Techniques d'enquête, décembre 2006
Vol. 32, N° 2, pp. 185-196
Statistique Canada, N° 12-001-XPB au catalogue

Sondage indirect : Les fondements de la méthode généralisée du partage des poids

Jean-Claude Deville et Pierre Lavallée [1]

Résumé

Lorsqu'on veut sélectionner un échantillon, il arrive qu'au lieu de disposer d'une base de sondage contenant les unités de collecte souhaitées, on ait accès à une base de sondage contenant des unités liées d'une certaine façon à la liste d'unités de collecte. On peut alors envisager de sélectionner un échantillon dans la base de sondage disponible afin de produire une estimation pour la population cible souhaitée en s'appuyant sur les liens qui existent entre les deux. On donne à cette approche le nom de *sondage indirect*.

L'estimation des caractéristiques de la population cible étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités étudiées de la population cible. La méthode généralisée du partage des poids (MGPP) a été mise au point par Lavallée (1995) et Lavallée (2002) afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité enquêtée de la population cible.

Le présent article débute par une description du sondage indirect, qui constitue le fondement de la MGPP. En deuxième lieu, nous donnons un aperçu de la MGPP dans lequel nous la formulons dans un cadre théorique en utilisant la notation matricielle. En troisième lieu, nous présentons certaines propriétés de la MGPP, comme l'absence de biais et la transitivité. En quatrième lieu, nous considérons le cas particulier où les liens entre les deux populations sont exprimés par des variables indicatrices. En cinquième lieu, nous étudions certains liens typiques spéciaux afin d'évaluer leur effet sur la MGPP. Enfin, nous examinons le problème de l'optimalité. Nous obtenons des poids optimaux dans un sens faible (pour des valeurs particulières de la variable d'intérêt), ainsi que les conditions dans lesquelles ces poids sont également optimaux au sens fort et indépendants de la variable d'intérêt.

Mots clés : Sondage indirect; méthode généralisée du partage des poids; absence de biais; poids optimaux.

1. Introduction

En vue de sélectionner les échantillons nécessaires pour les enquêtes sociales ou économiques, il est utile de disposer de bases de sondage, c'est-à-dire de listes d'unités, offrant un moyen d'atteindre les populations cibles souhaitées. Malheureusement, il arrive qu'au lieu de posséder une liste contenant les unités de collecte souhaitées, on dispose d'une liste d'unités reliée d'une certaine façon à celle des unités de collecte. On peut par conséquent parler de deux populations U^A et U^B liées l'une à l'autre, où l'on souhaite produire une estimation pour U^B. Malheureusement, on ne dispose d'une base de sondage que pour U^A. On peut alors envisager de sélectionner un échantillon s^A dans U^A afin de produire une estimation pour U^B en s'appuyant sur la correspondance qui existe entre les deux populations. On parle alors de *sondage indirect*.

L'estimation des caractéristiques d'une population cible U^B étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités de la population cible visées par le sondage. La méthode généralisée du partage des poids (MGPP) a été mise au point par Lavallée (1995) et Lavallée (2002), et également présentée dans Lavallée et Caron (2001), afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité étudiée de la population cible U^B. Fondamentalement, ce poids d'estimation correspond à une moyenne pondérée des poids de sondage des unités de l'échantillon s^A. La MGPP est une extension de la méthode de partage des poids décrite par Ernst (1989) dans le contexte des enquêtes longitudinales auprès des ménages.

Le but du présent article est de décrire le sondage indirect, c'est-à-dire les fondements de la MGPP, et d'obtenir, par la MGPP, des poids optimaux produisant des estimations sans biais dont la variance est minimale. Nous commencerons par décrire le sondage indirect, ainsi que la MGPP dans un cadre théorique qui fera appel, notamment, à la notation matricielle. L'utilisation de cette notation pour la MGPP a été présentée antérieurement par Deville (1998). Puis, nous utiliserons ce cadre théorique en vue d'énoncer certaines propriétés générales associées à la MGPP, dont l'absence de biais et la transitivité. Cette dernière consiste à passer de la population U^A à une population cible U^C par l'intermédiaire d'une population U^B. En troisième lieu, nous montrerons la correspondance entre la formulation

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête (ENSAI/CREST), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, FRANCE. Courriel : deville@ensai.fr; Pierre Lavallée, Statistique Canada, Ottawa (Ontario), K1A 0T6, CANADA. Courriel : pierre.lavallee@statcan.ca.

matricielle et celle qui a été décrite dans Lavallée (1995), Lavallée (2002), ainsi que Lavallée et Caron (2001). En quatrième lieu, nous étudierons l'effet de diverses matrices de liens établissant la liaison entre U^A et U^B sur la précision des estimations obtenues par la MGPP. Enfin, nous examinerons le problème de l'optimalité. Nous obtiendrons des poids optimaux dans un sens faible (pour des valeurs particulières de la variable d'intérêt), ainsi que les conditions sous lesquelles ces poids sont également optimaux dans un sens fort et indépendants de la variable d'intérêt.

2. Sondage indirect

Comme nous l'avons mentionné dans l'introduction, le sondage indirect consiste à sélectionner un échantillon s^A dans une population U^A afin de produire une estimation pour une population cible U^B, en s'appuyant pour cela sur la correspondance qui existe entre les deux populations. Par exemple, supposons que nous voulions produire des estimations pour une population d'enfants (unités de collecte), mais que nous ne disposions d'une base de sondage que pour les parents. La population cible U^B est celle des enfants, mais nous devons sélectionner un échantillon de parents avant de pouvoir interviewer les enfants. Cette situation est illustrée à la figure 1.

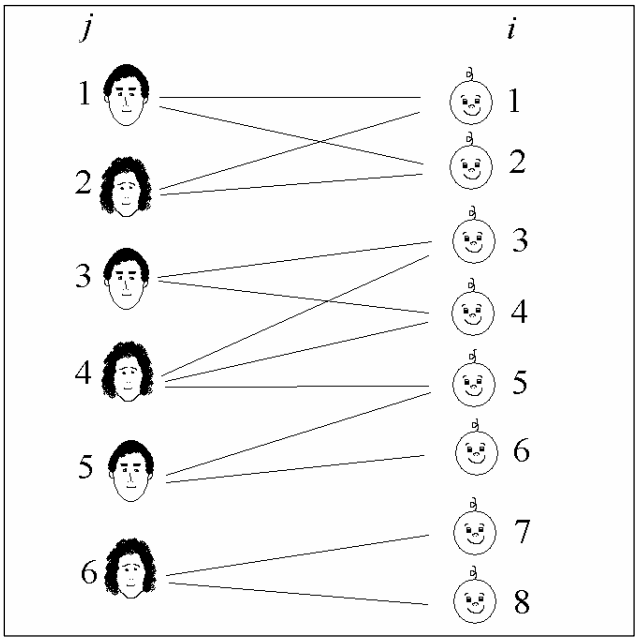

Figure 1. Population U^A de parents et population U^B d'enfants avec les liens entre les deux.

Soit U^A une population de N^A unités, où chaque unité est notée j. De même, soit U^B la population cible de N^B unités, où chaque unité est notée i. La correspondance entre les deux populations U^A et U^B peut être représentée par une *matrice de liens* $\boldsymbol{\Theta}_{AB} = [\theta_{ji}^{AB}]$ de taille $N^A \times N^B$, où chaque élément est $\theta_{ji}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ji}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités. Dans le cas de l'exemple susmentionné, la matrice de liens est donnée par

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \end{bmatrix}.$$

En sondage indirect, l'obtention de la *matrice de liens* $\boldsymbol{\Theta}_{AB} = [\theta_{ji}^{AB}]$ est une question cruciale. Dans le cas où deux unités $j \in U^A$ et $i \in U^B$ ne sont pas liées, nous fixons simplement que $\theta_{ji}^{AB} = 0$. Lorsqu'il existe un lien entre deux unités j et i, le choix de la valeur de $\theta_{ji}^{AB} > 0$ est important. Comme nous le verrons, il influe sur la précision des estimations émanant du sondage indirect. Dans plusieurs applications, les valeurs de θ_{ji}^{AB} pour les unités liées sont simplement fixées à 1. Naturellement, elles pourraient être choisies différentes de 1. Lavallée et Caron (2001) discutent de l'utilisation des poids de couplage obtenus à partir d'un processus de couplage d'enregistrements entre U^A et U^B pour attribuer des valeurs aux éléments θ_{ji}^{AB}. Les poids de couplage sont proportionnels à la probabilité que deux unités $j \in U^A$ et $i \in U^B$ soient liées. Puisque le choix de $\theta_{ji}^{AB} > 0$ pour les deux unités liées j et i peut influencer la précision des estimations, il est naturel de rechercher les valeurs de θ_{ji}^{AB} qui minimiseront la variance des estimations. Ce problème d'optimisation est examiné à la section 6 de l'article.

Dans le sondage indirect, nous sélectionnons l'échantillon s^A de n^A unités à partir de U^A selon un certain plan d'échantillonnage. Soit π_j^A la probabilité de sélection de l'unité j. Nous supposons que $\pi_j^A > 0$ pour tout $j \in U^A$. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{ji}^{AB} > 0$. Soit Ω^B l'ensemble des n^B unités de U^B identifié par les unités $j \in s^A$, c'est-à-dire $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}$. Pour chaque unité i de l'ensemble Ω^B, nous mesurons une variable d'intérêt y_i à partir de la population cible U^B. Soit $\mathbf{Y} = \{y_1, \ldots, y_{N^B}\}'$ le vecteur colonne de cette variable d'intérêt. D'un point de vue pratique, il est important de mentionner que, bien que la taille d'échantillon n^A soit habituellement déterminée d'avance, le nombre d'unités n^B est difficile à contrôler, car il dépend de l'échantillon sélectionné s^A et de la matrice de liens $\boldsymbol{\Theta}_{AB}$. Par conséquent, il s'avère difficile en général d'établir un budget

pour mesurer la variable d'intérêt y_i. Heureusement, dans la plupart des applications (par exemple, le cas parents-enfants susmentionné), le nombre de liens qui ont pour origine une unité donnée j de s^A est plus ou moins prévisible (par exemple, un parent a en général 1, 2 ou 3 enfants), ce qui facilite la détermination du nombre d'unités i de U^B qui, en dernière analyse, seront mesurées.

