Estimation et calcul de précision pour des échantillons rotatifs non chevauchants

Pierre LAVALLÉE

Statistique Canada, Division des méthodes d'enquêtes sociales

Introduction

Le sondage rotatif consiste à remplacer une partie de l'échantillon à chaque vague de consultation. Si la rotation est de 0%, on est dans le cas des panels. Avec une rotation de 100%, on renouvelle complètement l'échantillon à chaque vague de consultation; ce qui produit des échantillons rotatifs non chevauchants (RNC). On parle aussi dans ce cas d'échantillons distincts ou disjoints, d'échantillons avec renouvellement complet ou, en anglais, de *rolling samples*.

Kish (1981) a proposé l'utilisation d'échantillons RNC notamment dans le cadre des recensements. Il s'agit alors d'un recensement par étape qui consiste en P échantillons RNC qui sont conçus de sorte qu'en faisant la somme des données des P échantillons, on obtient un dénombrement complet de la population visée. Un des avantages de ce type de recensement est qu'il permet d'étaler les coûts du recensement au cours du temps, contrairement au recensement classique (par exemple, décennal) où les coûts sont concentrés sur les quelques années entourant l'année du recensement. Le recensement par étape a été décrit dans plusieurs autres articles comme celui de Kish (1990). Plus particulièrement, la France a opté pour ce type de plan de sondage pour son recensement de la population. La méthodologie du recensement a été décrite dans plusieurs articles, notamment de Durr et Dumais (2002), ainsi que de Bertrand, Chauvet, Christian et Grosbras (2004).

Les échantillons RNC sont aussi utilisés lorsqu'on cherche à obtenir des estimations obtenues par cumul. Par exemple, pour l'Enquête sur le camionnage pour compte d'autrui (ECCA) de Statistique Canada, on tire un échantillon à chaque année et un quart de cet échantillon est enquêté à chaque trimestre. Chaque échantillon trimestriel constitue alors un échantillon RNC et on obtient les estimations annuelles en sommant simplement les estimations trimestrielles. Pour plus de détail sur l'ECCA, on peut consulter Dumais et Lavallée (1990). Un autre exemple est celui de l'Enquête sur l'emploi, la rémunération et les heures (EERH) de Statistique Canada où on étudie actuellement la possibilité de produire des estimations trimestrielles sur la base d'une collecte mensuelle. Dans ce cas, chaque échantillon trimestriel serait divisé en trois de manière à enquêter auprès d'un tiers de l'échantillon mensuellement. Les estimations trimestrielles seraient alors produites en prenant la moyenne des estimations mensuelles.

Une dernière application des échantillons RNC est celle où on désire comparer deux ou plusieurs traitements. En assignant à chaque échantillon un traitement différent, on peut alors comparer l'effet des traitements. On remarque que parce que les échantillons sont non chevauchants, chaque unité se voit recevoir un et un seul traitement. Un exemple de ce type d'application est celui du choix du questionnaire de deux ou de six pages pour l'Enquête annuelle sur les heures et les gains (*Annual Survey of Hours and Earnings*, ASHE) du Royaume-Uni (voir Hidiroglou et Lavallée, 2005).

Il est à noter que la comparaison de traitements entre directement en conflit avec l'obtention d'estimations (ou de moyennes) par cumul en ce qui a trait à la précision des estimations désirées. En effet, les échantillons RNC permettent, entre autres, de produire des estimations de totaux cumulés (ou de moyennes cumulées) avec une plus grande précision qu'avec un panel ou des échantillons indépendants. On constate que les échantillons RNC sont corrélés entre eux et que cette corrélation est souvent négative; ce qui diminue la variance pour les estimations de totaux (ou moyennes) sur plusieurs vagues de consultation. Malheureusement, dans le cas des comparaisons, on parle d'estimations de différences et les covariances négatives ont alors l'effet d'augmenter la variance des estimations. Pour cette raison, le présent article se concentrera surtout sur le calcul d'estimations de totaux (ou de moyennes) au lieu de différences.

Le principal problème abordé dans cet article est l'estimation des covariances provenant de l'utilisation d'échantillons RNC. Ces covariances sont difficiles à estimer en pratique parce qu'une estimation sans biais requiert l'utilisation de données présentes sur au moins deux vagues, ce qui, par construction, n'est pas réalisable. On présentera tout d'abord différentes estimations pertinentes pour le sondage rotatif. Par la suite, on présentera une solution basée sur l'utilisation de l'imputation pour l'estimation des covariances (ou corrélations). On discutera finalement de l'application de la solution au recensement français.

