N° 12-001-XIF au catalogue

Techniques d'enquête

Juin 2006

Statistique Canada   Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit n° 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada
Division des méthodes d’enquêtes auprès des entreprises

Techniques d'enquête

Juin 2006

Publication autorisée par le ministre responsable de Statistique Canada

Juillet 2006

N^{o} 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Vol. 32, N° 1, pp. 127-133
Statistique Canada, N° 12-001 au catalogue

Effets de plan pour les échantillons à plans de sondage multiples

Siegfried Gabler, Sabine Häder et Peter Lynn [1]

Résumé

Dans certaines situations, le plan de sondage d'une enquête est assez complexe et comporte des plans fondamentalement différents pour divers domaines. L'effet de plan des estimations fondées sur l'échantillon total est une somme pondérée des effets de plan selon le domaine. Nous calculons les pondérations sous un modèle approprié et illustrons leur utilisation au moyen de données provenant de l'Enquête sociale européenne (European Social Survey ou ESS).

Mots clés : Stratification; mise en grappes; modèle des composantes de la variance; coefficient de corrélation intraclasse; probabilités de sélection.

1. Introduction

En recherche par sondage, l'application de plans de sondage complexes est fréquente. Ces plans possèdent des caractéristiques, telles la stratification, la mise en grappes et (ou) l'utilisation de probabilités d'inclusion inégales, qui donnent lieu à des « effets de plan ». L'effet de plan est une mesure qui représente l'effet du plan de sondage sur la variance d'une estimation. Fondé sur le plan de sondage, il est défini comme suit (voir Lohr 1999, page 239) :

$$deff\,(plan,\, statistique) = \frac{\text{V(estimation d'après le plan d'échantillonnage)}}{\text{V}\begin{pmatrix}\text{estimation d'après un eas avec le même} \\ \text{nombre d'unités d'observation}\end{pmatrix}}$$

où eas indique un échantillon aléatoire simple. Le recours à la mise en grappes et (ou) à des probabilités d'inclusion inégales produit habituellement des effets de plan dont la valeur est supérieure à 1,0; autrement dit, la variance d'une estimation est plus grande que celle de l'estimation établie d'après un échantillon aléatoire simple contenant le même nombre d'observations. La prise en compte des effets de plan est très importante lorsqu'on décide d'avance de la taille de l'échantillon d'une enquête. Ainsi, si l'on prévoit mener une enquête comparative entre plusieurs pays, il est très utile de disposer d'information sur les effets de plan pour ces pays. Il est alors possible de choisir les tailles d'échantillon nettes de façon que la précision des estimations soit approximativement uniforme. Pour cela, la taille d'échantillon qui serait nécessaire sous eas (taille effective d'échantillon) pour obtenir un degré donné de précision doit être multipliée par l'effet de plan prévu.

L'Enquête sociale européenne (ESS, voir www.europeansocialsurvey.com) est un programme d'enquête dans lequel les effets de plan sont pris en compte pour le calcul des tailles nettes d'échantillon, en cherchant à obtenir la même taille effective d'échantillon pour chaque pays (n_{eff} = 1 500). Des 22 pays qui ont participé au premier cycle de l'ESS, trois seulement, le Danemark, la Finlande et la Suède, ont utilisé un plan de sondage avec probabilités de sélection égales, sans mise en grappes (eas). Pour tous les autres, il a fallu prédire l'effet de plan avant l'étude. On peut utiliser, pour cela, une approche fondée sur un modèle (voir Gabler, Häder et Lahiri 1999) qui fait la distinction entre l'effet de plan dû à un échantillonnage avec probabilités d'inclusion inégales (1er terme) et l'effet de plan dû à la mise en grappes (2e terme) :

$$deff = m\frac{\sum_{i=1}^{I} m_i w_i^2}{\left(\sum_{i=1}^{I} m_i w_i\right)^2} \times [1 + (b^* - 1)\rho] = deff_p \times deff_c \qquad (1)$$

où m_i représente les répondants dans la i^{e} classe de probabilités de sélection, chacun recevant un poids de w_i, ρ est le coefficient de corrélation intragrappe et

$$b^* = \frac{\sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2}{\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2}$$

où b_c est le nombre d'observations dans la grappe $c\,(c = 1, \ldots, C)$ et w_{cj} est le poids de sondage de l'élément j dans la grappe c. (Il s'agit évidemment d'une simplification reposant sur l'hypothèse qu'il n'existe aucune association entre y et w_i, ou entre w_i et b^*, et ne tenant compte d'aucun effet de stratification, qui aura tendance à être avantageuse et modeste. Voir Lynn, Gabler, Häder et Laaksonen (2007, à paraître), ainsi que Park et Lee (2004) pour une discussion de la sensibilité des prédictions de *deff* à