Nous supposons que pour toute unité j de s^A, il est possible d'obtenir les correspondances pour $i = 1, ..., N^B$. Autrement dit, nous pouvons identifier tous les liens entre les deux populations par interview directe ou grâce à une source administrative pour toute unité j échantillonnée. En outre, pour toute unité i identifiée de U^B, nous supposons qu'il est possible d'obtenir les liens pour $j = 1, ..., N^A$ (comme l'a mentionné Lavallée (2002), il existe des cas où cette dernière contrainte est difficile à satisfaire en pratique. Si nous revenons à l'exemple des parents et des enfants, il pourrait être difficile pour un très jeune enfant, sélectionné par l'entremise de sa mère, de mentionner son père, si les parents sont divorcés. Afin de simplifier la discussion, nous supposerons que ce genre de problème d'identification de liens est négligeable). Par conséquent, il n'est pas nécessaire de connaître les valeurs des liens entre les populations complètes U^A et U^B. En fait, nous ne devons connaître les liens (et, par conséquent, les valeurs de θ_{ji}^{AB}) que pour les lignes j de $\mathbf{\Theta}_{AB}$, où $j \in s^A$, ainsi que pour les colonnes i de $\mathbf{\Theta}_{AB}$, où $i \in \Omega^B$.

Supposons que nous voulions estimer le total Y^B de la population cible U^B, où $Y^B = \sum_{i=1}^{N^B} y_i$. Nous pouvons aussi écrire $Y^B = \mathbf{1}'_B \mathbf{Y}$, où $\mathbf{1}_B$ est le vecteur colonne de 1 de taille N^B (notons que, pour simplifier, nous utilisons la notation $\mathbf{1}_B$ au lieu de $\mathbf{1}_{N^B}$). Maintenant, posons que $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ et que $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$. Nous avons $\mathbf{1}'_A \mathbf{\Theta}_{AB} = \{\theta_{+1}^{AB}, ..., \theta_{+N^B}^{AB}\}$. Nous définissons alors la *matrice de liens normalisée* $\tilde{\mathbf{\Theta}}_{AB} = \mathbf{\Theta}_{AB} [\text{diag}(\mathbf{1}'_A \mathbf{\Theta}_{AB})]^{-1}$, où $\text{diag}(\mathbf{v})$ est la matrice carrée obtenue en plaçant les éléments du vecteur ligne (ou du vecteur colonne) $\mathbf{v}$ sur la diagonale et 0 ailleurs. Notons que, pour que la matrice $\tilde{\mathbf{\Theta}}_{AB}$ soit bien définie, il faut que $[\text{diag}(\mathbf{1}'_A \mathbf{\Theta}_{AB})]^{-1}$ existe, ce qui n'est le cas que si et autrement si $\theta_{+i}^{AB} > 0$ pour tout $i = 1, ..., N^B$. Dans l'exemple des parents et des enfants, cela signifie que chaque enfant doit être lié à au moins un parent.

Résultat 1 :

La matrice de liens $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée si et seulement si

$$\tilde{\mathbf{\Theta}}'_{AB} \mathbf{1}_A = \mathbf{1}_B. \tag{2.1}$$

La preuve du résultat 1 découle directement de la définition d'une matrice de liens normalisée. Partant du résultat 1, nous obtenons directement le résultat 2 que l'on trouve aussi dans Deville (1998) :

Résultat 2 :

$$\begin{aligned} Y^B &= \mathbf{1}'_B \mathbf{Y} \\ &= \mathbf{1}'_A \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \frac{\theta_{ji}^{AB}}{\theta_{+i}^{AB}} y_i. \end{aligned} \tag{2.2}$$

Soit le vecteur colonne $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ de taille N^A. En considérant chaque ligne de $\mathbf{Z}$, nous définissons la variable $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$ pour chaque unité j de la population U^A et nous la mesurons pour chaque unité $j \in s^A$.

Pour estimer Y^B, nous voulons utiliser les valeurs de y_i mesurées à partir de l'ensemble Ω^B. Pour cela, nous utiliserons un estimateur de la forme :

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i \tag{2.3}$$

où w_i est le poids d'estimation de l'unité i de Ω^B, avec $w_i = 0$ pour $i \notin \Omega^B$. Soit $\mathbf{W}' = \{w_1, ..., w_{N^B}\}$. L'estimateur (2.3) peut être réécrit sous la forme

$$\hat{Y}^B = \mathbf{W}'\mathbf{Y}. \tag{2.4}$$

Habituellement, pour obtenir une estimation sans biais de Y^B, il suffit d'utiliser comme poids l'inverse de la probabilité de sélection π_i^B de l'unité i. Comme le mentionne Lavallée (1995) et Lavallée (2002), dans le cas du sondage indirect, il peut être difficile, voire impossible, de calculer cette probabilité. Il propose alors de recourir à la MGPP, qui est définie comme il suit.

Soit $\boldsymbol{\pi}^A = \{\pi_1^A, ..., \pi_{N^A}^A\}'$ et soit $\mathbf{\Pi}_A = \text{diag}(\boldsymbol{\pi}^A)$ la matrice diagonale de taille $N^A \times N^A$ contenant les probabilités de sélection utilisées pour le tirage de l'échantillon s^A. Similairement, soit $\mathbf{t}^A = \{t_1^A, ..., t_{N^A}^A\}'$ où $t_j^A = 1$ si $j \in s^A$, et 0 autrement. Soit $\mathbf{T}_A = \text{diag}(\mathbf{t}^A)$ la matrice diagonale de taille $N^A \times N^A$ contenant les variables indicatrices t_j^A. En partant de $Y^B = \mathbf{1}'_A \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y} = \mathbf{1}'_A \mathbf{Z}$, nous pouvons former directement l'estimateur d'Horvitz-Thompson en fonction du vecteur $\mathbf{Z}$:

$$\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{Z}. \tag{2.5}$$

Puisque $\mathbf{Z} = \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$, nous avons $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \tilde{\mathbf{\Theta}}_{AB} \mathbf{Y}$ et nous pouvons donc définir le vecteur colonne $\mathbf{W}$ de poids :

$$\mathbf{W} = \tilde{\mathbf{\Theta}}'_{AB} \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{1}_A. \tag{2.6}$$

Le vecteur $\mathbf{W}$ est de taille N^B et, pour chaque $i = 1, ..., N^B$, nous avons $w_i = \sum_{j=1}^{N^A} t_j^A \tilde{\theta}_{ji}^{AB} / \pi_j^A$. Les poids w_i de ce vecteur sont obtenus par la MGPP, comme le décrit Lavallée (2002).

3. Propriétés de la MGPP

3.1 Absence de biais

Comme l'a mentionné Ernst (1989), pour obtenir un estimateur sans biais, il suffit que $E(\mathbf{W}) = \mathbf{1}_B$. Par construction, puisque l'estimateur (2.5) est un estimateur d'Horvitz-Thompson, cette condition est directement satisfaite et, par conséquent, la MGPP produit des estimations sans biais.

Partant de cette discussion, nous pouvons aussi obtenir le résultat suivant :

Résultat 3 :

Le vecteur de poids $\mathbf{W}$ donné par (2.6) fournit des estimations sans biais si et seulement si la matrice $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée.

Démonstration :

Partant de (2.6), nous avons

$$E(\mathbf{W}) = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A. \qquad (3.1)$$

En utilisant le résultat 1, nous obtenons directement $E(\mathbf{W}) = \mathbf{1}_B$ et les estimations sont donc sans biais. Maintenant, supposons que $E(\mathbf{W}) = \mathbf{1}_B$. D'après (3.1), nous devons avoir $\tilde{\mathbf{\Theta}}'_{AB}\mathbf{1}_A = \mathbf{1}_B$ et, par conséquent, $\tilde{\mathbf{\Theta}}_{AB}$ est une matrice de liens normalisée.

3.2 Variance

Comme l'estimateur (2.5), est un estimateur d'Horvitz-Thompson, nous obtenons directement le résultat suivant :

Résultat 4 :

La variance de $\hat{Y}^B$ est donnée par

$$\begin{aligned}\mathrm{Var}(\hat{Y}^B) &= \mathbf{Z}'\mathbf{\Delta}_A\mathbf{Z} \\ &= \mathbf{Y}'\mathbf{\Delta}_B\mathbf{Y} \qquad (3.2)\end{aligned}$$

où $\mathbf{\Delta}_A = [(\pi^A_{jj'} - \pi^A_j\pi^A_{j'})/\pi^A_j\pi^A_{j'}]_{N^A\times N^A}$ est une matrice définie non négative de taille $N^A \times N^A$ et où $\pi^A_{jj'}$ est la probabilité de sélection conjointe des unités j et j' dans U^A, et où $\Delta_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}$.

Pour une preuve de la variance de l'estimateur d'Horvitz-Thompson, voir Särndal, Swensson et Wretman (1992).

3.3 Transitivité

Supposons que nous voulions produire des estimations pour une population cible U^C que l'on ne peut atteindre que par l'entremise de la population U^B. Nous supposons que la population cible U^C contient N^C unités, chacune notée k. La correspondance entre les deux populations U^B et U^C peut être représentée par la matrice de liens $\mathbf{\Theta}_{BC} = [\theta^{BC}_{ik}]$ de taille $N^B \times N^C$, où chaque élément $\theta^{BC}_{ik} \geq 0$. Autrement dit, l'unité i de U^B est reliée à l'unité k de U^C à condition que $\theta^{BC}_{ik} > 0$, sinon; il n'existe aucun lien entre les deux unités.