1. Plan de sondage et estimation

Soit U, une population constituée de N unités pouvant être observée à un temps donné. Cette population est destinée à être enquêtée à plus d'une occasion et on supposera qu'elle est constante dans le temps, ou du moins durant une période totale T. Notons que si l'on pend comme exemple le recensement français, cette hypothèse est tout à fait valable étant donnée que l'unité d'échantillonnage est géographique et que le nombre total d'unités géographiques ne change pas au cours du temps. Si l'on prend l'ECCA, considérer la population comme étant constante pose problème, mais ce dernier est contourné en définissant la population comme étant l'ensemble des entreprises de transport présentes au début de l'année. Si le nombre d'unité peut varier au cours du temps, il est alors utile de définir la population U comme l'union de toutes les populations U_t des périodes t couvertes durent l'étude, c'est-à-dire $U = \bigcup_{t \in T} U_t$ (voir Deville, 2000).

De la population U, on tire un échantillon s de taille n à partir d'un plan de sondage $p(s)$ quelconque où $p(s) > 0$. Soit $\pi_k > 0$, la probabilité de sélection de l'unité k de s. L'échantillon s est alors divisé en T sous-échantillons s_t de tailles n_t, où $s = \bigcup_{t=1}^{T} s_t$ et $s_t \cap s_{t'} = \varnothing$ pour $t \neq t'$. En supposant que les T sous-échantillons s_t sont de tailles égales, la probabilité de sélection π_{tk} de chaque unité k de s_t est alors $\pi_{tk} = \pi_k / T$. Pour chaque unité k de chaque sous-échantillon s_t, on mesure une variable d'intérêt y_{tk}.

On peut voir la division en T parties comme la mesure de la variable d'intérêt y_{tk} à différents temps t, et pour des unités différentes k. C'est le cas du recensement français et de l'ECCA. Rappelons que l'on peut aussi voir les T parties comme la mesure de l'effet de T traitements sur une variable donnée, comme pour le choix du questionnaire pour l'ASHE. Notons qu'une autre

façon de procéder pour obtenir les sous-échantillons s_t serait de diviser la population aléatoirement en *P* panels (*P*>*T*) et d'utiliser *p* panels différents pour chaque échantillon s_t. Une telle procédure est décrite dans Hidiroglou, Choudhry et Lavallée (1991).

On peut s'intéresser à l'estimation des quantités suivantes :

- $Y_t = \sum_{k=1}^{N} y_{tk}$: total au temps *t*;
- $\tau = \sum_{t=1}^{T} Y_t$: total cumulé pour l'ensemble de la période *T*;
- $\bar{\tau} = \sum_{t=1}^{T} Y_t / T$: moyenne sur la période *T*;
- $\Delta_{t,t'} = Y_t - Y_{t'}$: différence entre les totaux des temps t et t'. Rappelons cependant que ce paramètre est d'intérêt secondaire dans le présent article.

Il importe ici de préciser le sens des quantités τ et $\bar{\tau}$. La quantité τ représente le total cumulé de la variable d'intérêt *y* au cours du temps. Dans le contexte de l'ECCA, on parle ici, par exemple, du revenu annuel obtenu par la somme des quatre revenus trimestriels. Dans le contexte du recensement français, une telle statistique n'a pas d'intérêt proprement dit parce que le total de la population de la France pour cinq années consécutives n'a pas de sens en soi. Dans le contexte du recensement français, on s'intéresse alors plus à la quantité $\bar{\tau}$ qui représente la population moyenne de la France pour cinq années consécutives.

Soit $w_{tk} = 1/\pi_{tk}$, le poids de sondage associé à l'unité *k* du sous-échantillon s_t. On a $\sum_{k=1}^{n_t} w_{tk} \approx N$, et ainsi, chaque sous-échantillon s_t nous ramène à la population *U*. On peut alors estimer Y_t en utilisant l'estimateur $\hat{Y}_t = \sum_{k=1}^{n_t} w_{tk} y_{tk}$. En disposant des estimations $\hat{Y}_t$ obtenues à partir de chaque sous-échantillon s_t, on a $\hat{\tau} = \sum_{t=1}^{T} \hat{Y}_t = \sum_{t=1}^{T} \sum_{k=1}^{n_t} w_{tk} y_{tk}$ et

$$Var(\hat{\tau}) = \sum_{t=1}^{T} Var(\hat{Y}_t) + \sum_{t=1}^{T} \sum_{t' \neq t} Cov(\hat{Y}_t, \hat{Y}_{t'}) \qquad (1)$$

L'estimation des variances $Var(\hat{Y}_t)$ s'obtient à partir des estimateurs habituels reliés au plan de sondage $p(s)$ utilisé pour la sélection de l'échantillon *s*. Comme mentionné plus tôt, l'estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ s'avère plus problématique parce qu'il n'y a pas d'unités communes entre les échantillons s_t et $s_{t'}$.

