Nº 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 nº 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

Nº 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. 97-106
Statistique Canada, N° 12-001 au catalogue

Calcul de la taille de l'échantillon pour l'estimation pour petits domaines

Nicholas Tibor Longford [1]

Résumé

Nous décrivons une approche générale de détermination du plan d'échantillonnage des enquêtes planifiées en vue de faire des inférences pour de petits domaines (sous-domaines). Cette approche nécessite la spécification des priorités d'inférence pour les petits domaines. Nous établissons d'abord des scénarios de répartition de la taille de l'échantillon pour l'estimateur direct, puis pour les estimateurs composite et bayésien empirique. Nous illustrons les méthodes à l'aide d'un exemple de planification d'un sondage de la population suisse et d'estimation de la moyenne ou de la proportion d'une variable pour chacun des 26 cantons.

Mots clés : Efficacité; estimation pour petits domaines; priorité d'inférence; répartition de la taille de l'échantillon.

1. Introduction

Le plan d'échantillonnage est un instrument essentiel à la production d'estimations efficaces et d'autres formes d'inférence au sujet d'une grande population, lorsque les ressources disponibles ne permettent pas de recueillir l'information pertinente pour chaque membre de la population. Dans ce contexte, nous interprétons l'efficacité comme étant la combinaison optimale d'un plan d'échantillonnage et d'un estimateur d'un paramètre de population θ. Par optimale, nous entendons que l'erreur quadratique moyenne est minimale, quoique le développement présenté dans l'article puisse être adapté à d'autres critères. Le groupe de plans de sondage possibles est délimité par les ressources et celles-ci sont habituellement exprimées en fonction d'une taille fixe d'échantillon. Cette approche n'est pas toujours appropriée, parce que les coûts moyens par sujet ne sont pas nécessairement les mêmes pour tous les plans d'échantillonnage. Toutefois, si nous considérons une gamme limitée de plans, nous pouvons ignorer ce point.

Le problème de l'établissement du plan d'échantillonnage afin d'estimer efficacement une grandeur unique est bien compris et des solutions existent pour bon nombre de spécifications utilisées fréquemment. La plupart comportent un problème d'optimisation univarié sous contraintes. L'établissement du plan d'échantillonnage pour l'estimation de plusieurs paramètres est considérablement plus complexe, parce que le problème comprend plusieurs facteurs, habituellement un pour chaque paramètre. Il est essentiel d'optimiser le plan simultanément pour tous les facteurs, parce que les objectifs d'inférence efficace au sujet des paramètres cibles peuvent être conflictuels. Par exemple, dans l'estimation pour petits domaines, l'allocation d'une part plus généreuse de la taille de l'échantillon à un petit domaine doit être compensée par une allocation moins généreuse à un ou à plusieurs autres.

Au cours des dernières décennies, la production de statistiques pour des petits domaines est devenue un important sujet de recherche en méthodologie d'enquête (Fay et Herriot 1979; Platek, Rao, Särndal et Singh 1987; Ghosh et Rao 1994; Longford 1999; Rao 2003), étant donné l'intérêt grandissant des organismes gouvernementaux, du secteur de la publicité et du marketing et de celui de la finance et des assurances pour ce genre d'information. À l'heure actuelle, de nombreuses enquêtes à grande échelle sont conçues en vue de produire des estimations de niveau national, mais sont parfois utilisées après coup pour faire des inférences au sujet de petits domaines. Cela n'aurait pas d'inconvénient si les plans d'échantillonnage optimaux pour l'inférence sur petits domaines et l'inférence nationale étaient les mêmes. Nous montrons dans le présent article qu'il n'en est pas ainsi et que le plan d'échantillonnage peut effectivement être ciblé pour l'estimation pour petits domaines, en tenant compte de l'objectif de production d'estimations efficaces de paramètres de niveau national. Pour éviter le cas banal, supposons que les populations des petits domaines soient de taille inégale. Nous appliquons les méthodes au problème de la planification d'inférences au sujet des 26 cantons de la Suisse; la taille de la population de ces cantons varie de 15 000 (Appenzell-Innerrhoden) à 1,23 million (Zürich). La population de la Suisse se chiffre à 7,26 millions d'habitants.

La littérature traitant de la planification des enquêtes pour l'estimation pour petits domaines est peu abondante. L'une des contributions importantes est celle de Singh, Gambino et Mantel (1994). Dans l'une des approches dont discutent ces auteurs, la taille prévue de l'échantillon de l'Enquête sur la population active du Canada est divisée en deux. Une partie est répartie optimalement en vue de la production d'estimations de niveau national (domaine) et l'autre est répartie optimalement en vue de l'estimation pour petits domaines (sous-domaines). Pour ce dernier objectif, des

1. Nicholas Tibor Longford, Departament d'Económia i Empresa, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005 Barcelone, Espagne. Courriel : NTL@SNTL.co.uk.

sous-échantillons de même taille sont attribués à chaque petit domaine, lorsque les variances dans les sous-domaines sont égales, que la correction pour population finie peut être ignorée et que les coûts d'enquête par sujet sont les mêmes pour tous les sous-domaines, mais aussi quand les paramètres visées par l'inférence sont les moyennes de petit domaine. Si l'on veut estimer des totaux de population, l'équirépartition de l'échantillon entre les sous-domaines n'est pas efficace, parce qu'elle pénalise l'estimation pour les petits domaines les plus peuplés. Même si l'on estime des proportions ou des taux (pourcentages), les variances intradomaine dépendent de la proportion de population, quoique la dépendance soit faible lorsque toutes les proportions sont loin de zéro et de l'unité. Pour des travaux plus récents sur les plans d'échantillonnage pour l'estimation pour petits domaines, voir Marker (2001).

La section suivante décrit l'approche proposée, fondée sur la minimisation de la somme pondérée des variances d'échantillonnage (erreurs quadratiques moyennes) des estimateurs prévus, avec les pondérations spécifiées de façon à refléter les priorités d'inférence. Nous l'appliquons pour commencer à l'estimation directe de paramètres au niveau du petit domaine. Puis, nous l'étendons afin d'intégrer l'objectif de production d'estimations nationales et, enfin, l'estimation composite à la section 3. La section 4, qui conclut l'article, contient une discussion.

La présente section se termine par une description de la notation utilisée dans la suite de l'article. Nous supposons que les paramètres de population au niveau du petit domaine θ_d, $d = 1, ..., D$, sont estimées par $\hat{\theta}_d$ avec des erreurs quadratiques moyennes (EQM) v_d respectives qui sont des fonctions des tailles des sous-échantillons dans les petits domaines n_d; $v_d = v_d(n_d)$. La taille globale de l'échantillon est dénotée par n et nous supposons qu'elle est fixe. Les tailles de population sont dénotées par N (globale) et N_d (pour le petit domaine d). Par souci de concision, nous dénotons $\mathbf{n} = (n_1, ..., n_D)^\mathsf{T}$. La plupart des paramètres de population θ sont des fonctions d'une seule variable, comme la moyenne, le total et ainsi de suite. La variable peut être enregistrée directement durant le sondage ou construite d'après une ou plusieurs variables directes. Bien que notre développement ne soit pas limité à ce genre de paramètres, la justification est plus simple en ce qui les concerne. Nous disons qu'un estimateur de θ_d est *direct* s'il s'agit d'une fonction de la variable étudiée sur les sujets du petit domaine d seulement.