Nous pouvons maintenant utiliser le sondage indirect par *transitivité*. Pour cela, nous sélectionnons un échantillon s^A à partir de la population U^A et commençons par identifier l'ensemble Ω^B de U^B. À partir de cet ensemble Ω^B, nous identifions alors les unités de U^C qui y sont associées, afin de former l'ensemble $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ et } \theta^{BC}_{ik} > 0\}$ d'unités devant être mesurées à partir de U^C. Une question importante est celle de savoir si, lorsqu'elle est appliquée dans le contexte du sondage indirect par transitivité, la MGPP est également transitive. Autrement dit, l'application de la MGPP de U^A à U^B, puis de U^B à U^C équivaut-elle à son application directe de U^A à U^C ?

Pour commencer, considérons le sondage indirect allant de U^A directement à la population cible U^C. Passer de la population U^A à U^B, puis à U^C revient à définir la matrice de liens $\mathbf{\Theta}_{AC} = [\theta^{AC}_{jk}]$ de taille $N^A \times N^C$ par $\mathbf{\Theta}_{AC} = \mathbf{\Theta}_{AB}\mathbf{\Theta}_{BC}$. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{AC}_{jk} > 0$, pour obtenir l'ensemble $\vec{\Omega}^C = \{k \in U^C \mid \exists\, j \in s^A \text{ et } \theta^{AC}_{jk} > 0\}$. Nous mesurons la variable d'intérêt y_k à partir de la population cible U^C. En appliquant la MGPP, nous obtenons, d'après (2.6), les poids suivants :

$$\vec{\mathbf{W}}_C = \tilde{\mathbf{\Theta}}'_{AC}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A \qquad (3.3)$$

où $\tilde{\mathbf{\Theta}}_{AC} = \mathbf{\Theta}_{AC}\,[\mathrm{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AC})]^{-1}$.

Considérons maintenant l'utilisation du sondage indirect en deux étapes. Pour chaque unité j sélectionnée dans s^A, nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{AB}_{ji} > 0$. Comme auparavant, nous avons $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta^{AB}_{ji} > 0\}$. Pour chaque unité i de l'ensemble Ω^B, nous identifions alors les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta^{BC}_{ik} > 0$. Nous avons alors l'ensemble $\Omega^C = \{k \in U^C \mid \exists\, i \in \Omega^B \text{ et } \theta^{BC}_{ik} > 0\}$. Partant de (2.6), nous obtenons le vecteur colonne $\mathbf{W}_B$ de poids associés aux unités de la population U^B :

$$\mathbf{W}_B = \tilde{\mathbf{\Theta}}'_{AB}\mathbf{T}_A\mathbf{\Pi}^{-1}_A\mathbf{1}_A. \qquad (3.4)$$

Pour chaque unité i de l'ensemble Ω^B, nous avons alors un poids non nul w^B_i. Or, l'ensemble Ω^B peut être considéré comme un échantillon d'unités qui sont utilisées dans un processus de sondage indirect pour identifier l'ensemble Ω^C. Par similarité avec l'échantillonnage indirect allant de l'échantillon s^A à la population cible U^B, l'application de la MGPP dans le contexte du sondage indirect allant de l'ensemble Ω^B à la population cible U^C produit les poids suivants :

$$\mathbf{W}_C = \tilde{\mathbf{\Theta}}'_{BC}\,\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B \qquad (3.5)$$

où $\tilde{\boldsymbol{\Theta}}_{BC} = \boldsymbol{\Theta}_{BC}[\,\mathrm{diag}(\mathbf{1}'_B \boldsymbol{\Theta}_{BC})]^{-1}$ et $\mathbf{T}_B = \mathrm{diag}(\mathbf{t}_B)$ avec $\mathbf{t}_B = (t_1^B, ..., t_{N^B}^B)'$ et $t_i^B = 1$ si $i \in \Omega^B$, et 0 autrement. Comme les poids $w_i^B = 0$ pour $i \notin \Omega^B$, nous avons $\mathbf{T}_B\,\mathrm{diag}(\mathbf{W}_B) = \mathrm{diag}(\mathbf{W}_B)$. Par conséquent, nous obtenons

$$\mathbf{W}_C = \tilde{\boldsymbol{\Theta}}'_{BC}\,\mathrm{diag}(\mathbf{W}_B)\mathbf{1}_B. \tag{3.6}$$

En remplaçant $\mathbf{W}_B$ par (3.4) dans l'équation (3.6), nous obtenons

$$\begin{aligned}\mathbf{W}_C &= \tilde{\boldsymbol{\Theta}}'_{BC}\,\mathrm{diag}(\tilde{\boldsymbol{\Theta}}'_{AB}\,\mathbf{T}_A\,\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A)\mathbf{1}_B \\ &= \tilde{\boldsymbol{\Theta}}'_{BC}\,\tilde{\boldsymbol{\Theta}}'_{AB}\,\mathbf{T}_A\,\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A. \end{aligned}\tag{3.7}$$

Puisque $\tilde{\boldsymbol{\Theta}}'_{BC}\tilde{\boldsymbol{\Theta}}'_{AB}\mathbf{1}_A = \tilde{\boldsymbol{\Theta}}'_{BC}\mathbf{1}_B = \mathbf{1}_C$, d'après le résultat 1, la matrice $\tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$ est une matrice de liens normalisée. Par conséquent, la MGPP est transitive, du moins dans un certain sens. Autrement dit, les poids $\mathbf{W}_C$ peuvent être obtenus en une seule étape en utilisant la matrice de liens normalisée $\tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$ dans la MGPP. Maintenant, pour que la MGPP soit parfaitement transitive, les poids $\mathbf{W}_C$ donnés par (3.7) devraient être exactement les mêmes que les poids $\vec{\mathbf{W}}_C$ donnés par (3.3). En comparant les équations (3.3) et (3.7), nous obtenons le résultat suivant :

Résultat 5 :

L'application de la MGPP de U^A à U^B, puis de U^B à U^C est transitive si et seulement si

$$\tilde{\boldsymbol{\Theta}}_{AC} = \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}. \tag{3.8}$$

Malheureusement, la condition (3.8) n'est pas vérifiée en général. En fait, il est relativement facile de produire des exemples où $\tilde{\boldsymbol{\Theta}}_{AC} \neq \tilde{\boldsymbol{\Theta}}_{AB}\tilde{\boldsymbol{\Theta}}_{BC}$.

4. Une propriété structurelle de la MGPP

À la présente section, nous insistons sur le fait que, dans le cas du sondage indirect, le processus d'échantillonnage dépend uniquement des liens entre les deux populations U^A et U^B. Outre le fait d'être nulles ou non, les valeurs des θ_{ji}^{AB} proprement dites n'interfèrent pas avec le processus d'échantillonnage. Par ailleurs, les valeurs des θ_{ji}^{AB} **jouent** un rôle dans les poids (et donc l'estimateur) produits par la MGPP. Nous développons cette notion dans les paragraphes qui suivent.

L'échantillonnage indirect associe à chaque échantillon s^A dans U^A un échantillon Ω^B dans U^B, nommément $\Omega^B = \{i \in U^B \mid \exists\, j \in s^A \text{ et } \theta_{ji}^{AB} > 0\}$. Donc, une fonction $f : s^A \to \Omega^B$ qui établit la correspondance entre l'échantillon s^A et l'échantillon Ω^B est déterminée de façon unique par l'ensemble de couples (j, i) avec $\theta_{ji}^{AB} > 0$. Soit $l_{ji}^{AB} = 1$ si $\theta_{ji}^{AB} > 0$, et 0 autrement. Il s'agit des éléments de la matrice d'incidence du graphe reliant U^A à U^B.

Supposons qu'on nous donne une fonction ϕ partant de l'ensemble de sous-ensembles de U^A vers l'ensemble de sous-ensembles de U^B. Comme f, supposons que ϕ satisfait la « propriété d'union » : $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, où s_1^A et s_2^A sont deux sous-ensembles de U^A.

Résultat 6 :

La fonction ϕ est déterminée sans équivoque par une *matrice de liens zéro-un*.

Démonstration :

Nous pouvons le démontrer comme il suit : prenons $s_j^A = \{j\}$ pour une unité j dans U^A. Alors, $\phi(s_j^A)$ est un ensemble dans U^B. Soit $l_{ji}^{AB} = 1$ si l'unité i de U^B appartient à $\phi(s_j^A)$, et 0 autrement. En vertu de la propriété d'union, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ et l'ensemble de l_{ji}^{AB} définit la *matrice de liens zéro-un* $\mathbf{L}_{AB} = [l_{ji}^{AB}]$ de taille $N^A \times N^B$, qui définit précisément la fonction ϕ.

Cela nous donne une relation d'équivalence entre les matrices de liens, associées à une propriété plus profonde. Soit p^A un plan d'échantillonnage sur U^A (c'est-à-dire une loi de probabilité sur l'ensemble de sous-ensembles de U^A). La fonction f induit un plan d'échantillonnage sur U^B par $p^B(\Omega^B) = \sum_{s^A:\,\Omega^B = f(s^A)} p^A(s^A)$. Comme le plan est induit par f, il ne dépend pas de la matrice de liens particulière $\boldsymbol{\Theta}_{AB}$ définissant la fonction, mais est plutôt une caractéristique de la classe d'équivalence par la voie de la matrice de liens zéro-un $\mathbf{L}_{AB}$. Par conséquent, l'estimateur d'Horvitz-Thompson en U^B dépend uniquement de cette classe. Il y a donc un certain intérêt à choisir dans cette classe une matrice $\boldsymbol{\Theta}_{AB}$ ayant, dans un certain sens, une caractéristique optimale (voir la section 6).

