Techniques d'enquête, juin 1979
Vol. 5, nº 1
Statistique Canada, nº 12-001-X au catalogue

Étude empirique d'une méthode améliorée d'estimation de la variance de réponse corrélée[1]

A. MacLeod et K.P. Krótki[2]

Résumé

Deux méthodes d'estimation de la variance de réponse carrelée d'un estimateur d'enquête sont examinées à partir d'une comparaison théorique et d'une étude empirique. On examine ensuite la variance de ces estimateurs et les effets des observations détachées. Enfin, un estimateur amélioré est défini et évalué.

Mots-clés : Variance de réponse corrélée; estimation de variance; valeur aberrante.

1 Introduction

La variance de réponse corrélée (VRC) est une composante de la variance globale ou totale d'un estimateur d'enquête. Sa définition est basée sur la considération qu'un processus d'enquête est théoriquement répétable, de sorte qu'il existe une série d'essais indépendants de l'enquête. L'écart de réponse désigne la différence entre la réponse d'une personne dans l'essai réel de l'enquête et la valeur attendue de la réponse de cette personne, l'espérance porte sur tous les essais conceptuels de l'enquête. La VRC désigne alors une mesure de la variabilité de la réponse causée par la corrélation de ces écarts de réponse pour toute paire de personnes dans l'essai donné du processus d'enquête. Il a été démontré empiriquement que, dans un contexte où les agents recenseurs aident les répondants à remplir le questionnaire, la VRC est une composante importante de la variance totale [3].

Pour estimer la VRC de diverses variables, deux méthodes ont été employées. La première, appelée ci-après l'ancienne méthode, est décrite dans [2]. Il s'agit de la méthode sur laquelle Statistique Canada a fondé ses estimations publiées de la variance totale pour les recensements de 1971 et de 1976. L'autre méthode, appelée ci-après la nouvelle méthode, est définie et obtenue dans [4].

1. Adapté d'un article présenté lors de la réunion annuelle de l'American Statistical Association, Section on Survey Research Methods, 13 au 16 août 1979, Washington, D.C., États-Unis.
2. A. MacLeod et K.P. Krótki, Division des méthodes de recensement d'enquêtes-ménages, Statistique Canada.

Le présent article vise à examiner et à comparer ces deux méthodes, tant en ce qui concerne les estimateurs eux-mêmes (section 2) que les estimations correspondantes qui sont obtenues (section 3). Ces estimations seront fournies et examinées pour les données recueillies à la fois sur une base de l'ensemble de la population et à partir d'un échantillon. En outre, la variance des deux estimateurs est présentée dans la section 4, et la section 5 traite des effets des valeurs aberrantes. La section 6 consiste à mettre au point un estimateur amélioré dont la variance est inférieure à celle de l'ancien ou du nouvel estimateur. L'article se termine par la section 7 qui contient des commentaires récapitulatifs, une évaluation globale de la méthode améliorée et quelques idées de travaux de recherche.

2 Une comparaison théorique des estimateurs

Étant donné qu'un processus d'enquête ne peut pas, en pratique, être répété un certain nombre de fois de manière indépendante, une technique du réseau d'échantillons des agents recenseurs et des secteurs de dénombrement (SD)[1] est utilisée pour les deux estimateurs de VRC. Dans la pratique normale du recensement, un agent recenseur est affecté à un seul SD. Dans le cas du réseau d'échantillons, cependant, les SD voisins sont jumelés et, au sein de ces paires, les ménages de chaque SD sont répartis au hasard en deux groupes égaux. Un agent recenseur est ensuite affecté de manière aléatoire à l'un de ces deux groupes dans chaque SD de la paire, tandis que le second agent recenseur se voit confier l'autre groupe de chaque SD de la paire.

Pour l'ancien estimateur, seul un échantillon de paires de SD superposées est nécessaire. Le nouvel estimateur, en revanche, nécessite non seulement l'échantillon de paires de SD superposées, mais aussi un échantillon de paires de SD non superposées. En théorie, toutes les paires auxquelles la technique du réseau d'échantillons n'est pas appliquée pourraient être utilisées, au lieu d'un simple échantillon de ces paires. Cependant, des contraintes budgétaires pratiques imposent l'utilisation d'un échantillon.

1. Le SD est une unité spatiale constituée d'un groupe de ménages géographiquement contigus et assignée à un agent recenseur. En 1976, le Canada était divisé en 35 154 SD.

La variance de réponse corrélée pour un estimateur de recensement X établi à partir de données-échantillons[1] est donnée par l'expression suivante :

$$CRV\ (X) = \sum_{k=1}^{2P} N_k^2 \ \frac{(n_k-1)}{n_k}\ \rho_k\ \sigma_k^2 \qquad (2.1)$$

où : k correspond à l'indice de SD désignant le SD;

P est le nombre de paires de SD au Canada;

N_k est le nombre total de ménages dans le k^e SD;

n_k est le nombre de ménages de l'échantillon dans le k^e SD (n_k est approximativement égal à $N_n/3$).

