Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

CORRECTION DE LA NON-RÉPONSE DANS LE CADRE DU SONDAGE INDIRECT

Pierre Lavallée[1]

RÉSUMÉ

Il arrive en pratique qu'on ne dispose pas directement d'une liste contenant les unités de collecte désirées, mais plutôt d'une liste d'autres unités reliées d'une certaine façon à la liste des unités de collecte. On peut donc parler de deux populations U^A et U^B reliées entre elles où on désire produire une estimation pour U^B. Malheureusement, on dispose d'une base de sondage seulement pour U^A. On peut alors imaginer le tirage d'un échantillon s^A de U^A afin de produire une estimation pour U^B en se servant de la correspondance existante entre les deux populations. C'est ce qu'on peut désigner par *sondage indirect*. Afin d'associer une probabilité de sélection, ou un poids d'estimation, aux unités enquêtées dans la population cible, Lavallée (1995) a développé la *méthode généralisée du partage des poids* (MGPP). La MGPP permet d'obtenir un poids d'estimation qui correspond en gros à une moyenne des poids de sondage des unités de l'échantillon s^A.

Avec le sondage indirect, de la non-réponse totale peut être présente au sein de l'échantillon s^A tiré de U^A, ou au sein des unités identifiées pour être enquêtées au sein de U^B, c'est-à-dire des unités de collecte. Comme l'enquête des unités au sein de la population U^B se fait par grappe, on distingue deux types de non-réponse totale associé à U^B : la non-réponse de grappes et la non-réponse d'unités. Avec le sondage indirect, on retrouve aussi la non-réponse de liens. Ce type de non-réponse est associé à la situation où on ne peut établir si une unité donnée de U^B est liée à une autre unité de U^A, ce qui pose de sérieux problèmes d'estimation dans l'application de la MGPP.

MOTS CLÉS: Méthode généralisée du partage des poids; non-réponse; sondage en grappes

1. INTRODUCTION

Pour tirer les échantillons nécessaires aux sondages sociaux ou économiques, il est utile de disposer de bases de sondage, c'est-à-dire de listes d'unités destinées à cerner les populations cibles. Malheureusement, il arrive qu'on ne dispose pas directement d'une liste contenant les unités de collecte désirées, mais plutôt d'une liste d'unités reliées d'une certaine façon à la liste des unités de collecte. On peut donc parler de deux populations U^A et U^B reliées entre elles où on désire produire une estimation pour U^B. Malheureusement, on dispose d'une base de sondage seulement pour U^A. On peut alors imaginer le tirage d'un échantillon s^A de U^A afin de produire une estimation pour U^B en se servant de la correspondance existante entre les deux populations. C'est ce qu'on peut désigner par *sondage indirect*.

L'estimation d'un total (ou d'une moyenne) d'une population cible U^B de grappes en se servant d'un échantillon tiré d'une autre population U^A reliée d'une certaine façon à la première peut constituer un défi de taille et ce, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté d'associer une probabilité de sélection, ou un poids d'estimation, aux unités enquêtées dans la population cible. Afin de résoudre ce type de problème d'estimation, on a développé la *méthode généralisée du partage des poids* (MGPP).

La MGPP permet d'obtenir un poids d'estimation pour chaque unité enquêtée de la population cible U^B. Ce poids d'estimation correspond en gros à une moyenne des poids de sondage des unités de l'échantillon s^A. Lavallée (1995) a présenté pour la première fois la MGPP dans le cadre du problème de la pondération transversale d'enquêtes longitudinales auprès de ménages. La MGPP constitue une généralisation de la *méthode du partage des poids* décrite par Ernst (1989).

1 Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa (Ontario), K1A 0T6 CANADA. Courriel : pierre.lavallee@statcan.ca.

