faisaient partie de I. Ainsi, $p_{f(k)}^{(k)}$ est la probabilité qu'une tentative soit faite de retenir $A_{f(k)}$ dans le nouvel échantillon soit comme premier membre d'une paire ordonnée d'UPÉ de l'échantillon initial, soit comme UPÉ unique de l'échantillon initial dans S. En règle générale, plus $\pi_{f(k)}/p_{f(k)}^{(k)}$ est grand et plus cette tentative risque de porter fruit. Ainsi, la motivation de l'ordonnancement f des UPÉ individuelles est l'analogue de la motivation de l'ordonnancement des paires d'UPÉ dont nous avons déjà parlé.

Il nous reste à expliquer comment calculer $p_i^{(k)}$ pour $k \geq 2$. À cette fin, désignons par r le nombre de strates initiales qui ont des UPÉ en commun avec S, et par F_α, $\alpha = 1, \ldots, r$, une partition de $\{1, \ldots, n\}$ telle que i et j font partie du même F_α si et seulement si A_i et A_j faisaient partie de la même strate initiale. Soit

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r,$$
$$T \subset \{1, \ldots, n\}, \qquad (3.8)$$

$$p''_{i\alpha}(T) = P(i \in I \text{ et } I \cap F_\alpha \subset T), \quad \alpha = 1, \ldots, r,$$
$$T \subset \{1, \ldots, n\}, \quad i \in F_\alpha \cap T. \qquad (3.9)$$

On observe que

$$p'_\alpha(T) = 1 - \sum_{i \in F_\alpha \sim T} p_i + \sum_{\substack{i,j \in F_\alpha \sim T \\ i<j}} p_{ij}, \qquad (3.10)$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_\alpha \sim T} p_{ij}, \qquad (3.11)$$

et finalement, tel que l'ont établi Ernst et Ikeda (1994),

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^{r} p'_\ell(T_k), \quad k = 1, \ldots, n,$$
$$i \in F_\alpha \cap T_k. \qquad (3.12)$$

Ensuite, pour chaque $k = 1, \ldots, n-1$, l'ordonnancement $g_k(\ell)$, $\ell = 1, \ldots, n-k$, est défini de façon récursive par le choix de $g_k(\ell) \in T_{k\ell}$ qui satisfait à

$$\pi_{f(k),g_k(\ell)}/p_{f(k),g_k(\ell)}^{(\ell)} = \max\{\pi_{f(k),j}/p_{f(k),j}^{(\ell)} : j \in T_{k\ell}\},$$

où

$$T_{k1} = \{1, \ldots, n\} \sim \{f(1), \ldots, f(k)\},$$
$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell = 2, \ldots, n-k,$$
$$T_{k\ell}^* = T_{k\ell} \cup \{f(k)\}, \quad \ell = 1, \ldots, n-k,$$
$$p_{f(k),j}^{(\ell)} = P(f(k), \quad j \in I \text{ et } I \subset T_{k\ell}^*),$$
$$\ell = 1, \ldots, n-k, \quad j \in T_{k\ell}. \qquad (3.13)$$

Noter que $p_{f(k),j}^{(\ell)}$ désigne ainsi la probabilité conjointe que $f(k)$ soit le premier nombre entier de l'ordonnancement f dans I, qu'aucun des premiers nombres entiers $\ell - 1$ dans l'ordonnancement g_k ne soit dans I et que $j \in I$. En conséquence, $p_{f(k),g_k(\ell)}^{(\ell)}$ désigne la probabilité que $I^* = \{f(k), g_k(\ell)\}$. Par ailleurs, si $I_i = \{f(k), g_k(\ell)\}$, alors $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$, et le choix de $g_k(\ell)$ donne donc la valeur de $\pi_{f(k),g_k(\ell)}/p_i^*$ la plus grande parmi les éléments de $T_{k\ell}$, conformément à l'objectif déjà mentionné de l'ordonnancement des paires d'UPÉ.