L'entité $\rho_k\sigma_k^2$ est le terme clé de (2.1). Elle est définie comme la valeur moyenne de $E(x_{kh}-X_{kh})\,(x_{kh'}-X_{kh'})$, pour les paires de ménages $(h,h',h\neq h')$ dans le k^e SD qui sont dénombrées par le même agent recenseur. x_{kh} est la réponse réelle au recensement et X_{kh} est la moyenne de toutes les réponses conceptuelles de sorte que $(x_{kh}-X_{kh})$ est l'écart de réponse dans le h^e ménage du k^e SD. L'espérance porte sur toutes les réponses conceptuelles, y compris les tâches de l'agent recenseur, pour un ménage donné. Pour les données intégrales, $n_k=N_k$ et donc (2.1) se réduit à :

$$CRV\ (X) = \sum_{k=1}^{2P} N_k(N_k-1)\ \rho_k\ \sigma_k^2 \quad . \qquad (2.2)$$

En traitant d'abord des caractéristiques de l'échantillon, l'ancien estimateur est donné par :

$$\hat{CRV}(X) = \frac{P}{P_1} \sum_{k=1}^{2P_1} N_k^2 \ \frac{(n_k-1)}{n_k}\ (C_k-D_k) \qquad (2.3)$$

où p_1 est le nombre de paires de SD superposées dans l'échantillon, C_k est la somme des carrés entre les agents recenseurs et D_k est la somme des carrés parmi les agents recenseurs. Les expressions pour C_k et D_k sont données dans [2]. Sous certaines conditions, (C_k-D_k) est un estimateur sans biais de $\rho_k\sigma_k^2$.

1. Dans le Recensement de 1976, les données démographiques de base et la langue maternelle ont été recueillies sur la base de l'ensemble de la population. Les données sur la migration, l'éducation et l'activité sur le marché du travail sont basées sur un échantillon constitué du tiers de la population.

Dans le cas des données intégrales, les expressions pour C_k et D_k sont légèrement modifiées. L'estimateur de VRC dans cette situation est donné par :

$$\hat{CRV}(X) = \frac{P}{P_1} \sum_{k=1}^{2P_1} N_k(N_k-1)(C_k-D_k) \quad . \tag{2.4}$$

Il est démontré dans [2] que, sous certaines conditions, les deux estimateurs sont exempts de biais pour une variance de réponse corrélée, c'est-à-dire que (2.3) est un estimateur sans biais de (2.1) alors que (2.4) est un estimateur sans biais de (2.2).

Une moyenne pondérée de l'expression $(C_k - D_k)$ sur les SD superposés a été utilisée comme estimateur de la VRC et de la variance totale. Par exemple, Statistique Canada a utilisé une telle procédure pour les recensements de 1971 et de 1976. La méthodologie et les résultats sont abordés dans [5] et [6]. Le Bureau of the Census des États-Unis a également utilisé une combinaison pondérée de $(C_k - D_k)$ basée sur les données de 1960, comme il est indiqué dans [7]. Bailey, Moore et Bailar [1] ont calculé des moyennes d'expressions équivalentes à $(C_k - D_k)$ sur toutes les paires de zones d'affectation dans le cadre d'une enquête nationale sur la criminalité.

Dans la nouvelle méthode, les échantillons non superposés et superposés sont considérés séparément au niveau de la paire de SD. Pour chaque paire m, on détermine la quantité $A_m = (t_{m1} - t_{m2})^2$ où t_{mj} est le nombre d'unités d'une certaine catégorie dans le j^e SD de la m^e paire $(j = 1, 2)$. La moyenne de A_m est ensuite calculée séparément pour les paires non superposées et pour les paires superposées. Enfin, on obtient une différence pondérée de ces deux moyennes résultantes. Fellegi [4] a indiqué que cette différence est un estimateur biaisé de $\rho_m \sigma_m^2$, qui est la valeur moyenne de $E(x_{kh} - X_{kh})(x_{k'h'} - X_{k'h'})$, $(k = 1, 2;$ si $k = k'$, puis $h \neq h')$. L'espérance porte sur l'ensemble des réponses conceptuelles pour un ménage donné. La moyenne est calculée sur toutes les paires de ménages de la m^e paire de SD qui ont été interrogés par le même agent recenseur et sur les deux agents recenseurs de la paire de SD. Fellegi [4] prévoit que le biais de cet estimateur est faible.

Comme les estimations basées sur les formules (2.3) et (2.4) étaient déjà disponibles, on a décidé d'obtenir une forme du nouvel estimateur ayant

la même valeur attendue que les estimateurs de (2.3) et (2.4). Cela s'est avéré quelque peu compliqué puisque l'ancien estimateur est basé sur des calculs au sein de SD individuels alors que le nouvel estimateur est basé sur des calculs au sein de paires de SD. Le problème a été résolu en examinant la valeur attendue de l'ancien et du nouvel estimateur et en supposant que $\rho_k\sigma_k^2 = \rho_{k'}\sigma_{k'}^2 = \rho_m\sigma_m^2$, où k et k′ sont les deux SD de n'importe quelle paire m. Pour les données intégrales, (2.2) peut être réécrit comme suit :

$$CRV(X) = \sum_{m=1}^{P} [N_{m1}(N_{m1}-1) + N_{m2}(N_{m2}-1)]\ \rho_m\sigma_m^2$$

où m est l'indice désignant la paire considérée et N_{m1} est le nombre total de ménages privés dans l'i^{e} SD (1 = 1 ou 2) de la m^{e} paire. Pour les données-échantillons, (2.1) peut être réécrit de façon similaire.