2. DESCRIPTION

On sélectionne un échantillon s^A contenant m^A unités dans la population U^A contenant M^A unités selon un certain plan de sondage. Soit π_j^A, la probabilité de sélection de l'unité j où $\pi_j^A > 0$ pour toutes les unités $j \in U^A$. D'autre part, la population cible U^B contient M^B unités. Cette population est divisée en N grappes, où la grappe i contient M_i^B unités. On suppose qu'il existe un lien (ou une correspondance) entre les unités j de la population U^A et les unités k des grappes i de la population U^B. Ce lien est identifié par une variable indicatrice $l_{j,ik}$, où $l_{j,ik}=1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité $ik \in U^B$, et 0 dans les autres cas. Notons qu'il peut y avoir des cas où il n'existe pas de lien entre une unité j de la population U^A et les unités k des grappes i de la population U^B, c'est-à-dire $L_j^A = \sum_{i=1}^{N}\sum_{k=1}^{M_i^B} l_{j,ik} = 0$. D'autre part, il peut exister aucun, un ou plusieurs liens pour une unité k de la grappe i de la population U^B, c'est-à-dire qu'il est possible d'avoir $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik} = 0$, $L_{ik}^B = 1$, ou même $L_{ik}^B > 1$ pour les unités $ik \in U^B$. Pour utiliser la MGPP et que celle-ci soit sans biais, nous devons cependant satisfaire la contrainte suivante : Chaque grappe i de U^B doit posséder au moins un lien avec une unité j de U^A, c'est-à-dire $L_i^B = \sum_{k=1}^{M_i^B}\sum_{j=1}^{M^A} l_{j,ik} > 0$.

Pour chaque unité j sélectionnée dans s^A, on identifie les unités ik de U^B qui ont un lien non nul avec j, c'est-à-dire $l_{j,ik}=1$. Si $L_j^A = 0$ pour une unité j de s^A, il n'y a tout simplement pas d'unité de U^B identifiée par cette unité j, ce qui affecte l'efficacité de l'échantillon s^A mais n'introduit pas de biais. Pour chaque unité ik identifiée, on suppose qu'on peut établir la liste des M_i^B unités de la grappe i contenant cette unité. Chaque grappe i représente alors, en elle-même, une population U_i^B où $U^B = \bigcup_{i=1}^{N} U_i^B$. Soit Ω^B, l'ensemble des n grappes identifiées par les unités $j \in s^A$.

Une contrainte importante à laquelle est assujettie le processus d'enquête (ou de mesure) est de considérer **toutes** les unités appartenant à la même grappe. Autrement dit, si une unité est sélectionnée dans l'échantillon, alors toutes les unités de la grappe contenant l'unité sélectionnée seront enquêtées. Cette contrainte survient souvent dans les enquêtes pour deux raisons : (i) par souci d'économie et (ii) par nécessité de produire des estimations sur les grappes. On enquête ainsi auprès de toutes les unités k des grappes $i \in \Omega^B$ où on mesure une variable d'intérêt y_{ik} et le nombre de liens $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$ entre chaque unité ik et la population U^A.

Pour la population cible U^B, nous cherchons à estimer le total $Y^B = \sum_{i=1}^{N}\sum_{k=1}^{M_i^B} y_{ik}$. En appliquant la MGPP, nous voulons attribuer un poids d'estimation w_{ik} à chaque unité k d'une grappe enquêtée i. Pour estimer le total Y^B de la population cible U^B, on peut alors se servir de l'estimateur

$$\hat{Y}^B = \sum_{i=1}^{n}\sum_{k=1}^{M_i^B} w_{ik} y_{ik} \qquad (2.1)$$

où n est le nombre de grappes enquêtées et w_{ik}, le poids attribué à l'unité k de la grappe i.

Étapes de la MGPP :

Étape 1 : Pour chaque unité k des grappes i de Ω^B, on calcule le poids initial $w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik}\, t_j / \pi_j^A$, où $t_j=1$ si $j \in s^A$, et 0 sinon. On note qu'une unité ik n'ayant de lien avec aucune unité j de U^A possède automatiquement un poids initial nul.

Étape 2 : Pour chaque unité k des grappes i de Ω^B, on obtient le nombre total de liens $L_{ik}^B = \sum_{j=1}^{M^A} l_{j,ik}$. Cette quantité représente le nombre de liens entre les unités de U^A et

l'unité k de la grappe i de la population U^B. La quantité $L_i^B = \sum_{k=1}^{M_i^B} L_{ik}^B$ correspond alors au nombre total de liens de la grappe i.

Étape 3 : On calcule le poids final $w_i = \sum_{k=1}^{M_i^B} w'_{ik} \big/ L_i^B$.

Étape 4 : Enfin, nous posons $w_{ik} = w_i$ pour tous les $k \in U_i^B$.

Parce que les poids d'estimation issus de la MGPP sont les mêmes pour l'ensemble des M_i^B unités de chaque grappe i, on note que l'estimateur (2.1) peut s'écrire en fonction des grappes seulement. On a ainsi $\hat{Y}^B = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} = \sum_{i=1}^{n} w_i \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{n} w_i Y_i$, où $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$.