1. Siegfried Gabler et Sabine Häder, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Allemagne. Courriel : gabler@zuma-mannheim.de; Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, Royaume-Uni. Courriel : plynn@essex.ac.uk.

ces hypothèses; voir Lynn et Gabler (2005) pour une discussion de divers moyens de prédire $deff_c$).

Dans certains pays, les plans de sondage utilisés étaient encore plus compliqués, comprenant des plans fondamentalement différends dans chacun des deux domaines indépendants. Au Royaume-Uni, par exemple, il s'agissait du mélange d'un plan par grappes avec probabilités d'inclusion inégales (en Grande-Bretagne) et d'un échantillon sans mise en grappes (en Irlande du Nord). En Pologne, des échantillons aléatoires simples ont été sélectionnés dans un domaine (villes grandes et moyennes), tandis qu'un plan à deux degrés avec mise en grappes a été appliqué au deuxième domaine (toutes les autres régions). En Allemagne, un échantillon par grappes avec probabilités de sélection égales a été sélectionné dans chaque domaine (Allemagne de l'Ouest, y compris Berlin Ouest; Allemagne de l'Est), mais les fractions d'échantillonnage n'étaient pas les mêmes dans les deux domaines.

La question de la prédiction des effets de plan s'est donc posée pour ces échantillons à plan de sondage double. Comme nous le montrons plus loin, il ne s'agit pas simplement d'une combinaison convexe des effets de plan pour les divers domaines, à part dans des cas spéciaux. Nous présentons une solution générale pour les échantillons à plans de sondage multiples à la section 2, ainsi que des exemples d'application de cette solution afin de prédire les effets de plan avant le travail sur le terrain (section 3) et après le travail sur le terrain (section 4). À la section 5, nous concluons par une discussion.

2. Effets de plan pour les échantillons à plans de sondage multiples

Soit $\{C_1, \ldots, C_K\}$ une partition des grappes en K domaines. Alors, $C\bar{b} = \sum_{c=1}^{C} b_c = \sum_{k=1}^{K} \sum_{c \in C_k} b_c = \sum_{k=1}^{K} m_k = m$, où $m_k = \sum_{c \in C_k} b_c$ est le nombre d'observations dans le k^{e} domaine de grappes. Soit y_{cj} l'observation pour l'unité d'échantillonnage j dans la grappe $c\,(c = 1, \ldots, C;$ $j = 1, \ldots, b_c)$. L'estimateur habituel fondé sur le plan de sondage de la moyenne de population est

$$\bar{y}_w = \frac{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \sum_{k=1}^{K} \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} \bar{y}_w^{(k)}$$

où

$$\bar{y}_w^{(k)} = \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}\, y_{cj}}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}.$$

Nous postulons le modèle M1 suivant :

$$\left.\begin{aligned} E(y_{cj}) &= \mu \\ \operatorname{Var}(y_{cj}) &= \sigma^2 \end{aligned}\right\} \quad \text{pour} \quad c = 1, \ldots, C;\ j = 1, \ldots, b_c$$
$$\operatorname{Cov}(y_{cj}, y_{c'j'}) = \begin{cases} \rho_k \sigma^2 & \text{si } c = c' \in C_k;\ j \neq j' \\ 0 & \text{autrement} \end{cases} \quad k = 1, \ldots, K. \tag{2}$$

