Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

CONSTRUCTION DE CELLULES D'IMPUTATION POUR L'ENQUÊTE SUR LA POPULATION ACTIVE DU CANADA

D. Haziza, C. Charbonnier, O. S. Y. Chow, J. F. Beaumont[1]

RÉSUMÉ

En cas d'enquête à grande échelle, il est quasiment certain d'observer un certain niveau de non-réponse. En général, les organismes statistiques recourent à l'imputation pour corriger la non-réponse partielle. Une étape préalable courante est la création de cellules d'imputation. Dans le présent article, nous étudions la création de ces cellules par deux méthodes. La première s'inspire de celle utilisée par Eltinge et Yansaneh (1997) pour créer des cellules de pondération et la deuxième est la méthode appliquée à l'heure actuelle à l'Enquête sur la population active du Canada. À l'aide des données sur la population active, nous testons par simulation l'effet du taux de réponse, du mécanisme de réponse et des contraintes sur de la qualité de l'estimateur ponctuel dans le cas des deux méthodes.

MOTS CLÉS : Cellules d'imputation; imputation hot-deck; mécanisme de réponse uniforme; mécanisme de réponse négligeable; mécanisme de réponse non négligeable.

1. INTRODUCTION

1.1 Définition du problème

Malgré les efforts déployés par le personnel d'enquêtes en vue de maximiser le taux de réponse, il est quasiment certain d'observer un certain niveau de non-réponse lors des enquêtes à grande échelle. Essentiellement, les statisticiens d'enquête font la distinction entre deux catégories de non-réponse, à savoir la non-réponse totale ou unitaire (lorsqu'aucun renseignement n'est recueilli au sujet d'une unité échantillonnée) et la non-réponse partielle ou non-réponse à une question (lorsque l'absence d'information est limitée à certaines variables). En général, on applique des méthodes de correction par pondération pour tenir compte de la non-réponse totale, mais des méthodes par imputation pour compenser la non-réponse partielle. La correction par pondération vise essentiellement à augmenter les poids de sondage appliqués aux répondants afin de compenser pour les non-répondants, tandis que l'imputation est un processus qui consiste à produire une « valeur artificielle » pour remplacer une valeur manquante. Habituellement, tant pour la pondération que pour l'imputation, on commence par classer les répondants et les non-répondants dans des cellules formées en se basant sur les renseignements enregistrés pour la totalité des unités de l'échantillon. Deux raisons au moins justifient la création de cellules à la place d'une imputation directe de la valeur résultant de l'utilisation d'un modèle de régression : 1) il s'agit d'une méthode pratique lorsqu'il faut imputer des valeurs pour plus d'une variable à la fois et 2) la méthode est plus robuste en cas de spécification erronée du modèle.

[1] David Haziza, Cédric Charbonnier, Ophelia Chow et Jean-François Beaumont, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

En pratique, plusieurs méthodes sont utilisées pour créer les cellules d'imputation. Les méthodes courantes sont les suivantes : 1) pour les échantillons stratifiés, on peut utiliser des strates ou des groupes de strates comme cellules d'imputation, ce qui se fait couramment dans le cas des enquêtes-entreprises; 2) si les paramètres étudiés sont des moyennes ou des totaux par domaine et que l'on connaît les domaines étudiés au stade de la planification, on peut utiliser ces domaines comme cellules d'imputation; 3) les cellules d'imputation peuvent aussi être créées par combinaisons de simples variables auxiliaires telles que la province, l'âge, le sexe, etc.; 4) des logiciels à arbre de décision tels que Knowledge Seeker (Kass, 1980) et CART (Steinberg et Colla, 1995) peuvent aussi être utilisés pour former des cellules d'imputation homogènes.

L'objectif du présent article est d'examiner la formation de cellules d'imputation selon deux méthodes. La première (*méthode1*) est similaire à celle proposée par Eltinge et Yansaneh (1997) dans le cas des cellules de pondération. La deuxième (*méthode2*) est celle appliquée à l'heure actuelle à l'Enquête sur la population active (EPA) du Canada. Dans tout l'article, nous utilisons l'imputation par la méthode hot-deck pour compenser la non-réponse à une question *y*. Autrement dit, nous remplaçons une valeur manquante pour la question *y* par la valeur de *y* obtenue pour une unité sélectionnée au hasard à partir de l'ensemble de personnes qui ont répondu à cette question.