3. BIAIS ET VARIANCE

Afin de pouvoir calculer le biais et la variance de l'estimateur $\hat{Y}^B$, on présente le théorème suivant :

Théorème : Dualité de la forme de $\hat{Y}^B$ par rapport à U^A et à U^B

Soit $z_{ik} = Y_i / L_i^B$ pour tous les $k \in U_i^B$. On définit $Z_j = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik}$. L'estimateur $\hat{Y}^B$, donné par (2.1), peut alors également s'écrire sous la forme

$$\hat{Y}^B = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \, . \tag{3.1}$$

L'estimateur $\hat{Y}^B$ peut donc s'écrire en fonction des unités ik de U^B, ou en fonction des unités j de U^A. On remarque que l'estimateur $\hat{Y}^B$ n'est en fait qu'un estimateur d'Horvitz-Thompson où la variable d'intérêt est la variable Z_j. Cette constatation nous mène à deux importants corollaires: (i) l'estimateur $\hat{Y}^B$ est sans biais pour l'estimation de Y^B et (ii) la formule de la variance de l'estimateur $\hat{Y}^B$ est celle de l'estimateur d'Horvitz-Thompson avec la variable Z_j. Cette variance est donnée notamment dans Särndal, Swensson et Wretman (1992). Les démonstrations du théorème et des corollaires se trouvent dans Lavallée (2001).

4. NON-RÉPONSE

Dans les enquêtes, qu'il s'agisse de recensements ou de sondages, il arrive inévitablement que l'on ait de la non-réponse. De nombreux articles et ouvrages traitent de la non-réponse. Des bibliographies imposantes, quoique certainement pas exhaustives, se trouvent par exemple dans Droesbeke et Lavallée (1996), Hedges et Olkin (1983), ainsi que Bogeström, Larsson et Lyberg (1983). Comme le sujet est vaste, nous nous restreindrons à étudier le traitement de la *non-réponse totale*, en opposition à la *non-réponse partielle*. Comme la MGPP sert à l'obtention de poids d'estimation, nous nous pencherons sur l'ajustement de ces poids pour tenir compte de la non-réponse totale.

4.1 Types de non-réponse

Avec le sondage indirect, la non-réponse totale peut donc ici être présente au sein de l'échantillon s^A tiré de U^A, ou au sein des unités identifiées pour être enquêtées au sein de U^B, c'est-à-dire des unités de collecte. Comme l'enquête des unités au sein de la population U^B se fait par grappe, on distingue deux types de non-réponse totale associés au sondage en grappes (directs ou indirects) : la *non-réponse de grappes* et la *non-réponse d'unités*. La non-réponse de grappes est celle où aucune des unités de la grappe n'a répondu à l'enquête. La non-réponse d'unités est une non-réponse totale où une ou plusieurs unités de la grappe, mais pas toutes, n'ont pas répondu. Avec le sondage indirect, on retrouve une autre forme de non-réponse qui est la *non-réponse de liens*. Ce type de non-réponse est associé à la situation où on ne peut établir si une unité j de U^A est liée à une unité ik de U^B.

4.2 Probabilités de réponse

La notion de probabilité de réponse s'avère très utile pour l'ajustement des estimations pour tenir compte de la non-réponse totale. Dans un contexte général, soit δ_k, une variable indicatrice qui prend la valeur 1 si l'unité *k* répond aux questions de l'enquête, et 0 sinon. On suppose généralement que cette variable suit un processus de Bernoulli avec probabilité ϕ_k. Autrement dit, on suppose que chaque individu *k* de la population enquêtée possède une certaine probabilité ϕ_k de répondre à l'enquête, c'est-à-dire $P(\text{unité } k \text{ répond} \mid s) = P(\delta_k = 1 \mid s) = \phi_k$. De plus, pour deux unités *k* et k', les variables indicatrices δ_k et $\delta_{k'}$ sont considérées comme indépendantes. Ceci implique que la probabilité de réponse conjointe $\phi_{kk'}$ pour ces deux unités est donnée par $\phi_{kk'} = P(\delta_k = 1, \delta_{k'} = 1 \mid s) = P(\delta_k = 1 \mid s)P(\delta_{k'} = 1 \mid s) = \phi_k \phi_{k'}$. L'indépendance entre les variables indicatrices δ de deux unités *k* et k' découle de l'hypothèse que le choix de répondre ou non d'une unité *k* n'influencera pas celui d'une autre unité k'. Finalement, on a $E(\delta_k \mid s) = P(\delta_k = 1 \mid s) = \phi_k$ et $Var(\delta_k \mid s) = \phi_k(1 - \phi_k)$.

