

Article

Rentabilité des enquêtes avec échantillonnage en grappes répétées

par Stanislav Kolenikov et Gustavo Angeles

Juin 2011



Rentabilité des enquêtes avec échantillonnage en grappes répétées

Stanislav Kolenikov et Gustavo Angeles¹

Résumé

Nous analysons l'efficacité statistique et économique de diverses enquêtes avec échantillonnage en grappes pour lesquelles la collecte des données est effectuée à deux périodes, ou vagues, consécutives. Dans le cas d'un plan à échantillons indépendants, un échantillon en grappes est tiré de manière indépendante à chacune des deux vagues. Dans le cas d'un plan à panel de grappes, les mêmes grappes sont utilisées aux deux vagues, mais le tirage des échantillons dans les grappes est effectué indépendamment aux deux périodes. Dans un plan à panel d'unités d'observation, les grappes ainsi que les unités d'observation sont retenues d'une vague de collecte des données à l'autre. En supposant que la structure de la population est simple, nous calculons les variances sous le plan ainsi que les coûts des enquêtes réalisées selon ces divers types de plan. Nous considérons d'abord l'estimation de la variation de la moyenne de population entre deux périodes et nous déterminons les répartitions d'échantillon optimales pour les trois plans étudiés. Nous proposons ensuite un cadre de maximisation de l'utilité emprunté à la microéconomie en vue d'illustrer une approche possible pour choisir le plan dans laquelle nous nous efforçons d'optimiser simultanément plusieurs variances. La prise en compte simultanée de plusieurs moyennes et de leurs variances a tendance à faire pencher la préférence du plan à panel d'unités d'observation vers les plans à panel de grappes et à échantillons indépendants plus simples si le mode de collecte de données par panel est trop coûteux. Nous présentons des exemples numériques qui illustrent comment un concepteur d'enquête pourrait choisir le plan efficace sachant les paramètres de population et les coûts de collecte des données.

Mots clés : Étude longitudinale ; échantillon en grappes ; DHS ; NHIS.

1. Introduction

Afin d'analyser la dynamique des phénomènes sociaux, comportementaux ou de santé des populations, les chercheurs et les responsables de l'élaboration des politiques doivent obtenir des renseignements sur les caractéristiques de la population à de multiples périodes. Les enquêtes à plan de sondage complexe sont la source d'information utilisée le plus fréquemment pour les grandes populations, telles que celle d'un pays dans son ensemble. Outre les aspects habituellement pris en considération dans les enquêtes ponctuelles, comme la stratification et la mise en grappes, d'autres éléments peuvent être importants dans les enquêtes dont les données sont recueillies pour deux périodes ou plus. Dans ce genre d'enquête, le coût total et l'erreur d'enquête totale sont influencés par le chevauchement entre les échantillons consécutifs, l'érosion (informative) de l'échantillon, les effets de la durée de la présence dans l'échantillon ou du conditionnement, et d'autres facteurs dynamiques.

En vue d'estimer les variations au moyen d'enquêtes répétées, il est souvent souhaitable que la corrélation temporelle entre les unités observées soient élevée, ce qui peut être réalisé en administrant le questionnaire de l'enquête aux mêmes unités d'échantillonnage et/ou d'observation. Dans les enquêtes longitudinales, une visite est faite aux mêmes unités d'observation (personnes, ménages) pendant plusieurs périodes, éventuellement pendant un nombre indéfini de périodes (Panel Study of Income Dynamics (PSID) des

États-Unis, British Household Panel Study (BHPS) et d'autres). Un recueil de renseignements sur les études longitudinales peut être consulté sur le site Web de l'Institute for Social and Economics Research, à l'adresse <http://iser.essex.ac.uk/ulsc/keeptrack/index.php>. Dans les enquêtes à panel rotatif, les unités d'observation sont recrutées dans l'échantillon pour quelques périodes, puis sortent de l'échantillon et sont de nouveau étudiées à une période ultérieure. La Current Population Survey (CPS) des États-Unis (Binder et Hidiroglou 1988, Eckler 1955, Rao et Graham 1964) et un certain nombre d'enquêtes environnementales (Fuller 1999, McDonald 2003, Scott 1998) sont des exemples d'enquête à panel rotatif. Une autre option consiste à utiliser les mêmes unités primaires d'échantillonnage (UPE) aux différentes vagues, mais à sélectionner indépendamment les unités d'observation (unités secondaires d'échantillonnage, USE). Les enquêtes dont les données sont recueillies de cette façon comprennent les Demographic and Health Surveys (DHS) internationales et la National Health Interview Survey (NHIS) des États-Unis.

Nous nous concentrerons sur les enquêtes dont les données sont recueillies à deux périodes, ou vagues, en utilisant un plan d'échantillonnage en grappes à deux degrés à chaque vague de collecte des données. Nous considérons trois plans de sondage possibles, qui se distinguent par l'importance et la profondeur du chevauchement des unités d'échantillonnage au cours du temps. Le concepteur de l'enquête peut se limiter à réduire tout effet éventuel du

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. Courriel : kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. Courriel : gangeles@insp.mx.

chevauchement des échantillons, et tirer deux échantillons indépendants à deux périodes distinctes. Nous donnons à ce plan le nom de plan à *échantillons indépendants*. Il pourrait aussi juger utile de recycler les UPE d'une vague à l'autre. S'il estime qu'il est difficile de suivre les USE d'une vague à l'autre, les sous-échantillons dans les grappes peuvent être tirés indépendamment aux deux vagues de collecte des données. Nous donnerons à ce type de plan le nom de plan à *panel de grappes*. Si une très grande précision est essentielle, le plan entièrement longitudinal visera à repérer tous les individus qui ont répondu à la première vague et à leur demander de participer à une deuxième interview. Afin de distinguer ce plan de celui à panel de grappes, nous le nommerons plan à *panel d'unités d'observation*.

Un aspect particulier qui nous semble est important dans la gestion des enquêtes, mais qui n'est pas souvent abordé dans la littérature existante, est le coût de mise en œuvre (Groves 1989). Les modèles de coût classiques, tel celui utilisé pour établir le plan de répartition optimale de l'échantillon de Neyman-Tchuprow (Neyman 1938), peuvent être étendus afin d'y inclure des termes de coût pour la première visite à la grappe et à l'unité d'observation finale, ainsi que pour les visites subséquentes. En ce qui concerne les grappes, le coût est vraisemblablement plus faible à la deuxième visite. Il n'est plus nécessaire de créer de nouvelles cartes ni d'établir des listes. Les mêmes intervieweurs peuvent être choisis pour effectuer les interviews lors des vagues subséquentes de collecte des données. La coopération avec les dirigeants communautaires, parfois importante, par exemple chez certaines sociétés traditionnelles, a déjà été établie. L'effet de la collecte des données par panel au niveau individuel est moins évident. Un ménage interviewé lors des vagues précédentes qui a déménagé doit éventuellement être dépisté dans une région géographique différente, ce qui accroît le coût (moyen) d'interview du panel. La probabilité que des circonstances de ce genre se présentent augmente dans le cas typique d'intervalles plus longs entre les vagues de l'enquête dans les pays en développement : les intervalles entre les vagues de la DHS sont habituellement de cinq à sept ans. En revanche, si un mode d'interview moins coûteux peut être utilisé après la première vague (par exemple, une interview téléphonique au lieu d'une visite sur place), le coût de l'interview du panel diminue.

Dans le présent article, nous tenons compte de considérations statistiques ainsi qu'économiques en vue de choisir le plan de sondage approprié et ses paramètres. Nous supposons que le concepteur de l'enquête souhaite estimer la variation de la moyenne de population entre deux périodes, et/ou les moyennes proprement dites. À la section 2, nous présentons une population sommaire et calculons les variances sous le plan des moyennes, ainsi que et la différence

entre ces moyennes, pour les trois plans d'échantillonnage étudiés.

Afin d'intégrer les aspects économiques de la collecte des données, à la section 3, nous présentons un modèle de coût relativement simple pour une enquête avec échantillonnage en grappes répétées. Nous définissons et résolvons les problèmes d'optimisation pour obtenir les tailles d'échantillon optimales pour les trois plans pris en considération. En introduisant dans le modèle les estimations des paramètres statistiques (variances et autocorrélations) et des composantes de coût (coûts au niveau de la grappe et au niveau individuel), le concepteur d'enquête peut comparer les valeurs numériques des variances pour choisir le meilleur plan. À la section 4, nous illustrons cette approche et montrons que chacun des plans pris en considération peut être le meilleur, selon les valeurs des paramètres. Les résultats intuitifs (par exemple, le coût plus élevé de la collecte des données et les autocorrélations plus faibles des caractéristiques observées rendent les modes de collecte des données par panel moins intéressants) sont assortis d'une justification analytique et d'un support quantitatif.

Les sections 2 à 4 traitent de l'efficacité de l'estimation de la différence entre les moyennes seulement, mais des objectifs de collecte des données plus réalistes comprendraient l'estimation de caractéristiques contemporaines et de leurs variances. À cette fin, à la section 5, nous présentons un cadre de maximisation de l'utilité qui décrit le choix de scénarios d'échantillonnage qui s'offre au concepteur de l'enquête. Ce cadre fournit une fonction objectif agrégée qui combine plusieurs critères d'élaboration du plan. Les résultats sont de nouveau ceux attendus : si les modes de collecte des données par panel plus coûteux produisent des tailles d'échantillon plus petites, les estimations des moyennes sont moins efficaces que dans le cas de plans d'échantillonnage plus simples. Le seul moyen de justifier cette perte d'efficacité est d'obtenir une amélioration radicale de l'estimation de la différence, qui ne peut se produire que si les autocorrélations sont plus fortes. Les effets de ce genre sont également illustrés à la section 5. À la section 7, nous présentons nos conclusions. Les preuves sont données en annexe.

2. Variances sous le plan

Posons que la population est constituée de N grappes, ou UPE, aux deux périodes et que chaque grappe est constituée de M individus, ou USE. De cette population est tiré un EAS de $1 < n_t \leq N$ grappes au temps $t = 1, 2$, et un EAS de $1 < m_t \leq M$ individus dans chaque grappe présente dans l'échantillon au temps t . Soit l'indice i désignant les UPE et l'indice j désignant les USE. Donc, la mesure typique sera désignée Y_{ij} dans la population et

y_{ij} dans l'échantillon. Les totaux de population $T[\cdot]$ et leurs estimations $t[\cdot]$ peuvent être calculés comme il suit :

total de grappe :

$$T_{i\cdot}[Y] = \sum_{j=1}^M Y_{ij}, \quad t_{i\cdot}[y] = \frac{M}{m} \sum_{j=1}^M y_{ij},$$

total de population :

$$T_{\cdot\cdot}[Y] = \sum_{i=1}^N Y_{i\cdot}, \quad t_{\cdot\cdot}[y] = \frac{N}{n} \sum_{i=1}^N t_{i\cdot}[y]. \quad (2.1)$$

Les moyennes pour les unités d'observation sont

$$\begin{aligned} \bar{Y}_{i\cdot} &= \frac{1}{M} \sum_{j=1}^M Y_{ij} = \frac{T_{i\cdot}[Y]}{T_{i\cdot}[1]}, \quad \bar{y}_{i\cdot} = \frac{1}{m} \sum_{j=1}^M y_{ij} = \frac{t_{i\cdot}[y]}{t_{i\cdot}[1]}, \\ \bar{Y}_{\cdot\cdot} &= \frac{T_{\cdot\cdot}[Y]}{T_{\cdot\cdot}[1]} = \frac{\sum_{i=1}^N \sum_{j=1}^M Y_{ij}}{NM}, \quad \bar{y}_{\cdot\cdot} = \frac{t_{\cdot\cdot}[y]}{t_{\cdot\cdot}[1]} = \frac{\sum_{i=1}^N \sum_{j=1}^M y_{ij}}{nm}. \end{aligned} \quad (2.2)$$

La variance de Y et ses composantes intragrappe (indiquée par l'indice w pour *within*) et intergrappe (indiquée par b pour *between-cluster*) sont

$$S_t^2 = \frac{\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_{i\cdot})^2}{NM - 1}, \quad (2.3)$$

$$S_{twi}^2 = \frac{\sum_{j=1}^M (Y_{ij} - \bar{Y}_{i\cdot})^2}{M - 1}, \quad \bar{S}_{tw}^2 = \frac{1}{N} \sum_{i=1}^N S_{twi}^2, \quad (2.4)$$

$$S_{tb}^2 = \frac{\sum_{i=1}^N (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{N - 1}. \quad (2.5)$$

La caractéristique d'intérêt principal est la variation des moyennes,

$$D = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}, \quad (2.6)$$

estimée par

$$d = \bar{y}_{2\cdot} - \bar{y}_{1\cdot}. \quad (2.7)$$

Une propriété de cet estimateur intéressante pour les analystes et les utilisateurs des données est sa cohérence interne : l'estimateur de la différence est égal à la différence des estimateurs. Si les échantillons de périodes consécutives ne se chevauchent que partiellement, l'estimateur composite ou l'estimateur des MCG (Fuller 1999, Hansen, Hurwitz et Madow 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) sont plus efficaces.

Dans la suite de l'exposé, nous supposons que toutes les procédures d'échantillonnage correspondent à l'échantillonnage aléatoire simple sans remise. Pour la moyenne contemporaine, la variance est donnée par (Cochran 1977, Th. 10.1) :

$$V[\bar{y}_{\cdot\cdot}] = \left(1 - \frac{n}{N}\right) \frac{S_{tb}^2}{n} + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{tw}^2}{nm}. \quad (2.8)$$

Pour la simplicité et la clarté de l'exposé, nous formulerons souvent l'hypothèse de conditions symétriques :

$$S_{1wi}^2 = S_{2wi}^2 = S_{wi}^2, \quad \bar{S}_{1w}^2 = \bar{S}_{2w}^2 = \bar{S}_w^2, \quad S_{1b}^2 = S_{2b}^2 = S_b^2. \quad (2.9)$$

Les dérivations analytiques sont possibles sans ces hypothèses, mais deviennent extrêmement lourdes. En outre, il n'est pas raisonnable de penser que le concepteur d'enquête pourrait connaître les caractéristiques de la future population. Donc (2.9) devrait être considérée comme un modèle de travail raisonnable.

2.1 Plan à échantillons indépendants

Proposition 1. Soit le tirage sans remise de n_1 sur N grappes et de m_1 sur M unités d'observation dans les grappes au temps $t = 1$. Soit le tirage sans remise de n_2 sur N grappes et de m_2 sur M unités d'observation dans les grappes au temps $t = 2$, ce tirage étant indépendant de celui effectué au temps $t = 1$. Alors

$$\begin{aligned} V_{\iota}(d) &= \left(1 - \frac{n_1}{N}\right) \frac{S_{1b}^2}{n_1} + \left(1 - \frac{n_2}{N}\right) \frac{S_{2b}^2}{n_2} \\ &\quad + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n_1 m_1} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n_2 m_2}. \end{aligned} \quad (2.10)$$

Le résultat découle immédiatement de (2.8) en raison de l'indépendance des deux échantillons. L'indice inférieur ι de la variance désigne le « plan à échantillons indépendants ». Sous les conditions de symétrie de (2.9), si les tailles d'échantillon sont les mêmes aux deux périodes, $n_1 = n_2 = n$ et $m_1 = m_2 = m$, alors

$$V_{e,\iota}[d] = 2 \left(1 - \frac{n}{N}\right) \frac{S_b^2}{n} + 2 \left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{nm}, \quad (2.11)$$

où l'indice inférieur e, ι signifie « variances égales, plan à échantillons indépendants ».