1.2 Bref aperçu de l'Enquête sur la population active

Au Canada, l'enquête sur la population active (EPA) est une enquête par panel mensuelle qui fournit des renseignements détaillés sur les caractéristiques de l'emploi, telles que le chômage (aux niveaux provincial et national), le nombre d'heures travaillées, la description des tâches et la rémunération. Elle fournit aussi certains renseignements sur les caractéristiques de la population en âge de travailler, telles que l'âge, l'état matrimonial et le niveau de scolarité. L'échantillon de l'EPA compte environ 53 000 ménages (environ 130 000 répondants) du Canada. Chaque ménage fait partie de l'échantillon de l'enquête pendant six mois consécutifs et, chaque mois, un sixième de l'échantillon est renouvelé. Dans chaque ménage, chaque membre de 15 ans et plus est interviewé. Sont exclues de l'enquête les personnes qui vivent dans les réserves indiennes, les membres à temps plein des Forces canadiennes et les détenus des établissements carcéraux. Ces exclusions ne représentent que 2 % de la population de 15 ans et plus du Canada.

1.3 Données de simulation

Nous avons réalisé deux études par simulation pour évaluer la performance des deux méthodes. À partir des fichiers de l'EPA, nous avons sélectionné des personnes auxquelles on avait posé la question sur la rémunération en janvier 2001. Nous avons obtenu ainsi un ensemble de données couvrant 11 270 personnes et plus de 100 variables.

Puisque les questions sur le revenu sont délicates, le taux de réponse observé est souvent inférieur à celui enregistré pour d'autres variables. C'est pourquoi nous avons choisi d'étudier la variable *rémunération hebdomadaire*. Pour la simulation, nous avons considéré cet ensemble de données comme étant la population à partir de laquelle était générée la non-réponse. La moyenne de population « réelle » de la *rémunération hebdomadaire* calculée d'après le fichier complet était égale à 554,97 $. Nous avons comparé cette valeur « réelle » à l'estimateur imputé pour trois types de mécanisme de réponse : uniforme, négligeable et non négligeable (voir la section 2).

2. DÉFINITIONS ET RÉSULTATS THÉORIQUES

Soit U une population de taille N. L'objectif est d'estimer la moyenne de la population $\bar{Y} = \frac{1}{N}\sum_U y_i$ d'une variable étudiée *y*. Supposons que nous sélectionnions un échantillon s de taille *n* à partir de U conformément à un plan de sondage $p(s)$. À cause de la non-réponse à la question *y*, nous

observons uniquement le sous-ensemble s_r de s. Nous supposons que le mécanisme de réponse $q(s_r \mid s)$ est indépendant de s, autrement dit que la probabilité conditionnelle, $q(s_r \mid s)$, d'observer s_r est égale à $q(s_r)$. Représentons par a_i l'indicateur de réponse qui est égal à 1 si l'unité échantillonnée sélectionnée i est un répondant à la question y et égal à 0 autrement. Supposons que les a_i sont des variables aléatoires de Bernoulli indépendantes (p_i).

Kalton et Kasprzyk (1986) ont montré que plusieurs méthodes d'imputation (imputation par la moyenne, imputation hot-deck, imputation par quotient,...) peuvent être représentées en se servant du modèle suivant

$$m: y_i = \mathbf{z}_\mathbf{i}' \boldsymbol{\beta} + \varepsilon_\mathrm{i}, \tag{1}$$

où $\mathbf{z}$ est un vecteur de variables auxiliaires disponible pour toutes les unités comprises dans l'échantillon et $E_m(\varepsilon_i) = 0,\ E_m(\varepsilon_i\ \varepsilon_j) = 0,\ i \neq j, V_m(\varepsilon_i) = \sigma_i^2$. Nous supposons que $\sigma_i^2 = \boldsymbol{\lambda}'\mathbf{z}_\mathrm{i}$ pour un vecteur constant $\boldsymbol{\lambda}$. Notons que cette hypothèse limite peu la gamme de modèles d'imputation.