Le modèle M1 convient pour tenir compte de l'effet de grappe avec divers types de grappes et généralise une approche antérieure (voir, par exemple, Gabler et coll. 1999). Des modèles plus généraux sont décrits dans Rao et Kleffe (1988, page 62). Nous définissons l'effet de plan (par rapport au modèle) comme étant $deff = \operatorname{Var}_{M1}(\bar{y}_w) / \operatorname{Var}_{M2}(\bar{y})$, où $\operatorname{Var}_{M1}(\bar{y}_w)$ est la variance de $\bar{y}_w$ sous le modèle M1 et $\operatorname{Var}_{M2}(\bar{y})$ est la variance de la moyenne globale d'échantillon $\bar{y}$, définie comme étant $\sum_{c=1}^{C} \sum_{j=1}^{b_c} y_{cj} / m$, calculée sous le modèle M2 suivant :

$$\left.\begin{aligned} E(y_{cj}) &= \mu \\ \operatorname{Var}(y_{cj}) &= \sigma^2 \end{aligned}\right\} \quad \text{pour} \quad c = 1, \ldots, C;\ j = 1, \ldots, b_c$$
$$\operatorname{Cov}(y_{cj}, y_{c'j'}) = 0 \quad \text{pour tous les} \quad (c, j) \neq (c', j'). \tag{3}$$

Soulignons que le modèle M2 est approprié sous échantillonnage aléatoire simple et donne l'expression habituelle $\operatorname{Var}_{M2}(\bar{y}) = \sigma^2 / m$.

De façon fort semblable à Gabler et coll. (1999), nous notons que

$$\operatorname{Var}_{M1}\left(\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}\, y_{cj} \right) = \sigma^2 \sum_{k=1}^{K} \sum_{c \in C_k} \left\{ \sum_{j=1}^{b_c} w_{cj}^2 + \rho_k \sum_{j \neq j'}^{b_c} w_{cj} w_{cj'} \right\}. \tag{4}$$

Donc

$$deff = \sum_{k=1}^{K} \left(\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} \right)^2 \frac{m}{m_k} deff_k \tag{5}$$

où

$$deff_k = m_k \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}{\left(\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj} \right)^2} \times [1 + (b_k^* - 1)\rho_k] = deff_{pk} \times deff_{ck},$$

et

$$b_k^* = \frac{\sum_{c \in C_k} \left(\sum_{j=1}^{b_c} w_{cj} \right)^2}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}.$$

On peut voir que *deff* n'est pas une combinaison convexe des effets de plan spécifiques $\{deff_k\}$, sauf dans des cas particuliers. Nous considérons ici quatre scénarios raisonnables, chacun représentant une simplification du cas général. La combinaison ne devient convexe que dans deux de ces scénarios (1 et 4) :

Scénario 1 : *Même pondération pour toutes les unités*

Si $w_{cj} = 1$ pour tous c, j, alors l'expression (5) se simplifie comme suit :

$$deff = \sum_{k=1}^{K} \frac{m_k}{m} deff_k. \tag{6}$$

Scénario 2 : *Même pondération des unités dans chaque domaine*

Si $w_{cj} = w_k$ pour tous $c \in C_k, j$, alors l'expression (5) devient :

$$deff = \sum_{k=1}^{K} \left(\frac{m_k w_k}{\sum_{k=1}^{K} m_k w_k} \right)^2 \frac{m}{m_k} deff_k. \tag{7}$$

Scénario 3 : *Taille d'échantillon pondéré proportionnelle à la taille de la population du domaine*

Si

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{N_k}{N},$$

où N_k est la taille de population dans le domaine $k; N = \sum_{k=1}^{K} N_k$, alors l'expression (5) devient :

$$deff = \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \frac{m}{m_k} deff_k. \tag{8}$$

Scénario 4 : *Taille d'échantillon non pondéré proportionnelle à la taille de population du domaine*

Si

$$\frac{m}{m_k} = \frac{N}{N_k},$$

alors l'expression (8) devient :