Nous supposons que chaque estimateur direct envisagé est sans biais. Cette hypothèse n'est pas particulièrement restrictive, car la plupart des estimateurs directs sont des estimateurs naïfs ou étroitement reliés à ces derniers. Nous supposons que les tailles d'échantillon pour les petits domaines sont sous le contrôle du concepteur de l'enquête. Il en est ainsi pour les plans d'échantillonnage stratifiés dans lesquels les strates coïncident avec les petits domaines. À la section 4, nous discutons des plans d'échantillonnage pour lesquels ce genre de contrôle ne peut être exercé; ces plans sont particulièrement indiqués pour la subdivision du pays en un grand nombre (centaines) de petits domaines.

2. Plan optimal pour l'estimation directe

Nous résolvons le conflit entre les objectifs d'estimation efficace de paramètres au niveau du petit domaine θ_d en choisissant le plan d'échantillonnage à ce niveau qui minimise la somme pondérée des variances d'échantillonnage (EQM),

$$\min_{\mathbf{n}} \sum_{d=1}^{D} P_d v_d, \qquad (1)$$

sachant que la taille globale d'échantillon $n = \mathbf{n}^\mathsf{T} \mathbf{1}_D$ est fixe; $\mathbf{1}_D$ est le vecteur des unités de longueur D. Le coefficient P_d est nommé *priorité d'inférence*. Une valeur plus grande de P_d (par rapport aux valeurs $P_{d'}$, $d' \neq d$) implique qu'il est plus important de réduire v_d, parce que l'augmentation de la contribution du petit domaine d à la somme (1) est plus importante que pour les autres petits domaines.

Le problème d'optimisation (1) est résolu par la méthode des multiplicateurs de Lagrange, ou simplement par substitution de $n_1 = n - n_2 - ... - n_D$, si bien qu'il comporte alors $D-1$ variables fonctionnellement non corrélées. La solution satisfait la condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

En général, il n'est pas possible d'obtenir une expression analytique des tailles optimales des sous-échantillons n_d, mais si $v_d = \sigma_d^2 / n_d$, comme dans le cas de l'échantillonnage aléatoire simple à l'intérieur des petits domaines, la solution est proportionnelle à $\sigma_d \sqrt{P_d}$, c'est-à-dire

$$n_d^\dagger = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + ... + \sigma_D \sqrt{P_D}}.$$

Lorsque les variances intra domaine σ_d^2 sont égales, $\sigma_1^2 = ... = \sigma_D^2 = \sigma^2$, la solution se simplifie encore davantage; les tailles optimales d'échantillon sont proportionnelles à $\sqrt{P_d}$ et ne dépendent pas de σ^2.

Dans la plupart des contextes, il est difficile d'exprimer un ensemble approprié de priorités P_d et il est donc plus constructif de proposer une classe paramétrique commode de priorités $\mathbf{P} = (P_1, ..., P_D)^\mathsf{T}$ et d'illustrer son effet sur la répartition de la taille de l'échantillon. Nous proposons les priorités $P_d = N_d^q$ pour $0 \leq q \leq 2$. Si $q = 0$, l'inférence est de même importance pour chaque petit domaine. À mesure

que q augmente, une importance relativement plus grande est accordée aux petits domaines les plus peuplés. Lorsque $v_d = \sigma^2 / n_d$, la répartition optimale de la taille de l'échantillon pour $q = 2$, $n_d^{\dagger} = n N_d / N$ est proportionnelle aux tailles de population dans les petits domaines et le même plan d'échantillonnage est donc optimal pour les inférences calculées au niveau national et du petit domaine. Pour $q > 2$, la répartition de la taille de l'échantillon est encore plus généreuse à l'égard des petits domaines les plus peuplés, aux dépens de ceux qui le sont moins. Comme cette situation est contre-intuitive dans le contexte de l'estimation pour petits domaines, le choix d'un exposant $q > 2$ n'est probablement jamais approprié. Un exposant de priorité q négatif conviendrait pour une enquête dont le but est de se concentrer sur les petits domaines les moins peuplés. Naturellement, ce genre de plan est très inefficace pour l'estimation du paramètre θ de niveau national, surtout si les tailles de population des petits domaines sont très dispersées.

Les priorités d'inférence P_d peuvent être des fonctions d'autres paramètres que N_d. Par exemple, les tailles de certaines sous-population présentant un intérêt particulier, comme une minorité ethnique dans le petit domaine, peuvent être utilisées au lieu de N_d, P_d peut être défini différemment dans les diverses régions du pays, ou bien la formule pour le calculer peut-être outrepassée pour un petit domaine ou quelques-uns d'entre eux.

Dans certains rapports d'analyse de données d'enquête, une estimation n'est publiée que si elle est fondée sur un échantillon de taille suffisamment grande ou que son coefficient de variation (le ratio de l'erreur-type estimée à l'estimation) est inférieur à un seuil spécifié. Si une « pénalité » associée au fait de ne pas publier un paramètre est précisée, elle peut être intégrée dans la définition des priorités d'inférence. La difficulté qui risque de se poser est que la fonction objectif (1) soit discontinue et que l'on ne puisse plus appliquer les approches standard d'optimisation. La pénalité doit être déterminée minutieusement. Si elle est trop faible, elle est inefficace; si elle est trop élevée, la solution accordera la préférence à la publication d'estimations pour un aussi grand nombre de petits domaines que possible, mais avec, pour chacun, une taille d'échantillon ou une précision qui n'excède que de justesse le seuil fixé. Voir Marker (2001) pour une autre approche de ce problème.

La figure 1 illustre l'effet de l'exposant de priorité q sur la répartition de la taille de l'échantillon d'une enquête planifiée en Suisse dans le but d'estimer les moyennes de population d'une variable dans les 26 cantons, en supposant qu'ils ont tous la même variance intracanton σ^2. La taille globale prévue de l'échantillon est $n = 10\,000$. Dans n'importe quel volet, les courbes relient les tailles d'échantillon optimales pour chaque exposant q; elles sont tracées sur l'échelle linéaire (à gauche) et sur l'échelle logarithmique (à droite). Les tailles de population sont inscrites sur la barre horizontale au bas de chaque graphique. Sur l'échelle logarithmique, les courbes sont linéaires. Cette échelle produit aussi une répartition plus uniforme des tailles de population des cantons.