Pour calculer $p_{f(k),j}^{(\ell)}$, Ernst et Ikeda (1994) établissent que si $f(k) \in F_\alpha, j \in F_\beta$ alors

$$p_{f(k),j}^{(\ell)} = p_{f(k),j} \prod_{\substack{t=1 \\ t \neq \alpha}}^{r} p'_t(T_{k\ell}^*) \text{ si } \alpha = \beta,$$
$$= p''_{f(k),\alpha}(T_{k\ell}^*) p''_{j\beta}(T_{k\ell}^*) \prod_{\substack{t=1 \\ t \neq \alpha,\beta}}^{r} p'_t(T_{k\ell}^*) \text{ si } \alpha \neq \beta. \qquad (3.14)$$

Nous présentons ci-après les calculs utilisés pour obtenir l'ordonnancement de l'exemple sur lequel nous nous

somme penchés. Notons d'abord que $f(1) = 2$ puisque la valeur de π_i/p_i la plus grande s'obtient quand $i = 2$. Trouvons ensuite $g_1(1)$ qui, puisque $f(1) = 2$, est le $j \in \{1,3\}$ qui donne la valeur maximale de $\pi_{2j}/p_{2j}^{(1)}$. Pour déterminer ce j, posons d'abord $F_\alpha = \{\alpha\}$, $\alpha = 1,2,3$, et notons que $T_{11}^* = \{1,2,3\}$. Selon (3.14) avec $\alpha = 2$, $\beta = 1$, il s'ensuit que

$$p_{21}^{(1)} = p''_{22}\{1,2,3\}p''_{11}\{1,2,3\}p'_3\{1,2,3\} = p_2p_1 \cdot 1 = .45,$$

et on peut de la même façon obtenir que $p_{23}^{(1)} = .525$. Ainsi, $g_1(1) = 3$, puisque $.5/.525 > .3/.45$. Par conséquent, la première paire de l'ordonnancement est $\{f(1), g_1(1)\} = \{2,3\}$. Ensuite, $g_1(2) = 1$, puisque 1 est le seul nombre entier qu'il nous reste à utiliser dans l'ordonnancement g_1, et la seconde paire est de ce fait $\{f(1), g_1(2)\} = \{2,1\}$. Il n'est pas vraiment nécessaire de déterminer $f(2)$ puisque $\{1,3\}$ est la seule et donc la dernière paire qui reste, mais à des fins d'illustration, on peut observer que $T_2 = \{1,3\}$, $p_1^{(2)} = p''_{11}\{1,3\}p'_2\{1,3\}$ $p'_3\{1,3\} = p_1(1 - p_2) \cdot 1 = .15$ selon (3.12), et que de la même façon $p_3^{(2)} = p_3(1 - p_2) \cdot 1 = .175$. Donc, $f(2) = 3$ puisque $.7/.175 > .5/.15$. En conséquence, $g_2(1) = 1, f(3) = 1$.

3.1.3 Calcul de p_i^* et de c_{ij}

Nous expliquons ensuite le calcul des valeurs de p_i^*. Si I_i désigne la paire de nombre entiers $I_i = \{f(k), g_k(\ell)\}$ alors, comme noté antérieurement, $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$. En conséquence, p_i^* peut être calculé à l'aide de la formule (3.14) avec $j = g_k(\ell)$.

Si I_i est un ensemble à un seul élément $\{t\}$ pour un certain $t \in F_\alpha$, alors, comme l'ont établi Ernst et Ikeda (1994),

$$p_i^* = p''_{t\alpha}(\{t\}) \prod_{\substack{u=1 \\ u\neq\alpha}}^{r} p'_u(\varnothing). \tag{3.15}$$

Finalement, si $I_i = \varnothing$, alors

$$p_i^* = \prod_{u=1}^{r} p'_u(\varnothing).$$

Il reste simplement à expliquer le calcul des c_{ij} qui, selon (3.5) et (3.6), revient à calculer b_{it}, $i = 1, \ldots, \binom{n}{2} + n + 1$, $t = 1, \ldots, n$.