Pour estimer les probabilités de réponse ϕ_k, on a souvent recours à un modèle. Un modèle fréquemment utilisé en pratique (Särndal, Swensson et Wretman, 1992) est le modèle uniforme à l'intérieur de *groupes de réponse homogènes* (GRH), c'est-à-dire

$$\phi_{qk} = E(\delta_{qk} \mid s) = \beta_q \tag{4.1}$$

où *q*=1,...,*Q*, où *Q* est le nombre de GRH et où β_q est un effet (ou facteur) fixe à estimer. Le paramètre β_q correspond en fait à la probabilité de réponse espérée dans le groupe *q*. Avec ce modèle, on suppose que toutes les unités d'un même GRH ont la même probabilité de répondre. Les GRH peuvent être formés par un seul facteur ou par le croisement de plusieurs.

Pour estimer ϕ_{qk}, on peut utiliser la méthode du maximum de vraisemblance pondérée (ou pseudo-maximum de vraisemblance) avec les poids correspondant à $1/\pi_k$. L'estimateur découlant du modèle (4.1) est alors donné par le taux de réponse pondéré, c'est-à-dire

$$\hat{\phi}_{qk} = R_q = \frac{\sum_{k=1}^{n_{r,q}} 1/\pi_k}{\sum_{k=1}^{n_q} 1/\pi_k} = \frac{\hat{N}_{r,q}}{\hat{N}_q} \tag{4.2}$$

où n_q est le nombre d'unités de l'échantillon appartenant au GRH *q* et $n_{r,q}$ est le nombre d'unités répondantes de ce groupe.

4.3 Traitement de la non-réponse au sein de s^A

La non-réponse au sein de l'échantillon s^A constitue un cas classique de non-réponse. Ce type de non-réponse est celui que présente en fait la plupart des livres traitant de la théorie des sondages. En effet, par le théorème de la section 3, on a vu que l'estimateur $\hat{Y}^B$ provenant de l'application de la MGPP peut s'écrire sous la forme d'un estimateur d'Horvitz-Thompson qui est fonction des unités *j* de s^A. La non-réponse au sein de s^A se traite donc comme on traiterait la non-réponse dans le contexte où on a tiré l'échantillon s^A dans le but de produire une estimation pour une quantité reliée à la population U^A.

Des m^A unités de l'échantillon s^A, on suppose qu'un sous-ensemble s_r^A de m_r^A unités ont répondu aux questions de l'enquête. Soit Ω_r^B, l'ensemble des n_r grappes identifiées par les unités $j \in s_r^A$. On enquête auprès de toutes les unités *k* des grappes $i \in \Omega_r^B$ où on mesure la variable d'intérêt *y*.

En appliquant la MGPP, nous voulons normalement attribuer un poids d'estimation w_{ik} à chaque unité *k* d'une grappe enquêtée *i*. Pour estimer le total Y^B de la population cible U^B, on peut alors se servir de l'estimateur (2.1) qui est construit en supposant aucune non-réponse au sein de l'échantillon s^A. À partir du

théorème de la section 3, on peut réécrire l'estimateur (2.1) sous la forme (3.1) qui est fonction des unités j de s^A. Comme on ne dispose que du sous-échantillon s_r^A des unités répondantes, on doit utiliser un estimateur corrigé afin de tenir compte de la non-réponse. On peut alors utiliser l'estimateur suivant :

$$\hat{Y}^{NRA,B} = \sum_{j=1}^{M^A} \frac{t_j \delta_j^A}{\pi_j^A \phi_j^A} Z_j = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \phi_j^A} \tag{4.3}$$

où ϕ_j^A est la probabilité de réponse de l'unité j. La variable indicatrice $\delta_j^A = 1$ si l'unité j de s^A est répondante, et 0 sinon. À partir du théorème de la section 3, on peut démontrer que l'estimateur (4.3) est sans biais.