Pour les données-échantillons, le nouvel estimateur révisé s'est avéré être :

$$\hat{CRV}(X) = 2P\left\{ \sum_{m=P_1}^{P} W_m \frac{\left\{\frac{N_{m1}}{n_{m1}} t_{m1} - \frac{N_{m2}}{n_{m2}} t_{m2}\right\}^2}{p - P_1} - \sum_{m=1}^{P_1} W_m \frac{\left\{\frac{N_{m1}}{n_{m1}} t_{m1} - \frac{N_{m2}}{n_{m2}} t_{m2}\right\}^2}{P_1} \right\} \quad (2.5)$$

où :

$$W_m = \frac{N_{m1}^2\ (n_{m1} - 1)}{N_m^2\ n_{m1}} + \frac{N_{m2}^2\ (n_{m2} - 1)}{N_m^2\ n_{m2}} \quad .$$

Dans le cas des données intégrales, le nouvel estimateur révisé est :

$$\hat{CRV}(X) = 2P\ \frac{1}{P-P_1} \sum_{m=P_1+1}^{P} \frac{N_m^2(N_{m-1}) - 2N_{m1}N_{m2}}{N_m^2} (t_{m1}-t_{m2})^2$$

$$- \frac{1}{P_1} \sum_{m=1}^{P_1} \frac{N_m^2(N_{m-1}) - 2N_{m1}N_{m2}}{N_m^2} (t_{m1} - t_{m2})^2 \quad . \quad (2.6)$$

Pour des raisons de commodité algébrique, on suppose, sans perte de généralité, que dans (2.5) et (2.6) les premières paires de SD P_1 sont superposées. Toutes les utilisations futures des termes « nouvelle méthode » ou « nouvel estimateur » feront référence à (2.5) pour les données-échantillons et à (2.6) pour les données intégrales.

3 Une comparaison empirique des estimateurs

Les anciennes et nouvelles estimations sont basées sur des échantillons de 375 paires de SD superposées et de 564 paires de SD non superposées. Les échantillons ont été obtenus par un processus à deux étapes. Dans la première étape, 188 districts de commissaire au recensement (DCR)[1] ont été sélectionnés par échantillonnage avec PPT (probabilité proportionnelle à la taille), où la mesure de la taille était le nombre de SD par DCR. Ces DCR ont été choisis indépendamment parmi 16 strates formées par le recoupement de huit régions géographiques du Canada avec deux types de méthodes de dénombrement (ramassage et retour par la poste). Dans la deuxième étape, dans chaque DCR sélectionné, tous les SD ont d'abord été appariés. Ensuite, deux paires ont été sélectionnées au hasard pour l'échantillon superposé, puis trois paires supplémentaires pour l'échantillon non superposé. La sélection des SD parmi les DCR a été faite sans remplacement.

Pour être jumelés, deux SD devaient être contigus et du même type de dénombrement et devaient posséder des caractéristiques linguistiques, agricoles et relatives à la densité similaires. En outre, chaque SD d'une paire non superposée devait être un SD « à pleine charge ». Autrement dit, il a été attribué à un agent recenseur à qui on n'a attribué aucun autre SD. Ces critères de jumelage ont tendance à rendre semblables les deux SD d'une paire, de sorte que l'hypothèse de la section 2 selon laquelle $\rho_k\sigma_k^2 = \rho_{k'}\sigma_{k'}^2 = \rho_m\sigma_m^2$ pour chaque paire m est confortée.

En utilisant l'ancien estimateur, la variance de réponse corrélée a été calculée pour un grand nombre de catégories en fonction des données intégrales (l'âge, le sexe, la langue maternelle, l'état matrimonial) et des données-échantillons (l'éducation, la population active, la mobilité). Une sélection semblable a été faite pour le nouvel estimateur.

1. Un DCR compte en moyenne 21 SD.

La sélection des variables et des catégories a été faite de manière à porter sur un large éventail de résultats et de champs d’intérêt spécialisés. Les estimations de la VRC résultante pour les deux estimateurs sont données pour les données intégrales dans le tableau 1 et pour les données-échantillons dans le tableau 2. Les coefficients de variation correspondants sont également donnés. Le coefficient de variation désigne la racine carrée de la valeur absolue de l’estimation de la VRC pour cette catégorie divisée par le nombre de personnes au Canada appartenant à cette catégorie.

Tant pour les données intégrales que pour les données-échantillons, la majorité des anciennes et des nouvelles estimations de la VRC sont négatives. Cette négativité suit diverses tendances. Par exemple, presque toutes les estimations de la mobilité sont positives, alors que la plupart des estimations de la population active sont négatives. Toutefois, on peut observer quelques écarts intéressants par rapport à ces tendances. Par exemple, toutes les nouvelles estimations concernant la langue maternelle sont négatives, tandis que seules 2 des 5 anciennes estimations sont négatives. En revanche, 3 des 4 anciennes estimations pour les catégories d’âge sont négatives, mais toutes les nouvelles estimations pour l’âge sont positives.

Les estimations examinées selon l’ancienne méthode tendent à démontrer une relation inverse entre le coefficient de variation et la taille de la catégorie. Cette même tendance se retrouve, dans une moindre mesure, dans les résultats de la nouvelle méthode, mais les exceptions sont plus marquées. Par exemple, « langue maternelle : français » et « mobilité : personne ayant déménagé » ont tous deux des populations et des coefficients de variation assez importants, tandis que « chômeur » et « fréquentant l’école à plein temps » ont tous deux des populations et des coefficients de variation assez faibles.