$$deff = \sum_{k=1}^{K} \frac{N_k}{N} deff_k. \tag{9}$$

3. Application à la prédiction de *Deff*

Lors du premier cycle de l'ESS, le plan de sondage était une combinaison de deux plans différents pour 5 des 22 pays, à savoir le Royaume-Uni, la Pologne, la Belgique, la Norvège et l'Allemagne. Nous pouvons appliquer la formule générale (5) des effets de plan pour échantillon à plans multiples à chacun de ces cas, où $K = 2$. Pour certains d'entre eux, nous pouvons utiliser indifféremment l'une des expressions simplifiées (6) à (9). Ici, nous illustrons comment la formule serait utilisée pour prédire les effets de plan avant le travail sur le terrain en vue d'établir la taille d'échantillon nette (répondants) requise pour atteindre une précision d'estimation préétablie. Dans chaque cas, l'approche consiste à prédire $\{deff_k\}$ en utilisant (1) pour chaque *k*, puis à utiliser (5) pour prédire *deff*. Afin de prédire $\{deff_k\}$, nous utilisons les valeurs observées de $\{w_{cj}\}$ provenant de l'échantillon de répondants du premier cycle de l'ESS pour estimer b^*, m_i et w_i. En d'autres termes, nous pourrions considérer que ces prédictions sont faites pour une future enquête basée sur le même plan de sondage (par exemple, un futur cycle de l'ESS). À titre d'exemple, nous supposons que $\rho_k = 0{,}02\, \forall\, k$ avec un plan de sondage par grappes et $\rho_k = 0{,}00\, \forall\, k$ avec un plan sans mise en grappes (0,02 dans l'ESS dans les cas où l'on ne disposait pas d'estimations d'après des enquêtes antérieures). Ici, nous nous concentrons sur l'application de l'équation (5). Pour une description plus détaillée des plans de sondage, voir Häder, Gabler, Laaksonen et Lynn (2003). Nous choisissons comme exemples trois des pays participants à l'ESS, la Pologne, le Royaume-Uni et l'Allemagne, car ils ont utilisé des plans de sondage multiples dont les différences entre domaines ne sont pas les mêmes. Les plans de sondage utilisés par la Norvège et la Belgique étaient semblables à celui de la Pologne, avec probabilités d'inclusion égales pour toutes les unités, mais mise en grappes dans un domaine et non dans l'autre.

3.1 Pologne

En Pologne, le premier domaine couvrait la population des villes de 100 000 habitants et plus. Dans ce domaine, des personnes ont été sélectionnées par EAS d'après le registre de population (base de données PESEL) dans chaque région, avec application d'une fraction d'échantillonnage légèrement différente selon la région afin de refléter les différences attendues de taux de réponse. Ce domaine comprenait 42 villes qui représentaient environ 31 % de la population cible.

Le deuxième domaine correspondait au reste de la population, c'est-à-dire les personnes vivant dans les villes de 99 999 habitants et moins et dans les régions rurales. Cette partie de l'échantillon a été stratifiée et mise en grappes (158 grappes). L'échantillonnage de cette deuxième partie a été réalisé selon un plan à deux degrés où les UPE ont été sélectionnées avec probabilité proportionnelle à la taille. La définition de l'UPE n'était pas la même pour les régions urbaines que pour les régions rurales. Pour les premières, l'UPE correspondait à une ville, tandis que pour les secondes, il s'agissait d'un village. À la deuxième phase, une grappe de 12 répondants a été sélectionnée par eas dans chaque UPE.

Dans le premier domaine, $\rho_1 = 0$ et $deff_{c1} = 1$. La variation modérée des probabilités de sélection mène à $deff_{p1} = 1{,}005$ et, par conséquent, $deff_1 = deff_{c1} \cdot deff_{p1} =$ 1,005. Dans le deuxième domaine, l'effet de plan dû à la mise en grappes prévu est $deff_{c2} = 1{,}18$ (d'après la prédiction que $b^* = 10{,}07$) et $deff_{p2} = 1{,}01$, ce qui donne $deff_2 = deff_{c2} \cdot deff_{p2} =$ 1,19. La substitution de ces valeurs à $deff_k$ dans (5) donne un effet de plan prévu égal à $deff = 1{,}17$.

Le plan de sondage appliqué en Pologne ne diffère que légèrement du scénario 2 et, dans ce cas, nous voyons que l'expression plus simple, (7), donne une prédiction raisonnable si nous calculons la valeur approximative des pondérations comme suit. Le domaine 1 contient 37,3 % de l'échantillon brut et 31 % de la population cible. Donc

$$w_1 = \frac{N_1 / N}{n_1 / n} = \frac{0{,}310}{0{,}373} = 0{,}831$$

et

$$w_2 = \frac{N_2 / N}{n_2 / n} = \frac{0{,}690}{0{,}627} = 1{,}100,$$

respectivement, où n_k est la taille de l'échantillon sélectionné dans le domaine $k; \sum_{k=1}^{K} n_k$.

Maintenant, nous pouvons appliquer l'expression (7) pour calculer l'effet de plan prévu pour les estimations pour la Pologne : $deff = (0{,}194 \cdot 1{,}005) + (0{,}821 \cdot 1{,}19) = 1{,}17$.