Dans la littérature, l'estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ a été, pour ainsi dire, ignorée. Une des raisons pour cela est que, parce que les échantillons sont disjoints, on considère souvent ces covariances comme étant nulles (voir Kish, 1965, et Kish, 1999). Ceci n'est malheureusement pas le cas en général puisque, bien que les sous-échantillons s_t soient disjoints, ils ne sont pas indépendants. En effet, si une unité *k* se retrouve dans s_t, elle ne peut se retrouver dans $s_{t'}$ avec $t \neq t'$. Donc, les covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ ne sont pas nulles et, de plus, elles sont souvent négatives. C'est le cas notamment dans le contexte du sondage aléatoire simple sans remise (SASSR) comme l'a démontré Tam (1984).

Si on ignore les covariances, on obtient une surestimation de la variance (1), ce qui nous permet d'être conservateur dans les tests d'hypothèses. Notons que dans le cas de l'estimation de la différence $\Delta_{t,t'}$, cette surévaluation de la variance se transforme en sous-évaluation, ce qui est un problème plus sérieux. Dans certaines enquêtes, les fractions des sondages sont suffisamment élevées pour que la surestimation de la variance (1) pose problème. C'est le cas notamment pour l'ECCA où certaines strates sont pratiquement à tirage complet. Notons que ceci est aussi particulièrement vrai pour le recensement français où l'échantillon *s* correspond en fait à la population *U*, et où chaque sous-échantillon s_t correspond, *grosso modo*, à un cinquième de la population *U*. Ainsi, ignorer les covariances pour le recensement français mène certainement à une surévaluation de la précision des estimations.

Dans le contexte des échantillons RNC, certains auteurs proposent de simplement ignorer le fait que les variables y_{tk} sont mesurées à différents temps *t*. Dans ce cas, le fait que l'échantillon *s* ait été divisé aléatoirement en s_t, *t*=1,...,*T*, n'est plus tenu en compte et c'est alors comme si on s'intéressait à l'estimation de $\bar{\tau}$ plutôt que τ. L'estimateur de $\bar{\tau}$ est, dans ce cas, donné par $\hat{\bar{\tau}}_{comb} = \sum_{k=1}^{n} w_k y_k$ où $w_k = 1/\pi_k$ et $y_k = y_{tk}$ pour $k \in s_t$. La variance de $\hat{\bar{\tau}}_{comb}$ est simplement la variance habituelle reliée à l'utilisation du plan de sondage $p(s)$ pour la sélection de l'échantillon *s*. Notons que cette variance est alors sous-évaluée puisqu'elle ne tient pas compte du fait que la variable *y* a été mesurée différemment selon son appartenance aux sous-échantillons s_t. Ardilly (2004) a suivi cette approche afin d'isoler les covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ dans le cadre du calcul de précision des variables structurelles dans l'enquête « emploi » de l'INSEE. Dans le cas traité par Ardilly (2004), on a *T*=2. Une fois qu'on obtient l'estimation $\hat{Var}(\hat{\bar{\tau}}_{comb})$ selon cette approche, on procède alors simplement à partir de

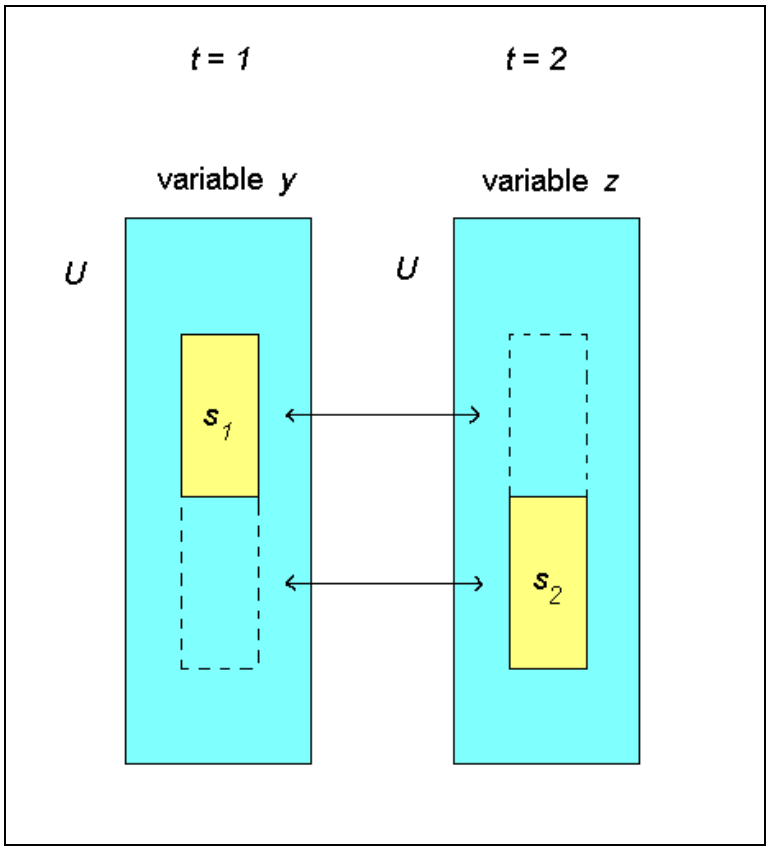
$$\hat{\hat{C}}ov(\hat{Y}_t, \hat{Y}_{t'}) = \frac{1}{2}\left[4 \times \hat{V}ar(\hat{\bar{\tau}}_{comb}) - \hat{V}ar(\hat{Y}_t) - \hat{V}ar(\hat{Y}_{t'})\right] \qquad (2)$$