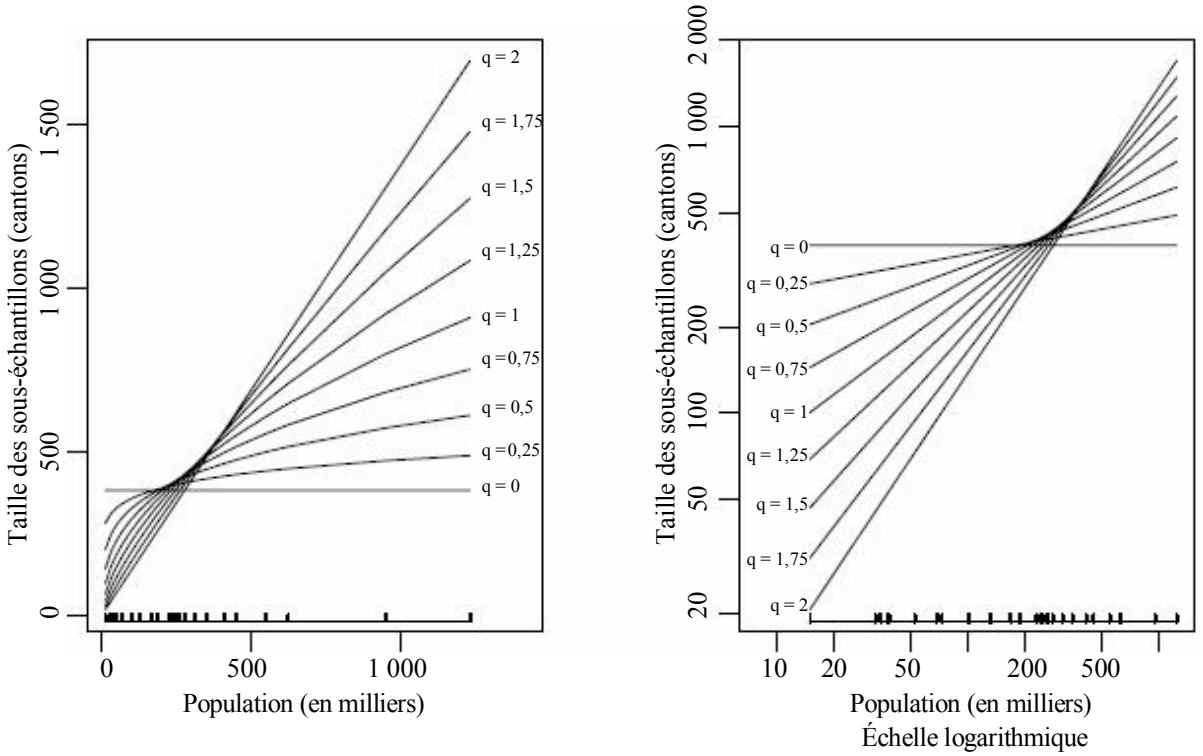

Figure 1. Répartition de la taille de l'échantillon entre les cantons suisses pour une gamme d'exposants de priorité q. Les tailles de population des cantons sont inscrites sur la barre horizontale au bas de chaque graphique.

Pour $q = 0$, une même taille d'échantillon est attribuée à chaque canton, soit $10\,000/26 = 385$, et pour $q = 2$, la répartition est proportionnelle à la taille de population du canton. Pour les valeurs intermédiaires de q, les tailles d'échantillon des cantons les moins peuplés sont augmentées par rapport à la répartition proportionnelle $(q = 2)$, au prix de l'attribution d'une taille réduite aux cantons les plus peuplés. Pour les cantons dont la population est supérieure à 250 000, environ 3 % du chiffre national de population, la taille des sous-échantillons dépend fort peu de la valeur de q.

2.1 Priorité accordée à l'estimation nationale

Comme les tailles de sous-échantillon au niveau du canton diffèrent de la répartition proportionnelle pour l'exposant de priorité $q < 2$, l'estimation optimale au niveau du canton est assortie d'une perte d'efficacité de l'estimateur national. Considérons l'estimateur stratifié

$$\hat{\theta} = \frac{1}{N}\sum_{d=1}^{D} N_d \hat{\theta}_d$$

de la moyenne nationale θ d'une variable, où $\hat{\theta}_d$ représente les estimateurs sans biais des moyennes intracanton de la même variable. En supposant que l'échantillonnage est stratifié avec échantillonnage aléatoire simple dans les strates (cantons) et que la valeur de $\hat{\theta}_d$ est fixée à la moyenne d'échantillon intrastrate,

$$\mathrm{var}(\hat{\theta}) = \frac{1}{N^2}\sum_{d=1}^{D} \frac{N_d^2}{n_d}(1 - f_d)\sigma_d^2,$$

où $f_d = n_d / N_d$ est la correction pour population finie.

La figure 2 représente la fonction qui relie l'erreur-type $\sqrt{\mathrm{var}(\hat{\theta})}$ à l'exposant de priorité q, calculée en supposant que $\sigma^2 = 100$. L'erreur-type est une fonction décroissante de q; elle diminue plus rapidement à $q = 0$ qu'à $q = 2$, où elle est relativement constante. Pour $q = 2$, les objectifs d'estimation au niveau du canton et au niveau national concordent, et $\sqrt{\mathrm{var}(\hat{\theta})} = 0{,}100$. Pour $q = 0$, $\sqrt{\mathrm{var}(\hat{\theta})} = 0{,}143$; dans ces conditions, l'optimalité de l'estimation pour petits domaines a sur l'estimation nationale un effet défavorable important, équivalant à la réduction de moitié de la taille de l'échantillon $(0{,}143/0{,}100 \doteq \sqrt{2})$. Pour une valeur négative de q, cet effet est encore plus prononcé.

Donc, nous pouvons répondre au besoin d'efficacité de l'estimateur national en augmentant la valeur de l'exposant de priorité. Par exemple, les parties ayant des intérêts concurrents en matière d'inférence pourraient négocier la perte d'efficacité de $\hat{\theta}$ qu'elles jugent acceptables et fixer ensuite l'exposant de priorité de façon à égaler cette perte. Ou bien, la perte pourrait être prise en considération lors de l'application du plan d'échantillonnage optimal pour l'estimation au niveau du petit domaine. Si elle est jugée excessive, q pourrait être augmenté jusqu'à l'obtention d'un équilibre entre les pertes d'efficacité de l'estimation nationale et de celle sur petits domaines.

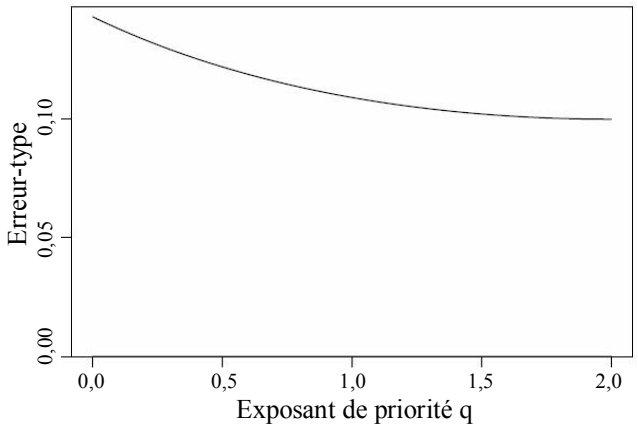

Figure 2. Erreur-type de l'estimateur national $\hat{\theta}$ de la moyenne d'une variable, sous forme de fonction de l'exposant q pour les priorités de l'estimation au niveau du canton.