3.2. Royaume-Uni

Au Royaume-Uni, le plan de sondage de l'ESS n'était pas le même pour la Grande-Bretagne (Angleterre, Pays de Galles, Écosse) que pour l'Irlande du Nord. En Grande-Bretagne, on a utilisé un plan stratifié à trois degrés avec probabilités de sélection inégales. À la première étape, 162 petits secteurs, appelés « secteurs de code postal » ont été sélectionnés systématiquement avec probabilité proportionnelle au nombre d'adresses dans le secteur, après stratification implicite selon la région et la densité de population. À la deuxième étape, 24 adresses ont été sélectionnées dans chaque secteur, ce qui a produit un échantillon avec probabilités égales d'adresses. À la troisième étape, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée au moyen d'une grille de Kish.

Pour l'Irlande du Nord, un échantillon aléatoire simple de 125 adresses a été tiré d'après la liste de propriétés privées de l'Agence d'évaluation foncière (Valuation and Land Agency). Puis, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée en utilisant une grille de Kish. Donc, l'échantillon du Royaume-Uni est mis en grappes dans un domaine, mais non dans l'autre. Dans les deux domaines, les probabilités de sélection sont inégales.

Pour la Grande-Bretagne, nous avons prédit $deff_{c1} = 1{,}20$ (d'après une prédiction de $b^* = 11{,}11$) et $deff_{p1} = 1{,}22$, de sorte que $deff_1 = 1{,}46$. Pour l'Irlande du Nord, nous obtenons les prédictions $deff_{c2} = 1$ (par définition) et $deff_{p2} = 1{,}27$, de sorte que $deff_{p2} = 1{,}27$. D'après l'expression (5), $deff = 0{,}978 \cdot 1{,}46 + \; 0{,}023 \cdot 1{,}27 = 1{,}460$. Il convient de souligner que les tailles des échantillons sélectionnés dans les deux domaines ont été choisies de façon à obtenir des tailles nettes d'échantillon à peu près proportionnelles aux tailles de population. Autrement dit, la simplification correspondant au scénario 4 est vérifiée approximativement. Si nous utilisons l'expression (9), nous obtenons $deff = N_1 / N \; deff_1 + N_2 / N \quad deff_2 = 0{,}97 \cdot 1{,}46 +$ $0{,}03 \cdot 1{,}27 = 1{,}457$, ce qui démontre que cette expression est une approximation raisonnable de (5) dans ce cas.

3.3. Allemagne

En Allemagne, des échantillons indépendants ont été sélectionnés dans deux domaines, c'est-à-dire l'Allemagne de l'Ouest, y compris Berlin Ouest et l'Allemagne de l'Est, y compris Berlin Est. Dans chaque domaine, on a sélectionné un échantillon par grappes avec probabilités égales, mais en utilisant une plus grande fraction d'échantillonnage pour l'Allemagne de l'Est.

À la première étape, 100 communautés (grappes) ont été sélectionnées pour l'Allemagne de l'Ouest et 50 pour l'Allemagne de l'Est avec probabilité proportionnelle à la taille de la population de la communauté (personnes de 15 ans et plus). Le nombre de communautés sélectionnées dans chaque strate a été déterminé par une méthode d'arrondissement contrôlé. Le nombre de points d'échantillonnage était de 108 à l'Ouest et de 55 à l'Est (pour certaines grandes collectivités, on a utilisé plus d'un point d'échantillonnage). À la deuxième étape, pour chaque point d'échantillonnage, on a sélectionné un nombre égal de personnes par échantillonnage aléatoire systématique, d'après les registres locaux des bureaux d'enregistrement des résidents.

Puisque le plan de sondage est autopondéré aussi bien pour l'Allemagne de l'Est que de l'Ouest, mais que la répartition est disproportionnelle, nous pouvons appliquer le scénario 2 et utiliser l'expression (7), où

$$w_1 = w_{\text{EST}} = \frac{N_{\text{EST}}}{N} \frac{n}{n_{\text{EST}}} = 0{,}567$$

et

$$w_2 = w_{\text{OUEST}} = \frac{N_{\text{OUEST}}}{N} \frac{n}{n_{\text{OUEST}}} = 1{,}257.$$

(Nous notons qu'il est courant, dans certaines enquêtes, de rééchelonner les pondérations afin que leur somme soit égale à la taille de population. Cette pratique n'aurait aucune incidence ici, car l'expression (5) ne comprend que des ratios de sommes de pondérations).