Pour le calcul de b_{it}, on observe que

$$b_{it} = 0 \quad \text{si} \quad I_i = \varnothing,$$

$$= 1 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t = v,$$

$$= 0 \quad \text{si} \quad I_i = \{v\} \quad \text{et} \quad t \neq v,$$

tandis que si $I_i = \{f(k), g_k(\ell)\}$ et $f(k) \in F_\alpha, g_k(\ell) \in F_\beta$, $t \in F_\gamma$, alors

$$b_{it} = 1 \quad \text{si} \quad t = f(k) \quad \text{ou} \quad t = g_k(\ell), \tag{3.16}$$

$$= 0 \quad \text{si} \quad t \notin T_{k\ell}^*, \tag{3.17}$$

$$= 0 \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \alpha = \beta, \tag{3.18}$$

$$= \frac{p_{f(k),t}}{p''_{f(k),\alpha}(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \alpha \neq \beta, \tag{3.19}$$

$$= \frac{p_{g_k(\ell),t}}{p''_{g_k(\ell),\beta}(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma = \beta \neq \alpha, \tag{3.20}$$

$$= \frac{p''_{t\gamma}(T_{k\ell}^*)}{p'_\gamma(T_{k\ell}^*)} \quad \text{si} \quad t \in T_{k\ell} \sim \{g_k(\ell)\} \quad \text{et} \quad \gamma \neq \alpha, \gamma \neq \beta. \tag{3.21}$$

Ernst et Ikeda (1994) démontrent comment les formules (3.16) à (3.21) ont été obtenues.

Lors de la mise en oeuvre du SIPP en conditions réelles, il a fallu modifier la méthode à taille réduite pour réaliser le chevauchement du plan d'échantillonnage du SIPP des années 1990 et de celui des années 1980. Ces modifications se sont avérées nécessaires par suite de changements intervenus dans les définitions des UPÉ d'une décennie à l'autre. En effet, certaines UPÉ du plan des années 1990 pouvaient recouper plus d'une UPÉ du plan des années 1980. Ces modifications sont décrites en détails dans l'article de Ernst et Ikeda (1994).

3.2 Modifications de la méthode à taille réduite pour d'autres plans

En règle générale, on considère n'importe quel plan d'échantillonnage initial sans remise de la forme m'-UPÉ par strate et n'importe quel plan d'échantillonnage final sans remise de la forme m-UPÉ par strate, où m', m sont des nombres entiers positifs quelconques. Même si la méthode à taille réduite de la section 3.1 n'a été présentée que pour le cas où $m = m' = 2$, elle peut en fait s'appliquer pour n'importe quelles valeurs de m, m'. Nous présentons brièvement ci-après les modifications nécessaires lorsque $m \neq 2$ ou $m' \neq 2$.

L'utilisation d'une valeur différente de m' ne demande le changement que d'une partie seulement des calculs. Par exemple, si $m = 2$ mais $m' \neq 2$, les calculs de $p_i^{(k)}$, $p_{f(k),j}^{(\ell)}$ et c_{ij} seront différents, mais leurs définitions resteront les mêmes.

Si $m = 3$, peu importe la valeur de m', on procède à l'ordonnancement de triplets de nombres entiers au lieu de paires dans $\{1, \ldots, n\}$. Si I consiste en au moins trois nombres entiers, les nouvelles probabilités de sélection sont conditionnées seulement sur le premier triplet de la

liste ordonnée de I. Autrement, les nouvelles probabilités de sélection sont conditionnées sur I lui-même. Ainsi, les nouvelles probabilités de sélection sont conditionnées sur $\binom{n}{3} + \binom{n}{2} + n + 1$ événements.