2.2 Plan à panel de grappes

Proposition 2. Soit n sur N grappes échantillonnées sans remise à la première période et utilisées aux deux périodes. Soit m sur M unités d'observation échantillonnées sans remise indépendamment aux deux périodes. Alors

$$\begin{aligned} V_c[d] &= \left(1 - \frac{n}{N}\right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^1 S_{1b} S_{2b}}{n} \\ &\quad + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2}{nm}, \\ \rho^1 &= \frac{1}{S_{1b} S_{2b} (N - 1)} \sum_{i=1}^N (\bar{Y}_{1i\cdot} - \bar{Y}_{1\cdot})(\bar{Y}_{2i\cdot} - \bar{Y}_{2\cdot}). \end{aligned} \quad (2.12)$$

Ici, l'indice inférieur c désigne le plan à panel de grappes (en anglais cluster) » et ρ^I est la corrélation intertemporelle, ou autocorrélation, des moyennes de grappe. L'indice supérieur I désigne le premier degré d'échantillonnage. Si ρ^I est positive, le plan à panel de grappes est plus efficace que le plan à échantillons indépendants pour des valeurs fixes de n et m . Sous les conditions de symétrie,

$$V_{e,c}[d] = 2\left(1 - \frac{n}{N}\right) \frac{S_b^2(1 - \rho^I)}{n} + 2\left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{mn}, \quad (2.13)$$

où l'indice inférieur e, c signifie « variances égales, plan à panel de grappes ».

2.3 Plan à panel d'unités d'observation

Proposition 3. Soit n sur N grappes et m sur M unités d'observation échantillonnées sans remise à la première période et utilisées aux deux périodes. Alors

$$V_o[d] = \left(1 - \frac{n}{N}\right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_{1w} \bar{S}_{2w} N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{1ij} - \bar{Y}_{1i.})(Y_{2ij} - \bar{Y}_{2i.}). \quad (2.14)$$

L'indice inférieur o désigne le « plan à panel d'unités d'observation ». Sous l'hypothèse des conditions symétriques,

$$V_{e,o}[d] = 2\left(1 - \frac{n}{N}\right) \frac{(1 - \rho^I) S_b^2}{n} + 2\left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II}) \bar{S}_w^2}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_w^2 N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{1ij} - \bar{Y}_{1i.})(Y_{2ij} - \bar{Y}_{2i.}) \quad (2.15)$$

où l'indice inférieur correspondant e, o signifie « variances égales, plan à panel d'unités d'observation ».

Ici, ρ^{II} est la corrélation intertemporelle, ou autocorrélation, des observations individuelles dans les grappes. L'indice supérieur II désigne le deuxième degré d'échantillonnage. Si ρ^{II} est positive, le plan à panel d'unités d'observation est plus efficace que le plan à panel de grappes pour des valeurs fixes de n et m .

Comment sont reliées les deux autocorrélations qui figurent dans (2.15) ? Conceptuellement, nous pouvons imaginer n'importe quel nombre de relations possibles entre elles. Introduisons un modèle de superpopulation

$$Y_{ij} = \mu_t + a_{it} + \varepsilon_{ij}, \quad E_\xi[a_{it}] = 0, \quad E_\xi[\varepsilon_{ij}] = 0, \quad (2.16)$$

dans lequel les termes a_{it} et ε_{ij} sont indépendants l'un de l'autre pour tout $s, t = 1, 2$. L'indice inférieur ξ désigne les espérances sous le modèle de superpopulation. Le cas où $\rho^I = 0$ et $\rho^{II} = 1$ s'observe quand les variations des moyennes de grappes ont lieu indépendamment entre les grappes ($E_\xi[a_{it} a_{st}] = 0$), mais que les individus gardent leur position à l'intérieur de la grappe, $\varepsilon_{1ij} = \varepsilon_{2ij}$. Le cas où $\rho^I = 1$ et $\rho^{II} = 0$ a lieu quand les effets aléatoires de grappes sont les mêmes aux deux périodes, $a_{1i} = a_{2i}$, tandis que les effets aléatoires individuels ne sont pas corrélés ($E_\xi[\varepsilon_{1ij} \varepsilon_{2ij}] = 0$). Ni l'une ni l'autre de ces situations n'est entièrement réaliste. Cependant, on peut sans doute s'attendre à ce que la dynamique individuelle soit une source plus importante de variation au cours du temps que la dynamique de grappes, de sorte que les relations $\rho^{II} \geq \rho^I \geq 0$ sont les plus plausibles. Nous examinerons dans les exemples numériques des sections 4 et 5 la mesure dans laquelle le choix du meilleur plan est sensible à la relation entre les deux corrélations.

3. Coûts pour des échantillons en grappes répétées

À la présente section, nous analysons la rentabilité des échantillons en grappes lorsque l'on veut estimer la différence entre deux moyennes d'échantillon pour deux périodes différentes.

Une discussion des coûts de l'échantillonnage en grappes est donnée dans Kish (1995, section 8.3B), Thompson (1992, section 12.5), ainsi que Lehtonen et Pahkinen (2004). Un exposé mathématique plus détaillé, avec les formules de variance corrigées pour les populations finies, peut être consulté dans Hansen et coll. (1953, volume II, section 6.11).

3.1 Notation et modèles de coûts

Supposons la structure de coût suivante, qui est une extension de celle de Kish (1995) pour les enquêtes répétées :

- c_1^I est le coût au niveau de la grappe au temps $t = 1$ pour les grappes qui sont utilisées à la première vague seulement ;
- c_2^I est le coût au niveau de la grappe pour une nouvelle grappe au temps $t = 2$;
- c_{12}^I est le coût au niveau de la grappe pour les grappes dans lesquelles les données sont recueillies aux deux périodes $t = 1$ et $t = 2$ (coût du panel d'UPE) ;
- c_1^{II} est le coût au niveau individuel au temps $t = 1$ pour les individus qui sont observés à la première vague seulement ;
- c_2^{II} est le coût au niveau individuel au temps $t = 2$ pour les individus qui sont observés à la deuxième vague seulement ;

- c_{12}^{II} est le coût au niveau individuel si l'unité est observée aux deux périodes dans le plan à panel d'unités d'observation (coût du panel d'USE) ;
- C_0 est le budget total alloué au travail sur le terrain aux deux périodes.

Les indices supérieurs en chiffres romains désignent le degré d'échantillonnage. Les indices inférieurs en caractères arabes correspondent à la vague à la quelle l'échantillon est tiré. Les coûts au niveau de la grappe comprennent le coût de l'échantillonnage des grappes, l'obtention des cartes des UPE, la collecte de données sur la collectivité, la formation locale des intervieweurs, etc. Les coûts au niveau individuel sont principalement ceux des interviews sur place auprès des unités d'observation finales. Le coût total C_0 est conçu comme le coût variable de l'enquête qui est directement relié au nombre d'unités échantillonnées. Le coût fixe, tel que le coût de la préparation du questionnaire et d'autres coûts au niveau organisationnel, ne fait pas partie de C_0 .

3.2 Plan à échantillons indépendants

Pour le plan à échantillons indépendants, la contrainte budgétaire est donnée par

$$C_0 = c_1^{\text{I}} n_1 + c_1^{\text{II}} n_1 m_1 + c_2^{\text{I}} n_2 + c_2^{\text{II}} n_2 m_2. \quad (3.1)$$

Les deux premiers termes sont les coûts de la première vague de collecte des données, et les deux derniers, ceux de la deuxième vague.

Proposition 4. Si les paramètres de configuration de l'enquête sont les mêmes aux deux périodes :

$$c_1^{\text{I}} = c_2^{\text{I}} = c^{\text{I}}, \quad c_1^{\text{II}} = c_2^{\text{II}} = c^{\text{II}}, \quad (3.2)$$

alors les tailles optimales d'échantillon et les variances résultantes sont données par

$$m = \sqrt{\frac{c^{\text{I}} \bar{S}_w^2}{c^{\text{II}} S_b^2 - \bar{S}_w^2 / M}},$$

$$n = \frac{C_0}{2\{c^{\text{I}} + [c^{\text{I}} c^{\text{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}},$$

$$V_{e,t}[d] = \frac{4\left[c^{\text{I}} + \sqrt{c^{\text{I}} c^{\text{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)}\right]}{C_0}$$

$$\times \left[S_b^2 + \left(\sqrt{\frac{c^{\text{II}} S_b^2 - \bar{S}_w^2 / M}{c^{\text{I}} \bar{S}_w^2}} - \frac{1}{M} \right) \bar{S}_w^2 \right] - \frac{2}{N} S_b^2. \quad (3.3)$$

Dans les équations (3.3), les tailles d'échantillon n et m sont traitées comme des variables continues. En pratique, le nombre entier le plus proche doit être utilisé, une valeur d'au moins 2 étant nécessaire pour estimer la composante de

variance appropriée, et les valeurs maximales étant N et M , respectivement.

Le nombre d'unités d'observation échantillonnées dans une grappe dépend uniquement des coûts relatifs au niveau de la grappe et au niveau de l'unité d'observation, $c^{\text{I}}/c^{\text{II}}$, et des variances relatives S_b^2/\bar{S}_w^2 , ou, ce qui est équivalent, la corrélation intragrappe. Des coûts d'interview c^{II} plus élevés empêchent le concepteur de l'enquête d'utiliser un plus grand nombre d'unités d'observation : une augmentation de c^{II} entraîne une diminution de m ainsi que de n . Un coût au niveau de la grappe plus élevé donne lieu à une redistribution des unités échantillonnées : quand c^{I} augmente, n diminue tandis que m augmente. Une plus grande variance intragrappe \bar{S}_w^2 nécessite le tirage d'un plus grand nombre d'unités d'observation m dans une grappe pour maintenir la précision globale. Une plus grande variance intergrappes S_b^2 nécessite l'échantillonnage d'un plus grand nombre n de grappes. Enfin, le budget total de l'enquête C_0 affecte le nombre de grappes n , mais non la taille m , du sous-échantillon. Par conséquent, la variance de d est inversement proportionnelle à C_0 .

Une situation non symétrique peut être traitée comme un sous-produit des conditions de premier ordre dérivées dans la preuve (voir l'annexe). Cependant, aucune solution analytique n'est disponible dans ce cas.

3.3 Plan à panel de grappes

Pour le plan à panel de grappes, la contrainte budgétaire est donnée par

$$C_0 = c_{12}^{\text{I}} n + c_1^{\text{II}} n m_1 + c_2^{\text{II}} n m_2. \quad (3.4)$$

Le premier terme est le coût au niveau de la grappe associé au plan d'échantillonnage, et les deux autres termes sont les coûts de la collecte des données au niveau individuel aux première et deuxième vagues, respectivement.

Proposition 5. Pour le plan à panel de grappes, les tailles d'échantillon sont données par

$$m_1 = 2C_0 / c_1^{\text{II}} \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2} + \sqrt{D} \right),$$

$$m_2 = \kappa m_1,$$

$$n = \frac{C_0}{c_{12}^{\text{I}} + c_1^{\text{II}} m_1 + c_2^{\text{II}} m_2},$$

$$\kappa = \sqrt{\frac{c_1^{\text{II}} \bar{S}_{2w}^2}{c_2^{\text{II}} \bar{S}_{1w}^2}}, \quad (3.5)$$

à condition que

$$D = \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2}\right)^2 + 8 \frac{(1 - \rho^I) S_b^2 C_0}{\bar{S}_{1w}^2 c_1^{II}} - 4 \frac{C_0}{c_1^{II} M} \left(1 + \frac{\bar{S}_{2w}^2}{\kappa \bar{S}_{1w}^2}\right) \geq 0.$$

La variance de l'estimateur de la différence est obtenue en entrant ces expressions dans (2.13). Sous les hypothèses de conditions symétriques aux deux vagues de l'enquête (2.9) et (3.2),

$$D = 4 - 8 \frac{C_0}{M c^{II}} + 8 \frac{(1 - \rho^I) S_b^2 C_0}{\bar{S}_w^2 c^{II}},$$

$$m_1 = m_2 = m$$

$$= \frac{C_0}{c^{II} + \sqrt{(c^{II})^2 - \frac{2c^{II}C_0}{M} + \frac{2(1 - \rho^I)S_b^2C_0c^{II}}{\bar{S}_w^2}}},$$

$$n = \frac{C_0}{c_{12}^I + 2c^{II}m}$$

$$= \frac{C_0}{c_{12}^I + 2C_0 \left[1 + \sqrt{1 - \frac{2c^{II}C_0}{M c^{II}} + \frac{2(1 - \rho^I)S_b^2C_0}{\bar{S}_w^2 c^{II}}} \right]},$$

et $V_{e,c}[d]$ peut être calculée à partir de (2.13).

Fait intéressant, le nombre d'USE dépend des coûts au niveau des USE c^{II} , mais non des coûts au niveau des UPE c_{12}^I . Une augmentation de la corrélation intragrappe, ou de S_b^2 , ou une diminution de S_w^2 , entraîne comme cela est à prévoir une diminution du nombre optimal d'USE et une augmentation du nombre optimal d'UPE. La dépendance des paramètres du plan à l'égard du budget de l'enquête C_0 n'est pas négligeable. Pour de très petites enquêtes, le nombre d'unités par grappe est proportionnel à C_0 , tandis que le nombre de grappes n'est pas affecté par C_0 . En effet, si la caractéristique étudiée présente une forte corrélation entre les périodes d'observation, il est préférable d'obtenir des estimations exactes des moyennes de grappe, desquelles découlerait une bonne exactitude de l'estimateur de la différence globale. Autrement dit, le premier terme de (2.13) est relativement petit en raison du coefficient de corrélation positif ρ^I , et le deuxième terme est inversement proportionnel à C_0 . Pour les grandes enquêtes, $D \propto C_0$, de sorte que le nombre d'unités par grappe ainsi que le nombre de grappes sont proportionnels à $\sqrt{C_0}$. Le premier terme de (2.13) est alors inversement proportionnel à $\sqrt{C_0}$, et le deuxième terme est inversement proportionnel à C_0 . Une augmentation du budget de l'enquête affectera tous les termes, quoique dans une mesure différente.

3.4 Plan à panel d'unités d'observation

Pour le plan à panel d'unités d'observation, la contrainte budgétaire est donnée par

$$C_0 = c_{12}^I n + c_{12}^{II} nm. \quad (3.6)$$

Le premier terme est le coût au niveau de la grappe et le deuxième terme, le coût des interviews individuelles.