5. Matrices de liens spéciales

Comme le montre les sections précédentes, la matrice de liens $\boldsymbol{\Theta}_{AB}$ dicte la forme de l'estimateur (2.4) donnée par la MGPP. À la présente section, nous décrivons certaines matrices de liens spéciales $\boldsymbol{\Theta}_{AB}$ qui correspondent à des cas extrêmes. Il est probable que tous ces cas ne seront pas observés en pratique, mais ils illustrent l'effet de la matrice de liens sur l'estimateur (2.4).

5.1 Matrice identité

Supposons que la matrice de liens $\boldsymbol{\Theta}_{AB}$ soit donnée par la matrice identité $\mathbf{I}$. En pratique, cela signifie que la relation entre la population U^A et la population cible U^B est bijective. Naturellement, cela implique que $N^A = N^B = N$ et que la matrice identité $\mathbf{I}$ est de taille $N \times N$.

Comme premier résultat, nous avons $\tilde{\boldsymbol{\Theta}}_{AB} = \mathbf{I}$. Par conséquent, le vecteur de poids (2.6) est donné par $\mathbf{W}' = (t_1^A / \pi_1^A, ..., t_{N^A}^A / \pi_{N^A}^A)$ et nous avons aussi $\mathbf{Z} = \tilde{\boldsymbol{\Theta}}_{AB}\mathbf{Y} = \mathbf{Y}$. Donc, l'estimateur $\hat{Y}^B$ donné par (2.5) n'est autre que l'estimateur d'Horvitz-Thompson $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{Y}$.

5.2 Un pour tous (à l'intérieur des grappes)

Considérons le cas où la population U^B est divisée en Γ grappes γ, chacune de taille N_γ^B. Ces grappes sont telles que chaque grappe γ de U^B est associée à exactement une unité j de U^A. Par conséquent, nous pouvons utiliser la lettre γ pour les unités j de U^A ainsi que pour les grappes de U^B. Notons aussi que $\Gamma = N^A$.

Cette situation correspond à une matrice de liens $\boldsymbol{\Theta}_{AB}$ de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule ligne. Soit le vecteur ligne $\mathbf{1}'_{B\gamma}$ de taille N_γ^B et contenant uniquement des 1. La matrice de liens $\boldsymbol{\Theta}_{AB}$ est alors définie comme étant

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}'_{B1} & & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots & \vdots \\ \mathbf{0} & & \mathbf{1}'_{B\gamma} & & \mathbf{0} \\ \vdots & \ddots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & & \mathbf{1}'_{B\Gamma} \end{bmatrix}. \qquad (5.1)$$

Nous pouvons aussi écrire $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$. En utilisant cette expression, nous avons $\text{diag}(\mathbf{1}'_A \boldsymbol{\Theta}_{AB}) = \text{diag}(\mathbf{1}'_A \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})) = \text{diag}(\{\mathbf{1}'_{B1}, ..., \mathbf{1}'_{B\Gamma}\})$ et donc $\tilde{\boldsymbol{\Theta}}_{AB} = \boldsymbol{\Theta}_{AB}$. À partir de l'équation (2.6), nous obtenons le vecteur colonne de poids $\mathbf{W}' = (t_1^A / \pi_1^A \mathbf{1}'_{B1}, ..., t_\Gamma^A / \pi_\Gamma^A \mathbf{1}'_{B\Gamma})$. Comme nous pouvons le voir, les éléments du vecteur colonne $\mathbf{W}$ ont les valeurs $t_\gamma^A / \pi_\gamma^A$ répétées dans chaque grappe γ de U^B. Partant de (2.4), nous obtenons

$$\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} \frac{t_\gamma^A}{\pi_\gamma^A} Y_\gamma^B \qquad (5.2)$$

où $Y_\gamma^B = \sum_{i=1}^{N_\gamma^B} y_i$.

5.3 Tous pour un (à l'intérieur des grappes)

Considérons le cas où la population U^A est divisée en Γ grappes γ, chacune de taille N_γ^A. Ces grappes sont telles que chaque grappe γ de U^A est associée à exactement une unité i de U^B. Par conséquent, nous pouvons utiliser la lettre γ pour les grappes de U^A ainsi que les unités i de U^B. Notons aussi que $\Gamma = N^B$.

Cette situation correspond à une matrice de liens $\boldsymbol{\Theta}_{AB}$ de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule colonne. Soit le vecteur colonne $\mathbf{1}_{A\gamma}$ de taille N_γ^A et contenant uniquement des 1. La matrice de liens $\boldsymbol{\Theta}_{AB}$ est alors définie comme étant

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \mathbf{1}_{A1} & & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \ddots & \vdots \\ \mathbf{0} & & \mathbf{1}_{A\gamma} & & \mathbf{0} \\ \vdots & \ddots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & & \mathbf{1}_{A\Gamma} \end{bmatrix}. \qquad (5.3)$$

Nous pouvons aussi écrire $\boldsymbol{\Theta}_{AB} = \text{diag}(\{\mathbf{1}_{A1}, ..., \mathbf{1}_{A\Gamma}\})$. En utilisant cette expression, nous avons $\tilde{\boldsymbol{\Theta}}_{AB} = \text{diag}(\{1/N_1^A \mathbf{1}_{A1}, ..., 1/N_\Gamma^A \mathbf{1}_{A\Gamma}\})$. D'après (2.6), nous obtenons le vecteur colonne des poids $\mathbf{W}' = (1/N_1^A \sum_{j=1}^{N_1^A} t_j^A / \pi_j^A, ..., 1/N_\Gamma^A \sum_{j=1}^{N_\Gamma^A} t_j^A / \pi_j^A)$. Donc, les éléments γ (ou i) du vecteur colonne $\mathbf{W}$ ont les valeurs moyennes $\sum_{j=1}^{N_\gamma^A} t_j^A / \pi_j^A N_\gamma^A$, $\gamma = 1, ..., \Gamma$. Partant de (2.4), nous obtenons $\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} y_\gamma / N_\gamma^A \sum_{j=1}^{N_\gamma^A} t_j^A / \pi_j^A$.

5.4 Échantillonnage inefficace

Supposons que certaines lignes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ ne contiennent que des zéros. Cela signifie que certaines unités de la population U^A ne sont associées à aucune unité de la population U^B. Alors, si de telles unités sont sélectionnées dans l'échantillon s^A, elles ne permettront d'identifier aucune unité de U^B, ce qui peut être considéré comme inefficace du point de vue de l'échantillonnage. De façon plus formelle, supposons que chacune des N^{1A} premières lignes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ contient au moins un $\theta_{ji} > 0$, et qu'elles forment la sous-matrice $\boldsymbol{\Theta}_1$. Supposons que les N^{0A} autres lignes de $\boldsymbol{\Theta}_{AB}$ ont $\theta_{ji} = 0$ pour $i = 1, ..., N^B$. Par conséquent, nous avons

$$\boldsymbol{\Theta}_{AB} = \begin{bmatrix} \boldsymbol{\Theta}_1 \\ \mathbf{0} \end{bmatrix}.$$

Comme premier résultat, nous obtenons

$$\tilde{\boldsymbol{\Theta}}_{AB} = \begin{bmatrix} \boldsymbol{\Theta}_1 [\text{diag}(\mathbf{1}'_{1A} \boldsymbol{\Theta}_1)]^{-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\Theta}}_1 \\ \mathbf{0} \end{bmatrix} \qquad (5.4)$$

où $\mathbf{1}_{1A}$ est le vecteur colonne de 1 de taille N^{1A}. Partant de l'équation (2.6), nous obtenons le vecteur colonne de poids $\mathbf{W} = [\tilde{\boldsymbol{\Theta}}'_1 \ \mathbf{0}'] \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{1}_A$. Soit $\boldsymbol{\Pi}_{1A} = \text{diag}(\{\pi_1^A, ..., \pi_{N^{1A}}^A\})$ la matrice diagonale de taille $N^{1A} \times N^{1A}$ et, de plus, soit $\mathbf{T}_{1A} = \text{diag}(\{t_1^A, ..., t_{N^{1A}}^A\})$ la matrice diagonale de taille $N^{1A} \times N^{1A}$. Nous obtenons alors

$$\begin{aligned} \mathbf{W} &= [\tilde{\boldsymbol{\Theta}}'_1 \ \mathbf{0}'] \mathbf{T}_A \boldsymbol{\Pi}_A^{-1} \mathbf{1}_A \\ &= \tilde{\boldsymbol{\Theta}}'_1 \mathbf{T}_{1A} \boldsymbol{\Pi}_{1A}^{-1} \mathbf{1}_{1A}. \end{aligned} \qquad (5.5)$$

Comme le montre (5.5), les poids dépendent uniquement des probabilités de sélection π_j^A des unités de U^A qui ont au moins un $\theta_{ji} > 0$ pour $i = 1, ..., N^B$. À partir de (2.4), nous obtenons finalement $\hat{Y}^B = \mathbf{1}'_{1A} \mathbf{T}_{1A} \boldsymbol{\Pi}_{1A}^{-1} \tilde{\boldsymbol{\Theta}}_1 \mathbf{Y}$.

5.5 Estimateur biaisé

Supposons que certaines colonnes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ ne contiennent que des zéros. Cela signifie que certaines unités de la population U^B ne sont associées à aucune unité de la population cible U^A. Rappelons que, pour que la matrice $\tilde{\boldsymbol{\Theta}}_{AB}$ soit bien définie, il faut que $\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})^{-1}$ existe. Comme nous le verrons, le cas qui nous occupe ne satisfait pas cette condition, ce qui mène à un estimateur biaisé du total Y^B.