Tableau 1
Estimations de la variance de réponse corrélée (VRC) selon l'ancien et le nouvel estimateur avec les coefficients de variation correspondants pour des catégories sélectionnées en fonction de données intégrales

Catégorie	Ancienne VRC Estimation	Coefficient de variation	Nouvelle VRC Estimation	Coefficient de variation
État matrimonial :				
Marié(e)	-322 041	$5{,}35 \times 10^{-5}$	25 951 715	$4{,}81 \times 10^{-4}$
Langue maternelle :				
Anglais	2 529 933	$1{,}13 \times 10^{-4}$	-49 979 989	$5{,}01 \times 10^{-4}$
Français	841 832	$1{,}56 \times 10^{-4}$	-51 555 178	$1{,}22 \times 10^{-3}$
Allemand	-677 412	$1{,}73 \times 10^{-3}$	-2 466 910	$3{,}30 \times 10^{-3}$
Italien	206 108	$9{,}34 \times 10^{-4}$	-7 485 919	$5{,}65 \times 10^{-3}$
Ukrainien	-336 429	$2{,}06 \times 10^{-3}$	-957 338	$3{,}47 \times 10^{-3}$
Âge :				
5 ans	-97 258	$8{,}29 \times 10^{-4}$	93 087	$8{,}11 \times 10^{-4}$
20 à 24 ans	-110 764	$1{,}56 \times 10^{-4}$	2 900 633	$7{,}98 \times 10^{-4}$
25 à 29 ans	325 555	$2{,}86 \times 10^{-4}$	3 175 864	$8{,}94 \times 10^{-4}$
35 ans	-147 409	$1{,}35 \times 10^{-3}$	15 733	$4{,}41 \times 10^{-4}$
Sexe :				
Masculin	-255 958	$4{,}45 \times 10^{-5}$	-7 608 908	$2{,}43 \times 10^{-4}$

Tableau 2
Estimations de la variance de réponse corrélée (VRC) selon l'ancien et le nouvel estimateur avec les coefficients de variation correspondants pour des catégories sélectionnées en fonction de données-échantillons

Catégorie	Ancienne VRC Estimation	Coefficient de variation	Nouvelle VRC Estimation	Coefficient de variation
Plus haut grade obtenu :				
Certificat d'études secondaires	2 344 591	$4{,}56 \times 10^{-4}$	11 293 155	$1{,}00 \times 10^{-3}$
Baccalauréat	-899 953	$1{,}20 \times 10^{-3}$	1 058 321	$1{,}30 \times 10^{-3}$
Maîtrise	-1 928 039	$8{,}24 \times 10^{-4}$	-147 189	$2{,}28 \times 10^{-3}$
Certificat non universitaire	-1 230 405	$6{,}05 \times 10^{-4}$	485 724	$3{,}80 \times 10^{-4}$
Dernière année d'études terminée				
N'ayant pas atteint la 5e année	932 373	$1{,}12 \times 10^{-3}$	1 398 493	$1{,}37 \times 10^{-3}$
5e à 8e année	1 648 121	$3{,}60 \times 10^{-4}$	-4 066 516	$5{,}66 \times 10^{-4}$
9e et 10e année	1 502 907	$3{,}68 \times 10^{-4}$	-3 689 625	$5{,}76 \times 10^{-4}$
Ne fréquentant pas l'école	4 197 176	$1{,}43 \times 10^{-4}$	10 793 519	$2{,}29 \times 10^{-4}$
Fréquentant l'école à temps plein	495 245	$4{,}20 \times 10^{-4}$	618 369	$4{,}69 \times 10^{-4}$
Chômeur	-2 171 495	$2{,}11 \times 10^{-3}$	-756 508	$1{,}24 \times 10^{-3}$
Mise à pied temporaire	-5 263 069	$2{,}24 \times 10^{-2}$	-88 831	$2{,}91 \times 10^{-3}$
En attente pour commencer un nouvel emploi	-2 761 460	$1{,}39 \times 10^{-2}$	-33 105	$1{,}53 \times 10^{-3}$
A cherché un emploi	-2 275 514	$3{,}15 \times 10^{-3}$	-609 951	$1{,}63 \times 10^{-3}$
Personne occupée	3 357 282	$1{,}92 \times 10^{-4}$	41 506 918	$6{,}74 \times 10^{-4}$
Inactif	4 241 919	$3{,}01 \times 10^{-4}$	-966 135	$1{,}46 \times 10^{-4}$
Mobilité (au cours des cinq dernières années) :				
Personne n'ayant pas déménagé	6 204 320	$2{,}28 \times 10^{-4}$	-122 571 179	$1{,}01 \times 10^{-3}$
Personne ayant déménagé	3 093 314	$1{,}71 \times 10^{-4}$	218 080 472	$1{,}43 \times 10^{-3}$
Non-migrant[a]	1 567 852	$2{,}51 \times 10^{-4}$	26 889 868	$1{,}04 \times 10^{-3}$

[a] Un non-migrant est une personne qui vit dans la même municipalité qu'il y a cinq ans.