Proposition 6. Pour le plan à panel d'unités d'observation, les tailles d'échantillon optimales sont données par

$$m = \sqrt{\frac{c_{12}^I (1 - \rho^{II}) \bar{S}_w^2}{c_{12}^{II} (1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}},$$

$$n = \frac{C_0}{c_{12}^I + \sqrt{\frac{(1 - \rho^{II}) \bar{S}_w^2 c_{12}^{II}}{(1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}}}. \quad (3.7)$$

La variance sous le plan de l'estimateur de la différence résultant est

$$V_{e,o}[d] = \frac{2}{C_0} \left\{ (1 - \rho^I) S_b^2 c_{12}^I \right.$$

$$+ (1 - \rho^{II}) \bar{S}_w^2 \sqrt{c_{12}^I c_{12}^{II} \frac{(1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}{(1 - \rho^{II}) \bar{S}_w^2}}$$

$$+ \left[(1 - \rho^I) S_b^2 - \frac{1}{M} (1 - \rho^{II}) \bar{S}_w^2 \right]$$

$$\times \sqrt{\frac{(1 - \rho^{II}) \bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1 - \rho^I) S_b^2 - (1 - \rho^{II}) \bar{S}_w^2 / M}}$$

$$\left. + (1 - \rho^{II}) \bar{S}_w^2 \left(c_{12}^{II} - \frac{c_{12}^I}{M} \right) \right\} - \frac{2(1 - \rho^I) S_b^2}{N}. \quad (3.8)$$

Les expressions des tailles d'échantillon (3.7) ressemblent à celles données à l'équation (3.3) pour le plan à échantillons indépendants, avec le coût de la collecte des données dans une seule vague remplacé par le coût de la collecte de données de panel, et les composantes de la variance S_b^2 et \bar{S}_w^2 remplacées par $(1 - \rho^I) S_b^2$ et $(1 - \rho^{II}) \bar{S}_w^2$. La taille de l'échantillon de deuxième degré m dépend uniquement du coût relatif aux niveaux de la grappe et de l'unité d'observation et du ratio des composantes de la variance augmentées des autocorrélations. Donc, comme dans le cas du plan à échantillons indépendants, la relation entre la taille de l'échantillon et la portée de l'enquête n'a lieu que par la voie de $n \propto C_0$, et la variance de la différence diminue de manière inversement proportionnelle à C_0 .

En étendant les relations entre les formes fonctionnelles des équations (3.3) et (3.8), nous pouvons établir les relations générales entre les deux plans :

Proposition 7. Si $M \gg 1$ et $N \gg 1$, alors $V_{e,i}[d] \geq V_{e,o}[d]$ si

$$2 \left(\sqrt{c_{12}^I S_b^2} + \sqrt{c_{12}^{II} \bar{S}_w^2} \right)^2$$

$$\geq \left[\sqrt{c_{12}^I (1 - \rho^I) S_b^2} + \sqrt{c_{12}^{II} (1 - \rho^{II}) \bar{S}_w^2} \right]^2. \quad (3.9)$$

Malheureusement, la variance sous le plan à panel de grappes que l'on peut obtenir en combinant les résultats de la proposition 5 avec (2.13) ne permet pas une comparaison aussi claire.

4. Exemple numérique

Afin d'illustrer comment les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts) influent sur le choix du plan le plus efficace, nous considérons un exemple numérique. Choisissons la configuration de base avec des conditions symétriques et posons que les valeurs des paramètres sont :

$$\begin{aligned} N &= 10\,000, & M &= 1\,000, & S_b &= 100, \\ S_w &= 400, & \rho^I &= 0,1, & \rho^{II} &= 0,35, \\ c_1^{II} &= c_2^{II} = 1, & c_{12}^{II} &= 3, & c_1^I &= c_2^I = 10, \\ c_{12}^I &= 18, & C_0 &= 20\,000. \end{aligned} \quad (4.1)$$

La structure de coût implique que le coût de la collecte de l'information initiale pour une grappe est le coût de dix interviews, tandis que le coût du suivi de la même grappe est celui de huit interviews seulement. Par ailleurs, obtenir la deuxième interview auprès de la même unité coûte deux fois plus cher qu'obtenir la première.

Étant donné ces paramètres, les tailles d'échantillon et les variances sous le plan sont :

$$\begin{aligned} m_{e,i} &= 12, & m_{e,c} &= 12, & m_{e,o} &= 8, \\ n_{e,i} &= 455, & n_{e,c} &= 476, & n_{e,o} &= 476, \\ m_{e,i}n_{e,i} &= 5\,460, & m_{e,c}n_{e,c} &= 5\,712, & m_{e,o}n_{e,o} &= 3\,808, \\ V_{e,i}[d] &= 99,86, & V_{e,c}[d] &= 91,37, & V_{e,o}[d] &= 90,20. \end{aligned} \quad (4.2)$$

Le plan à panel d'unités d'observation est 1,2 % plus efficace que le plan à panel de grappes, et 10,7 % plus efficace que le plan à échantillons indépendants. Cependant, il produit une taille d'échantillon appréciablement plus petite, égale aux deux tiers seulement de celle de l'échantillon sous le plan à panel de grappes et à 70 % de la taille de l'échantillon sous le plan à échantillons indépendants.

Naturellement, ces résultats sont fortement liés aux paramètres de la population et à la structure de coût. Pouvons-nous décrire des profils généraux de changement des variances et, donc, de l'efficacité relative des divers plans quand ces paramètres varient ? Dans (4.2), les variances sont dérivées de 13 paramètres donnés dans (4.1), et il est difficile de formuler des énoncés pertinents au sujet de tous ces paramètres simultanément. Plus bas, nous essayerons de donner des coupes transversales bidimensionnelles de cet espace à 13 dimensions et de présenter des illustrations graphiques de la variabilité des variances sous le plan, et donc les domaines d'optimalité de chaque plan, en faisant varier deux paramètres à la fois. Nous fournissons les graphiques des variances pour les plans concernés (habituellement, le plan à panel de grappes est représenté par des courbes en **pointillé** ; le plan à panel d'unités d'observation, par des courbes en **tirets** et le plan à échantillons indépendants, par des courbes en **tirets et points**. Dans la plupart des

graphiques, le plan indépendant n'est pas affecté par les variations des paramètres qui définissent les axes des graphiques et est donc ignoré). Nous montrons également l'efficacité relative des divers plans en représentant les domaines de l'espace des paramètres en jaune/gris clair si le plan à échantillons indépendants est le plus efficace, en vert/gris moyen si le plan à panel de grappes est le plus efficace, et en violet/gris foncé si le plan à panel d'unités d'observation est le plus efficace (le code R utilisé pour générer les graphiques est disponible au <http://web.missouri.edu/~kolenikovs/SMJ2011/>).

La figure 1 illustre comment les variances sous le plan, et donc le plan le plus efficace, varient en fonction des coûts des panels d'UPE et d'USE, c_{12}^I et c_{12}^{II} . Manifestement, ces variations n'ont pas d'incidence sur la variance du plan à échantillons indépendants, qui sert de référence. En outre, les variations de c_{12}^{II} n'influencent pas la performance du plan à panel de grappes, qui correspond aux droites d'iso-variance verticales en **pointillé** dans le graphique de gauche. Les courbes descendantes en **tirets** sont les courbes d'iso-variance pour le plan à panel d'unités d'observation. Il convient de souligner que le coin inférieur gauche du graphique correspond à la situation de gratuité dans laquelle la deuxième vague de collecte des données ne coûte rien : les coûts de panel sont égaux aux coûts de la période unique, $c_{12}^I = c_1^I$, $c_{12}^{II} = c_1^{II}$. Quand les coûts de la collecte des données de panel sont extrêmement élevés (coin supérieur droit du graphique), le plan à échantillons indépendants est le plus efficace. Le point où les trois plans ont la même variance correspond à $c_{12}^I = 22$, $c_{12}^{II} = 3,05$; autrement dit, le coût de la deuxième interview est 2,05 fois plus élevé que celui de la première, et les coûts au niveau de la grappe sont 20 % plus élevés à la deuxième vague qu'à la première. Néanmoins, une autocorrélation positive justifie la réduction de la taille de l'échantillon du plan à panel d'unités d'observation comparativement au plan à échantillons indépendants. Si le coût du panel au niveau de la grappe est plus faible, mais que le coût de la deuxième interview est plus élevé, le plan à panel de grappes est le plus efficace. Si les deuxièmes interviews sont plus coûteuses, le plan le plus efficace est le plan à panel d'unités d'observation. Ce dernier domaine comprend notre cas de référence avec $c_{12}^I = 18$ et $c_{12}^{II} = 3$.

La figure 2 montre les variations des variances sous le plan associées aux variations des autocorrélations ρ^I , ρ^{II} . La variance sous le plan à échantillons indépendants n'est pas affectée par cette variation, et celle sous le plan à panel de grappes n'est pas affectée par les variations de ρ^{II} . Le plan à panel d'unités d'observation est plus efficace quand l'autocorrélation des USE est plus forte, $\rho^{II} > 0,34$. Sinon, le plan à panel de grappes donne une variance plus faible.

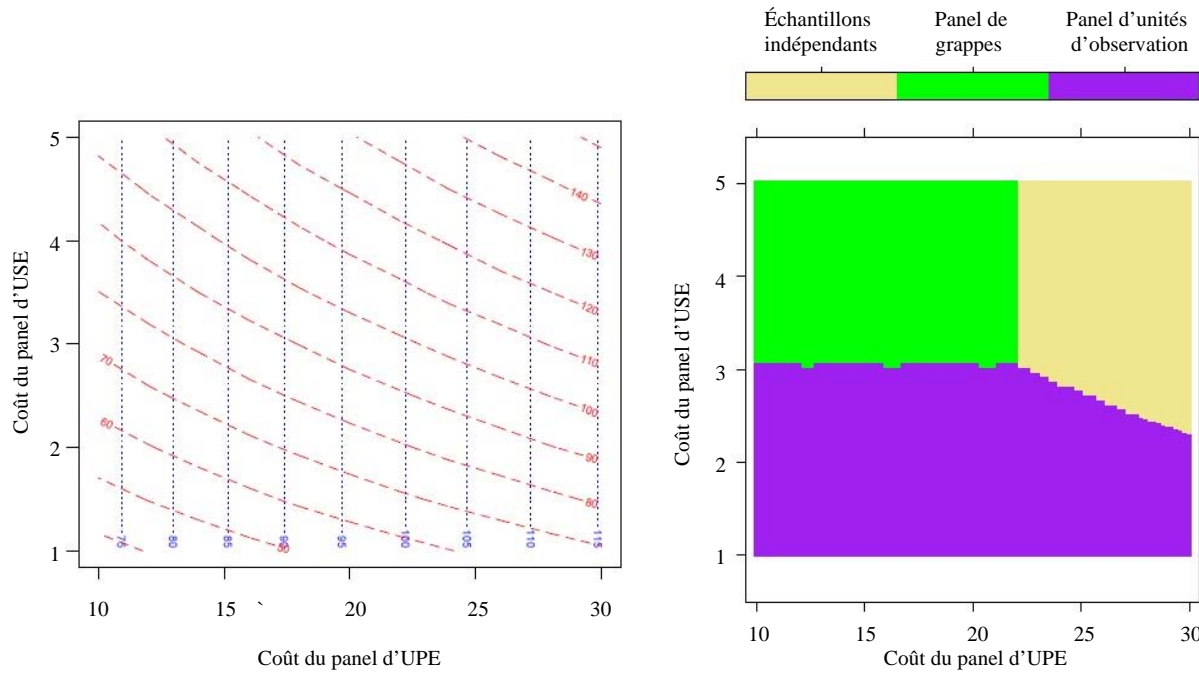


Figure 1 Variances sous le plan en fonction des coûts de collecte des données c_{12}^I, c_{12}^{II} . À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé) et $V_{e,o}[d]$ (tiret) ; $V_{e,t} = 99,86$; à droite : domaines d'optimalité des trois plans

À la figure 3, nous examinons l'effet du coût au niveau de la grappe et de l'autocorrélation sur le choix du plan. Les combinaisons d'un coût élevé de collecte des données à la deuxième vague et d'une faible autocorrélation des UPE, qui figurent dans le coin supérieur gauche du graphique, font du plan à échantillons indépendants celui qui est le plus intéressant. Autrement, le plan à panel d'unités d'observation est celui qui convient le mieux. Notons que les lignes de contour pour les plans à panel d'unités d'observation sont très proches les unes des autres et que les différences de variance entre les deux plans sont inférieures à 2 % dans l'espace entier des paramètres de ce graphique.

À la figure 4, nous examinons l'effet du coût au niveau de l'unité d'observation et de l'autocorrélation sur le choix du plan. Ni la variance sous le plan à échantillons indépendants ni celle sous le plan à panel de grappes n'est affectée par la variation des paramètres présentés dans ce graphique. La variance sous le plan à échantillons indépendants est égale à 99,86, tandis que celle sous le plan à panel de grappes est de 91,37, de sorte que nous comparons le plan à panel d'unités d'observation à ce dernier seulement. Une forte autocorrélation ($\rho^{II} \geq 0,6$) peut justifier le coût très élevé de la deuxième interview (jusqu'à quatre fois plus élevé que celui de la première), mais dans le coin supérieur gauche du graphique, qui correspond aux faibles autocorrélations et aux coûts de panel élevés, le plan à panel de grappes est celui qui donne les meilleurs résultats.

La figure 5 relie les variances sous le plan aux coûts d'enquête au niveau de la grappe. L'axe horizontal donne le

coût à la première période, c_1^I , et l'axe vertical, le coût supplémentaire à la deuxième période, quand les données sont recueillies selon un mode de panel, $c_{12}^I - c_1^I$. Pour le plan à échantillons indépendants, nous ignorons l'axe vertical, car ce paramètre ne figure pas dans le plan. En outre, en raison de (4.1), $c_1^I = c_2^I$. Le plan à panel d'unités d'observation est uniformément meilleur que le plan à panel de grappes pour toutes les combinaisons de paramètres sur ce graphique, quoique l'écart entre les variances ne dépasse pas 2 %. Dans le coin supérieur gauche, le coût additionnel du mode de collecte des données par panel est prohibitif, et le plan à échantillons indépendants est celui qui donne les meilleurs résultats.

La figure 6 illustre la relation entre le plan le plus efficace et le budget total de l'enquête, ainsi que le coût du mode panel de collecte des données au niveau de la grappe. Pour $C_0 > 10\,000$, le plan à panel d'unités d'observation est le meilleur si $c_{12}^I < 22,7$, c'est-à-dire si le coût additionnel du mode de collecte par panel des données au niveau de la grappe n'excède pas 127 % du coût initial de collecte au niveau de la grappe à la première vague. Curieusement, pour certaines configurations isolées des paramètres dans de petites enquêtes, le plan à panel de grappes peut donner de meilleurs résultats que le plan à panel d'unités d'observation qui domine le reste du graphique. La différence entre les variances sous le plan pour le plan à panel de grappes et le plan à panel d'unités d'observation est inférieure à 4 % pour toutes les combinaisons de paramètres dans ce graphique.