De façon plus formelle, supposons que chacune des N^{1B} premières colonnes de la matrice de liens $\boldsymbol{\Theta}_{AB}$ contient au moins un $\theta_{ji} > 0$, et supposons qu'elles forment la sous-matrice $\boldsymbol{\Theta}_1$, différentes de celles de la section précédente. Supposons que les N^{0B} autres colonnes de $\boldsymbol{\Theta}_{AB}$ ont $\theta_{ji} = 0$ pour $j = 1, ..., N^A$. Nous avons par conséquent $\boldsymbol{\Theta}_{AB} = [\boldsymbol{\Theta}_1, \mathbf{0}]$.

De cette définition, il découle directement que

$$[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^{-1} = [\text{diag}([\mathbf{1}'_A\boldsymbol{\Theta}_1, \mathbf{1}'_A\mathbf{0}])]^{-1} = \begin{bmatrix} \text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1}. \quad (5.6)$$

Puisque cette matrice est singulière, $[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^{-1}$ n'existe pas. Il serait peut-être possible d'utiliser une *inverse généralisée* comme solution de ce problème. Rappelons que, pour une matrice carrée donnée $\mathbf{A}$, la matrice $\mathbf{A}^-$ est une inverse généralisée de $\mathbf{A}$ à condition que $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ (Searle 1971). Une inverse généralisée possible de (5.6) est

$$[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^- = \begin{bmatrix} [\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_1)]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \quad (5.7)$$

Avec cette inverse généralisée, nous avons la matrice de liens normalisée suivante $\tilde{\boldsymbol{\Theta}}_- = \boldsymbol{\Theta}_{AB}[\text{diag}(\mathbf{1}'_A\boldsymbol{\Theta}_{AB})]^- = [\tilde{\boldsymbol{\Theta}}_1, \mathbf{0}]$. Partant de l'équation (2.6), nous pouvons obtenir le vecteur colonne $\mathbf{W}_-$ de poids :

$$\mathbf{W}_- = \begin{bmatrix} \tilde{\boldsymbol{\Theta}}'_1\mathbf{T}_A\boldsymbol{\Pi}_A^{-1}\mathbf{1}_A \\ \mathbf{0}' \end{bmatrix}. \quad (5.8)$$

Comme le montre l'expression (5.8), les poids sont nuls pour les unités i de la population cible U^B pour lesquels $\boldsymbol{\Theta}_{AB}$ contient $\theta_{ji} = 0$ pour $j = 1, ..., N^A$. Partant de (2.4) et en utilisant $\mathbf{W}_-$ au lieu de $\mathbf{W}$, nous obtenons $\hat{Y}^B_- = \mathbf{1}_A\mathbf{T}_A\boldsymbol{\Pi}_A^{-1}\tilde{\boldsymbol{\Theta}}_1\mathbf{Y}_1$ où $\mathbf{Y}_1 = \{y_1, ..., y_{N^{1B}}\}'$ est le sous-vecteur construit d'après les N^{1B} premiers éléments de $\mathbf{Y}$. Puisqu'en général, $E(\hat{Y}^B_-) = \mathbf{1}'_A\tilde{\boldsymbol{\Theta}}_1\mathbf{Y}_1 \neq \mathbf{1}'_A\mathbf{Y} = Y^B$, cet estimateur est biaisé pour le total Y^B.

6. Optimalité

L'optimalité est un aspect important de la MGPP. Comme nous l'avons montré au résultat 3, l'estimateur $\hat{Y}^B$ obtenu par cette méthode fournit des estimations sans biais à condition que la matrice $\tilde{\boldsymbol{\Theta}}_{AB}$ soit une matrice de liens normalisée. Étant donné que la variance (3.2) de cet estimateur dépend de cette matrice, il devrait exister au moins une matrice $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$ telle que la variance de l'estimateur $\hat{Y}^B$ soit minimale. Autrement dit, nous aimerions trouver les valeurs que les éléments θ_{ji}^{AB} plus grands que 0 devraient prendre pour obtenir l'estimateur de $\hat{Y}^B$ le plus précis.

Kalton et Brick (1995) ont été les premiers à examiner ce problème d'optimalité. Ils ont obtenu des résultats pour la situation simplifiée où $N^A = 2$ et où s^A est obtenu par échantillonnage avec probabilité égale. Ils ont conclu qu'il fallait utiliser $\theta_{ji}^{AB,\text{opt}} = 1$ lorsque $\theta_{ji}^{AB} > 0$ et $\theta_{ji}^{AB,\text{opt}} = 0$ lorsque $\theta_{ji}^{AB} = 0$. Lavallée (2002) et Lavallée et Caron (2001) ont obtenu des résultats du même genre par des simulations. Dans cette section, nous présentons de nouveaux résultats sur l'optimalité de la MGPP.

6.1 Factorisation

La factorisation est le problème inverse de la transitivité. Elle consiste à trouver une population U^G et des matrices de liens normalisées $\tilde{\boldsymbol{\Theta}}_{AG}$ et $\tilde{\boldsymbol{\Theta}}_{GB}$ telles que $\tilde{\boldsymbol{\Theta}}_{AB} = \tilde{\boldsymbol{\Theta}}_{AG}\tilde{\boldsymbol{\Theta}}_{GB}$. Cet exercice simplifie considérablement la recherche d'une matrice de liens normalisée optimale $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$.

Considérons que la population U^G est formée de grappes et que la factorisation est réalisée dans les contextes « un pour tous (à l'intérieur des grappes) » (de U^A à U^G) et « tous pour un (à l'intérieur des grappes) » (de U^G à U^B) présentés aux sections 5.2 et 5.3. Nous pouvons décrire cette situation de façon très générale comme il suit. Soit une population U^G contenant autant d'unités qu'il y a de liens partant des unités j de U^A. La taille de la population N^G est alors donnée par le nombre d'éléments θ_{ji}^{AB} de $\boldsymbol{\Theta}_{AB}$ dont la valeur est supérieure à 0. Chaque unité g de U^G peut être conceptualisée comme étant l'extrémité d'une « flèche » partant d'une unité j de U^A. Partant de ce graphe, il n'existe qu'une seule matrice de liens $\boldsymbol{\Theta}_{AG}$ de taille $N^A \times N^G$ assurant l'absence de biais, à savoir $\boldsymbol{\Theta}_{AG} = [\theta_{jg}^{AG}]$, où $\theta_{jg}^{AG} = 1$ s'il existe un lien (ou une « flèche ») partant de l'unité j de U^A vers l'unité g de U^G, et $\theta_{jg}^{AG} = 0$ autrement. Notons que, par construction, chaque unité g de U^G est liée, au plus, à une unité j de U^A et, donc, que $\tilde{\boldsymbol{\Theta}}_{AG} = \boldsymbol{\Theta}_{AG}$. Cela correspond à la situation « un pour tous dans les grappes » présentée à la section 5.2. Le sondage indirect de U^A à U^G est en fait un sondage en grappes type et fait aboutir la MGPP à l'estimateur d'Horvitz-Thompson habituel (voir Lavallée 2002). Dans le cas de l'exemple des parents et des enfants, le résultat de cette factorisation serait donné par la figure 2.

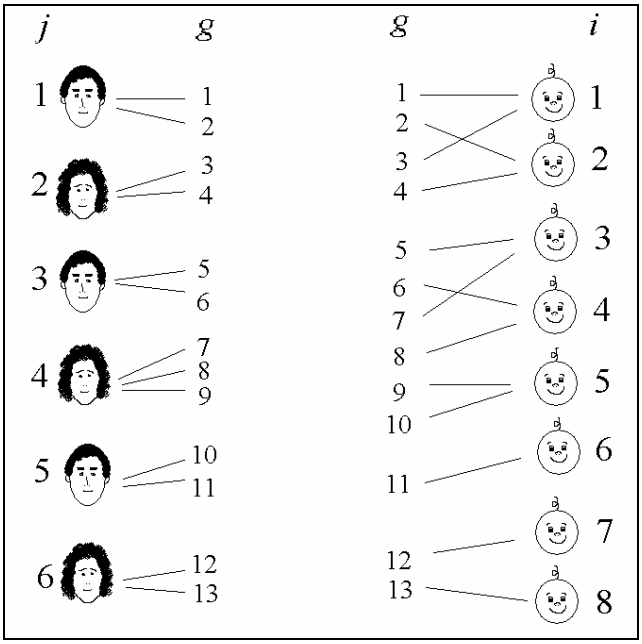

Figure 2. Résultat de la factorisation des populations parents-enfants.

Si nous considérons le graphe allant de U^G à U^B, nous pouvons construire la matrice de liens $\mathbf{\Theta}_{GB}$ de taille $N^G \times N^B$ comme suit. Étant donné la définition de la population U^G, chaque unité g de U^G est liée à exactement une unité i de U^B. Notons que le sondage indirect dans ce contexte peut être considéré comme un échantillonnage de grappes (c'est-à-dire les unités i de U^B) à partir de leurs éléments (c'est-à-dire les unités g de U^G). Il peut également être considéré comme étant le cas « tous pour un à l'intérieur des grappes » présenté à la section 5.3. Soit $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_G\mathbf{\Theta}_{GB})]^{-1}$ la matrice de liens normalisée obtenue à partir de $\mathbf{\Theta}_{GB}$. Nous avons $\text{diag}(\mathbf{1}'_G\mathbf{\Theta}_{GB}) = \text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})$, et, par conséquent, $\tilde{\mathbf{\Theta}}_{GB} = \mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1}$.

Or,

$$\begin{aligned}\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB} &= \mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB} \\ &= \mathbf{\Theta}_{AG}\mathbf{\Theta}_{GB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1} \\ &= \mathbf{\Theta}_{AB}[\text{diag}(\mathbf{1}'_A\mathbf{\Theta}_{AB})]^{-1} \\ &= \tilde{\mathbf{\Theta}}_{AB}. \end{aligned} \tag{6.1}$$

Donc, en utilisant cette construction, la matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$ reliant U^A à U^B peut toujours être factorisée en deux matrices $\tilde{\mathbf{\Theta}}_{AG}$ et $\tilde{\mathbf{\Theta}}_{GB}$.