Pour appliquer la MGPP avec un ajustement pour la non-réponse au sein de s^A, il suffit remplacer $1/\pi_j^A$ par $1/\pi_j^A \phi_j^A$ à l'étape 1 de la MGPP donnée à la section 2. Après avoir appliqué la MGPP ajustée pour la non-réponse, nous obtenons le poids d'estimation w_{ik}^{NRA} qui entre dans l'estimateur

$$\hat{Y}^{NRA,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} w_{ik}^{NRA} y_{ik} \tag{4.4}$$

L'estimateur $\hat{Y}^{NRA,B}$ n'est utile en pratique que si la valeur des probabilités de réponse ϕ_j^A est connue pour toutes les unités j de s_r^A. Puisque les ϕ_j^A sont vraisemblablement inconnues, on cherche alors à les estimer de manière à utiliser l'une des deux formes suivantes :

$$\hat{\hat{Y}}^{NRA,B} = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \hat{\phi}_j^A} \qquad \text{ou} \qquad \hat{\hat{Y}}^{NRA,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} \hat{w}_{ik}^{NRA} y_{ik} \tag{4.5}$$

où le poids $\hat{w}_{ik}^{NRA}$ est obtenu en remplaçant ϕ_j^A par $\hat{\phi}_j^A$. Pour obtenir $\hat{\phi}_j^A$, on peut s'inspirer du modèle (4.1) qui prend ici la forme : $\phi_{qj}^A = \beta_q^A$. L'estimateur basé sur ce modèle nous mène à l'utilisation du taux de réponse $R_q^A = (\sum_{j=1}^{m_{r,q}^A} 1/\pi_{qj}^A)/(\sum_{j=1}^{m_q^A} 1/\pi_{qj}^A)$. On obtient alors

$$\hat{\hat{Y}}^{NRA,B} = \sum_{j=1}^{m_r^A} \frac{Z_j}{\pi_j^A \hat{\phi}_j^A} = \sum_{q=1}^{Q} \frac{\sum_{j=1}^{m_{r,q}^A} Z_{qj}/\pi_{qj}^A}{\sum_{j=1}^{m_{r,q}^A} 1/\pi_{qj}^A} \left(\sum_{j=1}^{m_q^A} 1/\pi_{qj}^A \right) \tag{4.6}$$

où m_q^A est le nombre d'unités de l'échantillon s^A appartenant au GRH q et $m_{r,q}^A$ est le nombre d'unités répondantes de ce groupe. En regardant l'estimateur (4.6), on constate que ce dernier n'est autre qu'un estimateur par quotient dans le contexte d'un sondage en deux phases. Dans Särndal, Swensson et Wretman (1992), on trouve la démonstration voulant que l'estimateur (4.6) soit asymptotiquement sans biais, sous l'hypothèse du modèle (4.1). Lavallée (2001) donne la formule de variance de l'estimateur (4.6).

4.4 Traitement de la non-réponse de grappes

Pour aborder la non-réponse de grappes, on se replace dans le contexte où on sélectionne un échantillon s^A de m^A unités. Contrairement à la section 4.3, on suppose que l'ensemble des m^A unités de l'échantillon a répondu aux questions de l'enquête. Suivant le processus d'enquête, on tente d'enquêter auprès de toutes les unités k des grappes i de Ω^B. Malheureusement, pour certaines grappes entières, on ne peut obtenir de données. On est alors dans un cas de non-réponse de grappes. On suppose qu'il n'y a pas de grappe où seul un sous-ensemble non nul des unités a répondu. Soit Ω_r^B, l'ensemble des n_r grappes répondantes.

Soit δ_i^B, une variable indicatrice qui prend la valeur 1 si la grappe i répond aux questions de l'enquête, et 0 sinon. Comme à la section 4.2, on suppose que chaque grappe i de U^B possède une probabilité Φ_i^B de

répondre à l'enquête, c'est-à-dire $P(\text{grappe } i \text{ répond} \mid \Omega^B) = P(\delta_i^B = 1 \mid \Omega^B) = \Phi_i^B$. De plus, pour deux grappes i et i', les variables indicatrices δ_i^B et $\delta_{i'}^B$ sont indépendantes.

En appliquant la MGPP, nous voulons attribuer un poids d'estimation w_{ik}^{NRG} à chaque unité k d'une grappe répondante i. Pour estimer le total Y^B de la population cible U^B, on peut alors se servir de l'estimateur

$$\hat{Y}^{NRG,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} w_{ik}^{NRG} y_{ik} = \sum_{i=1}^{n} \delta_i^B \sum_{k=1}^{M_i^B} w_{ik}^{NRG} y_{ik} \tag{4.7}$$

Pour obtenir le poids w_{ik}^{NRG} à partir de la MGPP, nous utilisons la probabilité de réponse Φ_i^B de chaque grappe $i \in \Omega_r^B$. Il suffit alors de remplacer w_i par w_i / Φ_i^B à l'étape 4 de la MGPP donnée à la section 2.