où $\hat{V}ar(\hat{Y}_t)$ et $\hat{V}ar(\hat{Y}_{t'})$ sont deux estimateurs habituels des variances $Var(\hat{Y}_t)$ et $Var(\hat{Y}_{t'})$. Il est à noter que parce que $\hat{V}ar(\hat{\bar{\tau}}_{comb})$ sous-estime la variance $Var(\hat{\bar{\tau}})$, la covariance estimée $\hat{\hat{C}}ov(\hat{Y}_t, \hat{Y}_{t'})$ peut s'avérer négative, ce qui est tout à fait possible, compte tenu que les sous-échantillons s_t et $s_{t'}$ sont disjoints.

2. Une solution basée sur l'utilisation de l'imputation

Nous présentons dans cette section une autre approche pour l'estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ de l'équation (1). Afin de simplifier la discussion, nous supposerons que la sélection de l'échantillon *s* est faite par SASSR. Notons qu'ainsi, chaque sous-échantillon s_t correspond aussi à un SASSR. De plus, nous supposerons que *T*=2, c'est-à-dire que l'échantillon *s* n'est divisé qu'en deux sous-échantillons s_1 et s_2 de taille n_1 et n_2, respectivement. Afin de simplifier la notation, nous utiliserons la lettre *y* pour la mesure de la variable d'intérêt (y_{1k}) au temps *t*=1, et la lettre *z* pour la mesure de la variable d'intérêt (y_{2k}) au temps *t*=2. Ceci est illustré à la figure1.

Figure 1. Sélection de s_1 et s_2



Avec T=2, dans le contexte du SASSR, la variance (1) devient

$$\begin{aligned} Var(\hat{\tau}) &= Var(\hat{Y}) + Var(\hat{Z}) + 2Cov(\hat{Y},\hat{Z}) \\ &= N^2\left(1-\frac{n_1}{N}\right)\frac{S_y^2}{n_1} + N^2\left(1-\frac{n_2}{N}\right)\frac{S_z^2}{n_2} - 2N\,S_{yz} \end{aligned} \qquad (3)$$

où $S_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})^2$, $S_z^2 = \frac{1}{N-1}\sum_{k=1}^{N}(z_k - \bar{Z})^2$, et $S_{yz} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})(z_k - \bar{Z})$.

Notons que la covariance $Cov(\hat{Y},\hat{Z}) = -N\,S_{yz}$ a été obtenue à partir des résultats de Tam (1984).

Pour estimer S_{yz}, on doit disposer d'unités communes entre les sous-échantillons s_1 et s_2, ce qui n'est pas le cas puisqu'on ne dispose pas de valeurs de y pour $k \in s_2$, ni de valeurs de z pour $k \in s_1$. On propose donc d'imputer les données « manquantes » aux deux sous-échantillons. Soit $\hat{y}_k$, la valeur imputée de y pour $k \in s_2$, et soit $\hat{z}_k$, la valeur imputée de z pour $k \in s_1$. On définit ainsi

$$y_k^* = \begin{cases} y_k & \text{pour } k \in s_1 \\ \hat{y}_k & \text{pour } k \in s_2 \end{cases} \qquad \text{et} \qquad z_k^* = \begin{cases} \hat{z}_k & \text{pour } k \in s_1 \\ z_k & \text{pour } k \in s_2 \end{cases}$$

À partir des valeurs imputées, on estime S_{yz} de la manière suivante :

$$\begin{aligned}\hat{S}_{yz} &= \frac{1}{n-1}\sum_{k=1}^{n}(y_k^* - \bar{y}^*)(z_k^* - \bar{z}^*) \\ &= \frac{1}{n-1}\left[\sum_{k=1}^{n_1}(y_k - \bar{y}^*)(\hat{z}_k - \bar{z}^*) + \sum_{k=1}^{n_2}(\hat{y}_k - \bar{y}^*)(z_k - \bar{z}^*)\right]\end{aligned} \tag{4}$$

où $\bar{y}^* = \sum_{k=1}^{n} y_k^* / n$ et $\bar{z}^* = \sum_{k=1}^{n} z_k^* / n$. Un estimateur de la variance (3) est ainsi donné par