Un estimateur à données imputées de $\bar{Y}$, représenté par $\bar{y}_I$, est donné par

$$\bar{y}_I = \frac{1}{\sum_s w_i} \sum_s w_i \tilde{y}_i \tag{2}$$

où w_i est le poids de sondage lié à l'unité i, $\tilde{y}_i = \begin{cases} y_i & \text{si l'unité échantillonnée } i \text{ est un répondant} \\ y_i^* & \text{si l'unité échantillonnée } i \text{ est un non - répondant} \end{cases}$

et y_i^* représente la valeur imputée pour remplacer la valeur manquante y_i. Nous pouvons utiliser le modèle (1) pour justifier la forme de la valeur imputée pour l'unité i, qui est donnée par $y_i^* = \mathbf{z}_i'\hat{\mathbf{B}}_\mathbf{r}$, où $\hat{\mathbf{B}}_\mathbf{r} = \left(\sum_s w_i a_i \mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/\sigma_i^2\right)^{-1} \sum_s w_i a_i \mathbf{z}_\mathbf{i} y_i/\sigma_i^2$. En nous servant des y_i^*, nous pouvons représenter le biais asymptotique dans les conditions du plan d'échantillonnage et du mécanisme de réponse par

$$\text{Biais}(\bar{y}_I) \approx \frac{1}{N} \sum_U \mathbf{z}_\mathbf{i}'(\mathbf{B}_p - \mathbf{B}), \tag{3}$$

où $\mathbf{B}_p = \left(\sum_U p_i \mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/\sigma_i^2\right)^{-1} \sum_U p_i \mathbf{z}_\mathbf{i} y_i/\sigma_i^2$ et $\mathbf{B} = \left(\sum_U \mathbf{z}_\mathbf{i}\mathbf{z}_\mathbf{i}'/\sigma_i^2\right)^{-1} \sum_U \mathbf{z}_\mathbf{i} y_i/\sigma_i^2$. Notons que $\mathbf{B}$ est un estimateur du paramètre du modèle $\boldsymbol{\beta}$, autrement dit $\mathbf{B}$ est une estimation que nous aurions obtenue si nous observions la population finie entière. Notons aussi que, d'après (3), l'estimateur à données imputées (2) est approximativement non biaisé si $\mathbf{B}_p - \mathbf{B} \approx 0$. Dans les conditions du modèle (1), $\mathbf{B}_p - \mathbf{B}$ converge vers zéro selon le modèle si, selon le modèle, la covariance de la probabilité de réponse p_i et du résidu du modèle $(y_i - \mathbf{z}_\mathbf{i}'\boldsymbol{\beta})$ est nulle pour toute unité i de la population.
En cas d'imputation par la moyenne ou par la méthode hot-deck, (3) se réduit à

$$\text{Biais}(\bar{y}_I) \approx \frac{1}{N\bar{P}} \sum_U (p_i - \bar{P})(y_i - \bar{Y}), \tag{4}$$

où $\bar{P} = \frac{1}{N}\sum_U p_i$. En vertu de (4), nous notons que l'estimateur à données imputées (2) est approximativement non biaisé si, selon le modèle, la covariance de la probabilité de réponse p_i et de la

moyenne de population $\overline{Y}$ est nulle pour toute unité *i* de la population ou que, pour la population, la covariance des variables p et y est nulle. Cette exigence est manifestement irréaliste en pratique, puisque la probabilité de réponse dépend probablement de variables potentiellement corrélées avec la variable étudiée *y*. Par conséquent nous disons qu'$\overline{y}_I$ est un « estimateur non corrigé ». Afin de réduire le biais, on procède souvent à une division de la population en *C* cellules d'imputation U_c, $\left(\bigcup_{c=1}^{C} U_c = U\right)$ qui entraîne une subdivision correspondante de l'échantillon *s* en cellules s_c, $\left(\bigcup_{c=1}^{C} s_c = s\right)$. Puis, on procède indépendamment à l'imputation dans chaque cellule et on obtient un estimateur « corrigé » représenté par