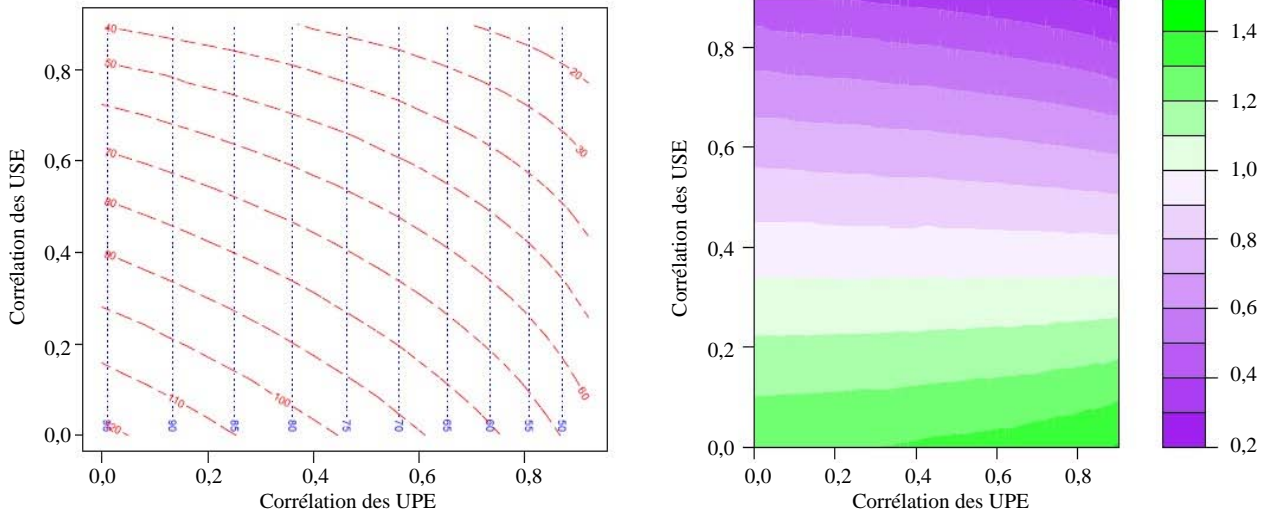


Figure 2 Variances sous le plan en fonction des corrélations de population ρ^I, ρ^{II} . À gauche : Lignes de contour de $V_{e,c}[d]$ (pointillé) et de $V_{e,o}[d]$ (tiret); $V_{e,t} = 99,86$; à droite : ratio $V_{e,o}[d]/V_{e,c}[d]$

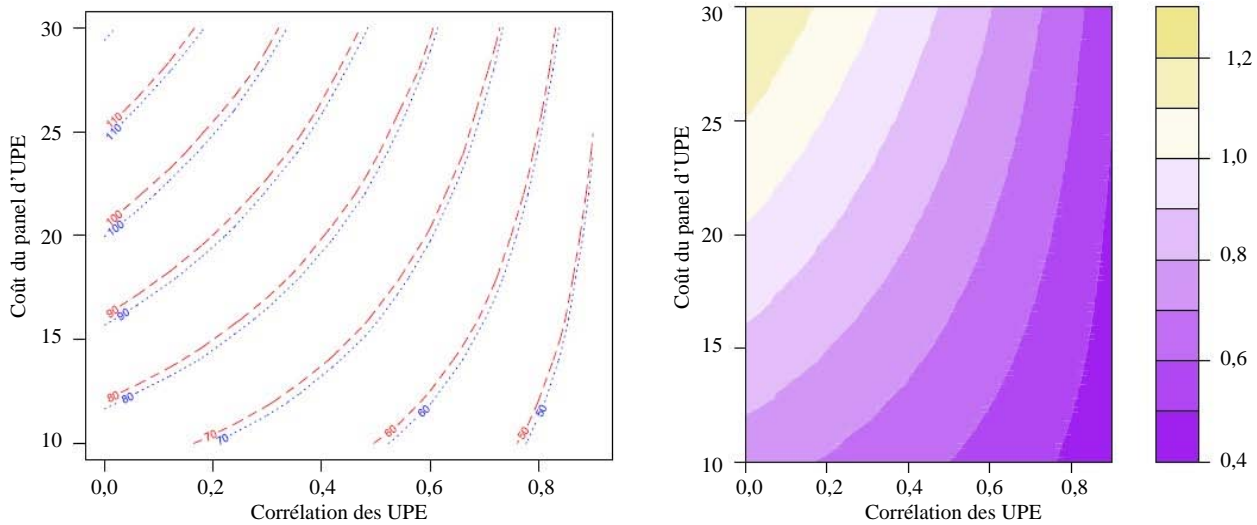


Figure 3 Variances sous le plan en fonction de l'autocorrélation au niveau de la grappe ρ^I et du coût c_{12}^I . À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé) et de $V_{e,o}[d]$ (tiret); $V_{e,t} = 99,86$; à droite : ratio $V_{e,o}[d]/V_{e,t}[d]$

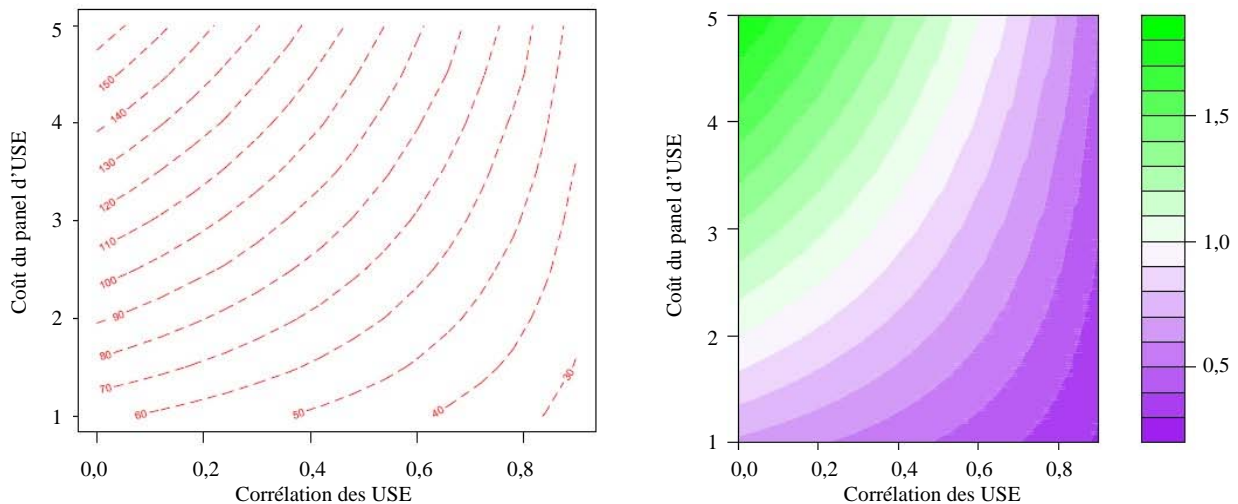


Figure 4 Variances sous le plan en fonction de l'autocorrélation au niveau des unités d'observation ρ^{II} et du coût c_{12}^{II} . À gauche : lignes de contour de $V_{e,o}[d]$ (tiret); $V_{e,t} = 99,86$; $V_{e,c}[d] = 91,37$; à droite : ratio $V_{e,o}[d]/V_{e,c}[d]$

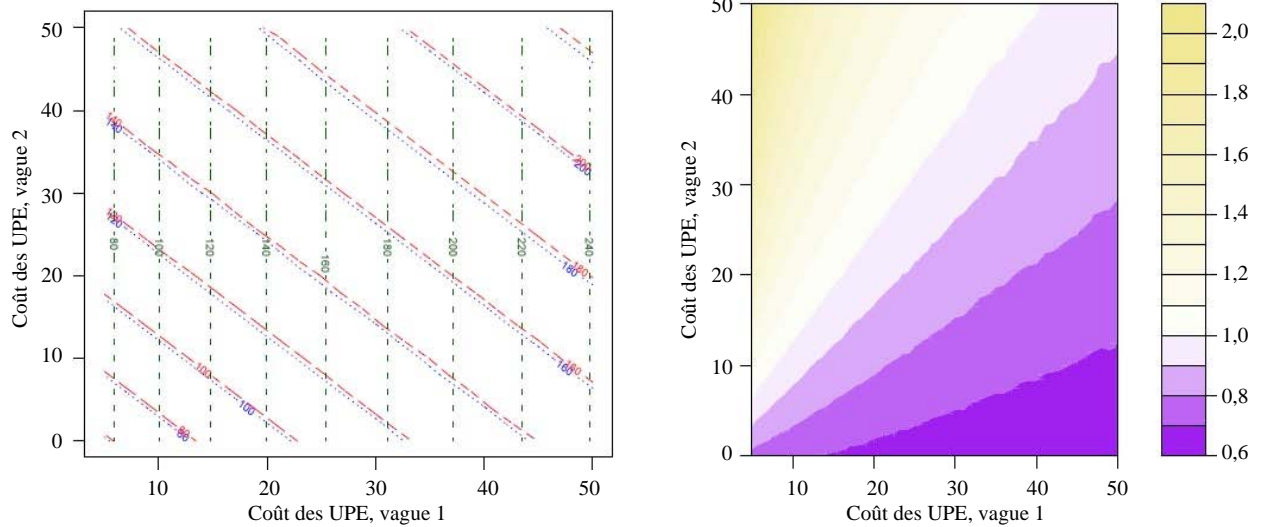


Figure 5 Variations sous le plan en fonction des coûts au niveau de la grappe à la première vague, c_1^I , et à la deuxième vague, $c_{12}^I - c_1^I$. À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé), $V_{e,o}[d]$ (tiret) et $V_{e,t}[d]$ (tiret et point) ; à droite : ratio $V_{e,o}[d]/V_{e,t}[d]$

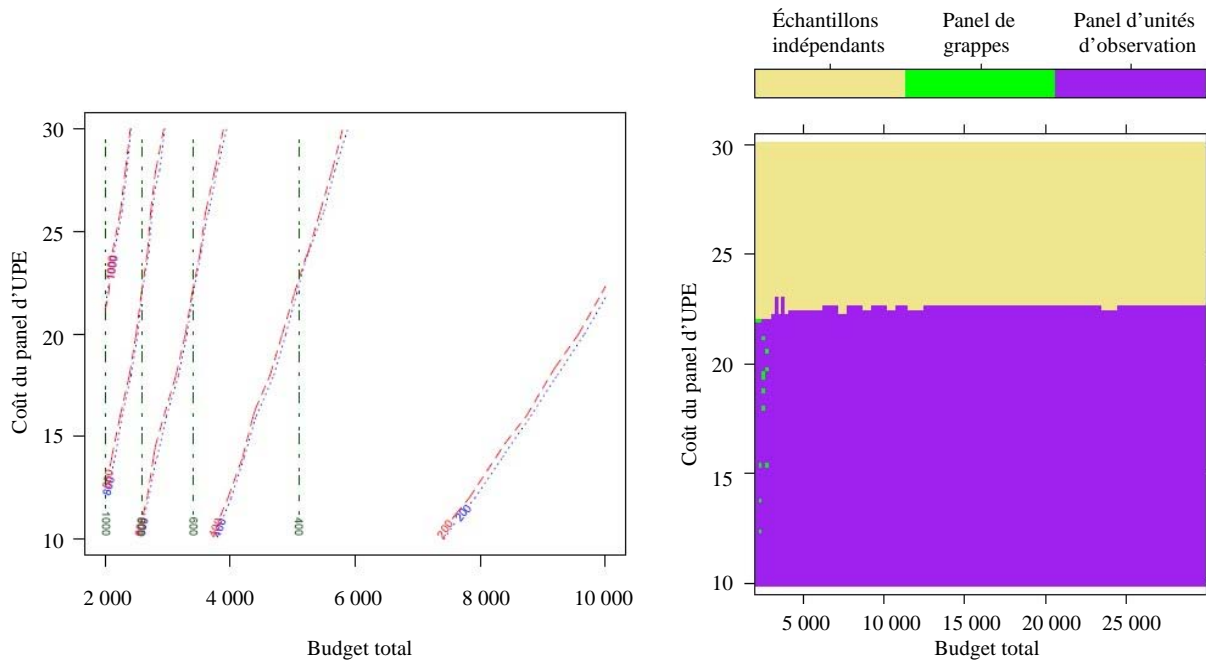


Figure 6 Variations sous le plan en fonction du budget total C_0 et du coût du panel d'UPE c_{12}^{II} . À gauche : lignes de contour de $V_{e,c}[d]$ (pointillé), $V_{e,o}[d]$ (tiret) et $V_{e,t}[d]$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

Dans l'ensemble, cet exemple numérique montre que, selon les paramètres de la population et les coûts de la collecte des données, chacun des trois plans peut être le plus efficace. De faibles corrélations et des coûts élevés à la deuxième vague ont tendance à favoriser le plan à échantillons indépendants. Vu que les six paramètres de population initiaux et les cinq paramètres de coût pourraient ne pas être représentatifs d'un grand nombre d'enquêtes répétées, une analyse de sensibilité semblable à celle présentée ici

pourrait être nécessaire pour toute enquête particulière qu'un statisticien doit concevoir.

5. Plan de sondage avec critères multiples

Jusqu'à présent, nous avons limité notre analyse à l'estimation de la différence entre les moyennes d'une seule variable à deux vagues de collecte des données. La plupart des enquêtes à grande échelle sont conçues pour recueillir

des données sur plusieurs caractéristiques et de nombreux utilisateurs s'intéressent à la production d'estimations contemporaines. Afin de tenir compte des contraintes d'exactitude associées à ces diverses variables et estimations, le concepteur d'enquête doit avoir à l'esprit plusieurs variances quand il choisit le plan à mettre en œuvre. Il s'agit d'un problème d'optimisation à critères multiples et aucun plan individuel ne sera le meilleur pour tous les problèmes d'estimation possibles. Dans le présent contexte, le plan à panel d'unités d'observation peut donner de bonnes estimations de la variation quand l'autocorrélation des UPE ainsi que celle des USE est élevée, mais risque de produire une petite taille d'échantillon si le coût du suivi des UPE ainsi que des USE est élevé. Des estimations d'une grande précision pour n'importe quelle période pourraient être obtenues en passant au plan à panel de grappes, voire même au plan à échantillons indépendants.

Dans cette situation, il est possible de comparer les différents plans en utilisant l'argument microéconomique classique de maximisation de l'utilité sous les contraintes budgétaires (Mas-Colell, Whinston et Green 1995). Dans le contexte des plans de sondage, l'utilité pour le concepteur d'enquête augmente avec la précision des estimations d'après les données de l'enquête, ou ce qui est équivalent, diminue quand les variances sous le plan augmentent. Une forme fonctionnelle simple est donnée par la fonction d'utilité de Cobb-Douglas :

$$U(\text{plan}) = V_{\text{plan}}^{-\alpha_1} [\bar{y}_1] V_{\text{plan}}^{-\alpha_2} [\bar{y}_2] V_{\text{plan}}^{-\alpha_3} [d]. \quad (5.1)$$

Ici, α_1 , α_2 et α_3 sont des constantes positives qui décrivent les poids relatifs des trois variances sous le plan dans le processus décisionnel. Dans (5.2), les variances $V[\bar{y}_1]$ et $V[\bar{y}_2]$ sont les variances des moyennes dans les enquêtes avec échantillonnage en grappes données par (2.8). La variance de l'estimateur de la différence correspond à (2.10), (2.12) ou (2.14), selon le plan choisi. Le problème que doit alors résoudre le concepteur d'enquête consiste à maximiser (5.1) sous les contraintes budgétaires particulières au plan (3.1), (3.4) ou (3.6). La maximisation est effectuée sur les paramètres du plan (mode de collecte des données, nombre de grappes à chaque période, nombre d'unités d'observation à chaque période), sachant les caractéristiques de la population (variances et autocorrélations) et le processus de collecte des données (coûts).