Pour obtenir l'ordonnancement voulu des triplets de nombres entiers, on établit d'abord l'ordonnancement des $f(1), \ldots, f(n)$ et des $g_k(1), \ldots, g_k(n-k)$ comme dans le cas où $m = 2$. Ensuite, pour chaque $k = 1, \ldots, n-2$, $\ell = 1, \ldots, n-k-1$, un ordonnancement $h_{k\ell}(1), \ldots, h_{k\ell}(n-k-\ell)$ de $\{1, \ldots, n\} \sim \{f(1), \ldots, f(k), g_k(1), \ldots, g_k(\ell)\}$ est établi d'une manière semblable à celle utilisée pour $g_k(1), \ldots, g_k(n-k)$. Par exemple, en définissant $h_{k\ell}(v)$ pour $v \geq 2$, $p^{\{\ell\}}_{f(k),j}$ de la définition de $g_k(\ell)$ est remplacé par

$$P\big(f(k), g_k(\ell), j \in I \quad \text{et} \quad I \subset (T^*_{k\ell} \cup g_k(\ell)) \sim \{h_{k\ell}(1), \ldots, h_{k\ell}(v-1)\}\big).$$

Un ordonnancement linéaire des triplets distincts de $\{1, \ldots, n\}$ est alors déterminé en représentant chaque triplet par un triplet ordonné unique ayant la forme $(f(k), g_k(\ell), h_{k\ell}(v))$. Un deuxième triplet $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ précède le premier si et seulement si $k' < k$, ou $k' = k$ et $\ell' < \ell$, ou $k' = k$ et $\ell' = \ell$ et $v' < v$.

Pour les valeurs de $m \geq 4$, des ordonnancements de m-uplets seraient définis selon une procédure semblable, et les nouvelles probabilités de sélections seraient conditionnées sur $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ événements.

Pour $m = 1$, les nouvelles probabilités de sélection sont conditionnées sur le premier membre de l'ordonnancement $f(1), \ldots, f(n)$ dans I si $I \neq \varnothing$, ou sur $\varnothing$ si $I = \varnothing$.

À noter que si $m > m'$, il est possible qu'au moins une partie des m-uplets ordonnancés ne soient pas des sous-ensembles de I, auquel cas ces sous-ensembles seraient exclus de l'ordonnancement et de l'ensemble d'événements sur lesquels les nouvelles probabilités de sélection sont conditionnées. Si aucun m-uplet ne peut être un sous-ensemble de I, alors les nouvelles probabilités de sélection sont conditionnées sur I lui-même.

Il n'est pas nécessaire de limiter les événements initiaux utilisés dans le problème de transport uniquement aux sous-ensembles de I de taille m ou moindre. Par exemple, si $m = 2$ et si $\binom{n}{3} + \binom{n}{2} + n + 1$ est suffisamment petit, on peut alors utiliser une méthode conditionnée sur des sous-ensembles de trois ou moins, ce qui donne un chevauchement attendu généralement plus grand. À l'inverse, si $\binom{n}{m} + \binom{n}{m-1} \ldots + n + 1$ est trop grand, les nouvelles probabilités de sélection peuvent être conditionnées sur des sous-ensembles de I de taille m'' ou moins, où $m'' < m$, en donnant par contre un chevauchement attendu généralement plus petit.

3.3 Rapports entre les chevauchements attendus de la méthode à taille réduite, de la méthode optimale et de la sélection indépendante

Désignons par Ω_I, Ω_R, Ω_O le chevauchement attendu avec la sélection indépendante, la méthode à taille réduite et la méthode optimale respectivement. Ernst et Ikeda (1994) examinent les rapports qui existent entre ces paramètres. Nous résumons brièvement ci-après les résultats de leur étude.

Il est admis au départ que $\Omega_I \leq \Omega_R \leq \Omega_O$ pour tout m, m', où m, m' sont tels que décrit à la section 3.2. En outre, pour le cas qui nous intéresse, $m = m' = 2$, les limites inférieures sont établies sur Ω_R, et les limites supérieures sont établies sur Ω_O et $\Omega_O - \Omega_R$.

