Techniques d’enquête, décembre 1978
Vol. 4, nº 2
Statistique Canada, nº 12-001-X au catalogue

Non-réponse et imputation

R. Platek et G.B. Gray[1]

Résumé

L’article analyse les problèmes posés par les mesures applicables, à diverses étapes de la planification d’une enquête, pour contrer la non-réponse, les répercussions de ces mesures sur l’erreur quadratique moyenne, ainsi que l’utilité pratique, les avantages et les inconvénients de ces mesures. Il examine aussi certaines questions théoriques touchant la complexité et les niveaux d’imputation. Il existe diverses méthodes d’imputation : par pondération, par reproduction et par substitution d’enregistrements. L’article traite aussi de certaines questions méthodologiques concernant le biais et la variance.

Mots-clés : Non-réponse; erreur quadratique moyenne; imputation; biais; variance.

1 Introduction

La fiabilité des estimations d’enquête dépend de nombreux facteurs, notamment l’effet des données manquantes et incohérentes ou incomplètes. Toute enquête, quelle que soit sa nature, doit composer avec un certain degré de non-réponse ou avec des réponses qui sont rejetées lors des procédures de contrôle des données. Il faut se poser la question suivante : « Que devrions-nous faire à propos d’un tel manque d’intégralité des données? » Bien sûr, on peut affirmer que lorsque l’ampleur des données lacunaires est inférieure à 1 %, il n’y a pas lieu de s’en inquiéter. Toutefois, en pratique, l’ampleur de la non-réponse est plutôt de 10 %, 15 % ou plus, selon le sujet.

En ne tenant pas compte de l’effet d’une non-réponse d’une telle ampleur, on risque d’obtenir des résultats d’enquête d’une qualité inacceptable, ce qui se traduira forcément par des totaux de population ne pouvant pas être estimés, puisqu’ils seraient basés sur des données partielles seulement. Par ailleurs, la fiabilité des moyennes et des proportions sera moins touchée que celle des totaux par la non-réponse et on peut également affirmer, non sans justification, que l’effet de la non-réponse sur les estimations nationales sera généralement plus faible qu’à certains niveaux infranationaux. Néanmoins, l’élimination et la réduction de l’effet de la non-réponse et des réponses invalides sont très importantes et elles devraient être entreprises à divers stades de

1. R. Platek, et G.B. Gray, Sous-division du développement d'enquêtes-ménages, Statistique Canada.

la conception d’enquête, ainsi que sur le terrain. Cependant, malgré ces efforts, des non-réponses et des lacunes subsisteront dans les données et, dans pratiquement toutes les enquêtes, une forme ou une autre de rajustement ou d’imputation pour tenir compte de la non-réponse devra être envisagée.

L’imputation peut se définir comme l’affectation de données à des champs vides (y compris la non-réponse totale) ou le remplacement de données invalides conformément à certaines règles. Il n’existe pas de méthode d’imputation impartiale connue, à moins que plusieurs suppositions soient faites au sujet des non-répondants et des répondants. Il semble toutefois que certaines méthodes soient peut-être plus efficaces que d’autres.

2 Traitement de la non-réponse

(i) Planification et développement d’enquêtes

Au stade de la planification, une connaissance de l’effet de la non-réponse sur l’erreur quadratique moyenne des données d’enquête entraînera certainement un plan d’enquête comportant le moins de non-réponse possible. Par conséquent, l’un des facteurs les plus importants de la planification d’une enquête est une décision à l’égard du niveau de tolérance à la non-réponse, et un concepteur d’enquêtes expérimenté peut estimer assez précisément le niveau de réponse pour une enquête en particulier auquel on peut s’attendre dans diverses conditions d’enquête. Certains diront que pour certaines enquêtes, lorsque seules des estimations nationales sont requises et que les caractéristiques des non-répondants ne sont pas très différentes de celles des répondants, un taux de non-réponse (20-30 %) peut être toléré même s’il entraînera une hausse de l’échantillonnage et peut-être de la variance de la réponse. Les mêmes arguments peuvent être appliqués aux enquêtes dont l’objectif est de donner une idée des tendances et des proportions. Toutefois, pour les enquêtes dont les estimations doivent être précises et sont requises à différents niveaux infranationaux, le taux de non-réponse devrait être maintenu aussi bas que 5 % ou moins, et les zones de non-réponse marquée dans les régions locales devraient également être évitées.

Le coût de l’enquête est un autre facteur qui aura une incidence sur bien des composantes de l’élaboration d’enquête, y compris la non-réponse. Il est important d’équilibrer les autres facteurs en fonction

du coût, afin d’obtenir un taux de non-réponse suffisamment bas pour atteindre les objectifs de l’enquête. Il faut également savoir que dans des limites raisonnables, il est parfois préférable d’accepter un échantillon un peu plus petit qu’on l’avait prévu au départ et de transférer les ressources aux procédures appropriées de collecte de données, de suivi et d’estimation. Ce serait particulièrement avantageux lorsque le concepteur de l’enquête soupçonne d’importantes différences entre les répondants et les non-répondants pour ce qui est de leurs caractéristiques.

Hormis l’intuition et l’expérience, qui jouent certainement un rôle important dans la planification et le développement des enquêtes, on peut mentionner un certain nombre de facteurs qui sont importants à la conception des enquêtes. Ces facteurs peuvent être classés dans trois groupes :

Groupe I
- a) taille de l’échantillon
- b) stratification
- c) degré de corrélation intra-grappe
- d) répartition de l’échantillon
- e) méthode de sélection

Groupe II
- a) base de sondage
- b) méthode d’interview
- c) sélection, formation et contrôle du personnel
- d) longueur et formulation du questionnaire
- e) nature délicate des questions
- f) type de région où se déroule l’enquête
- g) faisabilité des rappels et nombre d’entre eux
- h) publicité

Groupe III
- a) contrôle et imputation
- b) estimation
- c) estimation de la variance et autres analyses de données.

Toutes ces activités ont certainement une incidence sur l’erreur quadratique moyenne à divers degrés. Il est vrai que souvent, en pratique, nous n’avons pas assez de données sur l’effet de la plupart des facteurs. Cependant, puisque ces facteurs n’ont pas tous la même

importance, un examen des composantes les plus importantes de l'erreur quadratique moyenne serait très utile. Supposons que l'erreur quadratique moyenne puisse être décomposée comme suit :

$$EQM = V_S + V_R + V_{CR} + (B_S + B_R)^2$$

où

V_S = variance d'échantillonnage
V_R = variance de réponse
V_{CR} = variance de réponse corrélée
B_S = biais d'échantillonnage
B_R = biais de réponse.

La variance d'échantillonnage (V_S) et le biais d'échantillonnage (B_S) sont touchés par tous les facteurs du groupe I, par les procédures d'estimation et aussi par l'ampleur de la non-réponse. Plus la non-réponse est importante, plus l'effet est grand sur la variance d'échantillonnage et le biais. Par exemple, étant donné que la variance d'échantillonnage des estimations est inversement proportionnelle au taux de réponse dans le cas d'un échantillon aléatoire simple, les estimations basées sur ce genre d'échantillon avec un taux de réponse de 80 % auront une variance d'échantillonnage de 12,5 % plus élevée que la variance des estimations correspondantes avec un taux de réponse de 90 %. Dans les échantillons en grappes à plusieurs étapes, la même corrélation se maintient à peu près, mais elle touche principalement la dernière étape de l'échantillonnage. La corrélation entre le biais et l'ampleur de la non-réponse, bien qu'elle soit peut-être plus importante, est moins évidente, puisqu'elle dépend à la fois de l'ampleur de la non-réponse et des caractéristiques des répondants et des non-répondants. Lorsque l'on examine la non-réponse, il faut tenir compte du fait qu'une réduction de la non-réponse sur le terrain ne garantit pas nécessairement une réduction du biais. En fait, si les procédures de réduction de la non-réponse ne sont pas bien réfléchies et exécutées, le biais risque de ne pas être réduit et pourrait même être amplifié.

Dans certaines enquêtes, les conditions d’enquête peuvent avoir une incidence sur la variance d’échantillonnage et le biais d’échantillonnage. Par exemple, la formulation du questionnaire et/ou la formation des intervieweurs peuvent faire en sorte que des valeurs extrêmes légitimes soient éliminées. Une faible variance d’échantillonnage accompagnée d’un fort biais d’échantillonnage peut s’ensuivre. La variance d’échantillonnage artificiellement faible peut se produire parce que la variance entre les unités des réponses prévues sans valeurs extrêmes sera plus faible que la variance entre les valeurs réelles ayant des valeurs extrêmes. Les valeurs extrêmes aux pôles de la valeur moyenne ne seront pas nécessairement équilibrées, ce qui pourrait donner lieu à un fort biais d’échantillonnage. Par conséquent, les conditions d’enquête peuvent avoir une incidence sur la variance d’échantillonnage et le biais d’échantillonnage.

Les composantes hors échantillon de l’erreur quadratique moyenne (V_R, V_{CR}, B_R^2) qui comprennent également la non-réponse sont touchées à divers degrés par tous les facteurs du groupe II. En outre, l’erreur quadratique moyenne est également touchée par certains facteurs du groupe I. Par exemple, le regroupement peut avoir une incidence sur la variance corrélée à peu près de la même façon qu’il touche la variance d’échantillonnage, puisque les ménages dans les grappes peuvent produire de plus fortes corrélations des erreurs de réponse que les ménages plus distancés. Étant donné que l’estimation dépend des valeurs observées, qui sont elles-mêmes assujetties à l’erreur non due à l’échantillonnage, et puisque chaque procédure d’estimation distincte met en cause une fonction différente, la variance non due à l’échantillonnage sera également touchée par la procédure d’estimation.