$$\overline{y}_{I,c} = \sum_{c=1}^{C} w_c' \overline{y}_c \tag{5}$$

où $w_c' = \dfrac{\sum_{s_c} w_i}{\sum_{s} w_i}$ est une mesure de l'importance relative de la cellule *c* et $\overline{y}_c = \dfrac{1}{\sum_{s_c} w_i} \sum_{s_c} w_i \tilde{y}_i$ est l'estimateur à données imputées pour la cellule *c*, *c* = 1,…, *C.* En cas d'imputation par la moyenne ou par la méthode hot-deck, le biais de l'estimateur corrigé est donné par

$$\text{Biais}(\overline{y}_{I,c}) \approx \frac{1}{N} \sum_{c=1}^{C} \overline{P}_c^{-1} \sum_{U_c} (p_i - \overline{P}_c)(y_i - \overline{Y}_c), \tag{6}$$

où $\overline{P}_c = \dfrac{1}{N_c} \sum_{U_c} p_i$, $\overline{Y}_c = \dfrac{1}{N_c} \sum_{U_c} y_i$ et N_c est le nombre d'unités dans U_c. Il s'ensuit que, dans (6), le biais est approximativement nul si la covariance des variables p et y dans la population est approximativement nulle pour chaque cellule. En pratique, on s'efforce de réaliser cette condition en construisant des cellules homogènes en ce qui concerne la probabilité de réponse p_i ou la question y_i, ou les deux.

En général, les statisticiens d'enquête décrivent trois catégories de mécanismes de réponse : uniforme, négligeable et non négligeable. Un mécanisme de réponse uniforme est un mécanisme pour lequel $p_i = P(i \in s_r) = p \quad \forall i \in U$. De façon informelle, un mécanisme de réponse est dit négligeable en ce qui concerne un modèle s'il dépend de certaines variables auxiliaires du modèle, mais non de l'erreur du modèle. Par exemple, un mécanisme pour lequel la probabilité de réponse dépend de certaines variables auxiliaires, mais non de la variable étudiée, c'est-à-dire $p_i = P(i \in s_r) = p(\mathbf{z}_i)$ est négligeable en ce qui concerne le modèle (1). Autrement dit, si le mécanisme de réponse est indépendant de la loi de distribution du modèle, alors il est négligeable en ce qui concerne ce modèle. Un mécanisme de réponse non négligeable est un mécanisme qui ne peut être ignoré en ce qui concerne un modèle particulier. Par exemple, si la probabilité de réponse dépend de la variable étudiée, c'est-à-dire $p_i = P(i \in s_r) = p(\mathbf{z}_i, y_i)$, alors le mécanisme de réponse n'est pas négligeable en ce qui concerne le modèle (1). Notons qu'en cas de mécanisme de réponse uniforme, $p_i = p$ et (4) est nul. Par conséquent, si l'on pense que le mécanisme de réponse est uniforme pour l'ensemble de la population, il est inutile de créer des cellules d'imputation pour réduire le biais. Cependant, il pourrait être utile de le faire afin de réduire la variance de l'estimateur à valeurs imputées. Notons aussi qu'en cas de réponse non négligeable, l'équation (5) ne sera généralement pas nulle, puisque, dans ce cas, la covariance de p et y dans la population n'est vraisemblablement pas nulle.

3. MÉTHODE PROPOSÉE

3.1 Description de la méthode

À la présente section, nous commençons par décrire la *méthode1* proposée pour créer des cellules d'imputation. Cette méthode est semblable à celle utilisée par Eltinge et Yansaneh (1997) dans le cas de cellules de pondération. Comme nous l'avons mentionné à la section 2, l'objectif est de diviser l'échantillon de telle façon que, dans les cellules, les unités (répondants et non-répondants) soient homogènes en ce qui concerne l'un des deux critères, à savoir les probabilités de réponse p_i ou les valeurs de la question y_i. Les étapes de la création des cellules peuvent alors être décrites comme suit.