$$\begin{aligned}\hat{Var}(\hat{\tau}) &= \hat{Var}(\hat{Y}) + \hat{Var}(\hat{Z}) + 2\hat{Cov}(\hat{Y},\hat{Z}) \\ &= N^2\left(1-\frac{n_1}{N}\right)\frac{s_y^2}{n_1} + N^2\left(1-\frac{n_2}{N}\right)\frac{s_z^2}{n_2} - 2N\hat{S}_{yz}\end{aligned} \tag{5}$$

où $s_y^2 = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(y_k - \bar{y})^2$, $\bar{y} = \frac{1}{n_1}\sum_{k=1}^{n_1} y_k$, $s_z^2 = \frac{1}{n_2-1}\sum_{k=1}^{n_2}(z_k - \bar{z})^2$ et $\bar{z} = \frac{1}{n_2}\sum_{k=1}^{n_2} z_k$.

On peut chercher à mesurer la performance de l'estimateur (4) en calculant l'espérance de cet estimateur en supposant qu'on utilise le « bon » modèle d'imputation. De façon générale, on peut supposer que le modèle utilisé pour imputer les *y* et les *z* « manquants » mène à une imputation sans biais, c'est-à-dire $E_\xi(\hat{y}_k) = E_\xi(y_k) = \mu_{yk}$ et $E_\xi(\hat{z}_k) = E_\xi(z_k) = \mu_{zk}$. On suppose aussi que, en général, $Var_\xi(\hat{y}_k) = \sigma_{yk}^2$, $Cov_\xi(\hat{y}_k, \hat{y}_{k'}) = \sigma_{ykk'}$, $Var_\xi(\hat{z}_k) = \sigma_{zk}^2$, $Cov_\xi(\hat{z}_k, \hat{z}_{k'}) = \sigma_{zkk'}$ et $Cov_\xi(\hat{y}_k, \hat{z}_{k'}) = \sigma_{yk,zk'}$ sont non nulles. En prenant l'espérance par rapport au modèle, on obtient

$$E_\xi(\hat{S}_{yz}) = E_\xi(\tilde{S}_{yz}) \tag{6}$$

où $\tilde{S}_{yz} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \tilde{y})(z_k - \tilde{z})$ avec $\tilde{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$ et $\tilde{z} = \frac{1}{n}\sum_{k=1}^{n} z_k$. Ainsi, on obtient que si $E_\xi(\hat{y}_k) = E_\xi(y_k)$ et $E_\xi(\hat{z}_k) = E_\xi(z_k)$, l'estimateur $\hat{Cov}(\hat{Y},\hat{Z}) = -N\hat{S}_{yz}$ est sans biais.

Puisque l'on procède à l'estimation de S_{yz} à partir de valeurs imputées, on peut penser étendre cette méthode à l'estimation de S_y^2 et de S_z^2. On définit alors

$$\begin{aligned}\hat{S}_y^2 &= \frac{1}{n-1}\sum_{k=1}^{n}(y_k^* - \bar{y}^*)^2 \\ &= \frac{1}{n-1}\left[\sum_{k=1}^{n_1}(y_k - \bar{y}^*)^2 + \sum_{k=1}^{n_2}(\hat{y}_k - \bar{y}^*)^2\right]\end{aligned} \tag{7}$$

et de même,

$$\hat{S}_z^2 = \frac{1}{n-1}\sum_{k=1}^{n}(z_k^* - \bar{z}^*)^2 \tag{8}$$

Pour estimer $Var(\hat{\tau})$, on peut alors utiliser

$$\begin{aligned}\tilde{Var}(\hat{\tau}) &= \tilde{Var}(\hat{Y}) + \tilde{Var}(\hat{Z}) + 2\hat{Cov}(\hat{Y},\hat{Z}) \\ &= N^2\left(1-\frac{n_1}{N}\right)\frac{\hat{S}_y^2}{n_1} + N^2\left(1-\frac{n_2}{N}\right)\frac{\hat{S}_z^2}{n_2} - 2N\hat{S}_{yz}\end{aligned} \tag{9}$$

On peut de nouveau chercher à mesurer la performance de ce nouvel estimateur (9) en calculant l'espérance en supposant qu'on utilise le « bon » modèle d'imputation. En prenant l'espérance par rapport au modèle, on obtient

$$E_\xi(\hat{S}_y^2) = E_\xi(\tilde{S}_y^2) \tag{10}$$

et

$$E_\xi(\hat{S}_z^2) = E_\xi(\tilde{S}_z^2) \tag{11}$$

où $\tilde{S}_y^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \tilde{y})^2$ et $\tilde{S}_z^2 = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \tilde{z})^2$. Compte tenu des résultats (6), (10) et (11), l'estimateur de variance (9) est donc sans biais si $E_\xi(\hat{y}_k) = E_\xi(y_k)$ et $E_\xi(\hat{z}_k) = E_\xi(z_k)$. En supposant que l'on utilise le « bon » modèle d'imputation, les deux estimateurs (5) et (9) sont donc sans biais pour l'estimation de la variance $Var(\hat{\tau})$.