(ii) Étape de la collecte des données

La non-réponse peut être réduite par les efforts persistants des intervieweurs et par la motivation des non-répondants de devenir des répondants. Les efforts persistants sont habituellement sous forme de tentatives répétées de contacter un répondant et de recueillir des renseignements à son sujet. Au-delà d’un certain seuil, il est impossible d’essayer d’autres rappels, soit parce que l’enquête doit être terminée avant une date donnée, soit parce que les fonds sont insuffisants. Dans le cas des interviews téléphoniques, le coût se résume seulement aux

tentatives d'appels téléphoniques répétées, tandis que dans le cas des enquêtes envoyées par la poste, le coût est celui des rappels subséquents. Toutefois, dans le cas des interviews sur place où, pour des motifs de coûts, l'échantillon doit habituellement être regroupé pour réduire au minimum le temps de déplacement et la distance entre les appels successifs, les rappels répétés entraînent souvent une plus grande distance entre les ménages, et le coût par unité peut devenir déraisonnablement élevé, sans aucune réduction de la variance.

De plus, si la probabilité de non-réponse était la même pour chaque unité, les non-répondants deviendraient un sous-échantillon aléatoire de l'échantillon complet, et il n'y aurait pas de biais de non-réponse dans l'estimation (sauf un biais de l'estimation du ratio) lorsque les données sont pondérées par l'inverse du taux de réponse. Un léger biais de l'estimation du ratio peut s'ensuivre en raison de la variation de l'échantillon de répondants. Étant donné que dans la majorité des cas, la probabilité de non-réponse n'est pas connue, il faut faire tout son possible pour réduire au minimum l'ampleur de la non-réponse. Cependant, même si nous connaissions la probabilité de non-réponse, il pourrait quand même y avoir un biais de réponse dans l'estimation en fonction du sous-échantillon, tout comme ce serait le cas s'il n'y avait pas de non-réponse.

Une autre importante composante de la non-réponse est celle des refus, que l'on peut seulement prévenir, dans bien des cas, en motivant ces non-répondants à répondre. Cependant, il est possible que les répondants qui étaient initialement réticents à répondre puissent entraîner de plus grandes erreurs de réponse que ceux qui étaient disposés à coopérer. Par conséquent, bien que nous ayons réduit l'erreur d'imputation NR_{ε_i}, nous pourrions avoir augmenté l'erreur de réponse R_{ε_i} (Platek, Singh et Tremblay [7]). En se contentant de convertir chaque refus en répondant, on risque donc d'entraîner une fausse impression de sécurité en ce qui concerne la validité des réponses. Un intervieweur bien formé réussira certainement à motiver plus de personnes à répondre et à obtenir des réponses plus fiables qu'un intervieweur n'ayant pas reçu une formation adéquate.

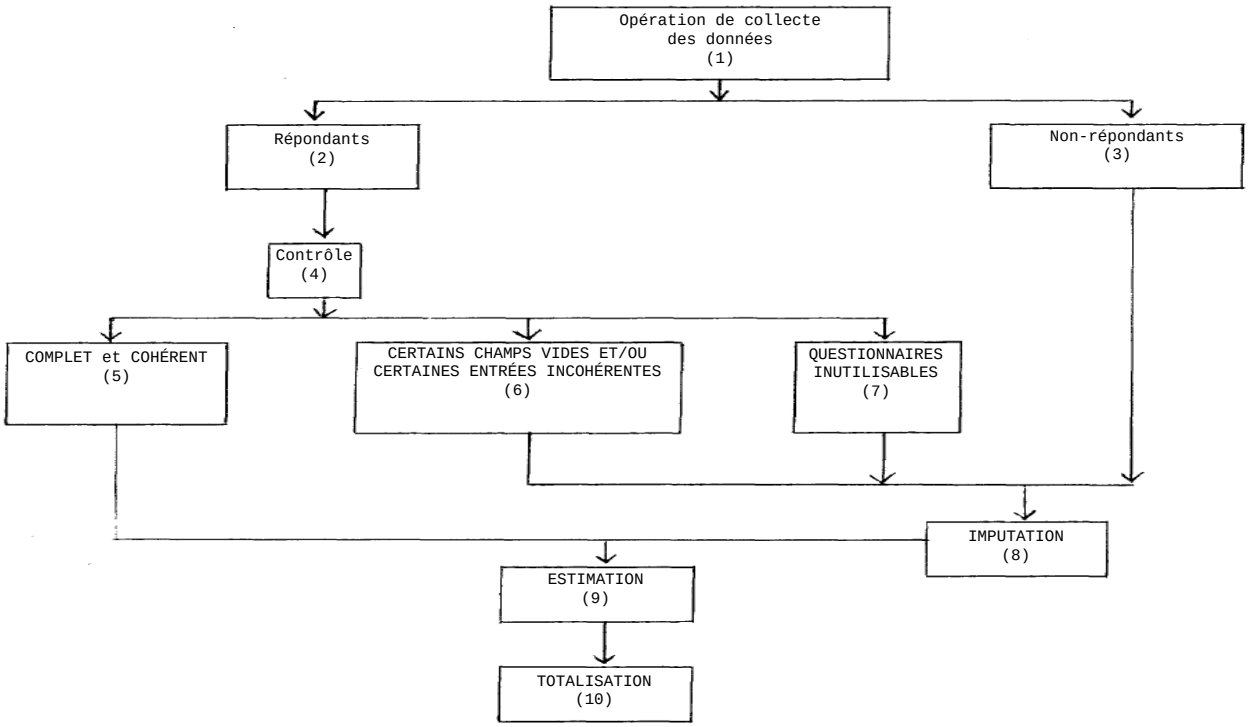
Une méthode pour composer avec la non-réponse à l’étape de la collecte des données consiste à substituer par d’autres unités non sélectionnées au préalable sur le terrain; par exemple, un voisin d’à côté. Malheureusement, cette méthode entraînerait un biais d’échantillonnage. Bien que n’importe quelle unité puisse être sélectionnée avec une probabilité connue en fonction du plan de sondage, la substitution par d’autres répondants non sélectionnés au préalable pour remplacer les répondants récalcitrants d’une quelconque manière non contrôlée, ou même d’une manière contrôlée, modifiera les probabilités d’inclusion à un degré tel qu’elles ne peuvent même pas être calculées. Un biais d’échantillonnage d’une ampleur inconnue existerait (puisque les probabilités de sélection sont inconnues pour plusieurs raisons), mais la variance d’échantillonnage pourrait être réduite à cause d’une augmentation de la taille réelle de l’échantillon. Toutefois, il n’y aurait probablement pas de réduction de l’erreur de réponse ou du biais de non-réponse. Même si les probabilités d’inclusion pouvaient être calculées, le biais de non-réponse subsisterait, puisque les unités non coopératives n’ont essentiellement aucune chance d’être incluses.

En plus des rappels ou de la substitution des unités sur le terrain, les intervieweurs peuvent appliquer (i) le double échantillonnage (sélection d’un sous-échantillon de non-répondants et déploiement d’efforts intensifs pour obtenir des réponses de ces unités), ou (ii) la méthode de Politz (prise en compte du « meilleur moment pour appeler » comme un des groupes de pondération). Ces méthodes sont également coûteuses et doivent faire l’objet d’une planification rigoureuse pour pouvoir être utilisées pour pallier la non-réponse.

3 Types de réponses et enjeux conceptuels de l’imputation

Étant donné que l’information circule de la collecte des données aux totalisations, les différents types de réponses peuvent être déterminés et sont présentés comme suit au graphique I.

Graphique 1 : Organigramme relatif à chaque unité d’échantillonnage



Ce diagramme manifestement simpliste du processus a pour unique objectif d’alimenter le débat pour les besoins du présent document. D’après le graphique 1, deux des trois groupes après l’étape du contrôle doivent faire l’objet d’un suivi quelconque avant l’estimation. Il s’agit des questionnaires inutilisables et des questionnaires qui renferment quelques champs vides et/ou entrées incohérentes. Les questionnaires inutilisables peuvent être classés dans la catégorie de la non-réponse totale, ou ils peuvent être associés aux ménages répondants ayant quelques champs vides ou incohérents. Il reste deux groupes qui doivent être examinés. Le premier groupe se compose des champs vides et/ou des réponses incohérentes, le deuxième groupe se compose des non-répondants. Les non-répondants (du moins dans les enquêtes-ménages, contrairement au recensement) sont habituellement pondérés d’une manière ou d’une autre. Les questionnaires lacunaires, par ailleurs, tombent dans deux catégories, comme les entrées incohérentes ou les champs vides sans raison valable.

Les entrées incohérentes peuvent être logiquement impossibles, ou encore plausibles, mais très peu probables. Il semble naturel que si les entrées sont logiquement impossibles et peuvent être détectées comme telles,

elles doivent être rajustées même si elles n’ont pas nécessairement une forte incidence sur les données. Le rajustement éviterait de mettre dans l’embarras les analystes spécialisés associés aux rapports publiés.

Dans le cas des entrées plausibles mais fort peu probables, on doit prendre une décision difficile : conserver des observations dans une répartition anormale, ou éliminer les valeurs extrêmes de la répartition, qui pourraient en fait représenter une situation réelle. Idéalement, il faut opter pour l’une ou l’autre de ces options en s’appuyant sur l’expérience relative aux mécanismes d’erreur et la nature de la répartition de fond en fonction de la connaissance du domaine. De toute manière, il faut pouvoir cerner les cas problématiques, c.-à-d. qu’il faut avoir des règles de contrôle appropriées chaque fois que l’on découvre des événements impossibles ou fort peu probables, ainsi qu’une méthode pour y remédier.