1. En se servant de données auxiliaires disponibles pour toutes les unités échantillonnées, prédire les probabilités de réponse p_i et les valeurs de la variable étudiée y_i. On dispose alors de deux scores, $\hat{p}_i$ et $\hat{y}_i$, pour toutes les unités de l'échantillon (répondants et non-répondants).
2. Choisir l'un des deux critères ($\hat{p}_i$ ou $\hat{y}_i$) et, au moyen d'une méthode à quantiles égaux ou d'un algorithme de classification, diviser l'échantillon en cellules. Dans le présent article, nous utilisons une méthode à quantiles égaux pour diviser l'échantillon, ce qui signifie que les bornes des cellules sont déterminées par les quantiles j/k des populations $\hat{p}_i$ ou $\hat{y}_i$ pour $j = 1,...,k-1$. À l'heure actuelle, nous étudions l'utilisation d'algorithmes de classification.
3. Dans chaque cellule, procéder à une imputation hot-deck et calculer la valeur de l'estimateur à données imputées intracellule $\bar{y}_c$.
4. Combiner ces estimateurs comme dans l'équation (5).

Notons que, dans le cas de la *méthode1*, les cellules sont disjointes, si bien que l'on ne peut utiliser un donneur que dans la cellule à laquelle il appartient.

3.1.1 Modélisation des probabilités de réponse et de la variable *rémunération hebdomadaire*

Lors de la construction de cellules d'imputation, la première étape consiste à prédire les probabilités de réponse p_i et les valeurs de la variable étudiée y_i. Le lecteur est invité à consulter Chow (2001) pour une description détaillée de la construction des modèles.

Puisque $p_i = P(a_i = 1)$ et que a_i est une variable binaire, nous avons prédit les valeurs de p_i par régression logistique. Nous avons considéré les variables qui suivent comme étant des prédicteurs importants :
province, âge, statut d'étudiant, classification type des industries, année où la personne a commencé à travailler, classification type des professions, état matrimonial, horaire fixe/variable, propriétaire/locataire du logement, catégorie de logement, emploi rémunéré par pourboires/commissions, taille du ménage, catégorie de travailleurs, sexe et *situation d'activité.*

Nous avons utilisé un modèle de régression linéaire pour prédire les valeurs de la variable continue étudiée *rémunération hebdomadaire*. Nous avons considéré les variables qui suivent comme étant des prédicteurs importants :
situation d'emploi à temps plein/temps partiel, classification type des professions, âge, classification type des industries, année où la personne a commencé à travailler, niveau de scolarité, statut d'étudiant, état matrimonial, sexe, taille de l'entreprise, situation de l'emploi permanent/temporaire et *situation syndicale*.

3.1.2 Création des cellules

Pour illustrer la *méthode1*, nous avons généré la non-réponse conformément au mécanisme de réponse *négligeable1* décrit à l'annexe. Nous avons fixé le taux de réponse global à 0,7 (valeur légèrement inférieure à la valeur la plus extrême observée dans le cas de l'EPA pour la variable *rémunération*

hebdomadaire). En nous servant des prédicteurs choisis (voir la section 3.1.1), nous avons obtenu deux scores ($\hat{p}_i$ et $\hat{y}_i$) pour chaque unité de l'échantillon. Puis, au moyen d'une méthode à quantiles égaux, nous avons divisé l'échantillon en cellules de tailles égales.

Le tableau 1 montre les estimations obtenues après une imputation aléatoire par la méthode hot-deck dans chaque cellule, ainsi que les différences $\bar{y}_{I,c} - \bar{y}_{I,1}$ mesurant l'effet sur l'estimation de l'utilisation de c cellules plutôt que 1 cellule et les différences $\bar{y}_{I,c} - \bar{y}_{I,c-1}$ mesurant l'effet sur l'estimation de l'utilisation d'une cellule supplémentaire. À mesure qu'augmente le nombre de cellules, les estimations s'approchent rapidement de la moyenne de population réelle. Lorsque l'on utilise 5 cellules d'imputation, l'erreur relative estimée n'est plus que de 0,91 % et l'utilisation d'un nombre plus grand de cellules ne produit qu'une très faible amélioration. Si nous examinons la différence $\bar{y}_{I,c} - \bar{y}_{I,1}$, qui mesure l'effet de l'utilisation de c cellules plutôt que 1 cellule, nous constatons de nouveau qu'après la création de 5 cellules d'imputation, l'effet de l'utilisation de 6 cellules ou plus se stabilise autour de 70. La différence entre l'utilisation de c cellules et c-1 cellules montre aussi une diminution rapide de l'avantage lié à l'utilisation d'une cellule d'imputation supplémentaire.