Par exemple, désignons par μ_2 la probabilité qu'il existe au moins deux éléments dans I, et désignons par μ_1 la probabilité que I soit un ensemble à un seul élément. Soit

$$\lambda = \min\big\{\min\{\pi_i/p_i : i = 1, \ldots, n\}, \min\{\pi_{ij}/p_{ij} : i, j = 1, \ldots, n, i \neq j\}, 1\big\}.$$

Alors $\Omega_O \leq 2\mu_2 + \mu_1$, $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$, et $\Omega_O - \Omega_R \leq 2(1-\lambda)\mu_2 + (1 - \lambda/2)\mu_1$.

Malheureusement, ces limites ne sont pas toujours très strictes. Toutefois, dans certaines circonstances, elles peuvent être utiles. Par exemple, si $\pi_{ij} \geq p_{ij}$ pour tous les i,j et s'il existe une probabilité 1 que I contienne au moins deux éléments, il s'ensuit, compte tenu de ces limites, que $\Omega_R = \Omega_O = 2$.

Finalement nous présentons un exemple du pire cas envisageable pour Ω_R en rapport avec Ω_O lorsque m, $m' = 2$. On observe que Ω_O peut être égal à 2, tandis que Ω_R est arbitrairement près de 0. Ainsi, du moins en théorie, la méthode à taille réduite peut s'avérer inefficace. En pratique toutefois, comme nous le démontrerons dans la prochaine section, Ω_R est beaucoup plus près de Ω_O que de Ω_I, au moins en ce qui concerne le SIPP.

4. APPLICATION DE LA MÉTHODE À TAILLE RÉDUITE AU SIPP

Nous présentons ci-après les résultats des simulations du chevauchement du SIPP réalisées avant l'enquête, aux fins de la recherche et des essais, ainsi que les résultats du chevauchement du SIPP obtenus en conditions réelles. Pour en savoir plus sur cette question, consulter Ernst et Ikeda (1992b, 1994).

Pour la mise en oeuvre de la méthode de chevauchement à taille réduite, nous avons utilisé un logiciel d'optimisation du flux de coût minimal (minimum cost flow ou MCF) créé par Darwin Kingman et John Mote de la University of Texas, à Austin, qui nous a permis de résoudre le problème de transport requis. Un programme en FORTRAN a été créé pour préparer les données d'entrée et pour traiter les données de sortie du logiciel MCF.

Pour tester le logiciel avant la tenue de l'enquête, nous avons utilisé le programme pour effectuer le chevauchement de deux stratifications du SIPP de la région du Midwest fondées sur les données du recensement de 1970 avec la stratification du plan véritable utilisé dans la même région pour le SIPP au cours des années 1980. (À l'époque de la réalisation de ce test, les données du recensement de 1990 n'étaient pas encore disponibles.) Les stratifications de 1970 ont été produites en stratifiant les UPÉ sélectionnées

avec probabilité des SIPP réalisés dans la région du Midwest au cours des années 1980, en utilisant les données de 1970. Les deux stratification de 1970 ont donné une répartition des UPÉ sélectionnées avec probabilité en 31 strates, en utilisant différents groupes de variables de stratification. Les stratifications fondées sur les données de 1980 et de 1970 ont été assimilées à des stratifications "initiales" et "finales" aux fins de l'algorithme de chevauchement.

Au cours de l'enquête véritable, tel qu'il est mentionné dans la section 3.1 et comme l'expliquent en détails Ernst et Ikeda (1994), on a utilisé une modification de la méthode à taille réduite pour le chevauchement du plan du SIPP des années 1990 et de celui des années 1980 à cause d'une différence dans la définition des UPÉ entre les deux décennies. La méthode à taille réduite modifiée a été utilisée pour le chevauchement de 103 strates non autoreprésentatives finales (plan des années 1990) du SIPP.

Le chevauchement attendu a été calculé pour l'algorithme de chevauchement maximum à taille réduite, pour la sélection indépendante des UPÉ finales et pour une limite supérieure du chevauchement attendu avec la méthode optimale. On a calculé une limite supérieure au lieu du chevauchement optimal réel parce qu'il est impossible de calculer un chevauchement optimal pour la strate la plus grande. Aux fins de la simulation, la limite supérieure utilisée à été celle mentionnée dans la section 3.3, $\mu_2 + 2\mu_1$, tandis qu'aux fins de l'enquête SIPP, il a fallu utiliser une limite supérieure différente, décrite par Ernst et Ikeda (1994), à cause de la différence des définitions des UPÉ entre les années 1980 et 1990.