Un aspect insatisfaisant de ces approches est qu'elles compromettent l'objectif premier des priorités **P**, c'est-à-dire refléter l'importance relative des inférences au sujet de petits domaines distincts. Pour contourner cet inconvénient, nous associons $\hat{\theta}$ à une priorité, dénotée G, relative à une estimation pour petits domaines, et nous considérons l'estimation optimale de l'ensemble de D paramètres cibles au niveau du petit domaine θ_d en même temps que le paramètre cible nationale θ. Donc, nous minimisons la fonction objectif

$$\sum_{d=1}^{D} P_d v_d(n_d) + GP_+ v(\mathbf{n}),$$

où $v = \mathrm{var}(\hat{\theta})$ et $P_+ = \mathbf{P}^{\mathsf{T}}\mathbf{1}_D$. Le facteur P_+ est introduit pour améliorer l'effet des tailles absolues de P_d et du nombre de petits domaines sur la priorité relative G. Les priorités P_d peuvent être interprétées uniquement d'après leurs tailles relatives, car, pour toute constante $c > 0$, P_d et cP_d correspondent à des ensembles identiques de priorités pour l'estimation pour petits domaines dans (1).

Lorsque le plan d'échantillonnage dans chaque petit domaine est aléatoire simple et que $\hat{\theta}$ est l'estimateur stratifié standard, le minimum est atteint quand

$$\sigma_d^2 \frac{P_d'}{n_d^2} = \mathrm{const},$$

où $P_d' = P_d + GP_+ N_d^2 / N^2$. Les tailles optimales d'échantillon pour les petits domaines sont

$$n_d^* = n\frac{\sigma_d\sqrt{P_d'}}{\sigma_1\sqrt{P_1'} + \ldots + \sigma_D\sqrt{P_D'}}.$$

Cette solution correspond à un ajustement des priorités P_d par $GP_+N_d^2/N^2$. Notons que cet ajustement n'est ni additif, ni multiplicatif. L'accroissement de la priorité est plus important pour les petits domaines plus peuplés. Par conséquent, les tailles des sous-échantillons de petit domaine sont réduites davantage quand la priorité relative de l'estimation nationale est intégrée et que les priorités au niveau des petits domaines ne changent pas. La correction pour population finie n'a aucun effet sur n_d^*, parce qu'elle réduit chaque variance d'échantillonnage v_d et v d'une quantité qui ne dépend pas de **n**.

La priorité G peut être fixée en insistant sur le fait que la perte d'efficacité lors de l'estimation de la grandeur nationale θ n'excède pas un pourcentage donné ou qu'au plus, quelques-uns seulement des écarts absolus $|P_d' - P_d|$ ou des logarithmes des ratios $|\log(P_d'/P_d)|$ (voire aucun) ne soient très grands. Cependant, le problème analytique est facile à résoudre, de sorte que la gestion de l'enquête peut être présentée au moyen des plans d'échantillonnage qui sont optimaux pour une gamme de valeur G.

La variation de la taille des sous-échantillons en fonction de l'exposant q et de la priorité relative G est représentée graphiquement à la figure 3 pour les cantons le moins et le plus peuplés, Appenzell-Innerrhoden et Zürich, dans les volets A et C. Les volets B et D donnent la représentation des mêmes courbes qu'A et C, respectivement, sur l'échelle logarithmique. Ne pas tenir compte de l'objectif de production d'une estimation nationale correspond au cas où $G = 0$ et ne pas tenir compte de l'objectif de production d'une estimation pour petits domaines correspond au cas des valeurs très grandes de G. Tout au long de l'article, nous supposons que $n = 10\,000$ et que $\sigma^2 = 100$ pour tous les cantons.

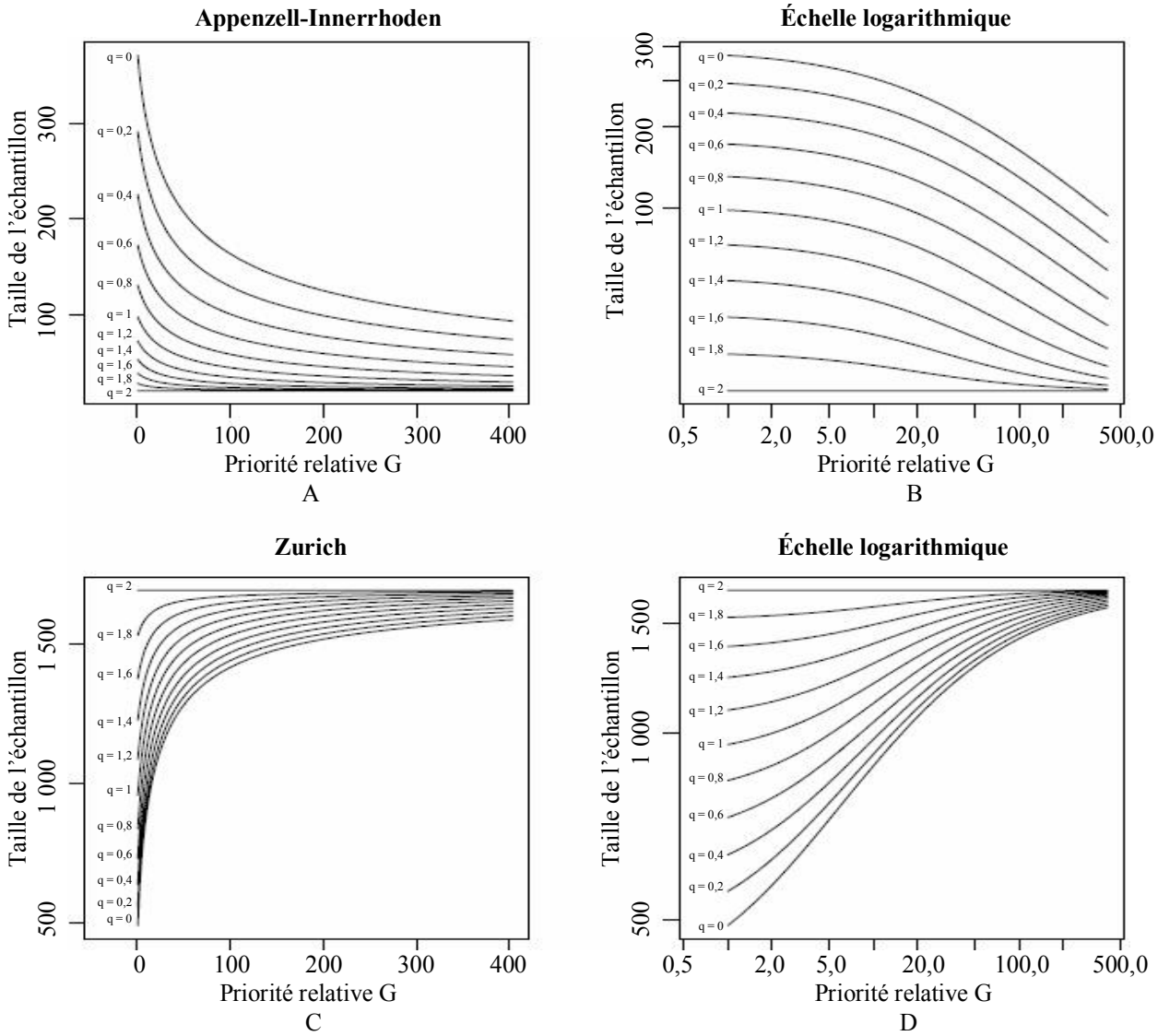

Figure 3. Tailles d'échantillon optimales pour l'estimateur direct $\hat{\theta}_d$ pour les combinaisons d'exposants de priorité q et de priorités relatives G pour les cantons le moins et le plus peuplés.