Il faut faire une distinction fondamentale entre le contrôle et l’imputation. Jetons un coup d’œil à l’ensemble de toutes les combinaisons de codes possible dans un questionnaire. Le contrôle peut se définir comme la division de cet ensemble en deux sous-ensembles mutuellement exclusifs : les combinaisons qui sont réputées acceptables, et celles qui sont inacceptables, ces dernières comprenant les questionnaires comportant des champs vides invalides et des entrées incohérentes. Par conséquent, le contrôle est essentiellement un diagnostic et, du point de vue opérationnel, il doit être défini par un ensemble de règles. L’imputation, quant à elle, s’apparente plutôt au traitement des données, même si les deux concepts se recoupent manifestement.

En ce qui concerne le contrôle, la détection des entrées logiquement impossibles et des champs vides invalides ne présente pas de problèmes conceptuels et, pour ce qui est de la détection des incohérences, il existe plusieurs possibilités. Par exemple, on peut comparer des paires de champs et décider que les deux sont incohérents, ce qui fait que l’un d’eux doit être modifié. On peut continuer cette procédure en comparant d’autres paires de champs (ou trois champs à la fois). Lorsque l’on détecte une incohérence en particulier, on peut soit imputer immédiatement un des champs en cause pour faire concorder ces champs, ou encore suivre le processus de contrôle en entier avant le début de

l’imputation. Cependant, en examinant deux ou trois champs à la fois, on ne tient pas compte de toutes les possibilités. Par exemple, si l’on fait concorder toutes les combinaisons de deux ou trois champs, il ne faut pas croire pour autant que l’enregistrement au complet sera cohérent. Un système élaboré à Statistique Canada est basé sur une approche qui consiste à relever toutes les incohérences avant qu’une mesure corrective soit prise. Ensuite, devant toutes les incohérences connues entre les champs de l’enregistrement donné, ainsi que toutes les impossibilités logiques et les champs vides invalides, on décide quels champs ou quel ensemble de champs, s’ils étaient corrigés, élimineraient toutes les incohérences dans l’enregistrement au complet.

Une fois qu’on a déterminé quels champs seront modifiés, la prochaine étape consiste bien sûr à effectuer l’imputation connexe. La situation la plus simple se produit lorsqu’il y a une seule valeur possible qui peut être imputée pour ce champ, de façon à ce que, après l’imputation, l’enregistrement sera cohérent. Parfois, il peut y avoir plus d’une valeur qui rendrait l’enregistrement cohérent. Si c’est le cas, on choisirait une valeur en particulier qui est plus prédominante dans le champ ou plus plausible. Un bon exemple de ce genre de situation se trouve dans l’Enquête sur la population active, où pendant les mois de l’automne au printemps, pour les personnes de 15 et 16 ans, si aucune caractéristique de la population active n’est entrée, on impute que ces personnes sont « aux études », même s’il n’est pas impossible du tout que ce ne soit pas le cas. Dans la mesure où la proportion de tels cas est suffisamment faible, l’effet sera une légère augmentation du biais. Par ailleurs, il y aura une certaine réduction de la variance, en plus de l’avantage de réduire la complexité de l’imputation.

Dans d’autres situations, lorsqu’on pourrait raisonnablement imputer toute une gamme de valeurs, d’autres critères s’imposent. Un critère possible serait de réduire au minimum l’erreur quadratique moyenne des estimations obtenues. Il faut toutefois se demander à quelles estimations se rapporte l’erreur quadratique moyenne? Compte tenu de la demande sans cesse croissante en microdonnées totalisées de plusieurs façons différentes et imprévues, on ne sait pas vraiment quelle erreur quadratique moyenne il faudrait réduire au minimum. De plus, on ne connaîtrait pas tous les types d’agrégats auxquels un enregistrement en particulier pourrait continuer dans différents types de totalisations.

Par conséquent, on aimerait utiliser d’autres critères pour produire l’entrée la plus appropriée pour un champ dans un enregistrement en particulier par rapport à l’autre information dans l’enregistrement. Autrement dit, comment peut-on prédire au mieux la valeur d’un champ en s’appuyant sur sa connaissance des autres champs dans l’enregistrement. Un bon exemple de ce genre d’imputation est l’utilisation des données du mois précédent dans l’Enquête sur la population active : pour une personne en particulier, il serait difficile de trouver une valeur mieux imputée, surtout dans les cas où les caractéristiques démographiques changent lentement. En l’absence d’information basée sur le passé, la meilleure valeur imputée peut découler d’une équation de régression quelconque. Toutefois, pour certaines enquêtes-ménages, l’application de la régression est quelque peu limitée en raison de la nature qualitative des variables. Par conséquent, on peut adopter comme critère raisonnable que la répartition après l’imputation devrait demeurer le plus près possible de la répartition avant l’imputation en ce qui concerne les répartitions marginales ou, de préférence et si c’est possible, en ce qui concerne les répartitions conjointes de toutes les variables à imputer.

Dans la plupart des cas, l’imputation pour la non-réponse au micro-niveau, plutôt qu’à un niveau agrégé quelconque, est en grande partie justifiée en raison de l’absence d’une bonne connaissance du type d’agrégats qui seront produits à partir du fichier de microdonnées. Cependant, dans certaines situations où l’on sait que l’on peut limiter les exigences de totalisation à l’avance, l’imputation au niveau de la personne n’est peut-être pas toujours nécessaire. C’est notamment le cas pour les enquêtes basées sur des échantillons aréolaires où les principales unités d’échantillonnage ne seront probablement pas fractionnées dans des totalisations subséquentes. Dans ce cas-ci, on ne peut guère faire mieux en ce qui concerne l’erreur quadratique moyenne des agrégats possibles que l’on produira, que d’imputer la moyenne de cette unité primaire d’échantillonnage (UPE)[1].

1. Bien que la correction de la pondération au niveau de l’UPE soit justifiée pour la non-réponse complète, elle serait inopportune pour la non-réponse partielle ou les champs dont les entrées ont été rejetées pour des motifs de révision. Si l’on effectuait la pondération au niveau des champs individuels, on ne pourrait pas recouper les données correctement, puisque les enregistrements auraient plus d’un poids.

4 Traitement et estimation

L'une des procédures les plus courantes pour tenir compte de la non-réponse à l'étape du traitement et de l'estimation est celle de la région d'équilibrage en fonction du plan de sondage, où les poids sont gonflés par l'inverse du taux de réponse. Dans une région d'équilibrage b, une estimation du total des caractéristiques est donnée par $\hat{X}_b = \sum_{i=1}^{n_b} x_i/\pi_i$, où x_i est la réponse, π_i = est la probabilité d'inclusion et n_b est la taille de l'échantillon dans les régions d'équilibrage. Si seulement m_b unités répondent, le poids π_i^{-1} est gonflé par l'inverse du taux de réponse, m_b/n_b, c.-à-d. par le facteur n_b/m_b.

Les régions d'équilibrage devraient préférablement être déterminées à l'étape de la planification, plutôt qu'à l'étape du traitement, et il pourrait s'agir de strates individuelles, de groupes de strates, d'une province, d'une unité primaire d'échantillonnage ou d'une grappe. Le choix d'une région d'équilibrage est très important, puisque les taux de non-réponse et le biais peuvent varier d'une région à une autre.

Par exemple, un important problème méthodologique consiste à déterminer une taille optimale ou en quelque sorte appropriée de la région d'équilibrage, où « appropriée » s'entend d'une taille adéquate pour assurer un taux de réponse suffisant afin de prévenir des poids excessifs, tout en garantissant les avantages découlant des mesures de l'homogénéité pour aider à réduire le biais de non-réponse. Il est facile de démontrer que l'inflation des poids de tous les enregistrements dans une région d'équilibrage pour neutraliser les non-répondants équivaut à la substitution par des valeurs moyennes de tous les répondants pondérés de chaque non-répondant dans la région. Si une caractéristique a un fort degré d'homogénéité (inversement proportionnel à la taille de la région), la pondération (ou la substitution par la valeur moyenne) dans les petites régions par rapport aux grandes régions aurait tendance à entraîner des valeurs moyennes qui ressemblent davantage à la valeur réelle des caractéristiques du non-répondant que ce n'aurait été le cas dans les grandes régions. Par conséquent, le biais de non-réponse aurait tendance à être plus faible dans le cas des petites régions d'équilibrage que dans le cas des grandes unités d'équilibrage. Qu'en est-il de la