Pour chaque domaine, nous avons prédit les effets de plan $deff_{c1} = 1{,}39$ et $deff_{c2} = 1{,}35$, respectivement (d'après les prédictions que $b^* = 20{,}56$ et 18,65, respectivement), si bien qu'il découle de (7) que

$$deff = 0{,}120 \cdot 1{,}39 + 0{,}991 \cdot 1{,}35 = 1{,}51.$$

Il convient de souligner que dans ce cas, toute combinaison convexe des effets de plan de domaine produira une prédiction de *deff* comprise entre 1,35 et 1,39. Par exemple, (6) donnerait $deff = 1{,}36.$ Ce résultat ne tient pas compte des différences de probabilité de sélection *entre* les domaines. Dans le cas du plan de sondage examiné ici, où la *seule* différence de plan de sondage entre les domaines est la différence de probabilité de sélection, *deff* pourrait aussi être prédit en prenant la combinaison convexe et en la multipliant par la prédiction de $deff_p$ pour le premier terme de l'expression (1), c'est-à-dire $deff = 1{,}36 \cdot 1{,}09 = 1{,}49.$ Cette méthode n'est toutefois équivalente que dans le cas particulier où les $\{deff_k\}$ sont égaux, et approximativement équivalente ici, où la variation est faible.

4. Application à l'estimation de *Deff*

Nous allons maintenant illustrer l'utilisation de l'expression (5) pour estimer les effets de plan après le travail sur le terrain. Nous présentons des estimations pour cinq variables démographiques/de comportement et un ensemble de 24 mesures d'attitude provenant du premier cycle de l'Enquête sociale européenne (ESS) pour les trois mêmes pays qu'à la section 3. Aux fins de comparaison, nous présentons aussi les estimations que l'on obtiendrait en utilisant les expression plus simples (6), (8) et (9). Les résultats montrent que les estimations de *deff* diffèrent considérablement selon la variable. Cette situation, qui est prévisible, reflète la variation de l'association de y avec les grappes et les probabilités de sélection. Mais ici, nous nous intéressons principalement aux différences entre les méthodes d'estimation pour une même variable.

Dans le cas de l'Allemagne, nous voyons que les estimateurs (6) et (9), qui ne tiennent pas compte de la variation de la pondération et des taux d'échantillonnage entre les deux domaines, sous-estiment *deff* pour toutes les variables. L'estimateur (8), qui repose uniquement sur l'hypothèse que les taux de réponse sont égaux dans chaque domaine, produit des estimations fort semblables à (5). Pour la Pologne, les trois estimateurs simplifiés sous-estiment *deff*, quoique (6) pourrait produire des résultats légèrement meilleurs que les deux autres. Pour ce qui est du Royaume-Uni, nous obtenons le résultat remarquable que les quatre estimateurs produisent des estimations presque identiques pour chaque variable. L'hypothèse qui sous-tend (9) (et par conséquent également celle qui sous-tend (8)) tient pour le Royaume-Uni et, bien que les pondérations soient loin d'être égales, leur distribution est fort semblable dans chaque domaine. Il convient de souligner que (6) est vérifié sous une hypothèse plus faible que

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}} = \frac{m_k}{m},$$

c'est-à-dire que la part des pondérations dans chaque strate est égale à la part des unités d'échantillonnage. Il est frappant que ces relations entre les estimateurs soient convergentes pour l'ensemble des variables considérées.

5. Discussion et conclusion

L'expression (5) offre un moyen approprié de combiner les effets de plan pour des domaines pour lesquels les plans de sondage sont fondamentalement différents. Elle peut être appliquée en estimant l'effet de plan *deff* de la manière habituelle pour chaque domaine, puis en les combinant d'après les données sur la pondération et l'appartenance des unités d'échantillonnage au domaine. L'utilisation de (5) pour prédire les effets de plan *deff* avant qu'une enquête soit réalisée est à peine plus difficile. Elle nécessite la prédiction de la part des pondérations dans l'échantillon de répondants dans chaque domaine, en plus d'une méthode de prédiction des *deff* propres aux divers plans de sondage.