Dans le cas de chaque exposant $q < 2$, la courbe de répartition de la taille de l'échantillon $n_d(G)$ montre une diminution pour les cantons les moins peuplés et une augmentation pour les plus peuplés en direction de la représentation proportionnelle, $n_d = nN_d / N$, qui correspond à $q = 2$. Sur l'échelle linéaire, l'augmentation est assez rapide pour Zürich pour les faibles valeurs de q et de G, tandis que la réduction pour Appenzell-Innerrhoden est plus progressive. À mesure que la priorité relative G est réduite, la taille d'échantillon excédentaire est réaffectée de Zürich (et de quelques autres cantons peuplés) à plusieurs cantons moins peuplés.

La figure 4 représente graphiquement l'erreur-type « nationale » $\sqrt{\text{var}(\hat{\theta})}$ sous la répartition optimale de l'échantillon pour une matrice de valeurs de q et de G. Le graphique montre qu'une légère augmentation de G aux alentours de $G = 0$ réduit spectaculairement l'erreur-type de $\hat{\theta}$, tandis que pour les valeurs plus grandes de G, l'erreur-type ne varie que légèrement. Pour chaque G, un exposant de priorité plus élevé q est associé à une précision plus élevée de $\hat{\theta}$.

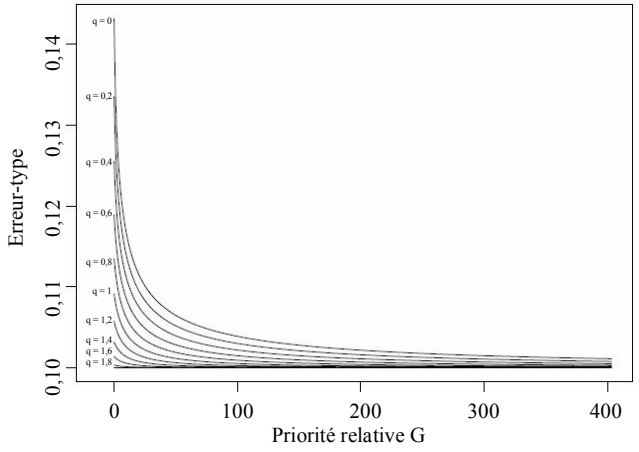

Figure 4. Erreur-type de l'estimateur national pour la répartition optimale sous une matrice de priorités données par q et G.

3. Estimation composite

L'utilisation la plus efficace des ressources disponibles pour réaliser une enquête s'obtient par combinaison optimale d'un plan d'échantillonnage et d'un ou de plusieurs estimateurs, si bien que le plan d'échantillonnage et (le choix de) l'estimateur devraient, dans des circonstances idéales, être optimisés simultanément. Ce problème est difficile à résoudre formellement dans la plupart des conditions, quoique certains estimateurs soient plus efficaces que leurs concurrents et que l'on considère une grande gamme de plans d'échantillonnage. Les estimateurs composites (Longford 1999, 2004) représentent l'une de ces classes d'estimateurs. Il s'agit de combinaisons convexes des estimateurs directs sur petits domaines et au niveau national,

$$\tilde{\theta}_d = (1 - b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}, \qquad (2)$$

avec des coefficients particuliers aux petits domaines b_d qui sont des estimations de l'optimum. La composition $\tilde{\theta}_d$ tire parti de la similarité des petits domaines et est particulièrement efficace lorsqu'ils présentent une faible variance interdomaines $\sigma_B^2 = D^{-1}\sum_d (\theta_d - \bar{\theta})^2$, où $\bar{\theta} = D^{-1}\sum_d \theta_d$. Cette variance est définie sur les D paramètres de population θ_d et n'est pas affectée par le plan d'échantillonnage. En pratique, il faut estimer σ_B^2. Lors de la planification d'une enquête, il est nécessaire d'utiliser des estimations provenant d'autres enquêtes auprès de la même population ou de populations apparentées, et de tenir compte de l'incertitude au sujet de σ_B^2, ce qui peut se faire par analyse de sensibilité, en recherchant les plans d'échantillonnage optimaux pour une gamme de valeurs plausibles de σ_B^2.

Si les écarts $\Delta_d = \theta_d - \bar{\theta}$ étaient connus, le coefficient optimal b_d dans (2) serait, approximativement, $b_d^* = \sigma_d^2 / (\sigma_d^2 + n_d \Delta_d^2)$. Puisque nous ne connaissons pas Δ_d (sinon, θ_d serait estimé avec une grande précision par $\bar{\theta} + \Delta_d$), nous remplaçons Δ_d^2 par sa moyenne sur les petits domaines, égale à σ_B^2, ce qui donne le coefficient $b_d = 1/(1 + n_d \omega_d)$, où $\omega_d = \sigma_B^2 / \sigma_d^2$ est le ratio de variance. La variance σ_B^2 doit aussi être estimée, mais, si le nombre de petits domaines est élevé, l'estimation est beaucoup plus précise que ne le sont la plupart des Δ_d^2.

Si les coefficients b_d sont estimés avec suffisamment de précision, l'estimateur composite $\tilde{\theta}_d$ est plus efficace que les deux estimateurs qui le constituent, $\hat{\theta}_d$ et $\hat{\theta}$. Si nous ne tenons pas compte de l'incertitude au sujet des variances intra et interdomaines, ni au sujet de la moyenne nationale $\bar{\theta}$ et de la corrélation entre les estimateurs (direct) au niveau national et sur petits domaines, l'EQM moyenne de $\tilde{\theta}_d$ est

$$\text{aEQM}(\tilde{\theta}_d) = \frac{\sigma_B^2}{1 + n_d\,\omega_d}, \qquad (3)$$

où « aEQM » dénote l'EQM dans laquelle Δ_d^2 est remplacé par σ_B^2, sa moyenne sur l'ensemble des petits domaines. Dans (3), aEQM est aussi une approximation de la variance conditionnelle de l'estimateur EBLUP de la moyenne au niveau du petit domaine fondée sur le modèle (empirique bayésien) à deux niveaux (Longford 1993, Goldstein 1995, Marker 1999 et Rao 2003). Voir Ghosh et Rao (1994) pour une revue reconnue de l'application de ces modèles à l'estimation pour petits domaines.