6.2 Optimalité forte : énoncé du problème

Comme nous l'avons mentionné plus haut, le problème d'optimalité examiné ici consiste à minimiser la variance (3.2) par rapport à la matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$. Par la factorisation présentée à la section 6.1, nous obtenons

$$\begin{aligned}\text{Var}(\hat{Y}^B) &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{AB}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \\ &= \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y} \end{aligned} \tag{6.2}$$

où $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}$.

Pour toute matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{AB}$, la factorisation présentée à la section 6.1 produit systématiquement le même premier facteur $\tilde{\mathbf{\Theta}}_{AG}$. Par conséquent, si nous recherchons une matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ qui minimise la variance (3.2), il suffit d'optimiser le deuxième facteur $\tilde{\mathbf{\Theta}}_{GB}$. Nous aimerions aussi que la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ produise des estimations sans biais.

Soit U_i^G la sous-population de U^G contenant les N_i^G liens vers l'unité i de U^B. Notons que les sous-populations U_i^G sont disjointes. Donc, sans perte de généralité, nous pouvons classer les liens allant de U^A à U^B de façon que, pour tout i, les liens vers l'unité i dans U^B soient indicés consécutivement. Soit, ensuite, $\tilde{\mathbf{\theta}}_{GB,\,i}$ le i^{e} vecteur colonne de la matrice $\tilde{\mathbf{\Theta}}_{GB}$, $i = 1, \ldots, N^B$. Par construction, le vecteur $\tilde{\mathbf{\theta}}_{GB,\,i}$ ne contient que des éléments non nuls pour les N_i^G liens vers l'unité i de U^B. Donc, si nous représentons par $\tilde{\mathbf{\theta}}_{GB,\,i}$ un vecteur colonne de taille N_i^G contenant les éléments non nuls de $\tilde{\mathbf{\theta}}_{GB,\,i}$, nous obtenons

$$\tilde{\mathbf{\theta}}_{GB,\,i} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{\theta}}_{GB,\,i} \\ \mathbf{0} \end{bmatrix}.$$

De même, soit $\dot{\mathbf{1}}_{G,\,i}$ le vecteur colonne de taille N^G contenant des valeurs 1 pour N_i^G éléments, et des valeurs 0 ailleurs. Si nous représentons par $\mathbf{1}_{G,\,i}$ un vecteur colonne de taille N_i^G contenant les valeurs 1, nous obtenons

$$\dot{\mathbf{1}}_{G,\,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{G,\,i} \\ \mathbf{0} \end{bmatrix}.$$

Afin que l'application de la MGPP pour passer de U^G à U^B soit sans biais, il faut que nous ayons $\tilde{\mathbf{\theta}}'_{GB,\,i}\mathbf{1}_{G,\,i} = 1$ pour toute i, ou de façon équivalente, $\tilde{\mathbf{\theta}}'_{GB,\,i}\dot{\mathbf{1}}_{G,\,i} = 1$. Ensemble, toutes ces considérations mènent au problème d'optimisation suivant :

Trouver une matrice $\tilde{\mathbf{\Theta}}_{GB,\text{opt}} = \{\tilde{\mathbf{\theta}}_{GB,\text{opt},1}, \ldots, \tilde{\mathbf{\theta}}_{GB,\text{opt},N^B}\}$ satisfaisant $\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}\dot{\mathbf{1}}_{G,\,i} = 1$ pour tout $i = 1, \ldots, N^B$, et minimisant la forme quadratique $\text{Var}(\hat{Y}^B) = \mathbf{Y}'\tilde{\mathbf{\Theta}}'_{GB}\mathbf{\Delta}_G\tilde{\mathbf{\Theta}}_{GB}\mathbf{Y}$.

Ce problème n'est rien d'autre que la minimisation d'une forme quadratique positive sous des contraintes linéaires. Il s'agit d'un problème assez typique et simple à résoudre. Il est bien connu qu'il existe toujours une solution et qu'elle est unique si l'expression (6.2) est définie positive, ou que le

sous-espace nul de $\tilde{\boldsymbol{\Theta}}_{GB}$ n'est pas inclus dans l'espace nul de $\boldsymbol{\Delta}_G$.

Le problème d'optimisation susmentionné peut être réécrit sous une forme différente. Soit $\boldsymbol{\Delta}_{G,ii'}$ la sous-matrice de $\boldsymbol{\Delta}_G$ correspondant aux éléments qui occupent les positions g et g' si g possède un lien avec l'unité i et que g' possède un lien avec l'unité i'. Ces matrices constituent une partition de $\boldsymbol{\Delta}_G$. Notons que les matrices $\boldsymbol{\Delta}_{G,ii}$ sont symétriques, définies positives et que $\boldsymbol{\Delta}'_{G,ii'} = \boldsymbol{\Delta}_{G,i'i}$. Sous ces notations, le problème d'optimisation peut s'écrire sous la forme :

Minimiser

$$\sum_{i=1}^{N^B}\sum_{i'=1}^{N^B} y_i y_{i'} \tilde{\boldsymbol{\theta}}'_{GB,i} \boldsymbol{\Delta}_{G,ii'} \tilde{\boldsymbol{\theta}}_{GB,i'} \tag{6.3}$$

sous les contraintes $\tilde{\boldsymbol{\theta}}'_{GB,i}\mathbf{1}_{G,i} = 1$ pour tout $i = 1, \ldots, N^B$.

La minimisation est réalisée pour les vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ qui satisfont

$$y_i \sum_{i'=1}^{N^B} \boldsymbol{\Delta}_{G,ii'} \tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} y_{i'} = \lambda_i \mathbf{1}_{G,i} \tag{6.4}$$

pour tout $i = 1, \ldots, N^B$ et où les λ_i représentent les multiplicateurs de Lagrange entrant dans la minimisation sous contraintes de (6.3). Comme le montre (6.4), le choix optimal $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ (et par conséquent $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$) dépend en général explicitement du vecteur $\mathbf{Y}$, ce qui n'est pas utile en pratique. Observons que l'ensemble des λ_i dépend aussi de la variable $\mathbf{Y}$. Ce qui apparaîtra plus explicitement à la section 6.3. Cette raison est celle pour laquelle, au lieu d'une optimalité forte, nous rechercherons une forme plus faible donnant une solution « optimale » $\tilde{\boldsymbol{\Theta}}_{GB,\text{opt}}$ (et $\tilde{\boldsymbol{\Theta}}_{AB,\text{opt}}$) indépendante de $\mathbf{Y}$.

6.3 Optimalité faible

Les équations (6.4) doivent être valides pour tout vecteur $\mathbf{Y}$. En particulier, une condition nécessaire est qu'elles doivent être vérifiées pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B et $y_{i'} = 0$ pour toutes les autres unités i' de U^B ($i' \neq i$). Cela nous donne les conditions nécessaires (une pour chacune de ces variables particulières) $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Si nous supposons que $\boldsymbol{\Delta}_{G,ii}$ est inversible, nous obtenons alors $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}$. Il peut être démontré qu'il s'agit aussi d'une condition suffisante. Maintenant, comme $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} = 1$, nous avons $\lambda_i = 1/\mathbf{1}'_{G,i}\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}$. Par conséquent, une condition nécessaire et suffisante pour que l'équation (6.4) soit satisfaite est que

$$\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \frac{\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}}{\mathbf{1}'_{G,i}\boldsymbol{\Delta}^{-1}_{G,ii}\mathbf{1}_{G,i}}. \tag{6.5}$$

Ce résultat correspond à une optimisation faible au sens suivant. Le poids w_i donné par (2.6) satisfait $E(w_i) = 1$ et de surcroît $E(w_i \mid i \in \Omega^B) = 1/\pi_i^B$ où π_i^B est la probabilité d'inclusion de l'unité i dans Ω^B, qu'il est généralement difficile, voire impossible, de calculer en pratique. Notons que l'estimateur d'Horvitz-Thompson est caractérisé par $\text{Var}(w_i \mid i \in \Omega^B) = 0$. L'optimisation faible obtenue ici revient à minimiser $\text{Var}(w_i \mid i \in \Omega^B)$ sur toutes les matrices de liens normalisées possibles $\tilde{\boldsymbol{\Theta}}_{GB}$, ou, de façon équivalente $\tilde{\boldsymbol{\Theta}}_{AB}$. Cette variance est strictement positive dans les cas où l'unité i de U^B peut recevoir plus qu'un seul poids pour divers échantillons s^A. En outre, si nous utilisons (6.3), le multiplicateur λ_i semble être la variance du poids w_i et est, par conséquent, toujours strictement positif (sauf, cas que nous excluons, quand l'unité i est sélectionnée avec un poids égal à 1).