3. Différentes méthodes d'imputation

Dans cette section, nous vérifions l'effet d'utiliser différentes méthodes d'imputation pour obtenir les valeurs « manquantes » nécessaires à l'estimation de la covariance $Cov(\hat{Y}, \hat{Z})$. On rappelle qu'en général, une méthode d'imputation est choisie de sorte que les valeurs imputées soient le plus près possible des vraies valeurs.

3.1 Imputation par la moyenne

Supposons que les données « manquantes » sont imputées par la moyenne. Dans ce cas, $\hat{y}_k = \bar{y}$ pour $k \in s_2$, et $\hat{z}_k = \bar{z}$ pour $k \in s_1$. En faisant le calcul à partir de (4), on obtient directement $\hat{S}_{yz} = 0$ et donc, $\hat{Cov}(\hat{Y}, \hat{Z}) = 0$. Ce résultat s'explique facilement par le fait que l'imputation des données de chaque sous-échantillon ne considère que les variables d'intérêt de ce sous-échantillon. Pour espérer avoir $\hat{Cov}(\hat{Y}, \hat{Z}) \neq 0$, il faut au minimum que l'obtention des $\hat{y}_k$ de $k \in s_2$ utilisent des variables auxiliaires mesurées pour $k \in s_2$ (par exemple, les z_k), et vice versa pour l'obtention des $\hat{z}_k$.

3.2 Imputation historique

En utilisant l'imputation historique, on a $\hat{y}_k = z_k$ pour $k \in s_2$, et $\hat{z}_k = y_k$ pour $k \in s_1$. On obtient alors $\bar{y}^* = \bar{z}^* = \frac{n_1\bar{y} + n_2\bar{z}}{n}$, et $\hat{S}_{yz} = \frac{1}{(n-1)}\left[(n_1 - 1)s_y^2 + (n_2 - 1)s_z^2 + \frac{n_1 n_2}{n}(\bar{y} - \bar{z})^2\right]$. Finalement,

$$\hat{Cov}(\hat{Y}, \hat{Z}) = -\frac{N}{(n-1)}\left[(n_1 - 1)s_y^2 + (n_2 - 1)s_z^2 + \frac{n_1 n_2}{n}(\bar{y} - \bar{z})^2\right] \tag{12}$$

On voit ici qu'avec l'imputation historique, la covariance $Cov(\hat{Y}, \hat{Z})$ est alors estimée à partir d'une moyenne pondérée des quantités s_y^2 et s_z^2, en plus d'un terme qui devient nul si les moyennes des *y* et des *z* s'avèrent identiques.

Il est intéressant de voir que si on utilise l'estimateur de variance (9), on peut démontrer qu'ici $\hat{S}_{yz} = \hat{S}_y^2 = \hat{S}_z^2 = \hat{S}^2$. Ainsi, en supposant que $n_1 = n_2 = n/2$, on obtient $\tilde{V}ar(\hat{\tau}) = 4N(N-n)\hat{S}^2/n$. Rappelons maintenant que dans le contexte des échantillons RNC, certains auteurs proposent de simplement ignorer le fait que les variables y_{tk} sont mesurées à différents temps *t*. L'estimateur de $\bar{\tau}$ est alors $\hat{\bar{\tau}}_{comb} = \sum_{k=1}^{n} w_k y_k$ où $w_k = 1/\pi_k$ et $y_k = y_{tk}$ pour $k \in s_t$. Puisque *T*=2, on a $\bar{\tau} = \tau/2$ et on obtient finalement $\hat{V}ar(\hat{\bar{\tau}}_{comb}) = \tilde{V}ar(\hat{\tau})$. Il s'avère donc qu'en utilisant l'imputation historique avec l'estimateur de variance (9), on se retrouve en pratique dans la situation où on ignore que les variables d'intérêt sont mesurées à différents temps *t*.

3.3 Imputation historique avec tendance

L'imputation historique avec tendance est couramment utilisée au niveau des enquêtes économiques. Avec cette méthode d'imputation, $\hat{y}_k = z_k \bar{y}/\bar{z}$ pour $k \in s_2$, et $\hat{z}_k = y_k \bar{z}/\bar{y}$ pour $k \in s_1$. Notons qu'en général, les tendances $\bar{y}/\bar{z}$ et $\bar{z}/\bar{y}$ sont calculées à l'intérieur de classes d'imputation. En remplaçant dans (4), on obtient au départ $\bar{y}^* = \bar{y}$ et $\bar{z}^* = \bar{z}$, et finalement,

$$\hat{C}ov(\hat{Y},\hat{Z}) = -\frac{N}{(n-1)}\left[(n_1-1)\frac{\bar{z}}{\bar{y}}s_y^2 + (n_2-1)\frac{\bar{y}}{\bar{z}}s_z^2\right] \qquad (13)$$

On constate ici qu'avec l'imputation historique avec tendance, la covariance $Cov(\hat{Y},\hat{Z})$ est alors estimée à partir d'une moyenne pondérée des quantités s_y^2 et s_z^2. Cette méthode d'imputation a notamment été employée pour l'ECCA (voir Dumais et Lavallée, 1990).