Les anciens coefficients de variation ont également tendance à être de moindre importance que les nouveaux. Il y a des exceptions à cette tendance, surtout lorsqu'on examine des données-échantillons; les plus

remarquables concernent le chômage. Néanmoins, lorsque l'on tient compte de l'ensemble des estimations de l'échantillon du tableau 2, les anciens coefficients sont le plus souvent plus petits que les nouveaux. Une raison possible de ce phénomène général est qu'il existe un effet non mesuré confondu avec toutes les nouvelles estimations de la VRC. La quantité $(t_{m1} - t_{m2})^2$ dans le nouvel estimateur est influencée partiellement et de manière indésirable par le fait que les nombres de ménages dans le premier et le deuxième SD de la paire m ne sont pas égaux entre eux. Cependant, avec l'ancien estimateur, puisque chaque agent recenseur a exactement le même nombre de ménages à dénombrer dans chaque SD, il n'y a pas d'effet indésirable de ce type sur C_k et D_k, et donc sur l'estimation globale.

Il est indiqué dans [7] qu'il existait des variances de réponse relativement élevées dans le recensement américain de 1960 pour les données relatives aux faibles niveaux de scolarité et au chômage (entre autres domaines). Un examen du tableau 2 a tendance à confirmer ce point de vue, dans la mesure où les coefficients de variation pour ces catégories sont parmi les plus élevés affichés dans les tableaux 1 et 2. Toutefois, il convient de souligner que de nombreuses autres estimations présentent des coefficients de variation du même ordre, notamment la langue maternelle (autre que l'anglais et le français) et de nombreuses catégories du « plus haut grade obtenu » et de la « mobilité ».

4 La variance des estimateurs de la variance de réponse corrélée

L'étape suivante de la présente étude consistait à tenter d'estimer la variance des estimateurs de la VRC afin de pouvoir associer une certaine mesure de fiabilité aux estimations de la VRC. La méthode du jackknife a été utilisée. Les 16 strates décrites dans la section 3 ont servi de strates pour l'estimation de la variance. En fait, 10 strates contenaient un nombre impair de DCR, ce qui a nécessité la suppression aléatoire de l'un d'eux dans chacune d'elles. Au total, 50 paires de SD ont été supprimées de l'échantillon (20 paires superposées et 30 paires non superposées).

Les variances estimées de l'ancien et du nouvel estimateur de la VRC sont données pour des catégories sélectionnées en fonction des données intégrales et des données-échantillons des tableaux 3 et 4, respectivement. Ces tableaux comprennent également le coefficient de variation qui désigne la racine carrée de la variance de l'estimateur de la VRC divisée par la valeur absolue de l'estimation de la VRC.

Les tableaux indiquent que les estimations de la variance du nouvel estimateur de la VRC sont presque toutes plus élevées que celles de l'ancien, la plupart des exceptions se produisant pour les estimations liées au chômage. Toutefois, la comparaison des anciens et des nouveaux coefficients de variation ne semble pas révéler de tendances claires. Même pour les codes d'une seule variable, il n'y a pas de tendance évidente pour les anciens coefficients. Il est admis que les nouveaux coefficients de la plupart des catégories de population active sont plus élevés que ceux des anciens et que les nouveaux coefficients de la plupart des catégories de mobilité sont plus faibles que ceux des anciens, mais il y a toujours des exceptions dans ces deux cas. L'observation la plus importante à tirer des estimations du coefficient de variation est donc leur grande ampleur. Beaucoup d'entre eux étant supérieurs à 1, il y a de fortes raisons de penser que la variance des anciennes et des nouvelles estimations de la VRC est élevée.

Tableau 3
Estimations de la variance (avec coefficients de variation) des estimateurs de l'ancienne et de la nouvelle variance de réponse corrélée (VRC) pour les données intégrales

Catégorie	Variance de l'ancienne VRC	Coefficient de variation (ancienne)	Variance de la nouvelle VRC	Coefficient de variation (nouvelle)
État matrimonial :				
Marié(e)	$4{,}355 \times 10^{11}$	2,049	$5{,}424 \times 10^{14}$	0,897
Langue maternelle :				
Anglais	$5{,}861 \times 10^{12}$	0,957	$1{,}612 \times 10^{16}$	2,540
Français	$7{,}119 \times 10^{11}$	1,002	$4{,}806 \times 10^{15}$	1,345
Allemand	$4{,}852 \times 10^{11}$	1,028	$1{,}206 \times 10^{13}$	1,408
Italien	$2{,}997 \times 10^{12}$	8,399	$5{,}626 \times 10^{14}$	3,169
Ukrainien	$1{,}107 \times 10^{11}$	0,989	$6{,}987 \times 10^{11}$	0,873
Âge :				
5 ans	$1{,}361 \times 10^{9}$	0,379	$7{,}818 \times 10^{9}$	0,950
20 à 24 ans	$8{,}922 \times 10^{10}$	2,697	$6{,}716 \times 10^{12}$	0,893
25 à 29 ans	$1{,}482 \times 10^{11}$	1,182	$8{,}043 \times 10^{12}$	0,893
35 ans	$4{,}995 \times 10^{8}$	0,152	$2{,}365 \times 10^{9}$	3,091
Sexe :				
Masculin	$3{,}267 \times 10^{11}$	2,233	$2{,}256 \times 10^{15}$	6,242

Tableau 4
Estimations de la variance (avec coefficients de variation) des estimateurs de l'ancienne et de la nouvelle variance de réponse corrélée (VRC) pour les données-échantillons