Nous avons montré à la section 4 qui précède que l'utilisation d'autres méthodes, plus simples, de combinaison des *deff* de domaine ne produit pas toujours de bonnes estimations. Plus précisément, l'utilisation d'une combinaison convexe aura tendance à causer une sous-estimation dont l'importance dépend de l'écart par rapport aux hypothèses qui sous-tendent les expressions simplifiées. Dans notre exemple empirique, les écarts étaient modérés, mais il est facile d'imaginer des plans de sondage où les variations des probabilités de sélection moyennes ou de la distribution des poids de sondage selon le domaine sont plus importantes. Nous devrions par conséquent recommander de n'utiliser les estimateurs (6) à (9) que si les hypothèses sont vraiment vérifiées ou que les données sur le plan de sondage nécessaires pour calculer (5) ne sont pas disponibles, auquel cas l'analyste devrait au moins tenir compte arbitrairement d'une sous-estimation en se basant sur sa connaissance du plan de sondage.

Tableau 1
Estimations de *Deff* pour les moyennes sous quatre estimateurs pour trois pays

	Allemagne				Royaume-Uni				Pologne			
Estimateur :	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)
Variables démographiques/de comportement												
N[bre] de personnes dans le ménage	1,87	1,85	1,87	1,74	1,66	1,66	1,66	1,66	1,51	1,43	1,41	1,42
N[bre] d'années d'études	3,25	2,80	3,25	2,88	2,81	2,79	2,80	2,79	1,77	1,66	1,63	1,64
Revenu net du ménage	2,46	2,15	2,46	2,19	2,82	2,80	2,80	2,80	2,16	2,00	1,95	1,98
Temps passé à regarder la TV	2,08	1,86	2,08	1,87	2,04	2,03	2,03	2,03	1,31	1,26	1,25	1,25
Temps passé à lire le journal	1,79	1,62	1,79	1,61	1,35	1,35	1,35	1,35	1,73	1,63	1,60	1,61
Mesures d'attitude												
Discrimination selon la race	1,16	1,03	1,16	1,04	1,92	1,92	1,92	1,92	1,02	1,01	1,01	1,01
Discrimination selon la religion	1,22	1,05	1,22	1,08	1,26	1,26	1,26	1,26	1,07	1,05	1,05	1,05
État de bonheur général	2,56	2,11	2,55	2,23	1,56	1,55	1,56	1,55	1,49	1,42	1,40	1,41
Confiance dans les autres	2,20	1,96	2,20	1,98	1,85	1,84	1,84	1,84	1,66	1,57	1,54	1,55
Confiance dans le Parlement européen	1,83	1,59	1,83	1,62	1,50	1,50	1,50	1,50	1,43	1,37	1,35	1,36
Confiance dans le système juridique	2,07	1,72	2,07	1,81	1,37	1,37	1,37	1,37	1,42	1,36	1,34	1,35
Confiance dans la police	1,92	1,63	1,92	1,69	1,24	1,24	1,24	1,24	1,24	1,20	1,19	1,19
Confiance dans les politiciens	1,75	1,62	1,75	1,59	1,38	1,38	1,38	1,38	1,63	1,54	1,51	1,53
Confiance dans le Parlement	1,64	1,48	1,64	1,48	1,45	1,45	1,45	1,45	1,13	1,10	1,10	1,10
Échelle gauche-droite	1,70	1,65	1,70	1,58	1,48	1,47	1,48	1,48	1,31	1,26	1,25	1,25
Satisfaction à l'égard de la vie	2,06	1,74	2,06	1,81	1,68	1,67	1,67	1,67	1,30	1,25	1,24	1,25
Satisfaction à l'égard du système d'éducation	3,03	2,89	3,03	2,79	1,37	1,37	1,37	1,37	1,40	1,34	1,32	1,33
Satisfaction à l'égard du système de santé	3,76	3,21	3,76	3,32	1,65	1,64	1,64	1,64	1,65	1,56	1,53	1,54
Attitude religieuse	1,94	1,75	1,94	1,75	1,57	1,56	1,56	1,56	1,73	1,63	1,60	1,61
Attitude à l'égard des immigrants	2,77	2,68	2,77	2,57	1,92	1,92	1,92	1,92	1,89	1,76	1,73	1,74
Appuie une loi contre la discrimination ethnique	2,82	2,85	2,82	2,66	1,73	1,72	1,72	1,72	2,57	2,36	2,29	2,33
Importance de la famille	2,17	1,99	2,17	1,97	1,19	1,19	1,19	1,19	1,21	1,17	1,17	1,17
Importance des amis	2,31	2,09	2,31	2,08	1,34	1,34	1,34	1,34	1,54	1,46	1,44	1,45
Importance du travail	2,20	2,16	2,20	2,05	1,90	1,89	1,89	1,89	1,69	1,59	1,57	1,58
Aide les personnes moins bien nanties	2,70	2,47	2,70	2,45	1,35	1,35	1,35	1,35	1,78	1,67	1,64	1,66
Respecte toujours la loi	2,43	2,21	2,43	2,20	1,53	1,52	1,52	1,52	2,11	1,96	1,91	1,93
Activisme politique	3,26	2,83	3,26	2,89	1,94	1,94	1,94	1,94	2,16	2,00	1,96	1,98
Libéralisme	2,28	2,18	2,28	2,10	1,78	1,77	1,78	1,78	1,75	1,64	1,61	1,63
Participation à des groupes	3,75	3,04	3,75	3,24	2,26	2,25	2,25	2,25	1,82	1,71	1,68	1,69