4. Le recensement de la population française

Dans la présente section, nous présentons la théorie décrite précédemment dans le cadre du recensement de la France. Il est à noter que nous ne prétendons pas pouvoir entrer dans les détails de ce projet. Nous nous contentons seulement de fournir des pistes qui pourraient s'avérer utiles pour le calcul de la précision des estimations issues du recensement. Pour des détails sur le recensement français, on peut consulter Bertrand, Chauvet, Christian et Grosbras (2004), ainsi que Durr et Dumais (2002).

Dans le cadre du recensement français, la population *U* est tout d'abord divisée en deux groupes: les communes de moins de 10 000 habitants, et les communes d'au moins 10 000 habitants. Pour le premier groupe, on effectue une sélection de régions géographiques appelées groupes de rotation. Pour le deuxième groupe, on sélectionne des adresses tirées du « répertoire d'immeubles localisés ». Bien que les unités de sélection soient différentes dans les deux groupes, on utilise ici tout de même l'indice *k* pour l'identification des unités.

Au niveau des petites communes (moins de 10 000 habitants), on compte couvrir en cinq ans l'ensemble du territoire français. La couverture complète des grosses communes (au moins 10 000 habitants) se fera sur une plus longue période, soit sept ans. Pour le présent article, nous supposerons toutefois que la couverture du territoire français sera fera sur cinq ans afin de ne pas compliquer inutilement la discussion. Pour le recensement français, on suppose ainsi que la population *U* est ainsi divisée en *T*=5 échantillons RNC s_t, où $U = \bigcup_{t=1}^{5} s_t$ et $s_t \cap s_{t'} = \varnothing$ pour $t \neq t'$.

Les échantillons s_t sont obtenus à partir d'un plan de sondage équilibré de manière à obtenir une certaine représentativité de la population (voir Deville et Tillé, 2000). Les données utilisées pour l'équilibrage proviennent des données du recensement de 1999. La probabilité de sélection π_{tk} de chaque unité k de s_t est alors approximativement de 1/5. Pour chaque unité k de chaque échantillon RNC s_t, on mesure une variable d'intérêt y_{tk}.

Il y a principalement deux statistiques d'intérêt pour le recensement français. Il s'agit du total $Y_t = \sum_{k=1}^{N} y_{tk}$ de la population au temps t, ainsi que la moyenne $\bar{\tau} = \sum_{t=1}^{5} Y_t / 5$ de la population sur la période de cinq ans. La moyenne $\bar{\tau}$ se veut en fait une mesure de la population française à l'année médiane T=3.

Soit $w_{tk} = 1/\pi_{tk}$, le poids de sondage associé à l'unité k du sous-échantillon s_t. Rappelons que l'on a $\sum_{k=1}^{n_t} w_{tk} \approx N$. On estime Y_t en utilisant l'estimateur $\hat{Y}_t = \sum_{k=1}^{n_t} w_{tk} y_{tk}$. Notons que parce qu'on ne dispose que d'un échantillon s_t de 1/5 de la population, l'estimation du total Y_t n'est destiné qu'à des estimations globales, c'est-à-dire nationales et régionales. En disposant des estimations $\hat{Y}_t$ obtenues à partir de chaque échantillon s_t, on a $\hat{\bar{\tau}} = \sum_{t=1}^{5} \hat{Y}_t / 5$ et

$$Var(\hat{\bar{\tau}}) = \frac{1}{25}\left[\sum_{t=1}^{5} Var(\hat{Y}_t) + \sum_{t=1}^{5}\sum_{t' \neq t} Cov(\hat{Y}_t, \hat{Y}_{t'})\right] \qquad (14)$$

parce qu'on dispose des cinq échantillons s_t, t=1,…,5, l'estimation de la moyenne $\bar{\tau}$ peut servir à produire des statistiques détaillées, c'est-à-dire au niveau communal.

L'estimation des variances $Var(\hat{Y}_t)$ s'obtient à partir des estimateurs habituels reliés au plan de sondage utilisé pour la sélection des échantillons s_t. Comme mentionné plus tôt, l'estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ pose beaucoup plus de problèmes parce qu'il n'y a pas d'unités communes entre les échantillons RHC s_t et $s_{t'}$. On propose donc d'imputer les données « manquantes » aux différents échantillons s_t. Supposons que la variable y_{tk} a été mesurée au temps t. Soit $\hat{y}_{t'k}$, la valeur imputée de y pour $k \in s_{t'}$, $t' \neq t$. On définit ainsi $y^*_{tk} = y_{tk}$ pour $k \in s_t$, et $y^*_{t'k} = \hat{y}_{t'k}$ pour $k \in s_{t'}$, $t' \neq t$. Comme on l'a vu plus tôt, on utilise alors les variables y^*_{tk} et $y^*_{t'k}$ dans l'estimateur habituel de la covariance $Cov(\hat{Y}_t, \hat{Y}_{t'})$.