Catégorie	Variance de l'ancienne VRC	Coefficient de variation (ancienne)	Variance de la nouvelle VRC	Coefficient de variation (nouvelle)
Plus haut grade obtenu :				
Certificat d'études secondaires	$1{,}088 \times 10^{12}$	0,445	$2{,}473 \times 10^{13}$	0,440
Baccalauréat	$4{,}388 \times 10^{11}$	0,736	$5{,}801 \times 10^{11}$	0,720
Maîtrise	$8{,}694 \times 10^{11}$	0,484	$9{,}277 \times 10^{9}$	0,654
Certificat non universitaire	$2{,}424 \times 10^{11}$	0,400	$7{,}545 \times 10^{12}$	5,655
Dernière année d'études terminée				
N'ayant pas atteint la 5e année	$3{,}535 \times 10^{12}$	2,017	$4{,}193 \times 10^{12}$	1,464
5e à 8e année	$3{,}623 \times 10^{12}$	1,155	$3{,}608 \times 10^{13}$	1,477
9e et 10e année	$1{,}834 \times 10^{12}$	0,901	$4{,}435 \times 10^{13}$	1,805
Ne fréquentant pas l'école	$1{,}013 \times 10^{13}$	0,758	$2{,}968 \times 10^{15}$	5,047
Fréquentant l'école à temps plein	$5{,}706 \times 10^{11}$	1,525	$4{,}688 \times 10^{12}$	3,501
Chômeur	$4{,}147 \times 10^{11}$	0,297	$4{,}515 \times 10^{11}$	0,888
Mise à pied temporaire	$1{,}747 \times 10^{13}$	0,794	$1{,}629 \times 10^{10}$	1,437
En attente pour commencer un nouvel emploi	$8{,}143 \times 10^{11}$	0,327	$1{,}764 \times 10^{9}$	1,269
A cherché un emploi	$2{,}620 \times 10^{11}$	0,225	$2{,}139 \times 10^{11}$	0,758
Personne occupée	$7{,}000 \times 10^{12}$	0,788	$7{,}739 \times 10^{14}$	0,670
Inactif	$1{,}357 \times 10^{13}$	0,868	$2{,}544 \times 10^{14}$	16,012
Mobilité (au cours des cinq dernières années) :				
Personne n'ayant pas déménagé	$4{,}309 \times 10^{13}$	1,058	$8{,}912 \times 10^{14}$	0,400
Personne ayant déménagé	$9{,}542 \times 10^{13}$	3,158	$2{,}110 \times 10^{13}$	0,402
Non-migrant	$7{,}176 \times 10^{12}$	1,709	$2{,}512 \times 10^{14}$	0,589

5 Les effets des valeurs aberrantes sur les estimations de la variance de réponse corrélée

Afin de déterminer avec précision certaines des causes des coefficients de variation élevés, une étude a été réalisée sur les effets des valeurs aberrantes sur les anciennes et nouvelles estimations de la VRC. Au départ, on soupçonnait fortement que quelques valeurs aberrantes influaient fortement sur certaines des nouvelles estimations de la VRC. Le premier problème d'une telle étude consistait à élaborer un critère systématique pour définir une valeur aberrante. Tandis que certaines valeurs aberrantes étaient plutôt évidentes, d'autres, potentielles, ne l'étaient pas autant. Cela signifie qu'une règle de décision était nécessaire.

Pour ce qui est de la nouvelle méthode, les valeurs de $(t_{m1} - t_{m2})$ ont été examinées pour les 564 paires non superposées et pour les 375 paires superposées (séparément). Toute paire pour laquelle $(t_{m1} - t_{m2})$ présentait plus de cinq écarts-types par rapport à la moyenne a été déclarée paire aberrante.

Pour ce qui est de l'ancienne méthode, les quantités C_k et D_k ont été examinées pour chacun des 750 SD superposés pour des caractéristiques données. On a constaté que les valeurs de C_k et D_k étaient plus dispersées autour de leur moyenne que les valeurs de $(t_{m1} - t_{m2})$. Le critère élaboré pour l'ancienne méthode était que, si l'un ou l'autre de C_k ou de D_k s'écartait de plus de huit écarts-types de la moyenne, alors le SD k serait appelé un SD aberrant (pour la caractéristique particulière).

L'étape suivante consistait à recalculer les estimations de la VRC pour certaines catégories en supprimant les valeurs aberrantes (et en réduisant en conséquence la taille de l'échantillon). Si l'on examine d'abord l'ancien estimateur, environ quatre SD ont été supprimés en moyenne pour chaque catégorie, mais on a observé peu de changement dans les estimations de la VRC (tableau 5). Sans les valeurs aberrantes, les VRC étaient légèrement plus proches de zéro, mais l'ordre (puissance de 10) des estimations était le même avec ou sans valeurs aberrantes pour pratiquement toutes les catégories. L'ancien estimateur ne semblait donc pas être très sensible aux valeurs aberrantes, et des études supplémentaires ont été omises.

La suppression des paires de SD aberrantes a eu un effet plus notable sur les estimations de la VRC. En moyenne, il y avait environ trois paires aberrantes non superposées et environ deux paires aberrantes superposées par catégorie, bien que ces chiffres soient légèrement supérieurs pour les catégories de langue maternelle et légèrement inférieurs pour les catégories d'âge. Ainsi, sans trop de surprise, les VRC n'ont pas beaucoup changé pour les catégories d'âge, mais pour la langue maternelle, la tendance des VRC à se rapprocher de zéro était assez prononcée, comme on peut le voir dans le tableau 5.