Tableau 1 : Estimations de la *rémunération hebdomadaire*

Nombre de cellules	Estimation	Différence ($\bar{y}_{I,c} - \bar{y}_{I,1}$)	Différence ($\bar{y}_{I,c} - \bar{y}_{I,c-1}$)	Erreur relative
1	627,35			0,1304
2	581,85	-45,49	-45,49	0,0484
3	568,64	-58,71	-13,21	0,0246
4	562,71	-64,64	-5,92	0,0140
5	560,03	-67,32	-2,68	0,0091
6	558,43	-68,92	-1,59	0,0062
7	557,83	-69,52	-0,59	0,0052
8	557,10	-70,25	-0,73	0,0038
9	556,47	-70,87	-0,62	0,0027
10	555,85	-71,50	-0,62	0,0016

3.1.3 Diagnostics à l'intérieur des cellules

Une fois les cellules créées, il peut être utile de procéder à certains tests diagnostiques. L'un de ces tests pourrait être l'évaluation du biais intracellule. Considérons les estimateurs suivants

$$\bar{y}_{w-c}^{(1)} = \frac{1}{\sum_{s_c} w_i \hat{p}_i^{-1}} \sum_{s_c} w_i \hat{p}_i^{-1} y_i \tag{7}$$

et

$$\bar{y}_{w-c}^{(2)} = \frac{1}{\sum_{s_c} w_i} \sum_{s_c} w_i \hat{y}_i \tag{8}$$

Il est facile de constater que l'estimateur intracellule (7) est approximativement non biaisé dans les conditions du plan d'échantillonnage et du mécanisme de réponse pour $\bar{Y}_c$, à condition que $\hat{p}_i = p_i$ pour toutes unités $i \in s_c$. Pareillement, l'estimateur intracellule (8) est approximativement non biaisé dans les conditions de conception du modèle pour $\bar{Y}_c$ à condition que $E_m(\hat{y}_i) = E_m(y_i)$, où $E_m(.)$ représente l'opérateur d'espérance en ce qui concerne le modèle donné en (1). Nous pouvons utiliser ces deux estimateurs pour mesurer le biais à l'intérieur des cellules $\hat{B}_c = \bar{y}_c - \bar{y}_{w-c}^{(j)}$ $j = 1,2$. Des valeurs élevées

de $\hat{B}_c$ pourraient être indicatrices d'un biais intracellule important dans $\bar{y}_c$ mais, comme l'ont fait remarquer Eltinge et Yansaneh (1997), elles pourraient aussi indiquer que les erreurs $\hat{p}_i - p_i$ (ou $\hat{y}_i - y_i$) sont grandes. Donc, il convient d'interpréter avec prudence les valeurs élevées de $\hat{B}_c$. Toutefois, si l'on considère les diagnostics (7) et (8) simultanément, une grande valeur de $\hat{B}_c$ uniquement dans la probabilité de réponse p_i ou la variable étudiée y_i est moins indicatrice d'un biais que si les deux valeurs sont élevées. Indépendamment de la cause, il convient d'examiner plus en profondeur toute valeur élevée de $\hat{B}_c$.

Le tableau 2 montre les estimations intracellule, ainsi que le biais intracellule lors de la création de une à quatre cellules d'imputation. Notons que nous n'avons utilisé que l'estimateur (7) à des fins diagnostiques. À mesure qu'augmente le nombre de cellules, les données deviennent plus homogènes et, par conséquent, le biais intracellule diminue. Nous observons au tableau 2 que, dans chaque cas, le biais dans la cellule est plus important pour la première cellule que pour les autres, peut-être à cause de problèmes de modélisation de la probabilité de réponse. L'utilisation d'une méthode plus perfectionnée pour créer les cellules, comme des algorithmes de mise en grappes, pourrait aussi permettre de réduire le biais intracellule observé pour la première cellule.