Les résultats des deux stratifications finales de la simulation étaient généralement semblables l'un à l'autre. La combinaison des résultats des deux a donné un chevauchement attendu moyen pour cet ensemble de 62 strates de 1.552, 1.569 et .480 UPÉ/strate pour la méthode à taille réduite, la limite supérieure du chevauchement optimal et la sélection indépendante respectivement. Pour la mise en oeuvre du SIPP en conditions réelles, les données correspondantes ont été de 1.523, 1.647 et .582 respectivement, alors que les chevauchements attendus correspondant d'UPÉ pour les 103 strates étaient 156.9, 169.6 et 59.9 respectivement. Ainsi, tant dans les simulations que dans l'enquête SIPP véritable, la méthode à taille réduite a donné des résultats raisonnablement proches de la limite supérieure de la méthode optimale.

Tableau 6

Temps machine pour la méthode à taille réduite

Nombre d'UPÉ	Temps machine (hh:mm:ss)
18	0:36
37	5:44
49	24:05
68	2:23:43

Le temps d'exécution de l'algorithme à taille réduite a été relativement court pour la plupart des strates. Nous présentons ci-après les temps machine de la simulation pour la strate finale avec différents nombres d'UPÉ. Nous avons utilisé un ordinateur Solbourne 5/605. Le nombre médian d'UPÉ dans une strate, pour le groupe entier de 62 strates, était de 17. La strate la plus grande contenait 68 UPÉ.

Nous avons également déterminé, pour l'enquête SIPP en conditions réelles, que 41 des 103 strates finales du chevauchement de la méthode modifiée à taille réduite auraient été incompatibles avec la méthode optimale. En effet, selon nos estimations, la taille maximale du problème de transport, en termes de nombre de variables, pour l'exécution du programme, était fixée à 4×10^6. Le nombre de variables pour la méthode optimale était inférieur à 4×10^6 pour toutes les strates (56) à $n \leq 14$, mais dépassait la limite pour 41 des 47 strates à $n \geq 15$, y compris deux à $n = 15$. La taille maximale du problème de transport avec la méthode optimale pour les 103 strates a été atteinte avec une strate à $n = 46$, pour laquelle il y avait 3.61×10^{12} variables. Par contre, il y avait 1.03×10^6 variables pour la même strate avec la méthode à taille réduite modifiée.

L'efficacité relative du chevauchement de la méthode à taille réduite comparativement à la méthode de Ernst (1986) présente également un intérêt. On pense en général que la méthode à taille réduite devrait produire un chevauchement plus grand dans les cas où les deux méthodes sont utilisables, puisque la méthode à taille réduite tire profit de l'indépendance de strate à strate du plan initial. Toutefois, même si la méthode de Ernst (1986) s'applique aux plans à deux UPÉ par strate, aucun logiciel n'a jamais été créé au Census Bureau (ni ailleurs à notre connaissance) pour en permettre l'utilisation avec ce type de plan puisqu'il n'y a pas encore eu d'application en conditions réelles pour un tel programme. En conséquence, il n'est pas possible de comparer directement ces deux méthodes avec les mêmes données. On peut toutefois faire une comparaison sommaire à partir des résultats de la méthode de chevauchement à taille réduite pour l'enquête SIPP et des résultats du chevauchement obtenus avec la méthode de Ernst (1986) pour le chevauchement des plans du CPS et du NCVS pour les années 1990 avec leurs contreparties respectives des années 1980. (Les plans des enquêtes CPS et NCVS des années 1980 et 1990 comptent tous une UPÉ par strate.)