En pratique, on cherche à estimer les probabilités Φ_i^B connues pour toutes les grappes i de Ω_r^B de manière à utiliser l'estimateur suivant :

$$\hat{\hat{Y}}^{NRG,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} \hat{w}_{ik}^{NRG} y_{ik} \tag{4.8}$$

où $\hat{w}_{ik}^{NRG} = w_{ik} / \hat{\Phi}_i^B$. Pour obtenir $\hat{\Phi}_i^B$, on peut s'inspirer de l'estimateur (4.2). Le modèle (4.1) prend ici la forme : $\Phi_{qi}^B = \beta_q^B$. Si on utilise ce modèle, on définit $\hat{\Phi}_{qi}^B$ de la manière suivante :

$$\hat{\Phi}_{qi}^B = R_q^B = \frac{\sum_{i=1}^{n_{r,q}} \sum_{k=1}^{M_{qi}^B} w_{qik}}{\sum_{i=1}^{n_q} \sum_{k=1}^{M_{qi}^B} w_{qik}} = \frac{\hat{M}_{r,q}^B}{\hat{M}_q^B} \tag{4.9}$$

où w_{qik} est le poids d'estimation provenant de la MGPP (en supposant aucune non-réponse) pour les unités k des grappes i appartenant au GRH q. Avec (4.9), l'estimateur $\hat{\hat{Y}}^{NRG,B}$ donné par (4.8) devient

$$\hat{\hat{Y}}^{NRG,B} = \sum_{i=1}^{n_r} \sum_{k=1}^{M_i^B} \frac{w_{ik}}{\hat{\Phi}_i^B} y_{ik} = \sum_{q=1}^{Q} \frac{\hat{M}_{r,q}^B}{\hat{M}_q^B} \sum_{i=1}^{n_q} \delta_{qi}^B \sum_{k=1}^{M_{qi}^B} w_{qik} y_{qik} \tag{4.10}$$

On peut voir l'estimateur (4.10) comme un estimateur par quotient dans le contexte d'un sondage en deux phases. Comme dans le cas de la non-réponse au sein de s^A, on peut démontrer que l'estimateur (4.10) est asymptotiquement sans biais, sous l'hypothèse du modèle (4.1). On retrouve la formule de variance de l'estimateur (4.10) dans Lavallée (2001).

4.5 Traitement de la non-réponse d'unités

Pour aborder la non-réponse d'unités, on se replace dans le contexte où on sélectionne un échantillon s^A de m^A unités. On suppose que l'ensemble des m^A unités de s^A a répondu aux questions de l'enquête. Pour chaque unité j de s^A, on identifie les unités ik de U^B qui ont $l_{j,ik}=1$. Suivant le processus d'enquête, on tente d'enquêter auprès de toutes les unités k des grappes i de Ω^B. Malheureusement, pour certaines des unités des grappes identifiées, on ne peut obtenir de données. On est alors dans un cas de non-réponse d'unités. On suppose ici qu'on possède une réponse pour au moins une unité de chaque grappe i de Ω^B. Soit $s_{r,i}^B$, l'ensemble des unités répondantes de la grappe i identifiée, et soit $M_{r,i}^B > 0$, la taille de cet ensemble.

Soit $\delta_{(i)k}^B$, une variable indicatrice qui prend la valeur 1 si l'unité k de la grappe i répond aux questions de l'enquête, et 0 sinon. On suppose que chaque unité k des grappes i de U^B possède une probabilité $\phi_{(i)k}^B$ de répondre à l'enquête, c'est-à-dire $P(\text{unité } k \in i \text{ répond} \mid \Omega^B) = P(\delta_{(i)k}^B = 1 \mid \Omega^B) = \phi_{(i)k}^B$. De plus, pour deux unités k et k' d'une grappe i, ou de deux grappes différentes, les variables indicatrices $\delta_{(i)k}^B$ et $\delta_{(i)k'}^B$ sont considérées comme indépendantes.