6.4 Forte optimalité indépendante de Y

L'optimalité faible est une condition nécessaire à l'optimalité forte indépendante du vecteur $\mathbf{Y}$ d'une variable d'intérêt. Elle donne la forme nécessaire des vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ dans (6.4). Pour obtenir les conditions suffisantes pour une forte optimalité indépendante de $\mathbf{Y}$, nous retournons aux équations (6.4). Ces dernières doivent être satisfaites pour tous les vecteurs $\mathbf{Y}$ et doivent par conséquent être satisfaites pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B, $y_{i'} = 1$, pour une autre unité i' de U^B, et $y_{i''} = 0$ pour toutes les autres unités i'' de U^B ($i'' \neq i' \neq i$). Dans ce cas, pour que les équations (6.4) soient satisfaites, il est nécessaire d'avoir les relations suivantes pour tout i et i' :

$$\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} + \boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} = \lambda_i^{ii'}\mathbf{1}_{G,i} \tag{6.6}$$

$$\boldsymbol{\Delta}_{G,i'i'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} + \boldsymbol{\Delta}_{G,i'i}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_{i'}^{ii'}\mathbf{1}_{G,i'}.$$

Comme nous devons nécessairement avoir une optimalité faible, nous avons $\boldsymbol{\Delta}_{G,ii}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i} = \lambda_i \mathbf{1}_{G,i}$. Partant de la première ligne de (6.6), nous obtenons alors

$$\begin{aligned}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} &= (\lambda_i^{ii'} - \lambda_i)\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\mathbf{1}_{G,i}.\end{aligned} \tag{6.7}$$

En multipliant les deux membres de (6.7) par $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}$, nous obtenons

$$\begin{aligned}\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'} &= \Phi_{ii'}\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} \\ &= \Phi_{ii'}\end{aligned}$$

puisque $\tilde{\boldsymbol{\theta}}'_{GB,\text{opt},i}\mathbf{1}_{G,i} = 1$. Soit $\boldsymbol{\Phi}$ la matrice contenant les éléments $\Phi_{ii'}$ hors de la diagonale et $\Phi_{ii} = \lambda_i$ sur la diagonale. En utilisant de nouveau (6.2), nous pouvons

montrer que la variance optimale (quand elle existe) a pour expression $\mathbf{Y}'\mathbf{\Phi}\mathbf{Y}$.

Démontrons que cet ensemble de conditions est également suffisant. Supposons que (6.7) est vérifiée. Notons que, pour $i = i'$, la condition (6.7) n'est rien d'autre que (6.5) qui donne les valeurs nécessaires pour les $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$. Il est maintenant simple de vérifier que (6.4) tient quelle que soit la valeur de $\mathbf{Y}$ et que nous avons obtenu l'optimalité forte. Les valeurs de λ_i dépendent de $\mathbf{Y}$, ainsi que de la variance $\mathrm{Var}(\hat{Y}^B)$, mais nous savons que les équations (6.4) ont toujours la même solution (6.5) qui ne dépend pas de $\mathbf{Y}$. Par conséquent, nous avons le résultat suivant :

Résultat 7 :

Les conditions $\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} = \Phi_{ii'}\mathbf{1}_{G,\,i}$ sont nécessaires et suffisantes pour qu'il existe une matrice de liens normalisée $\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}}$, ou de façon équivalente, $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}}$, qui permet d'obtenir une optimalité forte indépendante du vecteur $\mathbf{Y}$ de la variable d'intérêt. Les valeurs figurant dans les colonnes de cette matrice optimale forte sont données par (6.5) qui sont les vecteurs $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ obtenus à partir de l'optimalité faible.

Il convient de souligner que, puisque $\mathbf{\Delta}_{G,\,ii}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \lambda_i\mathbf{1}_{G,\,i}$, l'expression (6.7) peut s'écrire de façon équivalente sous la forme

$$\Phi^{**}_{ii'}\,\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i} = \mathbf{\Delta}^{-1}_{G,\,ii}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'} \qquad (6.8a)$$

ou

$$\Phi^{*}_{ii'}\,\mathbf{1}_{G,\,i} = \mathbf{\Delta}_{G,\,ii'}\mathbf{\Delta}^{-1}_{G,\,i'i'}\mathbf{1}_{G,\,i'} \qquad (6.8b)$$

où $\Phi^{**}_{ii'} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,\,i}\mathbf{\Delta}^{-1}_{G,\,ii}\mathbf{1}_{G,\,i})$ et $\Phi^{*}_{ii'} = (\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}\mathbf{\Delta}_{G,\,ii'}\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i'})(\mathbf{1}'_{G,\,i'}\mathbf{\Delta}^{-1}_{G,\,i'i'}\mathbf{1}_{G,\,i'})$. Dans certaines situations, ces expressions peuvent s'avérer plus faciles à utiliser que l'expression (6.7) énoncée dans le résultat 7.

6.5 Deux exemples

Nous présentons maintenant deux exemples qui illustrent la théorie que nous venons d'exposer sur l'optimalité faible et l'optimalité forte indépendante de $\mathbf{Y}$.

Exemple 1 : Échantillonnage de Poisson

Supposons que l'échantillon s^A soit sélectionné par échantillonnage de Bernoulli ou de Poisson. Dans ce cas, la matrice $\mathbf{\Delta}_A$ de taille $N^A \times N^A$ est donnée par $\mathbf{\Delta}_A = \mathrm{diag}(1/\pi^A_j - 1)$. Si nous considérons la factorisation de la section 6.1, nous avons $\mathbf{\Delta}_G = \tilde{\mathbf{\Theta}}'_{AG}\,\mathbf{\Delta}_A\,\tilde{\mathbf{\Theta}}_{AG} = \tilde{\mathbf{\Theta}}'_{AG}\,[\mathrm{diag}(1/\pi^A_j - 1)]\,\tilde{\mathbf{\Theta}}_{AG} = [\mathrm{diag}((1/\pi^A_j - 1)\mathbf{1}_{A,\,jj})]$, où $\mathbf{1}_{A,\,jj}$ est une matrice carrée de taille N^A_j, avec N^A_j égal au nombre de liens (ou « flèches ») ayant pour origine l'unité j de U^A. De $\mathbf{\Delta}_G$, nous extrayons les sous-matrices $\mathbf{\Delta}_{G,\,ii}$ qui sont, ici, diagonales. Chaque sous-matrice $\mathbf{\Delta}_{G,\,ii}$ est donnée par $\Delta_{G,\,ii} = \mathrm{diag}(1/\pi^A_g - 1)$, qui est de taille N^G_i. Notons que chaque valeur $(1/\pi^A_g - 1)$ correspond simplement à une unité j de U^A qui a été liée antérieurement à l'unité g de U^G, qui à son tour a été liée à l'unité i de U^B. Maintenant, partant de (6.5), nous obtenons directement les valeurs optimales $\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,i}$ qui minimisent $\mathrm{Var}(\hat{Y}^B)$, au sens faible. Ces valeurs sont données par les vecteurs

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{ \frac{\pi^A_1}{(1-\pi^A_1)\,\tau^G_i},\ \ldots,\ \frac{\pi^A_{N^G_i}}{(1-\pi^A_{N^G_i})\,\tau^G_i} \right\}$$

où

$$\tau^G_i = \sum\nolimits_{g=1}^{N^G_i} \pi^A_g/(1-\pi^A_g),\ i = 1,\ \ldots,\ N^B.$$

Les $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$ sont utilisés pour construire les vecteurs $\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i}$, puis la matrice $\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}} = \{\tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,1}, \ldots, \tilde{\boldsymbol{\theta}}_{GB,\,\mathrm{opt},\,N^B}\}$. Enfin, après avoir calculé la matrice optimale, $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}} = \mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB,\,\mathrm{opt}}$, nous obtenons les poids optimaux $\mathbf{W}_{\mathrm{opt}}$ en utilisant (2.6).

Il convient de souligner que, si les probabilités d'inclusion π^A_j sont égales, nous obtenons

$$\tilde{\boldsymbol{\theta}}'_{GB,\,\mathrm{opt},\,i} = \left\{ \frac{1}{N^G_i},\ \ldots,\ \frac{1}{N^G_i} \right\} = \frac{1}{N^G_i}\mathbf{1}_{GB,\,i},$$

où N^G_i est tout simplement le nombre d'unités de U^A liées à l'unité i de U^B. Autrement dit, dans le contexte de l'échantillonnage de Bernoulli (c'est-à-dire l'échantillonnage de Poisson avec probabilités égales), pour minimiser la variance $\mathrm{Var}(\hat{Y}^B)$, le choix des valeurs de $\theta^{AB}_{\mathrm{opt},\,ji}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B, et 0 autrement. Cela correspond aux résultats obtenus par Kalton et Brick (1995), Lavallée (2002), ainsi que Lavallée et Caron (2001).

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) sont satisfaites pour la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\,\mathrm{opt}}$ que nous avons obtenue par optimisation faible. Le cas échéant, cette matrice donne aussi une optimalité forte indépendante de la variable d'intérêt y_i. Premièrement, nous avons

$$\Delta^{-1}_{G,\,ii} = \mathrm{diag}\left(\frac{\pi^A_g}{1-\pi^A_g} \right).$$

En outre, chaque sous-matrice $\Delta_{G,\,ii'}$ de taille $N^G_i \times N^G_{i'}$ a plus ou moins une structure diagonale, mais « rembourrée » de zéro. Autrement dit, un élément typique de $\Delta_{G,\,ii'}$ est donné par $(1/\pi^A_g - 1)$ sur une partie de la diagonale si i et i' sont toutes deux liées à la même unité j de U^A

(c'est-à-dire liées à l'unité g de U^G provenant de la même unité j de U^A), et 0 autrement. Par conséquent, si deux unités i et i' ne sont pas liées aux mêmes unités de U^A, alors $\Delta_{G,ii'}$ est une matrice de zéros et les conditions (6.7), (6.8a) et (6.8b) sont automatiquement satisfaites. Si nous nous référons à la figure 1, les enfants $i=2$ et $i'=3$ de U^B ne sont pas apparentés aux mêmes parents j de U^A. Si la sélection des parents est faite par échantillonnage de Poisson ou de Bernoulli, la matrice $\mathbf{\Delta}_{G,23}$ de dimension 2×2 ne contiendra alors que des zéros, c'est-à-dire

$$\mathbf{\Delta}_{G,23}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}.$$

Donc, les relations (6.7), (6.8a) ou (6.8b) seront satisfaites avec $\Phi_{23}=0$, ce qui exprime le fait que les poids de i et de i' ne sont pas corrélés.