Tableau 2 : Estimations et biais intracellule

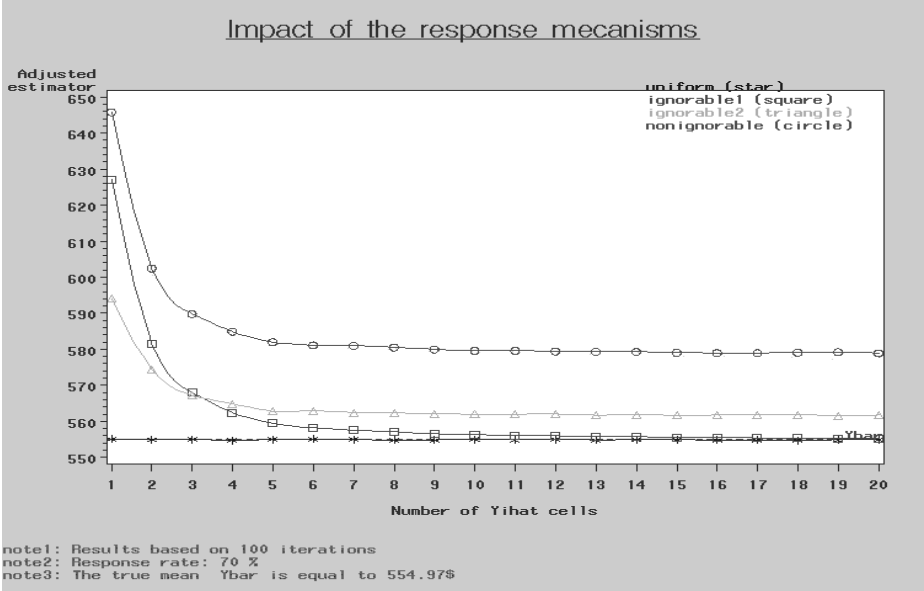
Nombre de cellules		$\bar{y}_c$	$\bar{y}_{w-c}$	$\hat{B}_c$
1	1	627,35	569,04	58,31
2	2,1	355,33	326,61	28,72
	2,2	808,38	795,07	13,30
3	3,1	254,62	235,51	19,10
	3,2	551,79	545,61	6,18
	3,3	899,34	890,38	8,95
4	4,1	199,17	184,05	15,18
	4,2	451,06	446,30	4,76
	4,3	646,84	643,19	3,64
	4,4	953,49	947,23	6,26

3.2 Résultats des simulations

Pour évaluer la *méthode1*, nous avons réalisé une étude par simulation en utilisant la population décrite à la section 1.3. Nous avons généré un nombre $R = 100$ de sous-ensembles de non-réponses conformément au mécanisme de réponse décrit à l'annexe. Pour chaque itération, nous avons fixé le taux de réponse à 70 %. Puis, nous avons prédit les probabilités de réponse $\hat{p}_i$ et les valeurs de la variable $\hat{y}_i$, et créé des cellules par la méthode des quantiles égaux. Dans chaque cellule, nous avons procédé à une imputation aléatoire par la méthode hot-deck et calculé les estimations de la moyenne de la population pour la *rémunération hebdomadaire*.

Rappelons que la moyenne réelle de population était égale à 554,97 $. En cas de réponse uniforme, l'estimateur est approximativement non biaisé quel que soit le nombre de cellules, comme il l'était prévu (voir le graphique 1). Dans les conditions des mécanismes de réponse négligeables (*négligeable1* et *négligeable2*), nous constatons qu'à mesure que le nombre de cellules augmente, le biais tend vers zéro. Dans le cas d'un mécanisme de réponse non négligeable, nous voyons que, même si l'utilisation de cellules d'imputation réduit considérablement le biais, l'estimateur corrigé demeure biaisé, comme nous nous y attendions (voir la section 2). De nouveau, notons que la plupart de la réduction du biais est réalisée à l'aide d'un assez petit nombre de cellules (5 à 10). Après cela, le biais reste assez stable et l'augmentation du nombre de cellules n'apporte aucune autre amélioration. Ce résultat qui, à première vue, semble étonnant, a également été observé par Eltinge et Yansaneh (1997) dans le cas des cellules de pondération. Nous avons obtenu des résultats comparables lorsque nous avons utilisé les probabilités de réponse comme critères pour la construction des cellules, ainsi qu'un taux de réponse de 0,5.