Dans le cadre d'un recensement de la population, l'imputation historique avec tendance semble appropriée pour l'imputation des variables y_{tk} « manquantes ». La tendance peut alors être obtenue à partir des données déjà collectées, ou sinon à partir de projections démographiques provenant d'une source externe comme, par exemple, les recensements précédents ou les données fiscales. Soit $\beta_{t't}$, le ratio démographique entre les temps t' et t. On a ainsi $\hat{y}_{t'k} = \beta_{t't} y_{tk}$ pour $k \in s_{t'}$, $t' \neq t$. Notons qu'en général, les ratios démographiques $\beta_{t't}$ sont calculés à l'intérieur de classes d'imputation.

L'utilisation de valeurs imputées pourra être utile afin de calculer la précision du recensement de la population française, compte tenu que ce dernier a été étalé sur une période de cinq ans et plus, au lieu d'un recensement ponctuel. Notons que si on utilise l'approche visant à ignorer que les données ont été recueillies selon des échantillons s_t à des temps différents, on obtient alors une variance $Var(\hat{\bar{\tau}})$ nulle, ce qui donne alors une surévaluation de la précision du recensement.

Remerciements

L'auteur tient à remercier Chantal Grondin pour avoir fait resurgir ce problème d'estimation de covariances dans le cadre du remaniement de l'EERH. L'auteur veut aussi grandement remercier Carl Särndal pour avoir, il y a quelques années, donné l'idée d'imputer les données « manquantes » des différents sous-échantillons RNC de l'ECCA pour obtenir une estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$. Une idée qui a fait du chemin... Finalement, l'auteur se doit de remercier Jean Dumais, David Haziza et Yves Morin pour leurs suggestions pour l'amélioration de cet article.

Bibliographie

[1] Bertrand, Ph., Chauvet, G., Christian, B., Grosbras, J.-M. (2004), « Les plans de sondages du recensement rénové de la population », In *Échantillonnage et méthodes d'enquêtes* (Ardilly, éditeur), Dunod, pp. 307-313.

[2] Deville, J.-C. (2000), « Échantillonnage et estimation pour les enquêtes continues : que mesure-t-on ? », Actes des Journées de méthodologie statistique, 4-5 décembre 2000, Tome 1, *INSEE Méthodes*, No. 100.

[3] Hidiroglou, M.A., Choudhry, G.H., Lavallée, P. (1991), « Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises », *Techniques d'enquêtes*, Vol. 17, No. 2, pp. 211-227.

[4] Deville, J.-C., Tillé, Y. (2000), « Échantillonnage équilibré par la méthode du cube et estimation de variance », Actes des Journées de méthodologie statistique, *INSEE Méthodes*, No. 100, décembre 2000.

[5] Durr, J.-M., Dumais, J. (2002), « La rénovation du recensement français », *Techniques d'enquête*, Vol. 28, No. 1, pp. 47-53.

[6] Dumais, J., Lavallée, P. (1990), « Une approche pour l'échantillonnage et l'estimation d'enquêtes trimestrielles : L'Enquête sur le camionnage pour compte d'autrui », *Recueil des textes des présentation du Colloque sur les méthodes et domaines d'application de la statistique*, Bureau de la Statistique du Québec, pp. 51-57.

[7] Hidiroglou, M.A., Choudhry, G.H., Lavallée, P. (1991), « Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises », *Techniques d'enquête*, Vol. 17, No. 2, pp. 211-227.

[8] Hidiroglou, M., Lavallée, P. (2005), « *Assignation aléatoire de questionnaires pilotes en présence de grappes* », Article présenté au Colloque francophone sur les sondages, Québec, mai 2005.

[9] Kish, L. (1965), « *Survey Sampling* », John Wiley and Sons, New York, 1965.

[10] Kish, L. (1981), « *Using Cumulated Rolling Samples* », Congressional Research Office, Washington,, 80-528-0.

[11] Kish, L. (1990), « Recensement pas étapes et échantillon avec renouvellement complet », *Techniques d'enquête*, Vol. 16, No. 1, pp.67-86, juin 1990.

[12] Kish, L. (1999), « Cumulating/Combining Population Surveys », *Survey Methodology*, Vol. 25, pp. 129-138.

[13] Tam, S.M. (1984), « On Covariances From Overlapping Samples », *The American Statistician*, Vol. 34, No. 4, pp. 288-289, novembre 1984.