Supposons que la précision des trois estimations \bar{y}_1 , \bar{y}_2 et d ont chacune la même importance pour le décideur, de sorte que $\alpha_1 = \alpha_2 = \alpha_3$. Afin d'obtenir une fonction objectif mesurée dans les unités de variance et sur la même échelle que les variances, il est commode de définir une variance multicritères

$$V_{\text{plan}} = (V_{\text{plan}}[\bar{y}_1] V_{\text{plan}}[\bar{y}_2] V_{\text{plan}}[d])^{1/3}, \quad (5.2)$$

et d'exprimer le problème d'optimisation comme étant la maximisation de cette expression.

La caractérisation analytique du plan qui optimise (5.2) devient relativement lourde. À la place, nous utilisons l'exemple numérique de la section précédente pour démontrer comment la prise en compte des autres objectifs de conception de l'enquête affecte le choix du plan. Nous devrions nous attendre à ce que, dans le cas des plans pour lesquels le coût du suivi est plus élevé ($c_{12}^I \geq c_1^I + c_2^I$, $c_{12}^{II} \geq c_1^{II} + c_2^{II}$), les plans plus simples soient sélectionnés plus fréquemment : le plan à panel de grappes peut être préféré au plan à panel d'unités d'observation, et le plan à échantillons indépendants peut être préféré au plan à panel de grappes. Pour les conditions de base (4.1), nous avons

$$\begin{aligned} V_{e,i}[\bar{y}] &= 49,93, & V_{e,c}[\bar{y}] &= 47,68, & V_{e,o}[\bar{y}] &= 61,69, \\ V_{e,i} &= 62,91, & V_{e,c} &= 59,23, & V_{e,o} &= 70,02, \end{aligned}$$

où les indices temporels de $y_{j..}$ sont omis. Le plan à panel d'unités d'observation est assez inefficace pour l'estimation de la moyenne propre à la période, car un plus petit nombre d'unités sont échantillonnées. En revanche, le plan à panel de grappes est le plus efficace, suivi de près par le plan à échantillons indépendants.

Les figures 7 à 12 sont la réplique des figures 1 à 6, respectivement. Puisque le meilleur plan en ce qui concerne V est maintenant le plan à panel de grappes, la plupart de ces graphiques indiquent la préférence pour ce plan. La figure 7 montre que, si l'on tient compte des variances des moyennes contemporaines, les plans à échantillons indépendants et à panel de grappes plus simple sont préférés pour une fraction plus importante de configuration des paramètres, et occupent une part plus grande du graphique dans la figure 1. Le point où les trois plans sont équivalents correspond à $c_{12}^I = 20,6$, $c_{12}^{II} = 2,27$, qui est plus proche de l'origine qu'à la figure 1, dans laquelle seule la variance de la différence était prise en compte.

La figure 8 révèle que le plan à panel d'unités d'observation n'est justifié que si les autocorrélations sont toutes deux supérieures à 0,6 (pour les valeurs données des variances de population et des coûts). Rappelons qu'à la figure 2, le plan à panel d'unités d'observation avait la préférence quand $\rho^{II} > 0,34$, le choix étant peu dépendant de ρ^I .

La figure 9 montre comment les corrélations et les coûts au niveau des UPE influencent le choix du plan. Le plan à panel d'unités d'observation est moins efficace que celui à panel de grappes pour toutes les combinaisons de paramètres dans ce graphique. Donc, le choix se fait entre le plan à échantillons indépendants et le plan à panel de grappes. Évidemment, si la collecte des données en mode de panel est coûteuse, le plan à échantillons indépendants sera préféré au plan à panel de grappes. Fait intéressant, la préférence pour un plan particulier n'est pas monotone en

ρ_{12}^I . Pour des valeurs de $\rho_{12}^I > 0,7$, la composante $V[d]$ dans (5.2) produit des plans contenant si peu de grappes que cela affecte suffisamment fort $V[\bar{y}]$ pour nuire à la fonction objectif complète. À cette valeur de l'autocorrélation du panel, le coût de panel maximal auquel le plan à panel de grappes demeure le plus efficace est $c_{12}^I = 24,4$, ce qui signifie que le coût au niveau de la grappe est 44 % plus élevé à la deuxième vague qu'à la première.

La figure 10 montre qu'une plus forte autocorrélation des mesures des USE peut justifier une augmentation modérée du coût associé à la collecte des données. Le coût plus élevé pour lequel le plan à panel d'unités d'observation demeure le plus efficace est $c_{12}^{II} = 2,75$, avec $\rho^{II} = 0,78$; cela signifie que le coût de la deuxième interview peut être 75 % plus élevé que celui de la première.

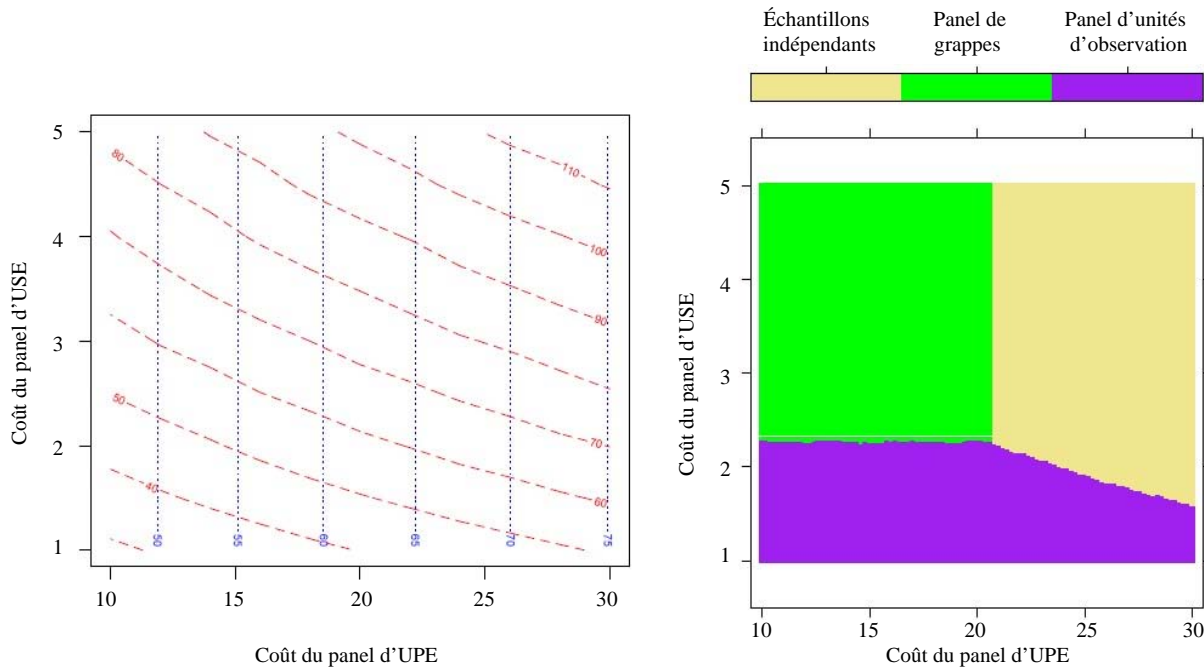


Figure 7 Variance sous le plan en fonction des coûts de collecte des données c_{12}^I, c_{12}^{II} . À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret); $V_{e,u} = 62,91$; à droite : domaines d'optimalité des trois plans

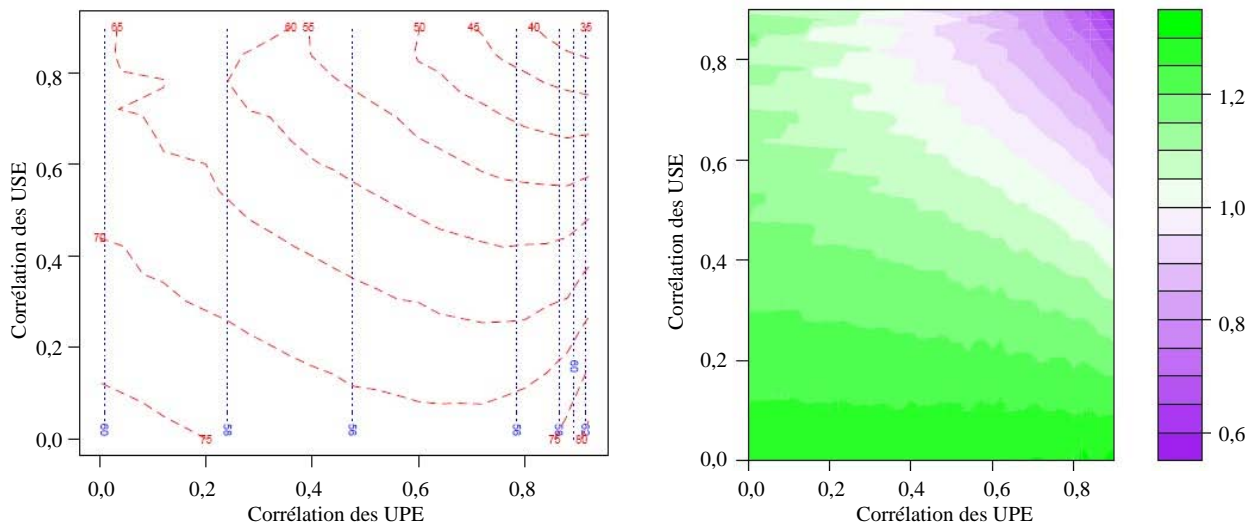


Figure 8 Variance sous le plan en fonction des autocorrélations ρ^I, ρ^{II} . À gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret); $V_{e,u} = 62,91$; à droite : ratio $V_{e,o}/V_{e,c}$

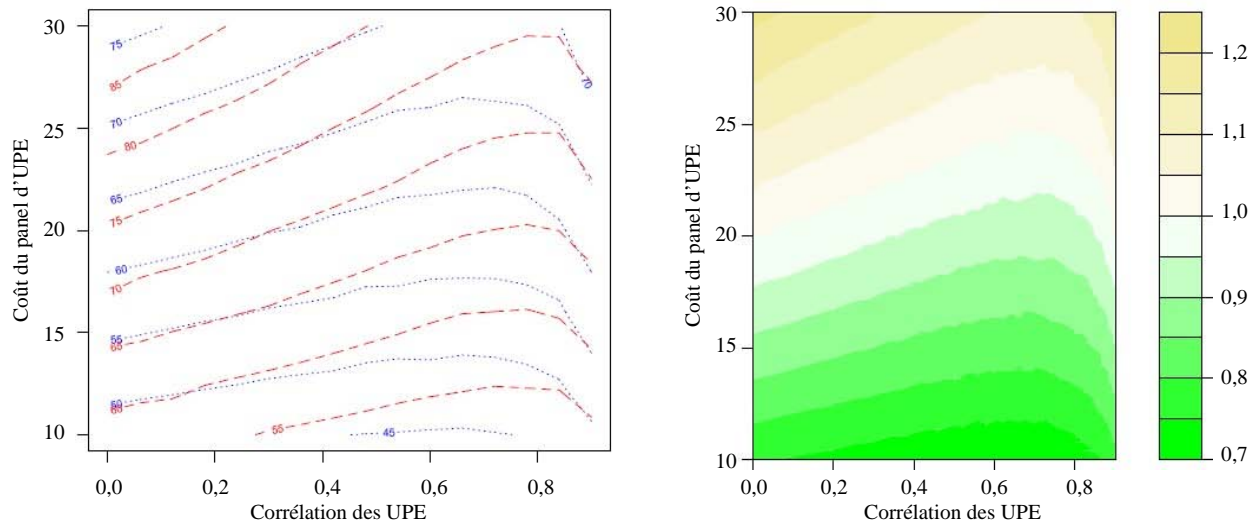


Figure 9 Variations sous le plan en fonction de l'autocorrélation au niveau de la grappe ρ^I et du coût c_{12}^I . à gauche : lignes de contour de $V_{e,c}$ (pointillé) et de $V_{e,o}$ (tiret) ; $V_{e,t} = 62,91$; à droite : ratio $V_{e,c}/V_{e,t}$

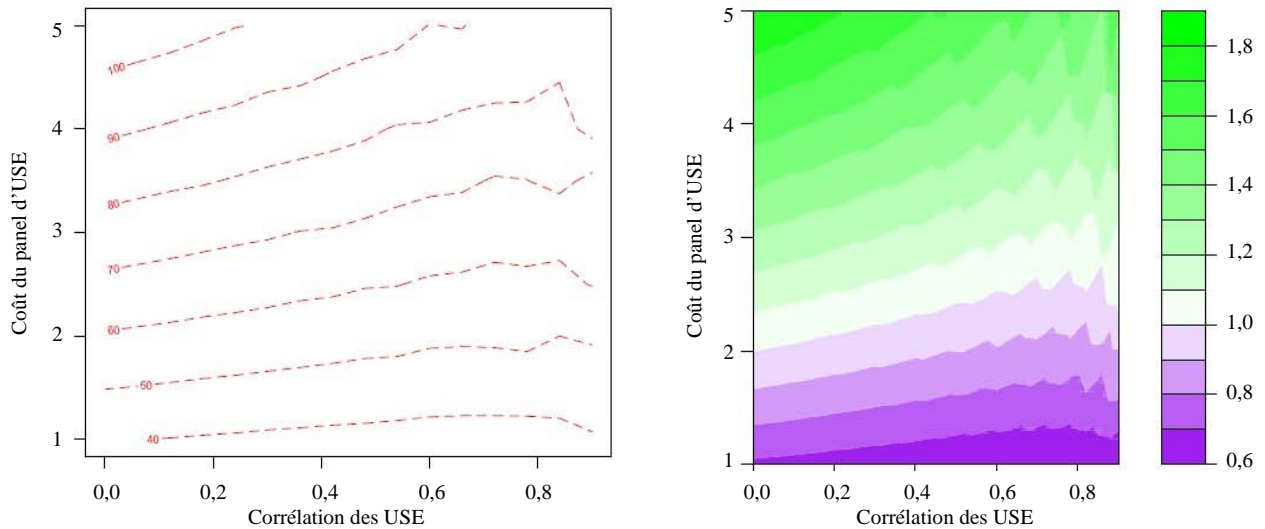


Figure 10 Variations sous le plan en fonction de l'autocorrélation au niveau des unités d'observations ρ^{II} et du coût c_{12}^{II} . À gauche : lignes de contour de $V_{e,o}$ (tiret) ; $V_{e,t} = 62,91$; $V_{e,c} = 59,23$; à droite : ratio $V_{e,o}/V_{e,c}$

La figure 11 est l'équivalent de la figure 5. La partie de gauche montre que le plan à panel d'unités d'observation est moins efficace que le plan à panel de grappes. La partie de droite montre que, si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 15 unités ce coût à la première vague, le plan à échantillons indépendants est plus efficace que le plan à panel de grappes.