Une autre méthode pour traiter les valeurs aberrantes au moyen du nouvel estimateur consistait à remplacer la valeur de $(t_{m1} - t_{m2})$ dans chaque paire aberrante par la moyenne appropriée, selon que la paire était non superposée ou superposée. Cette approche a également été tentée dans le cas présent et a donné des résultats très semblables à l'approche consistant à supprimer complètement les valeurs aberrantes, comme on peut le voir en comparant les colonnes (ii) et (iii) du tableau 5.

La variance des VRC réestimées a été déterminée pour les catégories du tableau 5 en utilisant le nouvel estimateur. Pour les tranches d'âge de

6 ans et de 4 à 6 ans, la variance a en fait légèrement augmenté lorsque les paires aberrantes ont été supprimées, mais la variance des autres catégories était plus faible lorsque les paires aberrantes avaient été supprimées. Ces diminutions de la variance étaient les plus notables pour les petits groupes de langue maternelle.

En conclusion, bien que la suppression des valeurs aberrantes ait eu plus d'effet sur le nouvel estimateur de la VRC que sur l'ancien en réduisant à la fois les estimations de la VRC et leurs variances, les effets n'étaient toujours pas significatifs. À l'exception des petits groupes de langue maternelle, les nouvelles VRC sont restées plus éloignées de zéro que leurs anciennes contreparties, et la variance des nouvelles VRC est restée plus grande que celle des anciennes VRC.

Tableau 5
Variance de réponse corrélée (VRC) et estimations de la variance (VRC) avec et sans valeurs aberrantes au moyen du nouvel estimateur

Catégories	(i) VRC, inclusion des valeurs aberrantes	(ii) VRC, suppression des valeurs aberrantes	(iii) VRC, remplacement des valeurs aberrantes	(iv) Variance (VRC), inclusion des valeurs aberrantes	(v) Variance (VRC), suppression des valeurs aberrantes
Langue maternelle :					
Anglais	-49 979 989	-38 256 600	-38 371 439	$1{,}612 \times 10^{16}$	$0{,}469 \times 10^{16}$
Français	-51 555 178	-3 928 355	-3 067 913	$4{,}806 \times 10^{15}$	$3{,}312 \times 10^{15}$
Allemand	-2 466 910	202 452	198 152	$1{,}206 \times 10^{13}$	$0{,}018 \times 10^{13}$
Italien	-7 485 919	555 478	541 320	$5{,}626 \times 10^{14}$	$0{,}023 \times 10^{14}$
Ukrainien	-957 338	29 849	32 289	$6{,}987 \times 10^{11}$	$1{,}666 \times 10^{11}$
Âge : 6 ans	-6 326	-14 027	-14 365	$4{,}264 \times 10^{9}$	$5{,}734 \times 10^{9}$
4 à 6 ans	-184 275	-136 416	-156 133	$7{,}301 \times 10^{11}$	$8{,}940 \times 10^{11}$
20 à 24 ans	2 900 633	2 045 793	2 031 966	$6{,}716 \times 10^{12}$	$4{,}190 \times 10^{12}$
25 à 29 ans	3 175 864	2 290 961	2 292 835	$8{,}043 \times 10^{12}$	$2{,}928 \times 10^{12}$

6 Un estimateur amélioré

Il a été suggéré dans Fellegi [4, p. 500] qu'une combinaison pondérée de l'ancien et du nouvel estimateur fournirait un estimateur amélioré. Un tel estimateur a été calculé au moyen de facteurs de pondération inversement proportionnels aux variances des estimateurs. Les résultats sont présentés dans le tableau 6. L'estimation améliorée a été calculée

uniquement pour les catégories en fonction des données-échantillons. Les résultats sont présentés en deux parties selon que l'ancienne VRC était négative ou positive. Le tableau présente la VRC améliorée, sa variance et le degré d'amélioration en prenant comme base la variance de l'ancienne VRC. Pour les sept premières catégories, tant l'ancienne VRC que la VRC améliorée sont négatives, et le réseau d'échantillons de la dernière colonne n'est pas évident. Néanmoins, on peut constater que dans 3 des 7 cas, une réduction importante a eu lieu et, dans 3 des 4 cas restants, une réduction substantielle a eu lieu. Pour les catégories dont les anciennes VRC sont positives, la réduction moyenne est de 15,4 % et pour les catégories individuelles, le taux de réduction varie de 0,3 % à 81,9 %. En conclusion, selon les catégories choisies et les données étudiées, il semble que des améliorations substantielles de la fiabilité des estimateurs peuvent être obtenues en ayant recours à la méthode améliorée.