Pour les estimateurs composites des moyennes de petit domaine, nous recherchons la répartition de l'échantillon qui minimise la fonction objectif

$$\sum_{d=1}^{D} P_d \, \mathrm{aEQM}(\tilde{\theta}_d) + GP_+ v.$$

La solution satisfait la contrainte

$$\frac{N_d^q \sigma_B^2 \omega_d}{(1+n_d\omega_d)^2} + GP_+ \frac{N_d^2}{N^2}\frac{\sigma_d^2}{n_d^2} = \text{const.} \qquad (4)$$

Cette équation ne possède pas de solution analytique commode, mais elle peut être résolue par application de scénarios itératifs. La valeur de n_1 détermine les autres tailles d'échantillon n_d, de sorte que l'optimisation correspond à une recherche unidimensionnelle. Si les tailles d'échantillon provisoires $\mathbf{n}$ fondées sur un ensemble de valeur de n_1 sont trop grandes, on réduit $\mathbf{n}^\top \mathbf{1}_D > n$, n_1 et on calcule les autres tailles d'échantillon n_d en résolvant (4). Notons que la solution dépend des variances σ_d^2 et σ_B^2. Le problème se simplifie quelque peu lorsque la variance est la même pour tous les petits domaines $\sigma^2 = \sigma_1^2 = ... = \sigma_D^2$. Alors, la solution de (4) dépend des variances uniquement par la voie du ratio $\omega = \sigma_B^2/\sigma^2$, parce que σ^2 est un facteur multiplicatif qui n'a aucun effet sur l'optimisation.

À titre d'exemple, supposons que $q = 1$ et $G = 10$ lors de la planification d'une enquête auprès de la population suisse, avec $n = 10\,000$, et en supposant que $\omega = 0{,}10$. Comme solution initiale, nous utilisons la répartition optimale pour l'estimation directe avec les mêmes valeurs de q et de G. Une itération met à jour la taille de l'échantillon de chaque canton et, dans les cantons, la mise à jour pour tous, sauf celui de référence sélectionné arbitrairement $d = 1$, est également itérative. La taille provisoire du sous-échantillon pour le canton de référence détermine la valeur courante de la constante dans le deuxième membre de (4). L'équation (4) est résolue, itérativement, pour chaque canton $d = 2, ..., D$, en utilisant la méthode de Newton. Dans l'application, le nombre d'itérations était inférieur à dix pour chaque canton. Enfin, la taille du sous-échantillon pour le canton de référence est ajustée par le facteur $1/D$ un multiple de la différence entre le total courant des tailles des sous-échantillons et le total cible n. La mise à jour des tailles d'échantillon des cantons est elle-même itérée, mais quelques itérations seulement sont nécessaires pour atteindre la convergence; par exemple, toutes les variations des tailles des sous-échantillons étaient inférieures à 1,0 après trois itérations et inférieures à 0,01 après huit itérations. La convergence est rapide, parce que la solution de départ est proche de la solution optimale; l'écart le plus important entre les deux tailles de sous-échantillon est celui observé pour Zurich, soit 20,0 (de 1199,5 au départ à 1219,5 après huit itérations). Pour Appenzell-Innerrhoden, la taille d'échantillon est réduite de 81,6 à 73,4. Des changements de moins d'une unité ont lieu pour cinq cantons dont la taille de population varie de 228 000 à 278 000. Notons qu'en pratique, les tailles des sous-échantillons seraient arrondies et éventuellement ajustées davantage afin de satisfaire aux diverses contraintes de gestion de l'enquête.

Pas de priorité accordée à l'estimation nationale

Si l'estimation nationale n'a aucune priorité, $G = 0$, l'équation (4) possède la solution explicite

$$n_d^* = \frac{n\omega + D}{\omega}\frac{N_d^{q/2}}{U^{(q)}} - \frac{1}{\omega},$$

où $U^{(q)} = N_1^{q/2} + ... + N_D^{q/2}$. Cette répartition est reliée à la répartition $n_d^\dagger$, $d = 1, ..., D$, qui est optimale pour l'estimation directe de θ_d, par l'identité

$$n_d^* = n_d^\dagger + \frac{1}{\omega}\left(\frac{DN_d^{q/2}}{U^{(q)}} - 1\right).$$

Donc, quand $q > 0$, la répartition optimale est plus dispersée dans le cas de l'estimation composite que dans celui de l'estimation directe. La taille de population au point d'équilibre est $N_T = (U^{(q)}/D)^{2/q}$; la taille du sous-échantillon pour les petits domaines ayant une taille de population $N_d < N_T$ est plus petite dans le cas de l'estimation composite que dans celui de l'estimation directe, et elle est plus grande pour les petits domaines dont la population est plus grande. (Pour $q = 0$, $n_d^* \equiv n/D$). Le degré de dispersion supplémentaire est inversement proportionnel à ω.

Si $\omega = 0$, les équations pour le plan d'échantillonnage optimal donnent lieu à une singularité. Dans ce cas, chaque θ_d est estimé efficacement par l'estimateur national $\hat{\theta}$, si bien que le plan optimal pour l'estimation composite coïncide avec le plan optimal pour l'estimateur national ($n_d^* = nN_d/N$). Pour $q > 0$, la répartition optimale donne des tailles d'échantillon négatives n_d^* quand

$$N_d < \left\{\frac{U^{(q)}}{n\omega + D}\right\}^{2/q}. \qquad (5)$$

Cette solution (formelle) n'a pas de sens. Une solution négative ne devrait pas être étonnante, car l'aEQM de (3) est une fonction analytique pour $n_d > -1/\omega_d$. Si les valeurs de $\omega > 0$ sont faibles, l'aEQM est une fonction décroissante à pente faible de la taille d'échantillon n_d. Une valeur négative de n_d^* indique qu'un « petit » canton ne vaut pas la peine d'être échantillonné, à cause de la faible priorité d'inférence P_d. Bien que l'accroissement de la taille de l'échantillon d'un canton plus peuplé d' puisse donner lieu à une réduction plus faible de l'aEQM que cela ne serait le cas pour un petit canton d, la priorité plus grande $P_{d'}$ augmente l'effet.

Priorité positive pour la moyenne nationale

Dans (3), l'aEQM ne tient pas compte de l'incertitude au sujet de la moyenne nationale θ, situation qui devient

critique lorsque l'un des cantons n'est pas représenté dans l'échantillon. Cette déficience de (3) peut être compensée en fixant la priorité relative G à une valeur positive.

La figure 5 résume l'effet de la priorité relative G et de l'exposant de priorité q sur les tailles d'échantillon optimales pour les cantons le moins et le plus peuplés, ainsi que le canton de Thurgau qui possède la 13e taille de population par ordre décroissant (médiane), soit 228 000. Chaque valeur de q, indiquée dans le titre, et de G, indiquée en utilisant différents types de lignes, est représentée pour un canton par un graphique de la taille d'échantillon optimale en fonction du ratio de variance ω. La limite de cette fonction lorsque $\omega \to +\infty$, égale à la taille d'échantillon optimale pour l'estimation directe, est marquée par une barre dans la marge de droite du volet en question. Pour $\omega = 0$, on obtient le plan d'échantillonnage optimal pour l'estimation de la moyenne nationale θ. Les volets A et B au haut de la figure correspondent à la taille d'échantillon globale $n = 10\,000$ et les volets C et D, à $n = 1\,000$.