variance? Plus les régions d’équilibrage sont petites, plus l’inflation des poids devient instable, la variation des taux de réponse devenant plus instable au sein de nombreuses petites régions d’équilibrage, au lieu d’un petit nombre de grandes régions d’équilibrage, et l’instabilité du poids aurait tendance à accroître la variance. De toute évidence, on fait un certain compromis pour ce qui est de la taille de la région d’équilibrage entre les petites régions pour tirer profit de la mesure de l’homogénéité et les petites régions pour assurer la stabilité des corrections de la pondération. Les valeurs extrêmes possibles des tailles des régions d’équilibrage sont l’échantillon complet au seuil supérieur et une taille de '1' au seuil inférieur. Cependant, dans ce dernier cas, on doit déterminer ce qui devrait être fait si cette unité ne répond pas. Au lieu de la substitution de la valeur moyenne, il faudrait recourir à l’analyse de la régression ou aux modèles de superpopulation (une approche entièrement différente de la substitution) ou encore s’en remettre aux valeurs historiques. Le choix de la taille de la région d’équilibrage dépend non seulement de la mesure de l’homogénéité, mais aussi du plan de sondage, de la taille de l’échantillon et du taux de réponse. Les enquêtes ayant de faibles taux de réponse nécessiteraient de plus grandes unités d’équilibrage que celles ayant des taux de réponse élevés. On pourrait utiliser de petites régions d’équilibrage et adopter une procédure quelconque pour les réduire jusqu’à ce que le taux de réponse atteigne un certain niveau acceptable (pas trop loin au-dessous du taux de réponse global), afin de réduire au minimum l’instabilité du poids. Toutefois, la réduction des régions d’équilibrage ajoute une dimension complexe à l’estimation de la variance, puisqu’il faudrait tenir compte des probabilités de réduire 1, 2, 3, 4, etc. régions d’équilibrage et le choix de 1, 2, 3 ou 4 régions d’équilibrage. Bien qu’une telle procédure soit entreprise dans le cadre de l’EPA, la nécessité de réduire est suffisamment rare pour ne pas justifier un traitement particulier pour les besoins de l’estimation de la variance. Par conséquent, si des calculs de la variance ou des analyses autres que les moyennes ou les totaux seulement sont envisagés, les régions d’équilibrage qui devraient rester stables sans trop de réduction devraient être intégrées au plan de sondage. Autrement dit, les taux de réponse devraient être suffisamment élevés et à forte probabilité pour éviter la nécessité de réduire les régions d’équilibrage si l’estimation

de la variance est envisagée. Cette démarche nous découragerait d'utiliser les petites régions pour neutraliser la non-réponse.

Au lieu de pondérer en fonction de l'inverse du taux de réponse dans une région d'équilibrage, on pourrait reproduire un nombre suffisant d'unités parmi les m_b répondants pour élever la taille apparente de l'échantillon au niveau initial de n_b unités. Cependant, on peut démontrer qu'une composante de variance supplémentaire se produit par rapport à celle survenue lorsqu'une pondération simple est appliquée et dans le cas de l'échantillonnage aléatoire simple, la variance d'échantillonnage est considérée séparément, et l'augmentation s'élèverait à environ 12 %, selon le taux de réponse (voir Hansen et coll. [3]). Le principal avantage de la reproduction par rapport à la pondération réside dans l'assurance que des poids intégraux plutôt que fractionnaires sont appliqués à chaque enregistrement. Dans certains types de données publiées, comme le nombre de personnes ayant une caractéristique donnée, les poids intégraux auraient tendance à éviter l'arrondissement des erreurs pour sous-classifier les données. Lorsque l'on estime les moyennes ou les proportions ou certains types de totaux quantitatifs comme le produit intérieur brut, l'utilisation de nombres entiers au lieu de poids fractionnaires n'offre aucun avantage. Hormis les commentaires dans le paragraphe qui précède, les problèmes méthodologiques relatifs à la pondération dans les régions d'équilibrage s'appliquent également à la reproduction dans les régions d'équilibrage.

Une autre méthode importante d'imputation de la non-réponse a trait à la substitution par des données historiques (les données du mois précédent) ou des données d'une source externe (fichiers administratifs, autres enquêtes, données du recensement). Une fois que la substitution par des données historiques ou d'une source externe a été effectuée dans la mesure du possible pour les non-répondants, la pondération ou la reproduction peut être touchée dans les régions d'équilibrage. Dans le cas de la pondération, on gonflerait le poids π_i^{-1} par le facteur $n_b/(m_b+m_b')$, où m_b répondants ont été obtenus comme auparavant et pour m_b' des (n_b-m_b) non-répondants, des enregistrements historiques ont substitué les données manquantes. Dans une telle méthode d'imputation,

la variance d'échantillonnage est réduite par rapport à celle qui se produit dans le plan de pondération, puisque nous avons augmenté la taille réelle de l'échantillon de m_b à quelque part entre m_b et (m_b+m_b') unités. L'échantillon accru, y compris les enregistrements imputés à partir d'enregistrements historiques, ne sera pas aussi bon que m_b+m_b', puisque les données historiques ou d'une source externe ne sont pas aussi bonnes que les renseignements à jour sur la réponse, à moins qu'il n'y ait pas eu de changement au niveau des caractéristiques des unités pour lesquelles la substitution par des données historiques a été entreprise.

Par ailleurs, on peut reproduire les répondants, c.-à-d. prendre un échantillon de répondants d'une taille équivalente au nombre de non-répondants et appliquer un poids de 2 au lieu de gonfler le poids pour tous les répondants. Cependant, on pourrait préférer éviter la reproduction des non-répondants pour lesquels une substitution par des renseignements historiques avait été effectuée, en optant plutôt pour sous-échantillonner $n_b - (m_b+m_b')$, disons m_b^* unités des m_b répondants à reproduire, afin de porter la taille apparente de l'échantillon de (m_b+m_b') à n_b unités dans la région d'équilibrage b. Le total estimatif pour la région d'équilibrage b serait

$$\hat{X}_b = \sum_{i=1}^{m_b} w_i\, x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} x_j'/\Pi_j, \qquad (4.1)$$

où x_j' est la valeur imputée pour l'unité j et $w_i = 1$ ou 2 (2 pour un sous-échantillon aléatoire de $n_b-m_n-m_b'$ unités parmi les m_b répondants). La valeur prévue de $\hat{X}_b$ par rapport à toutes les façons possibles de reproduire est

$$\hat{X}_b^* = (n_b-m_b')/m_b \sum_{i=1}^{m_b} x_i/\Pi_i + \sum_{j=m_{b+1}}^{m_b+m_b'} x_j'/\Pi_j. \qquad (4.2)$$

Par conséquent, $V(\hat{x}_b) = V(\hat{x}_b^*)$ + (composante supplémentaire de la variance en conséquence du sous-échantillonnage parmi les répondants). $\hat{x}_b^*$ n'est pas la même chose que l'estimation $n_b/(m_b+m_b')[\sum_{i=1}^{m_b} x_i/\pi_i + \sum_{j=m_{b+1}}^{m_b+m_b'} x_j/\pi_j]$ et la variance de $\hat{x}_b^*$ est également différente de celle de l'estimation où l'inflation des poids de $n_b/(m_b+m_b')$ est appliquée (voir l'annexe).

Les procédures d'estimation susmentionnées comprennent la pondération ou la reproduction des régions d'équilibrage dépendantes du plan de sondage. Si des données historiques ou d'une source externe sont disponibles pour certains des non-répondants, elles pourraient être utilisées à des fins d'imputation avant la pondération ou la reproduction dans les régions d'équilibrage. Au lieu des régions d'équilibrage, on pourrait utiliser des classes de pondération à des fins d'imputation, dont il est question dans le paragraphe qui suit.

Les classes de pondération se distinguent des régions d'équilibrage en ce qu'elles comprennent généralement des caractéristiques d'unités finales (p. ex., types de logements, groupes de revenus particuliers, etc.) plutôt que des régions géographiques, mais on pourrait probablement regrouper des régions en fonction de certaines caractéristiques distinctives qui ne sont pas liées au plan de sondage. Habituellement, on définit les classes de pondération ainsi que les régions d'équilibrage avant la procédure de collecte de l'enquête et on apporte des corrections par le biais d'une réduction si les taux de réponse sont inacceptablement faibles ou si l'échantillon est trop petit pour employer un type quelconque de correction des poids. Cependant, dans certaines procédures d'imputation, les classes de pondération sont définies après que les données d'enquête ont été recueillies, et une analyse des facteurs ou d'autres outils analytiques sont employés pour déterminer l'ensemble le plus efficace de classes de pondération. Une fois les classes de pondération déterminées, les procédures d'estimation sont essentiellement identiques à celles utilisées dans les régions d'équilibrage. Les

biais et les variances (du moins en ce qui concerne les probabilités d’inclusion individuelles et conjointes des unités finales et d’autres paramètres non liés au plan de sondage) sont identiques. Toutefois, en élargissant la variance pour tenir compte du plan de sondage en particulier, les variances des estimations relatives aux régions d’équilibrage et aux classes de pondération seront très différentes. Afin d’utiliser les classes de pondération à des fins d’imputation, il faut avoir une certaine connaissance des non-répondants (comme la tranche de revenu, la taille du ménage et le type de logement). Dans les faits, lorsque ces renseignements ne sont pas disponibles, la procédure ne peut pas être utilisée.

La formule d’estimation pour les méthodes d’imputation dont il est question ici peut être écrite comme suit.

$\hat{X} = \sum_b \hat{X}_b$ estime le total d’une caractéristique donnée,

où b représente soit la région d’équilibrage, soit la classe de pondération. L’estimation pour une région d’équilibrage ou une classe de pondération en particulier est obtenue par la formule suivante :

$$\hat{X}_b = \sum_{i=1}^{m_b} w_i x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} w_j x_j'/\Pi_j, \quad \text{où } \Pi_i \text{ ou } \Pi_j \text{ est} \qquad (4.3)$$

la probabilité d’inclusion et m_b est le nombre d’unités qui ont répondu des n_b unités dans la région d’équilibrage b.

x_i = valeur de réponse pour l’unité $i = X_i$ (valeur réelle) $+ {}^{R}\varepsilon_i$ (erreur de réponse) (4.4)

x_j' = valeur historique pour l’unité j (si disponible), étant donné que l’unité j n’a pas répondu. Parmi les $(n_b - m_b)$ unités non-répondantes

m_b' possèdent des enregistrements historiques dans la région d'équilibrage b.