Enfin, la figure 12 donne les variances en fonction du budget total de l'enquête et du coût de la collecte de données par panel. Le graphique ne révèle qu'une très légère dépendance à l'égard de C_0 , et le plan à échantillons indépendants est celui qui a la préférence si le mode de collecte par panel est trop coûteux, à savoir si le coût au niveau de la grappe à la deuxième vague dépasse de plus de 107 % le coût à la première vague.

Comme nous l'avions conjecturé au début de la présente section, l'intégration des variances des moyennes contemporaines dans la fonction objectif d'optimisation du plan déplace les préférences du concepteur d'enquête vers des plans plus simples permettant d'échantillonner un plus grand nombre d'unités d'observation finales. Le plan à panel d'unités d'observation ne se justifie maintenant que si les autocorrélations des UPE ainsi que des USE sont élevées, et que les coûts de collecte par panel sont raisonnablement faibles. En outre, le plan à panel de grappes n'est généralement justifié qu'en cas de réduction du coût au niveau de la grappe à la deuxième vague de l'enquête.

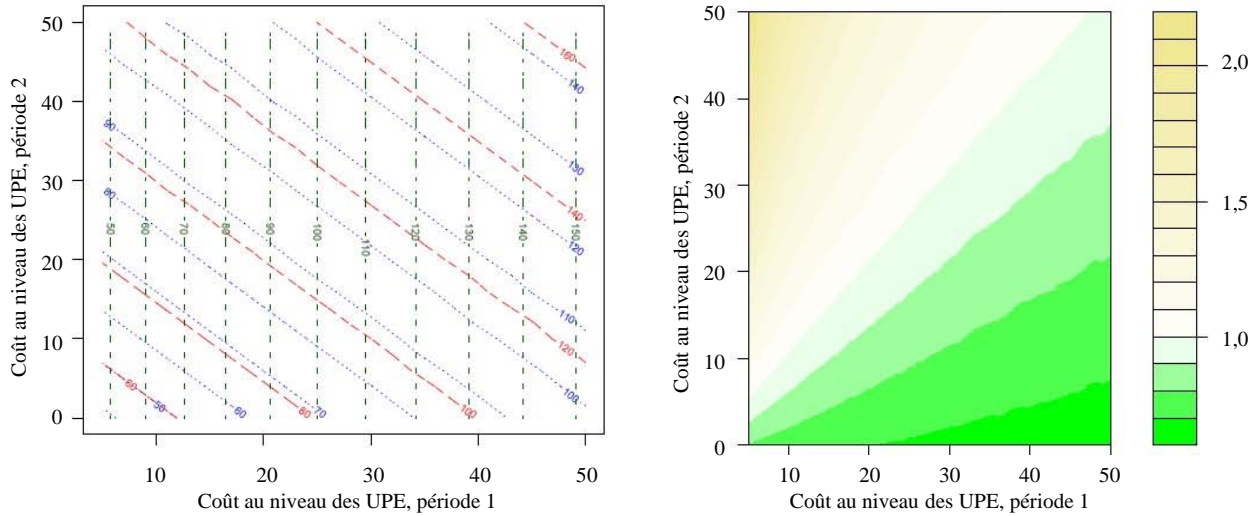


Figure 11 Variations sous le plan en fonction des coûts de collecte des données c_1^I, c_{12}^I . À gauche : lignes de contour de $V_{e,c}$ (pointillé), $V_{e,o}$ (tiret) et $V_{e,t}$ (tiret et point) ; à droite : ratio $V_{e,c}/V_{e,t}$

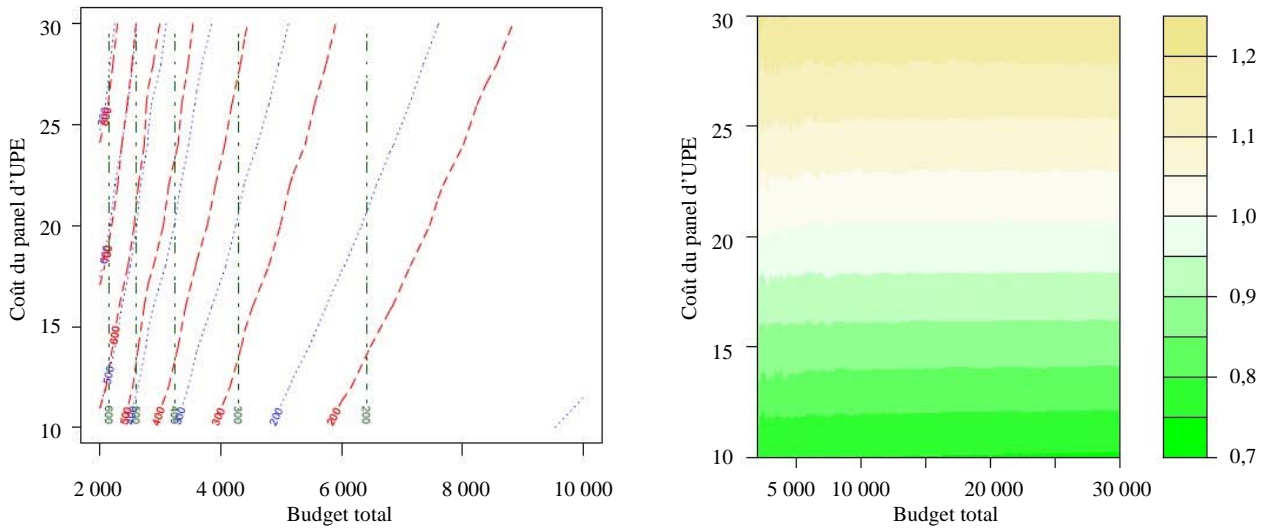


Figure 12 Variations sous le plan en fonction du budget total C_0 et du coût du panel d'UPE c_{12}^H . À gauche : lignes de contour de $V_{e,c}$ (pointillé), $V_{e,o}$ (tiret) et $V_{e,t}$ (tiret et point) ; à droite : domaines d'optimalité des trois plans

6. Extension à des vagues multiples

Si l'enquête à concevoir doit comporter plus de deux vagues de collecte des données, le concepteur de l'enquête pourrait étendre le cadre du problème de maximisation de l'utilité (5.1) en tenant compte des considérations qui suivent.

1. Un plus grand nombre de cibles d'inférence. Les variances possibles que le concepteur d'enquête doit éventuellement prendre en considération peuvent maintenant inclure les variances contemporaines $V[\bar{y}_1], V[\bar{y}_2], \dots, V[\bar{y}_T]$; les différences consécutives $V[\bar{y}_2 - \bar{y}_1], \dots, V[\bar{y}_T - \bar{y}_{T-1}]$ ou les estimateurs composite/MCG de la variation entre deux périodes adjacentes ; d'autres contrastes $V[\sum_t c_t \bar{y}_t]$,

$\sum c_t = 0$; la variance des taux de croissance linéaires provenant de la régression de \bar{y}_t sur t , estimée par les MCO ou les MCG, etc.

2. La possibilité d'une actualisation. En économie, on spécifie d'habitude les contraintes budgétaires faisant référence à l'avenir sous la forme $\sum_t x_t \delta^t$ où x_t est le montant dépensé au temps t , et $\delta < 1$ est le facteur d'actualisation associé aux taux d'intérêt. L'actualisation peut également être pertinente pour la fonction d'utilité, et les variances sous le plan qui se situe plus loin dans l'avenir peuvent avoir moins de poids dans le problème d'optimisation.
3. Des formes fonctionnelles inconnues pour les processus chronologiques associés à la variable d'intérêt. Le concepteur d'enquête doit avoir une bonne idée

de la structure de covariance de la série chronologique d'observations individuelles ainsi que de moyennes de grappes. Il est probable que les résultats seront sensibles au choix d'un modèle particulier. Dans la présente analyse, le problème est plus simple, car il suffit d'avoir un seul paramètre de corrélation pour chaque niveau. Le concepteur d'enquête pourrait devoir introduire plus de paramètres dans le modèle et, conséquemment, étudier la sensibilité du plan choisi à ces divers paramètres.

La complexité du problème, tel qu'il est exposé ci-dessus, peut devenir très rapidement difficile à maîtriser. Nous nous abstenons donc de présenter un traitement plus détaillé dans le présent article.

7. Discussion

Le présent article décrit l'analyse de diverses options de mise en œuvre d'enquêtes à échantillonnage en grappes répétées. Nous fournissons une expression analytique pour les variances sous le plan de l'estimateur par simple différence pour trois plans de sondage fréquemment utilisés (les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation). Nous dérivons également les tailles d'échantillon optimales pour l'estimation de la différence entre deux vagues de collecte des données.

Le concepteur d'enquête qui sait que la caractéristique d'intérêt persistera dans une certaine mesure au cours du temps choisira vraisemblablement l'un des plans à collecte par panel, à condition que les coûts des nouvelles visites aux grappes et/ou aux unités d'observation ne soient pas prohibitifs. La comparaison analytique entre les plans à échantillons indépendants et à panel d'unités d'observation est possible et est donnée par la proposition 7. Il convient de préciser que la variance sous le plan de la différence est d'ordre $< O(C_0^{-1})$ pour le plan à échantillons indépendants ainsi que pour le plan à panel d'unités d'observation, et qu'elle est d'ordre $O(C_0^{-1/2})$ pour le plan à panel de grappes, où C_0 est le budget total de l'enquête. Le plan à panel de grappes n'est donc viable que pour les petites enquêtes, tandis que le plan à échantillons indépendants ou à panel d'unités d'observation sera vraisemblablement choisi pour les enquêtes à grande échelle.

La structure de coût considérée à la section 3 est assez simpliste. Par exemple, les coûts d'échantillonnage de deuxième degré à la deuxième période pourraient différer selon que les unités sont échantillonnées à partir de nouvelles grappes ou à partir de grappes réutilisées. En outre, le coût peut dépendre de la taille de grappe M_i , car plus de temps et plus de ressources pourraient être nécessaires pour obtenir les cartes et recueillir les données au niveau de la

grappe dans le cas des grandes grappes. Au départ, nous avons l'intention de considérer des situations dans lesquelles le coût du panel d'USE était plus de deux fois plus élevé que le coût des interviews individuelles. Cependant, comme l'a suggéré l'un des examinateurs, ce coût pourrait être plus faible si les interviews de suivi sont effectuées selon un mode moins coûteux, tel qu'une interview téléphonique ou une enquête par la poste avec questionnaire à remplir soi-même au lieu d'une interview sur place. Le cas échéant, le plan à panel d'unités d'observation semble être le plus rentable des trois.

La structure de population est également sursimplifiée. Nous avons supposé que les grappes sont de taille équilibrée constante. Aucune unité de la population n'en sort et aucune nouvelle unité n'y apparaît. Dans de nombreuses situations pratiques, ces hypothèses sont assez contraignantes. Si la population évolue entre deux vagues de collecte des données, le concepteur d'enquête souhaitera inclure de nouvelles grappes à la deuxième vague, en utilisant les algorithmes de Ernst (1999). Les nouvelles grappes sont placées dans une strate distincte, d'où est tiré un échantillon en grappes. Dans le cas de la NHIS, cette approche est mise en œuvre en se servant comme base de sondage d'une liste de « permis de bâtir ». En outre, les effets de mesure dynamiques, tels que le conditionnement et le temps passé dans l'échantillon, donnent lieu à un biais de rotation de l'échantillon, de sorte qu'il pourrait être avantageux de prévoir au moins une certaine rotation des UPE.

En particulier, pour les études réalisées dans le cadre des DHS, le premier argument (couverture) est vraisemblablement plus important que le second (durée de la présence dans l'échantillon), étant donné l'intervalle de temps important entre les cycles de l'enquête (environ cinq ans). La non-réponse et la perte de couverture pourraient sans doute être ajoutées toutes deux au cadre courant comme sources de biais, entraînant ainsi l'optimisation de la moyenne quadratique de l'erreur totale d'enquête plutôt que de la variance sous le plan. Toutefois, la formulation de modèles convaincants de ce genre de biais pourrait être difficile.

Un autre problème qui se poserait si les grappes étaient de tailles différentes est celui de la plus grande gamme de plans applicables. Dans le présent article, nous avons supposé que l'on procédait à un échantillonnage aléatoire simple sans remise aux deux degrés d'échantillonnage. D'autres plans, tels que l'échantillonnage avec probabilités proportionnelles à la taille (PPT), peuvent aussi être utilisés. Pour d'autres plans que l'échantillonnage aléatoire simple, il faudrait se servir de l'estimateur d'Horvitz-Thompson et de l'estimateur de sa variance (Särndal, Swensson et Wretman 1992, Thompson 1997). Les calculs analytiques deviennent compliqués, mais des démonstrations numériques pratiques

semblables à celles de nos sections 4 et 5 peuvent encore être effectuées. Si la taille des grappes varie au cours du temps, l'obtention du plan optimal devient une cible mouvante et les plans qui sont optimaux pour les « anciennes » mesures de taille perdront de leur efficacité avec les « nouvelles ».

Dans des ébauches antérieures du présent article, nous avons analysé des plans intermédiaires dans lesquels une fraction non négligeable des unités est retenue et d'autres unités sont échantillonnées indépendamment. Le problème peut alors être considéré comme une minimisation de la variance sous des contraintes d'inégalité appliquées au degré de chevauchement $0 \leq \pi^I \leq 1$, $0 \leq \pi^{II} \leq 1$. La théorie générale de l'optimisation sous contraintes non linéaires fait en sorte qu'à condition que la variance de la variation de la moyenne de population D soit monotone en π^I et π^{II} , la solution optimale sera atteinte à l'un des sommets de l'espace des paramètres. Cela justifie notre intérêt pour les trois plans examinés dans le présent article. Ces plans correspondent aux sommets de l'espace des paramètres de coordonnées $(0, 0)$, $(1, 0)$ et $(1, 1)$ pour les plans à échantillons indépendants, à panel de grappes et à panel d'unités d'observation, respectivement. Le point $(0, 1)$ correspond à un plan impossible avec chevauchement complet des unités individuelles sans aucun chevauchement des grappes. Des calculs fastidieux montrent qu'il est possible de satisfaire les conditions de premier ordre dans certains cas intermédiaires également, mais ils correspondent à des maxima locaux de la variance. Bien que ces résultats puissent aussi être intéressants (en ce sens qu'ils fournissent une borne supérieure pour les variances sous le plan), nous ne les avons pas pris en considération dans le présent article. Dans les cas plus compliqués d'optimisation à critères multiples de la section 5, la monotonie n'est pas nécessairement vérifiée et d'autres plans que les trois cas extrêmes considérés dans l'article peuvent produire les valeurs optimales de la fonction objectif (5.2).