Tableau 6
Estimations au moyen de la méthode améliorée

Catégorie	Variance de réponse corrélée (VRC) améliorée	Variance de la VRC améliorée	Diminution de la variance par rapport à l'ancienne VRC (%)
Plus haut grade obtenu			
Baccalauréat	-56 602	$2{,}498 \times 10^{11}$	43,1
Maîtrise	-165 990	$0{,}092 \times 10^{11}$	98,9
Certificat non universitaire	-1 176 987	$2{,}348 \times 10^{11}$	3,1
Situation d'activité : chômeur	-1 494 059	$2{,}162 \times 10^{11}$	47,9
Mise à pied temporaire	-93 651	$0{,}002 \times 10^{13}$	99,9
En attente pour commencer un nouvel emploi	-39 002	$0{,}017 \times 10^{11}$	99,8
A cherché un emploi	-1 358 562	$1{,}178 \times 10^{11}$	55,0
Plus haut grade obtenu : certificat d'études secondaires	2 721 694	$1{,}042 \times 10^{12}$	4,2
Dernière année d'études terminée			
N'ayant pas atteint la 5e année	1 145 589	$1{,}918 \times 10^{12}$	45,7
5e à 8e année	1 126 646	$3{,}292 \times 10^{12}$	9,1
9e et 10e année	1 296 708	$1{,}761 \times 10^{12}$	4,0
Fréquentation scolaire			
Ne fréquentant pas l'école	4 219 613	$1{,}010 \times 10^{13}$	0,3
Fréquentant l'école à temps plein	508 604	$5{,}087 \times 10^{11}$	10,6
Situation d'activité			
Personne occupée	3 699 256	$6{,}937 \times 10^{12}$	0,9
Personne inactive	3 978 183	$1{,}288 \times 10^{13}$	5,1
Mobilité (au cours des cinq dernières années)			
Personne n'ayant pas déménagé	2 476 082	$4{,}110 \times 10^{13}$	4,6
Personne ayant déménagé	9 917 528	$1{,}728 \times 10^{12}$	81,9
Non-migrant[a]	2 271 132	$7{,}176 \times 10^{12}$	2,7

[a] Un non-migrant est une personne qui vit dans la même municipalité qu'il y a cinq ans.

7 Conclusion

On a démontré la façon dont un estimateur amélioré peut être obtenu en tant que combinaison pondérée de l’ancien et du nouvel estimateur. L’estimateur est amélioré dans la mesure où sa variance est inférieure à celle de l’un des deux estimateurs originaux. Pour certaines variables, la réduction de la variance et l’augmentation consécutive de la stabilité sont assez substantielles. Il n’y a pas de coûts supplémentaires sur le terrain pour déterminer l’estimateur amélioré. Il convient toutefois de souligner que l’estimateur amélioré nécessite que les SD non superposés (qu’il s’agisse d’un échantillon ou de la population entière) soient jumelés. D’après notre expérience, ce jumelage est un exercice long et fastidieux, surtout si l’on insiste sur les mêmes critères de jumelage que ceux qui ont été imposés aux SD superposés. En résumé, les coûts supplémentaires de la méthode améliorée sont minimes, et la stabilité de l’estimateur amélioré est considérablement plus élevée.

L’importance de l’augmentation de la stabilité ne doit pas être sous-estimée. L’un des problèmes que posent le calcul et la diffusion des estimations de la variance des réponses corrélées est que les estimations sont petites et ont une grande variance. Cela pose le problème de déterminer la façon de présenter aux utilisateurs des données de recensement les estimations de la variance des réponses corrélées d’une manière significative. Les avancées qui permettent de produire des estimateurs ayant une meilleure stabilité devraient se poursuivre.

L’étude des valeurs aberrantes dans le présent article était l’une de ces tentatives pour obtenir un estimateur plus stable. Le fait que la situation a très peu changé lorsque les valeurs aberrantes ont été supprimées est décevant, car, *a priori*, les valeurs aberrantes semblaient être une cause importante d’instabilité des estimations. Des travaux de recherche futurs pourraient permettre d’étudier ce problème de manière plus approfondie. Il a été suggéré que l’une des sources d’instabilité était la taille variable des SD. Pour compenser cela, on pourrait élaborer un nouveau modèle dans lequel la moyenne d’un SD, plutôt que le nombre total, serait utilisée comme unité d’analyse de base. Cette moyenne pourrait utiliser comme base le nombre de personnes ou de ménages dans un SD. Un certain développement théorique des formules est nécessaire dans ce cas-ci puisque l’unité de base de l’observation

deviendrait le rapport de deux variables aléatoires. Le modèle actuel a une complexité qui n'est pas prise en compte dans les estimateurs. Cette complexité fait référence au plan d'échantillonnage qui n'est pas un échantillon aléatoire simple, mais suppose une stratification et, plus important encore, un regroupement des SD dans les DCR. Les travaux à venir pourraient envisager d'intégrer ces complexités dans le calcul de la VRC. Enfin, cette étude empirique devrait être reproduite sur d'autres ensembles de données pour diverses variables afin de fournir une base plus solide pour l'évaluation de la méthode améliorée.

Bibliographie

[1] Bailey, L., Moore, T.F., et Bailar, B.A., "An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample", *JASA*, Vol. 73, 1978, pp. 16-23.

[2] Brackstone, G.J. et Hill, C.J., "The Estimation of Total Variance in the 1976 Census", *Techniques d'enquête*, Vol. 2, No. 2, 1976, pp. 195-208.

[3] Fellegi, I.P., "Response Variance and Its Estimation", *JASA*, Vol. 59, 1964, pp. 1016-1041.

[4] Fellegi, I.P., "An Improved Method of Estimating Correlated Response Variance", *JASA*, Vol. 69, 1974, pp. 496-501.

[5] Krótki, K.P. et Hill, C.J., "Estimation of Correlated Response Variance", ASA, *Proceedings of the Section on Survey Research Methods*, 1978, pp. 609-14.

[6] Krótki, I.P., "1976 Census Parametric Evaluation: Measurement of Total Variance for the 1976 Census - Methodology and Results Report", rapport final no. 18, Census Survey Methods Division, Statistique Canada, Novembre 1978.

[7] U.S. Bureau of the Census. Evaluation and Research Program of the U.S. Censuses of Population and Housing (1960): Effects of Interviewers and Crew Leaders. Series ER 60, numéro 7, Washington, D.C., 1968.