$x_j' = X_j$ (valeur réelle) $+ {}^R\varepsilon_j'$ (erreur de réponse de la valeur historique, par rapport à X_j) (4.5)

w_i et w_j sont des poids, qui conviennent à la méthode d'imputation, et les poids sont indiqués dans le tableau 1.

Tableau 1
Méthode d'imputation dans la région d'équilibrage/classe de pondération

		w_i	w_j
(a)/(c)	Pondération par l'inverse du taux de réponse m_b/n_b	n_b/m_b	0
(b)/(d)	Reproduction d'un sous-échantillon aléatoire de (n_b-m_b) unités de m_b répondants	2 pour (n_b-m_b) unités et 1 pour $(2m_b-n_b)$ unités	0 0
(e1)	Substitution par des enregistrements historiques pour m_b' des (n_b-m_b) non-répondants, suivie d'une pondération	$n_b/(m_n+m_b')$	$n_b/(m_b+m_b')$
(e2)	Substitution comme à (iii), suivi d'une reproduction des répondants seulement	2 pour $(n_b-m_b-m_b')$ unités et 1 pour $(2m_b+m_b'-n_b)$ unités	1

(c) et (d) désignent les classes de pondération, tandis que (a) et (b) représentent les régions d'équilibrage.

Dans le cas de la reproduction, nous avons présumé que le taux de réponse m_b/n_b est d'au moins 0,5. S'il est d'exactement 0,5, la reproduction et la pondération donneraient des estimations identiques. Si l'on suppose que $n_b/m_b = W_b+d_b$, où W_b est un nombre entier, et d_b, une fraction aux alentours de $0 \leq d_b < 1$, alors m_b serait fractionné en m_{b1} unités, sous-échantillonnées de façon aléatoire et nécessitant un poids de W_b

et $m_{b2} = (m_b - m_{b1})$ unités, nécessitant un poids de (W_b+1). Ainsi, $n_b = W_b m_b + d_b m_b = W_b m_{b1} + (W_b+1)\ m_{b2} = W_b m_b + m_{b2}$. Par conséquent, $m_{b2} = d_b m_b$ et $m_{b1} = (1-d_b)\ m_b$. Par conséquent, un sous-échantillon aléatoire de $d_b m_b$ répondants serait associé à un poids de (W_b+1) et les $(1-d_b)\ m_b$ répondants restants auraient un poids de W_b. Si $W_b = 1$, alors $n_b = m_b + d_b m_b$ ou $d_b m_b = (n_b - m_b)$ unités nécessiteraient un poids de 2, comme l'indique le tableau 1. Quelle que soit la valeur de W_b, la valeur prévue des estimations par la méthode (b) ou (d) dans tous les sous-échantillons possibles de $d_b m_b$ répondants qui pourraient se voir attribuer un poids de (W_b+1) au lieu de W_b correspond seulement à l'estimation par l'inflation du poids comme méthode (a) ou (c).

Dans le cas de la méthode (e2) (utilisation de données historiques ou d'une source externe, suivie de la reproduction des répondants), on confinerait fort probablement la reproduction à un sous-échantillon de m_b répondants au lieu des (m_b+m_b') unités qui ont soit répondu, soit utilisé des enregistrements historiques. En pareil cas, la valeur prévue conditionnelle dans tous les sous-échantillons aléatoires possibles d'unités attribuées aux fins de la reproduction, compte tenu de l'échantillon, n'est pas l'estimation par la méthode (e1), mais plutôt une estimation où $w_i = (n_b - m_b')/m_b$ pour les m_b répondants et $w_j = 1$ pour les m_b' non-répondants avec les enregistrements historiques disponibles.

5 Biais des estimations

Le biais de $\hat{X}_b$ conformément à la procédure d'imputation peut être obtenu facilement en trouvant tout simplement $E\ \hat{X}_b$. Étant donné que $\hat{X}_b$ est une estimation du ratio avec l'échantillon répondant m_b, une variable, et de même pour (m_b+m_b'), il existe un biais d'estimation du ratio en plus des biais de réponse et de non-réponse, mais nous n'en avons pas parlé dans le tableau 2, où les biais sont indiqués pour les estimations qui sont définies au tableau 1. Dans le tableau, α_i est la probabilité que l'unité i réponde, tandis que $\bar{\alpha}_b$ est le taux de réponse prévu dans la région d'équilibrage b et peut s'écrire comme suit : $\bar{\alpha}_b = E_b^* \alpha_i$. R_{B_i} indique le biais de réponse se rapportant à l'unité i, tandis que R_{B_i}'

indique le biais de la valeur historique, par rapport à la valeur réelle X_i. α_i' est la probabilité que l'unité i possède des données historiques et, enfin, Cov $\delta_i\delta_i'$ est la covariance entre la réponse ou la non-réponse ($\delta_i = 1$ ou 0) et l'existence ou l'absence de données historiques ($\delta_i' = 1$ ou 0).

On notera dans le tableau 2 que le biais est identique pour la pondération et la reproduction puisque, comme il a été mentionné précédemment, la valeur prévue de l'estimation au moyen de la reproduction à des fins d'imputation dans tous les sous-échantillons possibles d'unités à reproduire correspond simplement à l'estimation au moyen de l'inflation de poids. La valeur prévue globale des deux estimations est donc la même.

Le biais selon la méthode (e1) est facile à comparer avec le biais selon les méthodes (a) à (d). Le biais de non-réponse selon la méthode (e1) est donné par la formule $(\bar{\alpha}_b + \bar{\alpha}_b'')^{-1}\ En_b\ Cov_b^* (\alpha_i + \alpha_i'', X_i/\Pi_i)$, qui sera réduite au biais de non-réponse selon les méthodes (a) à (d) lorsque $\alpha_i'' = 0$ et $\bar{\alpha}_b'' = 0$. À mesure que les probabilités combinées $\alpha_i + \alpha_i''$ se rapprochent de un, la covariance de la population entre $\alpha_i + \alpha_i''$ et X_i/Π_i ou $Cov_b^* (\alpha_i + \alpha_i, X_i/\Pi_i)$ se rapproche de zéro. En fait, si $\alpha_i + \alpha_i''$ étaient égaux pour la totalité de i, la covariance serait de zéro et il n'y aurait pas de biais de non-réponse. Il en va de même pour les méthodes (a) à (d) si les α_i' étaient tous égaux. Le biais de non-réponse selon la méthode (e1) devrait être inférieur à celui des méthodes (a) à (d) en raison de la baisse prévue de la covariance de la population. Selon l'ampleur de la disponibilité des enregistrements historiques, $\alpha_i + \alpha_i''$ dépasseraient α_i et auraient très probablement une plus petite variance de population. Si $Cov_b^* (\alpha_i, X_i/\Pi_i) = r_{\alpha_i, X_i/\Pi_i} \sqrt{V_b^* (\alpha_i)\ V_b^* (X_i/\Pi_i)}$ et si $Cov_b^* (\alpha_i+\alpha_i', X_i/\Pi_i) = r_{\alpha_i+\alpha_i', X_i/\Pi_i} \sqrt{V_b^* (\alpha_i+\alpha_i')} \cdot \sqrt{V_b^* (X_i/\Pi_i)}$, alors $Cov_b^* (\alpha_i+\alpha_i', X_i/\Pi_i)$ serait certainement plus petite que $Cov_b^* (\alpha_i, X_i/\Pi_i)$ parce qu'il faudrait s'attendre à ce que $(\alpha_i + \alpha_i')$ soient plus proches de un et probablement moins variables parmi les unités que les α_i seulement, ce qui suppose que $V_b^* (\alpha_i + \alpha_i') < V_b^* (\alpha_i)$. Une diminution supplémentaire du biais de non-réponse se produirait selon les méthodes

(e1) par rapport aux méthodes (a) à (d) en raison du plus grand dénominateur $(\bar{\alpha}_b + \bar{\alpha}_b'')$ concernant la méthode (e1) comparativement à $\bar{\alpha}_b$ dans le dénominateur du biais pour les méthodes (a) à (d).

Un plus faible biais de non-réponse peut être neutralisé en partie par un plus grand biais de réponse en ce qui concerne la méthode (e1). Si les R_{B_i}' avaient environ la même ampleur que les R_{B_i} en moyenne, alors les biais de réponse seraient à peu près les mêmes, mais l'on s'attendrait à ce que les R_{B_i}' soient légèrement supérieurs aux R_{B_i}, puisque les données historiques ne seraient pas aussi proches des valeurs réelles que les réponses actuelles.