Les conditions d'égalité des variances (2.9) peuvent être relâchées au prix de la production d'expressions considérablement plus compliquées. Si les tailles d'échantillon sont maintenues fixes entre les deux vagues de l'enquête, les changements qui suivent peuvent être nécessaires dans toutes les formules pertinentes. Dans les expressions qui ne font pas intervenir les autocorrélations,

$$2S_b^2 \mapsto S_{1b}^2 + S_{2b}^2, \quad 2S_w^2 \mapsto \bar{S}_{1w}^2 + \bar{S}_{2w}^2, \quad (7.1)$$

tandis que dans les expressions dans lesquelles interviennent les autocorrélations,

$$\begin{aligned} 2(1 - \rho^I)S_b^2 &\mapsto S_{1b}^2 + S_{2b}^2 - 2\rho^I S_{1b}S_{2b}, \\ 2S_w^2(1 - \rho^{II}) &\mapsto \bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w}\bar{S}_{2w}. \end{aligned} \quad (7.2)$$

Qualitativement, les résultats seront les mêmes.

Le cadre multicritères de la section 5 permet d'attribuer des poids d'importance différents aux diverses variances d'intérêt. Des valeurs relativement grandes de α_1 , α_2 correspondent à l'importance plus grande des moyennes contemporaines, tandis que de plus grandes valeurs de α_3 correspondent à l'importance plus grande de l'estimation de la variation. Le problème original d'optimisation du plan pour $V[d]$ peut être considéré dans le contexte de (5.1) en posant que $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 1$. Ce cadre peut aussi être élargi afin d'inclure les plans destinés à mesurer plusieurs variables. Un défi supplémentaire associé à ce genre de configuration est que les autocorrélations peuvent différer d'une variable à l'autre. Certaines caractéristiques individuelles sont constantes au cours du temps (race, sexe), tandis que d'autres évoluent lentement (logement, dépenses, préférences politiques), et d'autres encore évoluent plus rapidement (revenus ou comportements).

Le présent article traite de trois plans d'échantillonnage et d'un estimateur particulier de la variation, à savoir la différence entre les estimations de la moyenne à deux périodes dans le temps. D'autres options existent en ce qui concerne tant les plans que les estimateurs. Par exemple, dans les plans à échantillon rotatif, une fraction des unités de la première vague est retenue et certaines nouvelles unités sont recrutées. Pour ce genre de plans, l'estimation composite (Hansen et coll. 1953, Patterson 1950, Rao et Graham 1964, Wolter 2007) dans laquelle sont pondérées différemment les contributions des unités indépendantes (celles retirées de l'échantillon après la première vague et celles nouvellement recrutées pour la deuxième vague) et les contributions des unités du panel (utilisées aux deux vagues) produirait des estimations plus efficaces. En général, ce sont des considérations non liées à l'échantillonnage, telles qu'une diminution du fardeau de réponse et la détérioration de la représentativité de l'échantillon en raison de l'évolution de la population, qui motive l'utilisation de ce genre de plans. Ces considérations peuvent être prises en compte dans le modèle de coût (par exemple, un plus grand nombre de rappels nécessaires pour convaincre une unité de répondre) ou dans le modèle de l'erreur totale d'enquête (en introduisant le biais de non-réponse ou de sous-dénombrement et en considérant l'erreur quadratique moyenne d'une estimation plutôt que sa variance sous le plan).

Remerciements

Les auteurs remercient Chris Skinner et John Eltinge de leurs discussions utiles, William Kalsbeek, de ses suggestions aux premières étapes de la rédaction de l'article, ainsi que le rédacteur associé et deux examinateurs de leurs commentaires. Nash Herndon et Oksana Loginova ont amélioré la rédaction. Un appui financier partiel a été fourni par

la U.S. Agency for International Development dans le cadre du Measure Evaluation Project du Carolina Population Center de l'Université de la Caroline du Nord à Chapel Hill, aux termes de l'entente de coopération GPO-A-00-03-00003-00. Les auteurs remercient aussi de leurs commentaires constructifs les participants aux Joint Statistical Meetings (2005) et au XXIII^e Symposium international sur les questions de méthodologie de Statistique Canada (2007).

Annexe

Dans les preuves qui suivent, les espérances, les variances et les covariances sont calculées par rapport au plan de sondage correspondant. Le premier degré de sélection sera désigné par l'indice supérieur I. Le deuxième degré de sélection sera désigné par l'indice supérieur II.

Preuve de la proposition 2. Désignons l'échantillon d'UPE par \mathcal{S}^I , l'échantillon d'USE à la première période par \mathcal{S}_{i1}^{II} , et l'échantillon d'USE à la deuxième période par \mathcal{S}_{i2}^{II} . Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn} \sum_{i \in \mathcal{S}^I} \left(\sum_{j \in \mathcal{S}_{i2}^{II}} y_{2ij} - \sum_{j \in \mathcal{S}_{i1}^{II}} y_{1ij} \right).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par E_I , et celles par rapport au deuxième degré d'échantillonnage par E_{II} , nous avons que la variance sous le plan de d est égale à

$$\begin{aligned} V[d] &= E_I V_{II}[d | \mathcal{S}^I] + V_I E_{II}[d | \mathcal{S}^I] \\ &= \frac{1}{m^2 n^2} E_I \left\{ \sum_{i \in \mathcal{S}^I} V_{II} \left[\sum_{j \in \mathcal{S}_{i2}^{II}} y_{2ij} - \sum_{j \in \mathcal{S}_{i1}^{II}} y_{1ij} \right] \right\} \\ &\quad + \frac{1}{m^2 n^2} V_I \left\{ \sum_{i \in \mathcal{S}^I} E_{II} \left[\sum_{j \in \mathcal{S}_{i2}^{II}} y_{2ij} - \sum_{j \in \mathcal{S}_{i1}^{II}} y_{1ij} \right] \right\} \\ &= \frac{1}{m^2 n^2} E_I \left\{ \sum_{i \in \mathcal{S}^I} V_{II} \left[\sum_{j \in \mathcal{S}_{i2}^{II}} y_{2ij} \right] + V_{II} \left[\sum_{j \in \mathcal{S}_{i1}^{II}} y_{1ij} \right] \right\} \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in \mathcal{S}^I} m \bar{Y}_{2i.} - m \bar{Y}_{1i.} \right] \\ &= \frac{1}{m^2 n^2} E_I \left[\sum_{i \in \mathcal{S}^I} \left(1 - \frac{m}{M} \right) m S_{2wi}^2 + \left(1 - \frac{m}{M} \right) m S_{1wi}^2 \right] \\ &\quad + \frac{1}{n^2} \left(1 - \frac{n}{N} \right) n (S_{1b}^2 + S_{2b}^2 - 2\rho^1 S_{1b} S_{2b}) \\ &= \frac{1}{m^2 n^2} nm \left(1 - \frac{m}{M} \right) m (S_{2w}^2 + S_{1w}^2) \\ &\quad + \frac{1}{n} \left(1 - \frac{n}{N} \right) (S_{1b}^2 + S_{2b}^2 - 2\rho^1 S_{1b} S_{2b}) \\ &= \left(1 - \frac{n}{N} \right) \frac{2S_b^2(1 - \rho^1)}{n} + \left(1 - \frac{m}{M} \right) \frac{2S_w^2}{mn}, \end{aligned}$$

où la dernière égalité repose sur l'hypothèse des conditions symétriques (2.9).

Preuve de la proposition 3. Désignons l'échantillon d'UPE par \mathcal{S}^I , et l'échantillon d'USE par \mathcal{S}_i^{II} . Alors

$$d = \bar{y}_{2..} - \bar{y}_{1..} = \frac{1}{mn} \sum_{i \in \mathcal{S}^I} \sum_{j \in \mathcal{S}_i^{II}} (y_{2ij} - y_{1ij}).$$

En désignant les espérances par rapport au premier degré d'échantillonnage par E_I , et celles par rapport au deuxième degré d'échantillonnage par E_{II} , nous avons que la variance sous le plan de d est égale à

$$\begin{aligned} V[d] &= E_I V_{II}[d | \mathcal{S}^I] + V_I E_{II}[d | \mathcal{S}^I] \\ &= \frac{1}{m^2 n^2} E_I \left[\sum_{i \in \mathcal{S}^I} V_{II} \sum_{j \in \mathcal{S}_i^{II}} (y_{2ij} - y_{1ij}) \right] \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in \mathcal{S}^I} E_{II} \sum_{j \in \mathcal{S}_i^{II}} (y_{2ij} - y_{1ij}) \right] \\ &= \frac{1}{m^2 n^2} E_I m \left[\sum_{i \in \mathcal{S}^I} \left(1 - \frac{m}{M} \right) (S_{2wi}^2 + S_{1wi}^2 - 2S_{2wi} S_{1wi} \rho^{II}) \right] \\ &\quad + \frac{1}{m^2 n^2} V_I \left[\sum_{i \in \mathcal{S}^I} m (\bar{Y}_{2ij} - \bar{Y}_{1ij}) \right] \\ &= \frac{1}{mn^2} n \left(1 - \frac{m}{M} \right) (\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}) \\ &\quad + \frac{1}{n^2} n \left(1 - \frac{n}{N} \right) (S_{1b}^2 + S_{2b}^2 - 2\rho^1 S_{1b} S_{2b}) \\ &= \left(1 - \frac{n}{N} \right) \frac{S_{1b}^2 + S_{2b}^2 - 2\rho^1 S_{1b} S_{2b}}{n} \\ &\quad + \left(1 - \frac{m}{M} \right) \frac{\bar{S}_{1w}^2 + \bar{S}_{2w}^2 - 2\rho^{II} \bar{S}_{1w} \bar{S}_{2w}}{mn} \\ &= 2 \left(1 - \frac{n}{N} \right) \frac{S_b^2(1 - \rho^1)}{n} \\ &\quad + 2 \left(1 - \frac{m}{M} \right) \frac{S_w^2(1 - \rho^{II})}{mn}, \end{aligned}$$

avec la dernière égalité vérifiée sous les conditions de symétrie

Preuve de la proposition 4. La fonction lagrangienne de minimisation de (2.11) sous la contrainte (3.1) est

$$\begin{aligned} L(n_1, m_1, n_2, m_2, \lambda) &= \\ &= \left(1 - \frac{n_1}{N} \right) \frac{S_b^2}{n_1} + \left(1 - \frac{n_2}{N} \right) \frac{S_b^2}{n_2} \\ &\quad + \left(1 - \frac{m_1}{M} \right) \frac{\bar{S}_w^2}{n_1 m_1} + \left(1 - \frac{m_2}{M} \right) \frac{\bar{S}_w^2}{n_2 m_2} \\ &\quad - \lambda (c_1^1 n_1 + c_1^{II} n_1 m_1 + c_2^1 n_2 + c_2^{II} n_2 m_2 - C_0). \end{aligned}$$

La résolution des conditions de premier ordre de cette fonction lagrangienne donne

$$\begin{aligned} -\lambda &= \frac{m_1 S_b^2 + \left(1 - \frac{m_1}{M}\right) \bar{S}_w^2}{n_1^2 m_1 (c_1^I + c_1^{II} m_1)} = \frac{m_2 S_b^2 + \left(1 - \frac{m_2}{M}\right) \bar{S}_w^2}{n_2^2 m_2 (c_2^I + c_1^{II} m_2)} \\ &= \frac{\bar{S}_w^2}{m_1^2 n_1^2 c_1^{II}} = \frac{\bar{S}_w^2}{m_2^2 n_2^2 c_2^{II}}. \end{aligned}$$

En utilisant ces conditions, nous avons

$$m^2 n^2 c^{II} \left[m S_b^2 + \left(1 - \frac{m}{M}\right) \bar{S}_w^2 \right] = n^2 m (c^I + c^{II} m) \bar{S}_w^2,$$

qui peut s'écrire sous la forme

$$\begin{aligned} 0 &= (c^I + c^{II} m) \bar{S}_w^2 - m c^{II} \left[m S_b^2 + \left(1 - \frac{m}{M}\right) \bar{S}_w^2 \right] \\ &= (c^I + c^{II} m) M \bar{S}_w^2 - m c^{II} [M m S_b^2 + (M - m) \bar{S}_w^2] \\ &= c^I M \bar{S}_w^2 + m c^{II} M \bar{S}_w^2 - m^2 c^{II} M S_b^2 - m c^{II} M \bar{S}_w^2 + m^2 c^{II} \bar{S}_w^2 \\ &= c^I M \bar{S}_w^2 + m^2 c^{II} (\bar{S}_w^2 - M S_b^2). \end{aligned}$$

D'où,

$$m = \sqrt{\frac{c^I \bar{S}_w^2}{c^{II} S_b^2 - \bar{S}_w^2 / M}}.$$

D'après le budget de l'enquête (3.1), nous trouvons que le nombre de grappes est

$$n = \frac{C_0}{2(c^I + m c^{II})} = \frac{C_0}{2\{c^I + [c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}.$$

En introduisant ces expressions dans (2.11) et en utilisant les relations d'égalité (2.9), nous obtenons la variance de l'estimateur sous la forme

$$\begin{aligned} V_{e,i}[d] &= 2 \left(1 - \frac{n}{N}\right) \frac{S_b^2}{n} + 2 \left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2}{mn} \\ &= 2 \left[1 - \frac{C_0}{2(c^I + m c^{II})N}\right] \frac{2(c^I + m c^{II}) S_b^2}{C_0} \\ &\quad + 4 \left(1 - \frac{m}{M}\right) \frac{\bar{S}_w^2 (c^I + m c^{II})}{m C_0} \\ &= 2 \left[\frac{2(c^I + m c^{II})}{C_0} - \frac{1}{N} \right] S_b^2 \\ &\quad + 4 \left(\frac{1}{m} - \frac{1}{M} \right) \frac{\bar{S}_w^2 (c^I + m c^{II})}{C_0} \\ &= \frac{4(c^I + m c^{II})}{C_0} \left[S_b^2 + \left(\frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 \right] - \frac{2}{N} S_b^2 \\ &= \frac{4 \left[c^I + \sqrt{c^I c^{II} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)} \right]}{C_0} \\ &\quad \times \left[S_b^2 + \left(\sqrt{\frac{c^{II} S_b^2 - \bar{S}_w^2 / M}{c^I}} - \frac{1}{M} \right) \bar{S}_w^2 \right] - \frac{2}{N} S_b^2. \end{aligned}$$

Preuve de la proposition 5. La fonction lagrangienne de minimisation de (2.13) sous la contrainte (3.4) est

$$\begin{aligned} L(n, m_1, m_2, \lambda) &= 2 \left(1 - \frac{n}{N}\right) \frac{(1 - \rho^I) S_b^2}{n} \\ &\quad + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n m_1} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n m_2} \\ &\quad - \lambda (c_{12}^I n + c_1^{II} n m_1 + c_2^{II} n m_2 - C_0). \end{aligned}$$