Si deux unités i et i' sont reliées à la même unité j de U^A, alors, si nous utilisons (6.7), le vecteur colonne $\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$ contient le scalaire $(\tau_{i'}^G)^{-1}=[\sum_{g=1}^{N_{i'}^G}\pi_g^A/(1-\pi_g^A)]^{-1}$ pour ses N_i^B premières composantes, et 0 pour les $N_{i'}^B-N_i^B$ composantes restantes (en supposant que $N_{i'}^B\geq N_i^B$). Comme la quantité $\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$ doit être égale à $\Phi_{ii'}\mathbf{1}_{G,i}$ pour satisfaire (6.7), elle doit contenir uniquement la valeur $\Phi_{ii'}$. Puisque $\Phi_{ii'}=\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}\mathbf{\Delta}_{G,ii'}\tilde{\mathbf{\theta}}_{GB,\text{opt},i'}$, cela se produira uniquement si le vecteur $\tilde{\mathbf{\theta}}_{GB,\text{opt},i}=[1]$, ce qui signifie qu'il n'existe qu'un seul lien vers l'unité i de U^B. Comme nous le voyons, il ne s'agit pas d'une condition qui sera satisfaite en général et, par conséquent, nous pouvons dire, dans le cas de l'échantillonnage de Poisson, il n'y aura généralement pas d'optimalité forte indépendante de $\mathbf{Y}$.

Pour conclure, nous pourrions dire que dans le cas de l'échantillonnage de Poisson ou de Bernoulli, les conditions (6.7), (6.8a) ou (6.8b) seront satisfaites en pratique uniquement si les unités de U^A sont liées à une seule unité de U^B, comme dans le cas de l'échantillonnage des ménages en utilisant une liste de personnes. Dans les autres cas, la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ obtenue par optimalité faible ne donnera vraisemblablement pas lieu à une optimisation forte indépendante de $\mathbf{Y}$.

Exemple 2 : Échantillonnage aléatoire simple

Supposons que l'on sélectionne l'échantillon s^A par échantillonnage aléatoire simple. Dans ce cas, la matrice $\mathbf{\Delta}_A$ de taille $N^A\times N^A$ est donnée par

$$\mathbf{\Delta}_A=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\left[\mathbf{I}_A-\frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right].$$

Si nous considérons la factorisation de la section 6.1, nous obtenons

$$\begin{aligned}\mathbf{\Delta}_G&=\tilde{\mathbf{\Theta}}'_{AG}\mathbf{\Delta}_A\tilde{\mathbf{\Theta}}_{AG}\\&=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\tilde{\mathbf{\Theta}}'_{AG}\left[\mathbf{I}_A-\frac{\mathbf{1}_A\mathbf{1}'_A}{N^A}\right]\tilde{\mathbf{\Theta}}_{AG}\\&=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\text{diag}(\mathbf{1}_{A,jj})-\frac{\mathbf{1}_G\mathbf{1}'_G}{N^A}\right]\end{aligned}\qquad(6.9)$$

où $\mathbf{1}_{A,jj}$ est une matrice carrée de taille N_j^A, avec N_j^A égal au nombre de liens (ou de « flèches ») ayant pour origine l'unité j de U^A. De $\mathbf{\Delta}_G$, nous extrayons les sous-matrices $\mathbf{\Delta}_{G,ii}$. Chaque sous-matrice $\mathbf{\Delta}_{G,ii}$ est donnée par

$$\mathbf{\Delta}_{G,ii}=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\mathbf{I}_{G,i}-\frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i}}{N^A}\right],$$

qui est de taille N_i^G. Alors, en utilisant un résultat matriciel que l'on peut trouver, entre autres, dans Jazwinski (1970), nous obtenons

$$\mathbf{\Delta}_{G,ii}^{-1}=\frac{(N^A-1)}{(N^A-n^A)}\frac{n^A}{N^A}\times\left[\mathbf{I}_{G,i}+\frac{1}{(N^A-N_i^G)}\mathbf{1}_{G,i}\mathbf{1}'_{G,i}\right].$$

Ensuite, partant de (6.5), nous obtenons directement les valeurs optimales

$$\tilde{\mathbf{\theta}}_{GB,\text{opt},i}=\frac{1}{N_i^G}\mathbf{1}_{G,i}$$

qui minimisent $\text{Var}(\hat{Y}^B)$, au sens faible, $i=1,\ldots,N^B$. Nous utilisons ces valeurs pour construire les vecteurs $\tilde{\mathbf{\theta}}'_{GB,\text{opt},i}$, puis la matrice $\tilde{\mathbf{\Theta}}_{GB,\text{opt}}=\{\tilde{\mathbf{\theta}}_{GB,\text{opt},1},\ldots,\tilde{\mathbf{\theta}}_{GB,\text{opt},N^B}\}$. Enfin, après avoir calculé la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}=\mathbf{\Theta}_{AG}\tilde{\mathbf{\Theta}}_{GB,\text{opt}}$, nous obtenons les poids optimaux $\mathbf{W}_{\text{opt}}$ en utilisant (2.6).

De nouveau, ce résultat est important, car il va directement dans le sens des résultats de Kalton et Brick (1995), de Lavallée (2002), et de Lavallée et Caron (2001). Autrement dit, dans le cas de l'échantillonnage aléatoire simple, le choix optimal de $\theta_{\text{opt},ji}^{AB}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B, et 0 sinon.

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) pour une optimalité forte indépendante de y_i sont satisfaites pour la matrice optimale $\tilde{\mathbf{\Theta}}_{AB,\text{opt}}$ que nous obtenons par optimisation faible. D'abord, chaque sous-matrice $\Delta_{G,ii'}$ de taille $N_i^G\times N_{i'}^G$ est donnée par

$$\mathbf{\Delta}_{G,ii'}=\frac{N^A}{n^A}\frac{(N^A-n^A)}{(N^A-1)}\times\left[\mathbf{H}_{G,ii'}-\frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right]$$

où $\mathbf{H}_{G,ii'}$ est une matrice diagonale de taille $N_i^G \times N_{i'}^G$ de valeur 1, « rembourrée » de zéros. En suivant exactement le même scénario que l'exemple 1, un élément type de $\mathbf{H}_{G,ii'}$ est donné par 1 si i et i' sont toutes deux liées à la même unité j de U^A (c'est-à-dire liées à l'unité g de U^G), et 0 sinon. Par conséquent, nous pouvons voir facilement dans quel cas les conditions (6.7), (6.8a) ou (6.8b) peuvent être satisfaites. En fait, comme toutes les composantes de $\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i}$ sont égales, $\boldsymbol{\Delta}_{G,ii'}\tilde{\boldsymbol{\theta}}_{GB,\text{opt},i'}$ est un vecteur proportionnel à la somme des lignes de $\boldsymbol{\Delta}_{G,ii'}$, c'est-à-dire la somme des lignes de

$$\left[\mathbf{H}_{G,ii'} - \frac{\mathbf{1}_{G,i}\mathbf{1}'_{G,i'}}{N^A}\right].$$

Mais (6.7) dit que ce vecteur doit avoir les mêmes composantes. Cela n'est possible que si et seulement si la matrice $\mathbf{H}_{G,ii'}$ ne contient que des zéros, ou qu'elle est de dimension 1×1, ce qui se produit lorsque i et i' sont chacune liées uniquement à un élément de U^A. Donc, comme pour l'échantillonnage de Poisson, une optimalité fort indépendante de $\mathbf{Y}$ n'a généralement pas lieu dans le cas de l'échantillonnage aléatoire simple.

7. Conclusion

Dans le présent article, nous avons discuté de l'utilisation du sondage indirect conjugué à la méthode généralisée du partage des poids (MGPP) élaborée pour produire des poids. Puis, nous avons démontré les propriétés suivantes de la MGPP : absence de biais, calcul de la variance et transitivité. Ensuite nous avons présenté une section sur l'utilisation de la MGPP lorsque les liens entre les populations U^A et U^B sont exprimés par des valeurs 1 et 0, c'est-à-dire qu'il existe un lien ou qu'il n'en existe pas. La section suivante a été consacrée aux résultats obtenus avec diverses formes de matrices de liens. Enfin, nous avons abordé le problème de l'optimalité, c'est-à-dire le choix des valeurs optimales pour exprimer les liens entre U^A et U^B de façon à minimiser la variance des estimations obtenues en appliquant la MGPP. Nous avons fait la distinction entre deux formes d'optimisation, à savoir l'optimisation faible et l'optimisation forte.

L'optimisation faible consiste à trouver les valeurs des liens qu'il convient d'utiliser pour minimiser, pour chaque unité, la variance des poids produits par la MGPP. La solution est toujours définie de façon unique, et est facile à calculer et à appliquer en pratique. L'optimisation faible est également une condition nécessaire de l'optimisation forte. L'optimisation forte consiste à trouver les valeurs des liens permettant de minimiser la variance de l'estimation du total de toute variable d'intérêt y. Elle n'existe pas pour tous les plans d'échantillonnage et type de liens entre les populations U^A et U^B. Elle dépend aussi de relations assez compliquées.

Nous recommandons d'utiliser l'optimisation faible, parce qu'elle coule de source et qu'elle est très facile à utiliser. En outre, si notre problème d'estimation peut être optimisé également au sens fort, nous aurons obtenu ce résultat par la voie de l'optimisation faible, même si nous ne l'avons pas démontré.

Remerciements

Les auteurs remercient toutes les personnes qui ont manifesté un intérêt pour le sondage indirect et particulièrement la MGPP. Elles ont motivé la rédaction de cet article qui dépasse le cadre de ce qui avait été écrit antérieurement à ce sujet.

Bibliographie

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. Dans *Panel Surveys* (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York : John Wiley & Sons, Inc. 135-159.

Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes, n°84-85-86, Actes des Journées de methodologie statistique des 17-18 mars 1998*, 63-82.

Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York : Academic Press.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 2, 171-187.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Searle, S.R. (1971). *Linear Models*. New York : John Wiley & Sons, Inc.