Tableau 2
Biais de l'estimation, conformément à la procédure d'imputation[1]

Méthode	Biais de l'estimation	
Pondération (a)/(c) et reproduction (b)/(d)	$\bar{\alpha}_b^{-1} En_b \, Cov_b^* (\alpha_i, X_i/\Pi_i)$	… biais de non-réponse
	$+ \bar{\alpha}_b^{-1} \sum_{i=1}^{N_b} \alpha_i R_{B_i}$	… biais de réponse
(e1) Substitution par des enregistrements historiques, puis pondération	$(\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} En_b \, Cov_b^*(\alpha_i, X_i/\Pi_i)$	… biais de non-réponse encouragé par l'utilisation de l'inflation des poids des répondants
	$+ (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} En_b \, Cov_b^*(\alpha_i'', X_i/\Pi_i)$	… biais de non-réponse, encouragé par la substitution par des enregistrements historiques
	$+ (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} \sum_{i=1}^{N_b} (\alpha_i R_{B_i} + \alpha_i'' R_{B_i}')$	… biais de réponse, encouragé respectivement par les répondants et les non-répondants ayant des données historiques
(e2) Substitution par des enregistrements historiques, puis reproduction ou pondération des répondants seulement	$\bar{\alpha}_b^{-1} En_b(1-\bar{\alpha}_b'') Cov_b^*(\alpha_i, X_i/\Pi_i)$	… biais de non-réponse encouragé par la reproduction
	$+ \bar{\alpha}_b^{-1} En_b(1-\bar{\alpha}_b'') \sum_{i=1}^{N_b} \alpha_i R_{B_i}$	… biais de réponse encouragé par les répondants
	$+ En_b \, Cov_b^*(\alpha_i'', X_i/\Pi_i)$	… biais de non-réponse, encouragé par la substitution par des enregistrements historiques
	$+ \sum_{i=1}^{N_b} \alpha_i'' R_{B_i}'$	… biais de réponse, encouragé par l'imputation par des enregistrements historiques

Pour (e1) et (e2), $\alpha_i'' = (1-\alpha_i)\alpha_i' - Cov \, \delta_i \, \delta_i'$ et $\bar{\alpha}_b'' = E_b^* \alpha_i'' = (1-\bar{\alpha}_b) \, \bar{\alpha}_b' - Cov_b^* \alpha_i \alpha_i' - E_b^* Cov \delta_i \delta_i'$

[1] Biais dérivé pour la méthode (e1) à l'annexe 1.

6 Variance des estimations

La variance de $\hat{X}_b$ telle que définie pour la méthode (e1) est partiellement dérivée à l'annexe 2 en considérant $\hat{X}_b$ comme un produit combiné et une expression d'un ratio et en utilisant le développement en séries de Taylor. Il en va de même pour $\mathrm{Cov}(\hat{X}_b, \hat{X}_c)$,

$$V(\hat{X}_b) \doteq \{X_b + (\bar{\alpha}_b + \bar{\alpha}_b'')^{-1}[En_b\ \mathrm{Cov}_b^*(\alpha_i + \alpha_i'', X_i/\Pi_i) + \sum_{i=1}^{N_b}(\alpha_i\ {}^R B_i + \alpha_i''\ {}^R B_i')]\}^2$$

$$\times \{V_s[\frac{n_b}{En_b} + \frac{\sum_{i=1}^{n_b}\Pi_i^{-1}[\alpha_i(X_i+{}^R B_i)+\alpha_i''(X_i+{}^R B_i')]}{\sum_{i=1}^{N_b}[\alpha_i(X_i+{}^R B_i)+\alpha_i''(X_i+{}^R B_i')]}$$

$$- \frac{\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')}{En_b(\bar{\alpha}_b+\bar{\alpha}_b'')}] \quad \text{… variance d'échantillonnage (s désignant un échantillon en particulier)}$$

$$+ E_s\ V[\frac{\sum_{i=1}^{n_b}\Pi_i^{-1}[\delta_i(X_i+{}^R\varepsilon_i)+\delta_i''(X_i+{}^R\varepsilon_i')}{\sum_{i=1}^{N_b}[\alpha_i(X_i+{}^R B_i)+\alpha_i''(X_i+{}^R B_i')]}$$

$$- \frac{\sum_{i=1}^{n_b}(\delta_i+\delta_i'')}{En_b(\bar{\alpha}_b+\bar{\alpha}_b'')}]\,|_s \quad \text{… variance non due à l'échantillonnage}\} \qquad (6.1)$$

où des formes développées de variances non dues à l'échantillonnage et de covariances peuvent être obtenues à l'annexe 2. De même, la formule $\mathrm{Cov}(\hat{X}_b, \hat{X}_c)$ peut être exprimée. Dans cette formule, $\delta_i'' = (1-\delta_i)\delta_i'$ et $\alpha_i'' = E\delta_i''$.

Dans le cas des méthodes (a) et (c), la formule 6.1 s'applique également à la totalité des α_i'', δ_i'' et $\bar{\alpha}_b''$ équivalant à zéro.

Pour les méthodes (b) et (d), la reproduction d'un sous-échantillon d'unités d'une manière aléatoire pour gonfler l'échantillon de m_b à n_b unités dans la région d'équilibrage b, la formule 6.1 donne une composante de variance. Il y a une composante supplémentaire, découlant de la variation du choix d'unités sous-échantillonnées à reproduire.

La composante de variance supplémentaire est donnée par :

$$E_s\, E\Big[\sum_{i=1}^{n_b} \Big(w_i - \frac{n_b}{m_b}\Big)\, \delta_i (X_i + {}^{R}\varepsilon_i)\, \pi_i^{-1}\Big]^2 \Big| s, \qquad (6.2)$$

où s est un échantillon donné de n_b unités et le deuxième E est retenu au lieu de toutes les réponses et non-réponses possibles dans un échantillon en particulier. Pour un taux de réponse donné m_b/n_b, $E\, w_i = n_b/m_b$ pour tous les répondants dans la région d'équilibrage b et on présume que le taux de réponse est d'au moins 0,5, ce qui fait que $w_i = 1$ ou 2. Dans le cas de srswor, en supposant que m_b et n_b sont constants, Hansen et coll. [3] ont démontré que la composante de variance supplémentaire causée par la reproduction au lieu de la pondération peut aller jusqu'à 12 % pour un taux de réponse d'environ 3/4. Des résultats semblables sont obtenus avec ppswor. Cependant, lorsque m_b et n_b varient, des études plus poussées sur l'expansion de 6.2 doivent être réalisées.

Il est difficile de comparer la variance de $\hat{X}_b$ au moyen de la méthode (e2) avec celle de la méthode (a) ou (c) à partir de la formule 6.1 sans substitution par des valeurs numériques. Intuitivement, on s'attendrait à ce que la variance au moyen de la méthode (e1) soit inférieure à celle de la méthode (a), l'ampleur de la diminution dépendant de la taille de la non-réponse utilisant des enregistrements historiques et de la corrélation entre les renseignements historiques et actuels. Les variances doivent être examinées, peut-être au moment de la reformulation de 6.1 en ce qui concerne les valeurs paramétriques moyennes de α_i, ${}^{R}\sigma_i^2$, α_i'', etc. dans la région d'équilibrage.

7 Conclusion

Les enjeux conceptuels englobent la difficulté de la non-réponse et les avantages et inconvénients de différentes méthodes pour y remédier. Des données empiriques seront nécessaires pour obtenir les paramètres dans les formules énoncées dans le présent document à des fins de comparaison. Un fait important à souligner concerne la composante de variance supplémentaire qui se produit au moment de la reproduction, contrairement à la pondération, où un taux de réponse donné est obtenu pour un échantillon donné. L'effet de la reproduction doit être examiné de plus

près, puisque la taille de l'échantillon et les taux de réponse varient tous les deux.

Une grande partie du développement méthodologique du biais et de la variance des estimations en fonction de différentes procédures d'imputation dépend de la connaissance de probabilités de réponse qui sont rarement connues dans la vraie vie. Certaines estimations des probabilités de réponse peuvent être obtenues à partir d'études longitudinales de profils de réponse dans le cas d'enquêtes continues; autrement, des études expérimentales spéciales auprès de non-répondants à l'extérieur de l'échantillon utilisé à des fins de publication peuvent être nécessaires pour obtenir des estimations approximatives des probabilités de réponse.

Il est très important de souligner que, conformément aux procédures habituelles d'imputation de la reproduction ou de la pondération, il y a un biais de non-réponse uniquement si les probabilités de réponse varient parmi les unités et s'il y a une corrélation entre les probabilités de réponse et les valeurs caractéristiques des unités. Cependant, le biais de réponse se produira que nous ayons ou pas une réponse complète.

8 Remerciements

Les auteurs sont sincèrement reconnaissants des commentaires et des suggestions de l'examinateur, du réviseur et d'A. Ashraf, méthodologiste principal de la Sous-division du développement d'enquêtes-ménages.

Bibliographie

[l] Fellegi, I.P. et Holt, D., "A Systematic Approach to Automatic Edit and Imputation", *Journal of the American Statistical Association* (1976), Vol. 71, pp. 17-35.

[2] Ghangurde, P.D. et Mulvihill, J., "Non-Response and Imputation in Longitudinal Estimation in LFS", rapport de Statistique Canada, Sous-division du développement d'enquêtes-ménages (Février 1978).

[3] Hansen, M.H., Hurwitz, W.N. et Madow, W.G., "Sample Survey Methods and Theory", Vol. 11, Theory, pp. 139-141, John Wiley and Sons, Inc. (1953).

[4] Nordbotten, S., "The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for lmproving the Quality of Statistics", Bulletin de l'International Statistical Institute, recueil de la 35e Session, Belgrade 41, (Septembre 1965), pp. 417-441.

[5] Platek R., "Imputation for Household Surveys in Statistics Canada", rapport préparé par l'European Statisticians' Conférence tenue à Genève, Mars 1978.

[6] Platek, R., "Some Factors affecting Non-Response", Techniques d'enquête, Statistique Canada, Vol. 3, No. 2 (Decembre 1977), pp. 191-214.

[7] Platek, R., Singh, M.P. et Tremblay, V., "Adjustment for Non-Response in Surveys", Techniques d'enquête, Statistique Canada, Vol. 3, No. 1 (Juin 1977), pp. 1-24.

[8] Platek, R. et Gray, G.B., "Imputation Methodology", Sous-division du développement d'enquêtes-ménages, document technique décrivant les biais et les variances des estimations, en utilisant différentes méthodes d'imputation.