Une question importante qui dépasse le cadre du présent article est celle de savoir comment traiter la non-réponse lors de la prédiction ou de l'estimation des effets de plan pour les échantillons à plans multiples. Les expressions présentées à la section 2 se rapportent au nombre d'observations (unités d'échantillonnage répondantes) dans chaque domaine, m_k, et les calculs présentés aux sections 3 et 4 sont fondés sur les nombres prévus et réels d'observations, respectivement. Cependant, l'interprétation naturelle des différences entre les quatre scénarios de la section 2 pourrait se faire en fonction du plan de sondage, où les pondérations sont les poids de sondage. Donc, le scénario 2, par exemple, aurait trait à un plan de sondage avec *probabilités de sélection égales* dans les domaines, mais où la fraction d'échantillonnage peut varier selon le domaine. Cependant, dans la plupart des applications réelles, il y aura une non-réponse qui pourrait fort bien différer entre les domaines, ainsi que dans les domaines, situation qui est souvent reflétée par un ajustement du poids de sondage. Donc, la simplification du scénario 2 ne serait applicable que si l'ajustement pour la non-réponse était constant dans les domaines, outre la sélection avec *probabilités égales* dans les domaines.

Le scénario 3, s'il est interprété uniquement par rapport au plan de sondage, devrait être vérifié pour tout plan de sondage bien spécifié dans lequel les domaines forment des strates explicites. L'expression (8) est par conséquent équivalente à l'expression (5), en l'absence de non-réponse. En présence de non-réponse, le scénario 3 exige que les taux de réponse (pondérés par les poids de sondage) soient égaux dans chaque domaine. De même, le scénario 4 demande que le taux d'inclusion net (produit du taux de couverture, de la fraction d'échantillonnage et du taux de réponse) soit le même dans chaque domaine, tandis qu'une interprétation basée sur le plan de sondage ne tiendrait pas compte de la composante du taux de réponse.

La recherche de moyens appropriés d'intégrer l'ajustement pour la non-réponse dans l'estimation de l'effet de plan et, en particulier, l'effet que cette correction pourrait avoir sur l'estimation dans le cas d'échantillons à plans multiples, semble être un domaine qui mérite d'être exploré lors de futures études.

Remerciements

Le troisième auteur remercie la ZUMA pour le poste de professeur invité qui lui a permis de trouver le temps et les conditions propices pour rédiger le présent article et remercie aussi de son soutien le UK Longitudinal Studies Centre de l'Université d'Essex, qui est financé par la subvention H562255004 de l'Economic and Social Research Council du Royaume-Uni.

Bibliographie

Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Häder, S., Gabler, S., Laaksonen, S. et Lynn, P. (2003). The sample. Chapître 2 dans *ESS 2002/2003: Rapport technique*. http://www.europeansocialsurvey.com.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis*. Pacific Grove: Duxbury Press.

Lynn, P., et Gabler, S. (2005). Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes. *Techniques d'enquête*, 31, 109-113.

Lynn, P., Gabler, S., Häder, S. et Laaksonen, S. (2007, a paraître). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics,* accepté.

Park, I., et Lee, H. (2004). Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe. *Techniques d'enquête*, 30, 205-216.

Rao, C.R., et Kleffe, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.