Dans le cas du CPS, la méthode de chevauchement a donné une augmentation moyenne du chevauchement attendu de .26 UPÉ par strate comparativement au résultat de la sélection indépendante; dans le cas du NCVS, elle a donné une augmentation moyenne du chevauchement attendu de .30 UPÉ par strate. À titre comparatif, on a obtenu une augmentation de .94 UPÉ par strate avec la méthode à taille réduite par rapport à la sélection indépendante pour le SIPP. Si les deux méthodes de chevauchement sont également efficaces, on pourrait alors s'attendre à une augmentation du chevauchement par strate à peu près deux fois plus importante pour le SIPP que pour le

CPS ou le NCVS puisque le SIPP utilise un plan à deux UPÉ par strate. Vue sous cet angle, la méthode à taille réduite fonctionne mieux que celle proposée par Ernst (1986). Toutefois, comme les stratifications ont été passablement différentes dans ces trois enquêtes, il est permis de douter de la validité d'une telle comparaison.

Pour l'exemple considéré dans les sections 2 et 3, il est possible de procéder à une comparaison valide des différentes méthodes de chevauchement puisque la valeur du chevauchement attendu dans Ernst (1986), soit 1.625, a été facilement calculée à la main. Les valeurs correspondantes du chevauchement pour la méthode à taille réduite et pour la méthode optimale sont 1.725 et 1.735 respectivement.

CONCLUSIONS

La méthode de chevauchement à taille réduite présentée dans le présent article répond en pratique à ses deux objectifs principaux. Elle réduit suffisamment la taille du problème de transport, comme le démontrent d'une part la taille du problème de transport dans les formules (3.1) à (3.3), et d'autre part le fait qu'elle a effectivement été utilisée dans le remaniement d'une enquête importante. En outre, cette méthode réalise la réduction de taille tout en donnant un chevauchement presque optimal, à tout le moins dans le cas du SIPP. Elle ne peut être utilisée que lorsque les UPÉ du plan initial sont sélectionnées indépendamment d'une strate à l'autre, mais lorsque cette condition est respectée, nous croyons qu'il s'agit de la meilleure méthode de chevauchement pour les grandes strates.

REMERCIEMENTS

Nous remercions M. Todd Williams pour son aide précieuse à la programmation. Merci également aux critiques et aux rédacteurs pour leurs commentaires constructifs. Les opinions exprimées dans le présent article sont celles des auteurs et ne sont pas nécessairement partagées par le Bureau of Labor Statistics ni par le Census Bureau.

BIBLIOGRAPHIE

ARAGON, J., et PATHAK, P.K. (1990). An algorithm for optimal integration of two surveys. *Sankhyā: The Indian Journal of Statistics*, 52, 198-203.

ARTHANARI, T.S., et DODGE, Y. (1981). *Mathematical Programming in Statistics*. New York: John Wiley and Sons.

CAUSEY, B.D., COX, L.H., et ERNST, L.R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80, 903-909.

ERNST, L.R. (1986). Maximizing the overlap between surveys when information is incomplete. *European Journal of Operational Research*, 27, 192-200.

ERNST, L.R. (1989). Further Applications of Linear Programming to Sampling Problems. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

ERNST, L.R., et IKEDA, M. (1992a). Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

ERNST, L.R., et IKEDA, M. (1992b). Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys. Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

ERNST, L.R., et IKEDA, M. (1994). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-93/02.

GLOVER, F., KARNEY, D., KLINGMAN, D., et NAPIER, A. (1974). A computation study on start procedures, basic change criteria and solution algorithms for transportation problems. *Management Sciences*, 20, 793-813.

KEYFITZ, N. (1951). Sampling with probabilities proportional to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.

PATHAK, P.K., et FAHIMI, M. (1992). Optimal integration of surveys. In *Essays in Honor of D. Basu*. Éds. M. Ghosh, et P.K. Pathak. Hayward, California: Institute of Mathematical Statistics, 208-224.

PERKINS, W.M. (1970). 1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Stata. Note de service à Joseph Waksberg, Bureau of the Census.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw Hill.