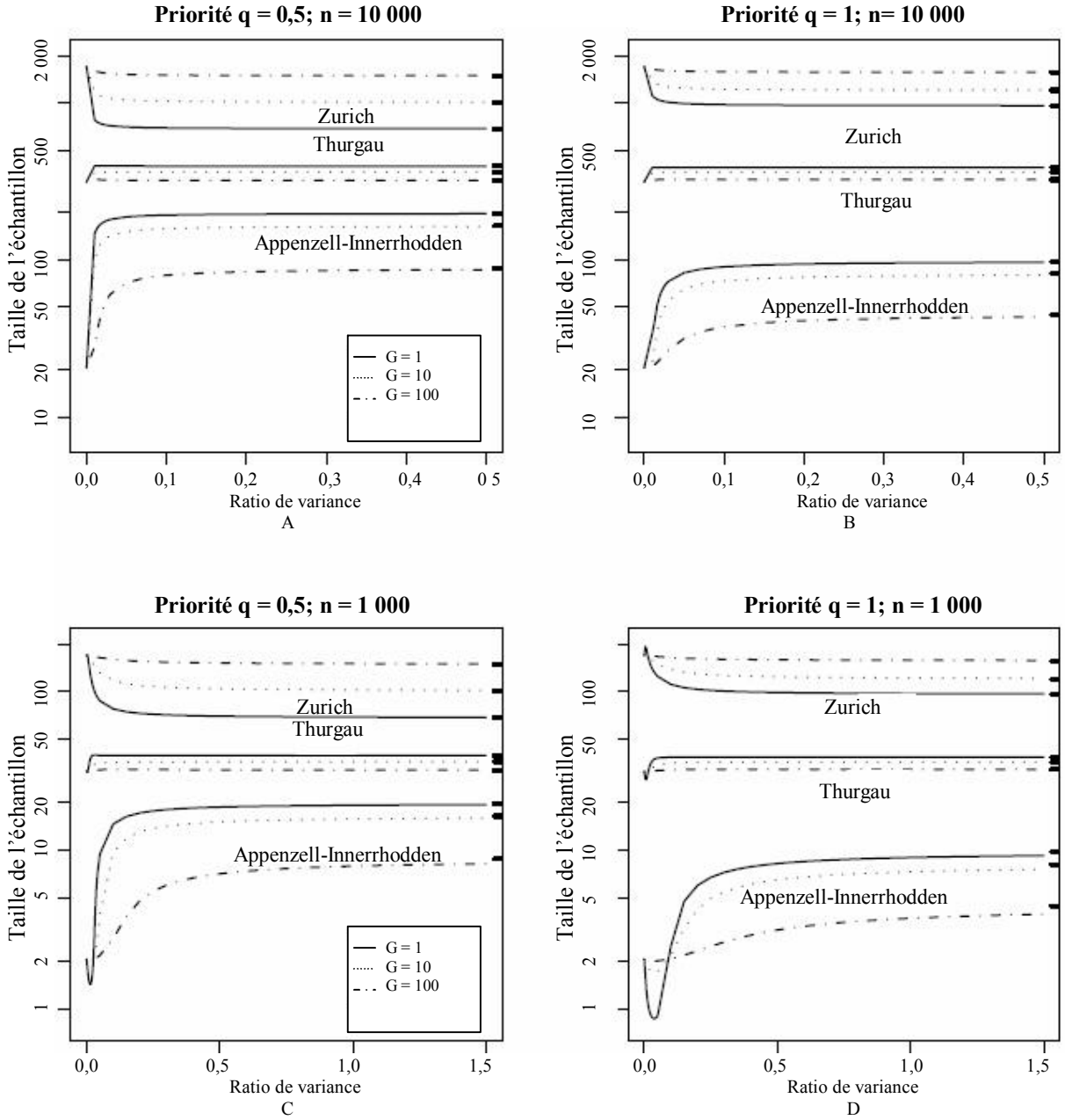

Figure 5. Tailles d'échantillon optimales pour l'estimation composite des moyennes de population pour trois cantons pour une gamme de rapports de variance ω, les exposants de priorité $q = 0,5$ et $q = 1,0$ et les priorités relatives $G = 1$, 10 et 100. Les tailles globales d'échantillon sont 10 000 (volets A et B) et 1 000 (volets C et D).

Le graphique montre que les tailles d'échantillon optimales sont presque constantes dans la fourchette $\omega \in (\omega^*, +\infty)$; ω^* augmente avec q, G et $1/n$. Il s'agit d'une conséquence de la taille d'échantillon relativement grande n, qui assure que les sous-échantillons de la plupart des cantons soient trop grands pour qu'un emprunt important d'information entre les cantons aient lieu, à moins que les cantons soient fort semblables $(\omega < \omega^*)$. La plupart des coefficients de rétrécissement $b_d = 1/(1 + n_d\,\omega)$ sont très petits. Lorsqu'une taille $n = 10\,000$ est prévue, pour les valeurs faibles de ω, la taille d'échantillon optimale augmente fortement pour les cantons les moins peuplés et chute brusquement pour les plus peuplés. La dispersion des tailles d'échantillon optimales augmente avec q et G, convergeant vers la répartition optimale pour l'estimation de la moyenne nationale θ, qui correspond à $\omega = 0$. Par contre, les tailles d'échantillon optimales sont discontinues à $\omega = 0$ quand $G = 0$; les solutions divergent vers $-\infty$ pour les cantons les moins peuplés.

Dans les volets C et D, pour $n = 1\,000$, la variation des tailles d'échantillon en fonction de ω persiste pour une plus grande fourchette de valeur de ω, parce que la portée de l'emprunt d'information entre les cantons est plus grande pour les tailles d'échantillon plus petites. Les tailles d'échantillon optimales ne sont pas des fonctions monotones de ω; pour les cantons les moins peuplés, on observe un creux pour les faibles valeurs de ω.. Le creux est plus prononcé pour les faibles valeurs de G et pour les grandes valeurs de q, c'est-à-dire lorsque les disparités entre les priorités des cantons sont grandes et que l'importance relative de l'inférence au sujet de la moyenne nationale est plus faible. Ce phénomène, quelque peu exagéré par l'échelle logarithmique de l'axe des ordonnées, est semblable au cas discuté pour $G = 0$. À cause des différences de priorité P_d, une faible réduction de l'aEQM pour un canton plus peuplé est préférable à une réduction plus importante pour un canton moins peuplé. Le creux existe aussi quand $n = 10\,000$, mais il est si peu profond et si étroit qu'il n'est pas visible dans les conditions de résolution du graphique. Notons que, dans les volets C et D, l'axe des abscisses possède une fourchette de valeurs de ω trois fois plus grande que dans les volets A et B.

Dans le contexte de l'enquête planifiée, il a été convenu qu'il était peu probable que la valeur de ω soit inférieure à 0,05. Par conséquent, le calcul des tailles d'échantillon a pu être fondé sur l'estimateur direct.