En appliquant la MGPP, nous voulons attribuer un poids d'estimation w_{ik}^{NRU} à chaque unité *k* répondante d'une grappe *i* de Ω^B. Pour estimer le total Y^B de la population cible U^B, on peut alors se servir de l'estimateur

$$\hat{Y}^{NRU,B} = \sum_{i=1}^{n}\sum_{k=1}^{M_{r,i}^B} w_{ik}^{NRU} y_{ik} = \sum_{i=1}^{n}\sum_{k=1}^{M_i^B} \delta_{(i)k}^B w_{ik}^{NRU} y_{ik} \tag{4.11}$$

L'obtention du poids w_{ik}^{NRU} peut se faire en établissant un parallèle avec un sondage indirect à deux degrés (Lavallée, 2001). En effet, on peut voir le processus de non-réponse d'unités comme le tirage d'un échantillon $s_{r,i}^B$ de $M_{r,i}^B$ unités obtenu à partir des M_i^B unités de chaque grappe *i* de Ω^B. En conséquence, on obtient $w_{ik}^{NRU} = w_i / \phi_{(i)k}^B$ pour tous les $k \in s_{r,i}^B$ où *i*=1,…,*n* , et où w_i est le poids d'estimation provenant de la MGPP (en supposant aucune non-réponse).

En pratique, on peut estimer les probabilités $\phi_{(i)k}^B$ de manière à utiliser l'estimateur suivant :

$$\hat{\hat{Y}}^{NRU,B} = \sum_{i=1}^{n}\sum_{k=1}^{M_{r,i}^B} \hat{w}_{ik}^{NRU} y_{ik} \tag{4.12}$$

où $\hat{w}_{ik}^{NRU} = w_i / \hat{\phi}_{(i)k}^B$. Pour obtenir $\hat{\phi}_{(i)k}^B$, on peut utiliser une approche qui consiste à considérer chaque ensemble $s_{r,i}^B$ d'unités répondantes individuellement. Les probabilités de réponse sont alors estimées à l'intérieur de chaque grappe *i*. Avec cette approche dite individuelle, le modèle (4.1) prend ici la forme : $\phi_{(qi)k}^B = \beta_{qi}^B$ et les GRH sont alors définis à l'intérieur de chaque grappe *i*. On définit $\hat{\phi}_{(qi)k}^B$ de la manière suivante :

$$\hat{\phi}_{(qi)k}^B = R_{qi}^B = \frac{M_{r,qi}^B}{M_{qi}^B} \tag{4.13}$$

où $M_{qi}^B = M_i^B$ et $M_{r,qi}^B = M_{r,i}^B$ pour $i \in q$. Avec (4.13), l'estimateur $\hat{\hat{Y}}^{NRU,B}$ donné par (4.12) devient

$$\hat{\hat{Y}}^{NRU,B} = \sum_{i=1}^{n}\sum_{k=1}^{M_{r,i}^B} \frac{w_i}{\hat{\phi}_{(i)k}^B} y_{ik} = \sum_{q=1}^{Q}\sum_{i=1}^{n_q} \frac{M_{qi}^B}{M_{r,qi}^B} \sum_{k=1}^{M_{r,qi}^B} w_{qi} y_{qik} \tag{4.14}$$

Cet estimateur n'est autre qu'un estimateur par quotient à l'intérieur de chaque UPE dans le cadre d'un sondage à deux degrés. L'estimateur $\hat{\hat{Y}}^{NRU,B}$ est asymptotiquement sans biais pour l'estimation de Y^B, sous l'hypothèse du modèle (4.1). On retrouve la formule de variance de l'estimateur (4.14) dans Lavallée (2001). En plus de l'approche individuelle, Lavallée (2001) décrit une autre approche, dite globale, qui consiste à considérer l'ensemble $s_r^B = \bigcup_{i=1}^{n} s_{r,i}^B$ des unités répondantes comme un tout.

4.6 Traitement de la non-réponse de liens

Le traitement de la non-réponse de liens s'aborde en se replaçant dans le contexte où on sélectionne un échantillon s^A de m^A unités. Suivant le processus d'enquête, pour chaque unité *j* de s^A, on identifie les unités *ik* de U^B qui ont $l_{j,ik}$=1. On suppose qu'on peut identifier **tous** les liens $l_{j,ik}$ associés à chaque unité *j* de s^A. Pour chaque unité *ik* identifiée, on suppose qu'on peut établir la liste des M_i^B unités de la grappe *i* contenant cette unité. On enquête auprès de toutes les unités *k* des grappes *i* de Ω^B. Bien qu'on puisse mesurer la variable d'intérêt *y* pour **toutes** les M_i^B unités de chaque grappe *i* de Ω^B, pour certaines unités *k*, la non-réponse de liens fait qu'on ne peut établir s'il y a un lien ou non entre ces unités *k* et une unité *j* de U^A. Autrement dit, pour **certaines** unités *k* d'une grappe $i \in \Omega^B$, il est impossible d'établir si $l_{j,ik} = 1$ ou $l_{j,ik} = 0$.