[9] Szameitat, K. et Zindler, H.J., "The Reduction of Errors in Statistics by Automatic Corrections", Bulletin de l'International Statistical Institute, Recueil de la 35e Session, Belgrade 41, (Septembre 1965), pp. 395-417.

Annexe 1

Biais de l'estimation dans la région d'équilibrage/classe de pondération

Examinons l'estimation $\hat{X}_b$ telle que définie par 4.3 en général, et en particulier pour le cas (e1) dans le tableau l, à savoir la substitution par des enregistrements historiques pour m_b' de $(n_b - m_b)$ non-répondants, suivie par la pondération.

Alors
$$\hat{X}_b = \frac{n_b}{m_b + m_b'} \left[\sum_{i=1}^{m_b} x_i/\Pi_i + \sum_{j=m_b+1}^{m_b+m_b'} x_j'/\Pi_j \right] \qquad \text{(A1.1)}$$

Pour dériver le biais de $\hat{X}_b$, définissons δ_i comme 1 ou 0 en fonction de la réponse ou de la non-réponse de l'unité i et $\delta_i' = 1$ ou 0 en fonction de la disponibilité des enregistrements historiques et de leur utilisation pour l'imputation ou pas. Alors $m_b = \sum_{i=1}^{n_b} \delta_i$ et $m_b' = \sum_{i=1}^{n_b} (1-\delta_i)\,\delta_i'$. Dans le cas des méthodes (a) à (d), la totalité de $\delta_i' = 0$ et donc $m_b' = 0$.

$$\text{Alors } \hat{X}_b = \frac{n_b}{\sum_{i=1}^{n_b} [(\delta_i+(1-\delta_i)\delta_i']} \sum_{i=1}^{n_b} \Pi_i^{-1} [\delta_i(X_i+{}^R\varepsilon_i)+(1-\delta_i)\delta_i'(X_i+{}^R\varepsilon_i')] \qquad \text{(A1.2)}$$

où x_i et x_j' définis par 4.4 et 4.5 respectivement ont été substitués.

Pour déterminer le biais, il suffit de dériver $E\hat{X}_b$ en fonction de Al.2. Il faut faire fi du biais d'estimation du ratio, ainsi que de la covariance entre n_b et le ratio avec $\sum_{i=1}^{n_b}$ comme numérateur et dénominateur, une covariance qui peut exister lorsque la taille de l'échantillon, n_b, est une variable.

$$\text{Alors } E\hat{X}_b \doteq \frac{En_b \sum_{i=1}^{N_b} [\alpha_i(X_i+{}^R B_i)+(\alpha_i'-\alpha_i\alpha_i'-\text{Cov } \delta_i\delta_i')(X_i+{}^R B_i')]}{\sum_{i=1}^{N_b} \Pi_i [\alpha_i+(1-\alpha_i)\alpha_i'-\text{Cov } \delta_i\delta_i']} \qquad \text{(A1.3)}$$

sachant que $E^R\varepsilon_i = {}^R B_i$ et $E^R\varepsilon_i' = {}^R B_i'$.

Nous n'avons pas présumé de l'indépendance entre δ_i et δ_i', puisque la présence d'un enregistrement historique peut être liée à la tendance à répondre ou à ne pas répondre. Donc, $E(1-\delta_i)\,\delta_i' = (1-\alpha_i)\alpha_i' - \text{Cov } \delta_i\delta_i'$.

On peut simplifier encore plus A1.3 en utilisant des paramètres « moyens », par exemple, $E_b^* T_i = \sum_{i=1}^{N_b} (\Pi_i/En_b) T_i = \bar{T}_b$, quelle que soit la valeur de T_i. D'autres expressions, comme $Cov_b^* (T_i U_i) = E_b^* T_i U_i - E_b^* T_i E_b^* U_i$ et $V_b^*(T_i) = Cov_b^*(T_i, T_i)$ peuvent être dérivées. E_b^* est une moyenne pondérée de valeurs de paramètres individuels, utilisant Π_i/En_b comme poids, sachant que $\sum_{i=1}^{N_b} \Pi_i = En_b$.

Par conséquent,

$$\sum_{i=1}^{N_b} \alpha_i X_i = \sum_{i=1}^{N_b} \Pi_i \alpha_i X_i/\Pi_i = En_b E_b^* \alpha_i X_i/\Pi_i$$

$$= En_b [E_b^* \alpha_i E_b^* X_i/\Pi_i + Cov_b^*(\alpha_i, X_i/\Pi_i)]$$

$$= En_b [\bar{\alpha}_b (En_b)^{-1} X_b + Cov_b^*(\alpha_i, X_i/\Pi_i)]$$

$$= \bar{\alpha}_b X_b + En_b Cov_b^*(\alpha_i, X_i/\Pi_i) \qquad \text{(A1.4)}$$

Maintenant, supposons que $(1-\alpha_i)\alpha_i' - Cov\delta_i\delta_i' = \alpha_i''$ (A1.4a)

$$E_b^* \alpha_i'' = (1-\bar{\alpha}_b)\bar{\alpha}_b' - Cov_b^*\alpha_i, \alpha_i' - E_b^* Cov\delta_i\delta_i' = \alpha_b'' \qquad \text{(A1.5)}$$

Un peu comme on l'a fait pour A1.4,

$$\sum_{i=1}^{N_b} \alpha_i'' X_i = \bar{\alpha}_b'' X_b + En_b Cov_b^*(\alpha_i'', X_i/\Pi_i), \quad \text{où} \quad E_b^* \alpha_i'' = \bar{\alpha}_b''.$$

Alors $\hat{EX}_b \doteq En_b [(\bar{\alpha}_b + \bar{\alpha}_b'') X_b + En_b Cov_b^*(\alpha_i + \alpha_i'', X_i/\Pi_i)$

$$+ \sum_{i=1}^{N_b} (\alpha_i R_{B_i} + \alpha_i'' R_{B_i}')]/En_b(\bar{\alpha}_b + \bar{\alpha}_b''), \text{ où le biais équivaut à}$$

$$(\bar{\alpha}_b + \bar{\alpha}_b'')^{-1} [En_b Cov_b^*(\alpha_i + \alpha_i'', X_i/\Pi_i) + \sum_{i=1}^{N_b} (\alpha_i R_{B_i} + \alpha_i'' R_{B_i}')] \qquad \text{(A1.6)}$$

Le biais en fonction des méthodes d'imputation (a) à (d) suit immédiatement en établissant que $\alpha''_i = 0$ et $\bar{\alpha}''_b = 0$ à A1.6.

Le biais de $\hat{X}_b$ en fonction de la méthode (e2) peut être dérivé de la même façon, sauf que la covariance entre δ_i et δ''_j a été omise afin de simplifier la formule.

Annexe 2

Sommaire du développement de la variance de l'estimation $\hat{X}$

$\hat{X} = \sum_b \hat{X}_b$, où b = région d'équilibrage ou classe de pondération

$V(\hat{X}) = \sum_b V(\hat{X}_b) + \sum_{b \neq c} Cov(\hat{X}_b, \hat{X}_c)$, et donc, la nécessité de dériver $V(\hat{X}_b)$ et $Cov(\hat{X}_b, \hat{X}_c)$. Une covariance peut exister entre les estimations en fonction de différentes régions d'équilibrage ou classes de pondération, conformément à la définition des régions d'équilibrage ou des classes de pondération, ainsi que du plan de sondage.

Nous allons aborder la méthode (e1), où les renseignements historiques sont substitués aux non-réponses, chaque fois que c'est possible ou approprié, avant de passer à la pondération ou à la reproduction des enregistrements pour gonfler l'échantillon au niveau requis. Nous allons d'abord aborder la question de la pondération, où $\hat{X}_b$ se définit comme en 4.3 ou en A1.1 ou A1.2 (la forme la plus pratique pour le développement du biais et de la variance).

$\hat{X}_b$ peut être considéré comme une expression complexe de la forme

$n_b y_b/z_b$ où n_b, y_b et z_b sont des variables. Dans certains plans de sondage, n_b reste constant, mais il n'a pas à être inclus dans les développements généraux.

Ensuite, au moyen du développement en séries de Taylor,

$$\begin{aligned}\mathrm{Cov}(n_b y_b/z_b,\ n_c y_c/z_c) \doteq\ & (En_b\ Ey_b/Ez_b)(En_c\ Ey_c/Ez_c)\\ & [\mathrm{Rel\ Cov}\ n_b,n_c + \mathrm{Rel\ Cov}\ y_b,y_c + \mathrm{Rel\ Cov}\ z_b,z_c\\ & + (\mathrm{Rel\ Cov}\ n_b,y_c + \mathrm{Rel\ Cov}\ n_c,y_b) - (\mathrm{Rel\ Cov}\ n_b,z_c + \mathrm{Rel\ Cov}\ n_c,z_b)\\ & - (\mathrm{Rel\ Cov}\ y_b,\ z_c + \mathrm{Rel\ Cov}\ y_c,\ z_b)]\end{aligned} \qquad \text{(A2.1)}$$

et Rel Var $(n_b y_b/z_b)$ suit immédiatement en établissant que $c=b$.

Pour dériver $V(\hat{X}_b)$ et $\mathrm{Cov}(\hat{X}_b,\hat{X}_c)$, il nous faut les expressions suivantes. Ici $(1-\delta_i)\delta_i'$ Comme défini en A1.2 sera abrégé par δ_i'' de sorte que $E\delta_i''/i = \alpha_i''$ comme défini en A1.4a.