4. Discussion

La méthode décrite dans le présent article permet de déterminer le plan d'échantillonnage optimal pour les conditions artificielles d'échantillonnage stratifié avec échantillonnage aléatoire simple dans des strates homoscédastiques. La spécification des priorités en ce qui concerne l'estimation pour petits domaines et l'estimation nationale est un élément essentiel de la méthode. En pratique, il peut être difficile de se mettre d'accord sur les priorités et certaines hypothèses peuvent être problématiques, en particulier celles de l'égalité des variances intrastrate et de l'échantillonnage aléatoire simple. La méthode peut être étendue à des estimateurs plus complexes, mais les valeurs de paramètres supplémentaires sont alors nécessaires. Une approche plus constructive consiste à considérer le plan d'échantillonnage optimal pour les conditions simplifiées en tant qu'approximation du plan d'échantillonnage qui est optimal pour les conditions plus réalistes. Même si le plan d'échantillonnage optimal était déterminé, il ne pourrait être appliqué littéralement, à cause des imperfections de la base de sondage et (éventuellement) de la non-réponse informative et non uniformément distribuée. Cependant, l'approche est applicable, en principe, à tout estimateur sur petits domaines pour lequel il existe une expression analytique exacte ou approximative de l'EQM. Cela inclut tous les estimateurs fondés sur les modèles bayésiens empiriques, auxquels l'estimateur composite est étroitement associé. Les poids de sondage peuvent être intégrés dans le calcul de la taille de l'échantillon s'ils sont connus ou que leurs distributions dans les petits domaines sont connues a priori, sous réserve de certaines approximations. Le calcul de la taille d'échantillon pour une grandeur (nationale) unique pose le même problème.

Bien que la solution numérique du problème pour l'estimation composite avec une priorité positive G soit simple et ne présente aucun problème de convergence, il est avantageux de disposer d'une solution analytique, afin de pouvoir étudier une gamme de scénarios. La proximité des solutions obtenues pour les estimations directe et composite donne à penser que la répartition optimale pour l'estimation directe pourrait également s'approcher de la situation optimale pour l'estimation composite avec des valeurs raisonnables de ω, disons, $\omega > 0,05$.

Diverses contraintes de gestion et d'organisation constituent un autre obstacle à l'application littérale d'un plan d'échantillonnage établi analytiquement. Dans les enquêtes-ménages, il est souvent préférable d'attribuer un quota (presque) complet d'adresses à chaque intervieweur, si bien que l'on accorde la préférence aux tailles d'échantillon qui sont des multiples du quota. Ces considérations et de nombreuses autres contraintes peuvent être intégrées dans le problème d'optimisation, quoiqu'elles soient souvent difficile à quantifier ou que le concepteur de l'enquête ne soit pas conscient de leur existence à cause d'une communication imparfaite. L'improvisation, après l'obtention d'un plan d'échantillonnage optimal pour des conditions plus

simples, pourrait être plus pratique. En outre, les priorités, ou l'opinion d'experts à leur sujet, peuvent évoluer au cours du temps, même pendant la réalisation de l'enquête et l'analyse des données. Les estimations associées à une erreur-type ou à un coefficient de variation supérieur à un seuil précisé sont souvent exclues des rapports analytiques. L'intention de les exclure peut être reflétée dans le calcul de la taille d'échantillon en considérant $\hat{\theta}$ comme étant l'estimateur de θ_d, c'est-à-dire en fixant l'EQM connexe à l'aEQM $\sigma_B^2 + \text{var}(\hat{\theta})$ correspondante ou à une autre (grande) valeur constante.

Bien que nous proposions une classe particulière de priorités pour les petits domaines, aucune difficulté conceptuelle ne se pose lorsque l'on utilise une autre classe. Elle pourrait dépendre de plusieurs grandeurs de population plutôt que de la taille de population uniquement. En principe, les priorités peuvent aussi être fixées individuellement pour les petits domaines, bien que cela ne soit pratique que si leur nombre est faible. Les priorités fondées sur la formule ou établies individuellement peuvent être combinées en ajustant les priorités, telles que $P_d = N_d^q$, pour quelques petits domaines afin de refléter leur rôle exceptionnel dans l'analyse.

Une analyse de sensibilité, en vue d'étudier les modifications du plan d'échantillonnage en fonction de diverses données d'entrée est essentielle à la compréhension de l'incertitude au sujet des paramètres estimés (le ratio de variance ω en particulier) et le caractère arbitraire, aussi limité qu'il soit, de l'établissement des priorités. Pour cela, il est préférable de disposer d'une solution analytiquement simple qui peut être exécutée de nombreuses fois, pour une gamme de conditions, plutôt qu'une solution plus complexe, dont les propriétés sont difficiles à étudier.

Les estimateurs composites multivariés exploitent la similarité non seulement entre les petits domaines, mais aussi entre les variables (auxiliaires), les périodes, les sous-populations, et ainsi de suite (Longford 1999 et 2005). L'aEQM de ces estimateurs dépend de la matrice de variances mise à l'échelle $\boldsymbol{\Omega}$, qui est le pendant multivarié de ω. Le calcul de la taille d'échantillon par cette méthode est difficile à appliquer directement, parce que, dans $\boldsymbol{\Omega}$, les variances et les covariances sont les unes et les autres essentielles à l'efficacité des estimateurs. Une approche plus constructive consiste à faire concorder la matrice $\boldsymbol{\Omega}$ avec un ratio ω qui peut être interprété comme étant la similarité des petits domaines après correction pour l'information auxiliaire, comme dans les méthodes bayésiennes empiriques.

Lorsque il est impossible d'exercer un contrôle sur les tailles d'échantillon affectées aux petits domaines, leur calcul demeure utile comme indication de la façon dont elles devraient être réparties *en moyenne*. En général, une réduction unitaire de la taille d'échantillon est associée à une perte plus importante de précision qu'un accroissement unitaire. Par conséquent, les plans dans lesquels la variance d'échantillonnage (estimée par rééchantillonnage) des tailles des sous-échantillons n_d (d fixé) est plus faible sont mieux adaptés à l'estimation pour petits domaines. Dans les plans d'échantillonnage où les grappes sont importantes, ces variances sont grandes parce que, dans le cas extrême, un petit domaine pourrait ne pas être représenté dans l'échantillon lors de certaines répliques et pourrait être surreprésenté plusieurs fois dans d'autres. En général, il est préférable d'utiliser de plus petites grappes pour l'estimation pour petits domaines, si cela n'augmente pas les coûts d'enquête et qu'il est possible de maintenir une taille globale d'échantillon fixe.

Remerciements

Je remercie le rédacteur en chef délégué et les examinateurs d'avoir proposé plusieurs améliorations, mais surtout de m'avoir fait découvrir une erreur dans une version antérieure du manuscrit. Je tiens aussi à mentionner mes discussions avec l'équipe polonaise du projet EURAREA.

Bibliographie

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Goldstein, H. (1995). *Multilevel Statistical Models.* Deuxième Édition. Edward Arnold, London, UK.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society*, Séries A, 162, 227-245.

Longford, N.T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society*, Séries A, 167, 341-373.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects. *Techniques d'enquête*, 27, 201-207.

Platek, R., Rao, J.N.K., Särndal, C.-E. et Singh, M.P. (Éds.) (1987). S*mall Area Statistics*. New York: John Wiley & Sons.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-23.