Soit $L_{r,i}^B$, le nombre total de liens qu'on a pu établir entre la grappe *i* et la population U^A. On note que $L_{r,i}^B \leq L_i^B$. De plus, parce qu'on suppose qu'on peut identifier **tous** les liens $l_{j,ik}$ associés à chaque unité *j* de s^A, on a $L_{r,i}^B > 0$ pour toutes les grappes *i* de Ω^B. En utilisant seulement le nombre total $L_{r,i}^B$ de liens établis, on produit une surestimation du total Y^B. Malheureusement, contrairement aux autres cas de non-réponse vus précédemment, il ne semble pas y avoir de solution « miracle » au problème de la non-réponse de liens. Une approche possible est de tenter de modéliser la quantité L_i^B à partir de variables auxiliaires. Avec l'estimation résultante $\hat{L}_i^B$, on peut alors construire l'estimateur

$$\hat{\hat{Y}}^{NRL,B} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} Y_i \frac{L_{j,i}}{\hat{L}_i^B} \tag{4.15}$$

où $L_{j,i} = \sum_{k=1}^{M_i^B} l_{j,ik}$. L'estimateur (4.15) découle du théorème de la section 3. On peut démontrer que l'estimateur $\hat{\hat{Y}}^{NRL,B}$ est asymptotiquement sans biais. L'aspect sans biais de cet estimateur dépend cependant de l'aspect sans biais de l'estimateur de L_i^B. En pratique, il n'est pas facile d'obtenir un estimateur sans biais de L_i^B. Ardilly et le Blanc (2000) ont été confrontés à un problème de non-réponse de liens lors de l'utilisation de la MGPP pour la pondération d'une enquête auprès de personnes sans domicile. Ils ont alors suggéré l'utilisation d'une hypothèse de régularité visant à imputer certains liens $l_{j,ik}$ à 1, ce qui revient, en fait, à modéliser la quantité L_i^B.

Une autre solution possible pour corriger le problème de surestimation est d'effectuer un calage sur marges. Bien qu'il offre une solution intéressante à la non-réponse de liens, il dépend toutefois de la disponibilité de variables auxiliaires $\mathbf{x}_{ik}^B$ corrélées avec la variable d'intérêt y_{ik}, ce qui n'est pas toujours le cas en pratique. La meilleure solution restera toujours celle de mesurer exactement L_i^B, ou sinon d'obtenir une estimation $\hat{L}_i^B$ le plus près possible de la quantité L_i^B.

BIBLIOGRAPHIE

Ardilly, P., le Blanc, P. (2000). "*Comment pondérer une enquête auprès des personnes sans domicile?*", article présenté au Deuxième Colloque Francophone sur les Sondages, Université Libre de Bruxelles, Belgique, 22-23 juin 2000.

Bogeström, B., Larsson, M., Lyberg, L. (1983). "Bibliography on Nonresponse and Related Topics", in *Incomplete Data in Sample Surveys* (Madow, W.G., Olkin, I., Rubin, D.B., Éditeurs), Vol. 2, Academic Press, New York, pp. 479-567.

Droesbeke, J.-J., Lavallée, P. (1996). "La non-réponse dans les enquêtes", *Methodologica*, No. 4, pp. 1-39.

Ernst, L. (1989). "Weighting issues for longitudinal household and family estimates", in *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Éditeurs), John Wiley and Sons, New York, pp. 135-159.

Hedges, L.V., Olkin, I. (1983). "Selected Annotated Bibliography", in *Incomplete Data in Sample Surveys* (Madow, W.G., Olkin, I., Rubin, D.B., Éditeurs), Vol. 2, Academic Press, New York, pp. 417-478.

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*, Vol. 21, No. 1, pp. 27-35.

Lavallée, P. (2001). "La Méthode généralisée du partage des poids (ou le Sondage indirect)", thèse de doctorat présentée à l'Université Libre de Bruxelles, Belgique, avril 2000.

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