Les conditions de premier ordre sont :

$$\begin{aligned} \frac{\partial L}{\partial n} &= -2 \frac{(1 - \rho^I) S_b^2}{n^2} - \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{n^2 m_1} \\ &\quad - \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{n^2 m_2} - \lambda (c_{12}^I + c_1^{II} m_1 + c_2^{II} m_2), \\ \frac{\partial L}{\partial m_1} &= -\frac{\bar{S}_{1w}^2}{n m_1^2} - \lambda c_1^{II} n, \\ \frac{\partial L}{\partial m_2} &= -\frac{\bar{S}_{2w}^2}{n m_2^2} - \lambda c_2^{II} n, \\ \frac{\partial L}{\partial \lambda} &= c_{12}^I n + c_1^{II} n m_1 + c_2^{II} n m_2 - C_0 = 0. \end{aligned}$$

En exprimant $-\lambda n$ d'après ces conditions, nous obtenons :

$$\begin{aligned} -\lambda n &= \frac{\bar{S}_{1w}^2}{m_1^2 n c_1^{II}} = \frac{\bar{S}_{2w}^2}{m_2^2 n c_2^{II}} = 2(1 - \rho^I) \frac{S_b^2}{C_0} \\ &\quad + \left(1 - \frac{m_1}{M}\right) \frac{\bar{S}_{1w}^2}{m_1 C_0} + \left(1 - \frac{m_2}{M}\right) \frac{\bar{S}_{2w}^2}{m_2 C_0}. \end{aligned}$$

Alors

$$\begin{aligned} \frac{1}{m_2} &= \frac{1}{m_1} \sqrt{\frac{c_2^{II} \bar{S}_{1w}^2}{c_1^{II} \bar{S}_{2w}^2}} \equiv \frac{1}{\kappa m_1}, \\ \frac{1}{m_1^2} \frac{(c_{12}^I + c_1^{II} m_1 + \kappa c_2^{II} m_1) \bar{S}_{1w}^2}{c_1^{II}} &= 2(1 - \rho^I) S_b^2 \\ &\quad + \left(\frac{1}{m_1} - \frac{1}{M} \right) \bar{S}_{1w}^2 + \left(\frac{1}{\kappa m_1} - \frac{1}{\kappa M} \right) \bar{S}_{2w}^2, \\ 0 &= [2(1 - \rho^I) S_b^2 \kappa c_1^{II} - \bar{S}_{1w}^2 \kappa c_1^{II} / M - \bar{S}_{2w}^2 c_1^{II} / M] m_1^2 \\ &\quad + [\bar{S}_{1w}^2 \kappa c_1^{II} + \bar{S}_{2w}^2 c_1^{II} - c_1^{II} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{II} \bar{S}_{1w}^2] m_1 - c_{12}^I \bar{S}_{1w}^2 \kappa, \\ D &= [\bar{S}_{1w}^2 \kappa c_1^{II} + \bar{S}_{2w}^2 c_1^{II} - c_1^{II} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{II} \bar{S}_{1w}^2]^2 \\ &\quad + 4[2(1 - \rho^I) S_b^2 \kappa - \bar{S}_{1w}^2 \kappa / M - \bar{S}_{2w}^2 / M] c_1^{II} c_{12}^I \bar{S}_{1w}^2 \geq 0, \\ m_1 &= \frac{c_1^{II} \bar{S}_{1w}^2 \kappa + \kappa^2 c_2^{II} \bar{S}_{1w}^2 - \bar{S}_{1w}^2 \kappa c_1^{II} - \bar{S}_{2w}^2 c_1^{II} \pm \sqrt{D}}{4(1 - \rho^I) S_b^2 \kappa c_1^{II} - 2 \bar{S}_{1w}^2 \kappa c_1^{II} / M - 2 \bar{S}_{2w}^2 c_1^{II} / M}. \end{aligned}$$

La solution pour $-\sqrt{D}$ donne une valeur négative de m_1 , et doit être écartée.

Les autres caractéristiques du plan sont

$$m_2 = \kappa m_1, \quad n = \frac{C_0}{c_{12}^I + m_1 c_1^{II} + m_2 c_1^{II}}, \quad \kappa = \sqrt{\frac{c_1^{II} \bar{S}_{2w}^2}{c_2^{II} \bar{S}_{1w}^2}}.$$

La variance de l'estimateur de la différence peut être calculée en utilisant (2.15).

Sous les conditions de symétrie, $\kappa = 1$, et

$$D = 4[2(1 - \rho^I)S_b^2 - 2\bar{S}_w^2/M] c^{II} c_{12}^I \bar{S}_w^2$$

est non négative, à moins que l'expression entre les crochets soit négative (ce qui ne peut se produire que si ρ^I est grande et que M est petite. Dans ce cas, une solution de coin $m = M$ est réalisée). En outre,

$$m = m_1 = m_2 = \sqrt{\frac{\bar{S}_w^2 c_{12}^I}{2[(1 - \rho^I)S_b^2 - \bar{S}_w^2/M] c^{II}}},$$

$$n = \frac{C_0}{c_{12}^I + 2mc^{II}} = \frac{C_0}{c_{12}^I + \sqrt{\frac{2\bar{S}_w^2 c_{12}^I c^{II}}{(1 - \rho^I)S_b^2 - \bar{S}_w^2/M}}},$$

$V_{e,o}[d]$

$$\begin{aligned} &= 2\left(1 - \frac{n}{N}\right) \frac{(1 - \rho^I)S_b^2}{n} + 2\left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II})\bar{S}_w^2}{nm} \\ &= \frac{2}{n} \left[(1 - \rho^I)S_b^2 + 2\left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II})\bar{S}_w^2}{m} \right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0} (c_{12}^I + 2mc^{II}) \\ &\quad \times \left[(1 - \rho^I)S_b^2 + 2\left(\frac{1}{m} - \frac{1}{M}\right) (1 - \rho^{II})\bar{S}_w^2 \right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0} (c_{12}^I + 2mc^{II}) \\ &\quad \times \left[(1 - \rho^I)S_b^2 + \frac{2}{m} (1 - \rho^{II})\bar{S}_w^2 - \frac{2}{M} (1 - \rho^{II})\bar{S}_w^2 \right] - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0} \left\{ c_{12}^I (1 - \rho^I)S_b^2 + 2(1 - \rho^{II})\bar{S}_w^2 \left[2c^{II} - \frac{c_{12}^I}{M} \right] \right. \\ &\quad \left. + \frac{2}{m} c_{12}^I (1 - \rho^{II})\bar{S}_w^2 + 2mc^{II} \left[(1 - \rho^I)S_b^2 - \frac{2}{M} (1 - \rho^{II})\bar{S}_w^2 \right] \right\} \\ &\quad - \frac{2(1 - \rho^I)S_b^2}{N} \\ &= \frac{2}{C_0} \left\{ c_{12}^I (1 - \rho^I)S_b^2 + 2(1 - \rho^{II})\bar{S}_w^2 \left[2c^{II} - \frac{c_{12}^I}{M} \right] \right. \\ &\quad \left. + 2(1 - \rho^{II})\sqrt{2[(1 - \rho^I)S_b^2 - \bar{S}_w^2/M] \bar{S}_w^2 c^{II} c_{12}^I} \right. \\ &\quad \left. + \sqrt{\frac{2\bar{S}_w^2 c_{12}^I c^{II}}{(1 - \rho^I)S_b^2 - \bar{S}_w^2/M}} \left[(1 - \rho^I)S_b^2 - \frac{2}{M} (1 - \rho^{II})\bar{S}_w^2 \right] \right\} \\ &\quad - \frac{2(1 - \rho^I)S_b^2}{N}. \end{aligned}$$

Preuve de la proposition 6. La fonction lagrangienne de minimisation de (2.15) sous la contrainte (3.6) est

$$L(n, m, \lambda) = 2\left(1 - \frac{n}{N}\right) \frac{(1 - \rho^I)S_b^2}{n} + 2\left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II})\bar{S}_w^2}{nm} - \lambda(c_{12}^I n + c_{12}^{II} nm - C_0).$$

Les conditions de premier ordre sont :

$$\begin{aligned} 0 &= \frac{\partial L}{\partial n} = -2 \frac{(1 - \rho^I)S_b^2}{n^2} \\ &\quad - 2\left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II})\bar{S}_w^2}{n^2 m} - \lambda(c_{12}^I + c_{12}^{II} m), \\ 0 &= \frac{\partial L}{\partial m} = -2 \frac{(1 - \rho^{II})\bar{S}_w^2}{nm^2} - \lambda c_{12}^{II} n, \\ 0 &= \frac{\partial L}{\partial \lambda} = c_{12}^I n + c_{12}^{II} nm - C_0. \end{aligned}$$

En exprimant $-\lambda n^2$ d'après ces conditions, nous obtenons :

$$\begin{aligned} -\lambda n^2/2 &= \frac{(1 - \rho^I)S_b^2}{c_{12}^I + c_{12}^{II} m} + \left(1 - \frac{m}{M}\right) \frac{(1 - \rho^{II})\bar{S}_w^2}{m(c_{12}^I + c_{12}^{II} m)} \\ &= \frac{(1 - \rho^{II})\bar{S}_w^2}{m^2 c_{12}^{II}}. \end{aligned}$$

D'où,

$$\begin{aligned} (1 - \rho^I)S_b^2 M m^2 c_{12}^{II} + (M - m)(1 - \rho^{II})\bar{S}_w^2 m c_{12}^{II} \\ - (1 - \rho^{II})M\bar{S}_w^2 (c_{12}^I + c_{12}^{II} m) &= 0, \\ [(1 - \rho^I)S_b^2 M c_{12}^{II} - (1 - \rho^{II})\bar{S}_w^2 c_{12}^{II}] m^2 \\ - [(1 - \rho^{II})M\bar{S}_w^2 c_{12}^I] &= 0, \end{aligned}$$

$$\begin{aligned} m &= \sqrt{\frac{(1 - \rho^{II})M\bar{S}_w^2 c_{12}^I}{[(1 - \rho^I)S_b^2 M - (1 - \rho^{II})\bar{S}_w^2] c_{12}^{II}}} \\ &= \sqrt{\frac{c_{12}^I}{c_{12}^{II}} \frac{(1 - \rho^{II})\bar{S}_w^2}{(1 - \rho^I)S_b^2 - (1 - \rho^{II})\bar{S}_w^2/M}}. \end{aligned}$$

D'après le budget de l'enquête (3.6),

$$n = \frac{C_0}{c_{12}^I + c_{12}^{II} m} = \frac{C_0}{c_{12}^I + \sqrt{\frac{(1 - \rho^{II})\bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1 - \rho^I)S_b^2 - (1 - \rho^{II})\bar{S}_w^2/M}}}.$$

Enfin, la variance de l'estimateur de la différence est

$$\begin{aligned}
& V_{e,o}[d] \\
&= \frac{2}{C_0} \left(c_{12}^I + \sqrt{\frac{(1-\rho^{II})\bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}} \right) \\
&\times \left[(1-\rho^I)S_b^2 \right. \\
&\quad \left. + \left(\sqrt{\frac{c_{12}^{II}(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}{(1-\rho^{II})\bar{S}_w^2}} - \frac{1}{M} \right) (1-\rho^{II})\bar{S}_w^2 \right] \\
&\quad - \frac{2(1-\rho^I)S_b^2}{N} \\
&= \frac{2}{C_0} \left\{ (1-\rho^I)S_b^2 c_{12}^I \right. \\
&\quad + (1-\rho^{II})\bar{S}_w^2 \sqrt{\frac{c_{12}^I c_{12}^{II} (1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}{(1-\rho^{II})\bar{S}_w^2}} \\
&\quad + \left[(1-\rho^I)S_b^2 - \frac{1}{M}(1-\rho^{II})\bar{S}_w^2 \right] \\
&\quad \times \sqrt{\frac{(1-\rho^{II})\bar{S}_w^2 c_{12}^I c_{12}^{II}}{(1-\rho^I)S_b^2 - (1-\rho^{II})\bar{S}_w^2 / M}} \\
&\quad \left. + (1-\rho^{II})\bar{S}_w^2 \left(c_{12}^{II} - \frac{c_{12}^I}{M} \right) \right\} \\
&\quad - \frac{2(1-\rho^I)S_b^2}{N}.
\end{aligned}$$

Preuve de la proposition 7. En ignorant les termes de correction pour population finie d'ordres $O(N^{-1})$ et $O(M^{-1})$, l'équation (3.3) peut s'écrire :

$$\begin{aligned}
V_{e,i}[d] &\approx \frac{4 \left(c^I + \sqrt{c^I c^{II} \bar{S}_w^2 / S_b^2} \right)}{C_0} \left[S_b^2 + \left(\sqrt{\frac{c^{II}}{c^I} \bar{S}_w^2 S_b^2} \right) \right] \\
&= \frac{4}{C_0} \left(c^I S_b^2 + c^{II} \bar{S}_w^2 + 2\sqrt{c^I c^{II} \bar{S}_w^2 S_b^2} \right) \\
&= \frac{4}{C_0} \left(\sqrt{c^I S_b^2} + \sqrt{c^{II} \bar{S}_w^2} \right)^2.
\end{aligned}$$

De même, l'équation (3.8) peut s'écrire

$$\begin{aligned}
V_{e,o}[d] &\approx \frac{2}{C_0} \left[(1-\rho^I)S_b^2 c_{12}^I \right. \\
&\quad \left. + 2\sqrt{c_{12}^I c_{12}^{II} (1-\rho^I)S_b^2 (1-\rho^{II})\bar{S}_w^2} + (1-\rho^{II})\bar{S}_w^2 c_{12}^{II} \right] \\
&= \frac{2}{C_0} \left[\sqrt{(1-\rho^I)S_b^2 c_{12}^I} + \sqrt{(1-\rho^{II})\bar{S}_w^2 c_{12}^{II}} \right]^2.
\end{aligned}$$

L'énoncé de la proposition 7 découle directement de ces deux expressions.

Bibliographie

- Binder, D.A., et Hidiroglou, M.A. (1988). Sampling in time. Dans *Handbook of Statistics*, (Éds., P.R. Krishnaiah et C.R. Rao), North Holland, Amsterdam, 6, 187-211.
- Cochran, W.G. (1977). *Sampling Techniques*, 3^e Éd., New York : John Wiley & Sons, Inc.
- Eckler, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26(4), 664-685.
- Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Rapport technique, U.S. Bureau of Labor Statistics.
- Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4(4), 331-345.
- Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York : John Wiley & Sons, Inc.
- Hansen, M., Hurwitz, W.N. et Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York : John Wiley & Sons, Inc.
- Kish, L. (1995). *Survey Sampling*, 3^e Éd., New York : John Wiley & Sons, Inc.
- Lehtonen, R., et Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys, Statistics in Practice*, 2^e Éd., New York : John Wiley & Sons, Inc.
- Mas-Colell, A., Whinston, M.D. et Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.
- McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *The Journal of the American Statistical Association*, 33, 101-116.
- Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Séries B*, 12(2), 241-255.
- Rao, J.N.K., et Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59(306), 492-509.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York : Springer.
- Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.
- Thompson, M.E. (1997). Theory of Sample Surveys. *Monographs on Statistics and Applied Probability*, New York : Chapman & Hall/CRC, 74.
- Thompson, S.K. (1992). *Sampling*, New York : John Wiley & Sons, Inc.
- Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2^e Éd., New York : Springer.