En_b ne peut pas être simplifié au-delà de $\sum_{i=1}^{N_b} \Pi_i$

$$V(n_b) = \sum_{i\neq j}^{N_b} \Pi_{ij} - En_b(En_b-1)$$

Ey_b est donné par [ ] en A1.6 et $Ez_b = En_b(\bar{\alpha}_b+\bar{\alpha}_b'')$, $\bar{\alpha}_b''$ étant donné en A1.5

Les expressions supplémentaires concernent les variances et les covariances, qui sont énoncées ci-après sans preuve, mais qui ont été développées par Platek et Gray [8].

$$V \sum_{i=1}^{n_b} \Pi_i^{-1}[\delta_i(X_i+{}^R\varepsilon_i) + \delta_i''(X_i+{}^R\varepsilon_i')]$$

$$= V_s\{\sum_{i=1}^{n_b} \Pi_i^{-1}[\alpha_i(X_i+{}^RB_i) + \alpha_i''(X_i+{}^RB_i')]\}$$

$$+ E_s\{V \sum_{i=1}^{n_b} \Pi_i^{-1}[\delta_i(X_i+{}^R\varepsilon_i) + \delta_i''(X_i+{}^R\varepsilon_i')]|s\ , \qquad \text{(A2.2)}$$

où s désigne un échantillon particulier de n_b unités.

La deuxième ligne, à savoir la composante de la variance non due à l'échantillonnage, est donnée par :

$$\sum_{i=1}^{N_b} V[\delta_i(X_i+{}^R\varepsilon_i) + \delta_i''(X_i+{}^R\varepsilon_i')]\Pi_i^{-1}$$

$$+ \sum_{i\neq j}^{N_b} \Pi_{ij}\Pi_i^{-1}\Pi_j^{-1}\ \mathrm{Cov}[\delta_i(X_i+{}^R\varepsilon_i) + \delta_i''(X_i+{}^R\varepsilon_i'), \delta_j(X_j+{}^R\varepsilon_j)+\delta_i''(X_j+{}^R\varepsilon_j')].$$

Ici, $V[\delta_i(X_i+{}^R\varepsilon_i)+\delta_i''(X_i+{}^R\varepsilon_i')] = \alpha_i\ {}^R\sigma_i^2 + \alpha_i''\ {}^R\sigma_i'^2$

$$+ 2(\alpha_i\alpha_i'' + \mathrm{Cov}\delta_i\delta_i'')r_{2ii''}\ {}^R\sigma_i\ {}^R\sigma_i' + V[\delta_i(X_i+{}^RB_i)+\delta_i''(X_i+{}^RB_i')]$$

et $\mathrm{Cov}[\delta_i(X_i+{}^R\varepsilon_i)+\delta_i''(X_i+{}^R\varepsilon_i'), \delta_j(X_j+{}^R\varepsilon_j)+\delta_j''(X_j+{}^R\varepsilon_j')]$

$$= (\alpha_i\alpha_j+\mathrm{Cov}\delta_i\delta_j)\ r_{2ij}\ {}^R\sigma_i\ {}^R\sigma_j + (\alpha_i\alpha_j'' + \mathrm{Cov}\delta_i\delta_j'')\ r_{2ij}''\ {}^R\sigma_i\ {}^R\sigma_j'$$

$$+ (\alpha_i''\alpha_j + \mathrm{Cov}\delta_i''\delta_j)\ r_{2ji}''\ {}^R\sigma_i'\ {}^R\sigma_j + (\alpha_i''\alpha_j'' + \mathrm{Cov}\delta_i''\delta_j'')\ r_{2ij}'\ {}^R\sigma_i'\ {}^R\sigma_j'$$

$$+ \mathrm{Cov}[\delta_j(X_i+{}^RB_i)+\delta_i''(X_i+{}^RB_i'), \delta_j(X_j+{}^RB_j)+\delta_j''(X_j+{}^RB_j')].$$

Dans la formule qui précède, $\mathrm{Cov}\ {}^R\varepsilon_i\ {}^R\varepsilon_j = r_{2ij}\ {}^R\sigma_i\ {}^R\sigma_j =$ la covariance entre les réponses actuelles de paires d'unités, $\mathrm{Cov}\ {}^R\varepsilon_i\ {}^R\varepsilon_j' = r_{2ij}''\ {}^R\sigma_i\ {}^R\sigma_j' =$

la covariance entre les réponses actuelles et historiques (également applicables pour $j=i$), et $\mathrm{Cov}\ {}^{R}\varepsilon_i'\ {}^{R}\varepsilon_j' = r'_{2ij}\ {}^{R}\sigma_i'\ {}^{R}\sigma_j' =$ la covariance entre les réponses historiques.

Si nous remplaçons α_i par $\alpha_i^2+V(\delta_i)$ et de même pour α_i'', les variances et covariances qui précèdent seraient décrites de façon symétrique.

$$
\begin{aligned}
& V\Big[\sum_{i=1}^{n_b}(\delta_i+\delta_i'')\Big] \\
& = V_s\Big[\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')\Big] \\
& + E_s\, V\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')\,|s \\
& = V_s\sum_{i=1}^{n_b}(\alpha_i+\alpha_i'') + \sum_{i=1}^{N_b}\Pi_i[V(\delta_i+\delta_i'')] \\
& + \sum_{i\neq j}^{N_b}\Pi_{ij}[\mathrm{Cov}(\delta_i+\delta_i'',\ \delta_j+\delta_j'')]\ . \qquad \text{(A2.4)}
\end{aligned}
$$

$$
\begin{aligned}
& \mathrm{Cov}\ \sum_{i=1}^{n_b}(\delta_i+\delta_i''),\ \sum_{i=1}^{n_b}\Pi_i^{-1}\,[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')] \\
& = \mathrm{Cov}\Big\{\sum_{i=1}^{n_b}(\alpha_i+\alpha_i''),\ \sum_{i=1}^{n_b}\Pi_i^{-1}[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]\Big\} \\
& + E_s\ \mathrm{Cov}\Big\{\sum_{i=1}^{n_b}(\delta_i+\delta_i''),\ \sum_{i=1}^{n_b}\Pi_i^{-1}[\delta_i(X_i+{}^{R}\varepsilon_i)+\delta_i''(X_i+{}^{R}\varepsilon_i')]\Big\}\,|s \qquad \text{(A2.5)}
\end{aligned}
$$

La deuxième ligne, à savoir la covariance non due à l'échantillonnage, est donnée par :

$$\sum_{i=1}^{N_b} \mathrm{Cov}(\delta_i+\delta_i''),\ [\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i')]/i$$

$$+ \sum_{i\neq j}^{N_b} \Pi_{ij}\,\Pi^{-1},\ \mathrm{Cov}(\delta_i+\delta_i'')\cdot[\delta_j(X_j+{}^{R}B_j)+\delta_j''(X_j+{}^{R}B_j')]\,|\,i,j \qquad \text{(A2.6)}$$

Pour les expressions de la covariance mettant en cause les régions d'équilibrage b et c, V_s est remplacé par Cov_s, $\sum_{i=1}^{N_b}$ n'existe pas et $\sum_{i\neq j}^{N_b}$ est remplacé par $\sum_{i=1}^{N_b}\sum_{j=1}^{N_c}$.

$$\mathrm{Cov}\ \{\, n_b,\ \sum_{i=1}^{n_b} \Pi_i^{-1}[\delta_i(X_i+{}^{R}B_i)+\delta_i''(X_i+{}^{R}B_i')]\}$$

$$= \mathrm{Cov}_s\{\, n_b,\ \sum_{i=1}^{n_b} \Pi_i^{-1}[\alpha_i(X_i+{}^{R}B_i)+\alpha_i''(X_i+{}^{R}B_i')]\} \qquad \text{(A2.7)}$$

et, enfin, $\mathrm{Cov}\ n_b,\ \sum_{i=1}^{n_b}(\delta_i+\delta_i'') = \mathrm{Cov}_s[n_b,\ \sum_{i=1}^{n_b}(\alpha_i+\alpha_i'')]$ (A2.8)

Maintenant $V(n_b y_b/z_b)$ peut être écrit à peu près comme suit :

$$\frac{(En_b)^2(Ey_b)^2}{(Ez_b)^2}\ V\left(\frac{n_b}{En_b}+\frac{y_b}{Ey_b}-\frac{z_b}{E_b}\right)$$

De même, $\mathrm{Cov}(n_b y_b/z_b,\ n_c y_c/z_c)$ peut être écrit à peu près comme suit :

$$\frac{En_b}{Ez_b}\frac{Ey_b}{}\quad\frac{En_c Ey_c}{Ez_c}\ \mathrm{Cov}\left(\frac{n_b}{En_b}+\frac{y_b}{Ey_b}-\frac{z_b}{Ez_b}\right),\ \left(\frac{n_c}{En_c}+\frac{y_c}{Ey_c}-\frac{z_c}{Ez_c}\right)$$

et grâce à une substitution partielle de la formule dérivée, nous pouvons obtenir $V(\hat{X}_b)$ comme indiqué en 6.1 et $\mathrm{Cov}\ (\hat{X}_b,\hat{X}_c)$.

Pour $V(\hat{X}_b)$ au moyen des méthodes (a) à (d), il suffit d’établir que δ_i'', α_i'', $\bar{\alpha}_b''$ équivalent à zéro. La totalité de $\text{Cov}\ \delta_i\delta_i''$, $\text{Cov}\ \delta_i\delta_j'' = 0$ pour les méthodes (a) à (d).