Graphique 1



4. MÉTHODE COURANTE

4.1 Description de la méthode

À la présente section, nous décrivons brièvement la méthode courante (*méthode2*) utilisée dans le cas de l'EPA pour créer des cellules d'imputation. Pour une description plus détaillée de la méthode, le lecteur est invité à consulter Lorenz (1996). Ici, les cellules sont formées par combinaisons de variables nominales. En premier lieu, on sélectionne les variables qui sont corrélées à la variable étudiée *rémunération hebdomadaire*, puis on les classifie par ordre d'importance, de la plus significative à la moins significative. Dans le cas de nos données, les variables classées par importance décroissante sont : *situation de travail à temps plein/temps partiel*, *classification type des professions*, *âge*, *classification type des industries*, *année où la personne a commencé à travailler*, *niveau de scolarité*, *situation d'étudiant*, *état matrimonial*, *sexe*, *taille de l'entreprise*, *situation d'emploi permanent/temporaire* et *situation syndicale*.

Pour s'assurer de la stabilité de l'estimateur ponctuel, on impose les deux contraintes suivantes :

1. le nombre minimal d'enregistrements donneurs dans une cellule donnée est fixé à *k* ;
2. dans une cellule donnée, le nombre d'enregistrements donneurs doit être supérieur au nombre d'enregistrements receveurs.

Les cellules d'imputation sont définies par les différentes combinaisons des variables énumérées plus haut. Puis, on procède à une imputation aléatoire par la méthode hot-deck dans toutes les cellules qui satisfont les contraintes susmentionnées. Cependant, puisque le nombre initial de cellules est très grand, il est vraisemblable qu'un nombre important de celles-ci ne satisferont pas les contraintes. Le cas échéant, la variable la moins significative (*situation syndicale*) est supprimée de la liste et les cellules sont de nouveau définies en combinant les variables restantes. De nouveau, on procède à l'imputation aléatoire par le méthode hot-deck dans toutes les cellules qui satisfont les contraintes. Si certaines cellules ne satisfont toujours pas ces contraintes, la variable la moins significative (*emploi permanent/temporaire)* est supprimée de la liste, et ainsi de suite. Ce processus se poursuit jusqu'à ce que chaque enregistrement receveur ait trouvé un donneur ou que l'on ait éliminé le nombre maximal permis de variables. Dans le cas

qui nous intéresse, aucun regroupement de cellules n'est permis au-delà de la variable *classification type des professions*. Notons que, contrairement à la *méthode1*, les cellules ne sont pas disjointes; par conséquent, un enregistrement donneur peut être utilisé plusieurs fois à diverses étapes du processus.

4.2 Étude par simulation

Nous avons réalisé une étude par simulation pour évaluer l'effet des mécanismes de non-réponse et des contraintes sur la qualité (le biais) de l'estimateur ponctuel de la moyenne de la population. Nous avons généré un nombre $R = 300$ de sous-ensembles de non-réponses conformément aux mécanismes de réponse décrits à l'annexe. Lors de chaque itération, les cellules d'imputation ont été créées conformément à la méthode décrite à la section 4.1 Nous avons fixé le taux global de réponse à 70 %.

Les tableaux 3, 4 et 5 donnent le biais relatif de l'estimateur ponctuel où le nombre minimal de donneurs k dans chaque cellule a été fixé à 3, 8 et 1, respectivement. Il est manifeste, lorsque l'on examine les résultats, que dans le cas de la réponse uniforme, l'estimation ponctuelle est approximativement non biaisée (moins de 0,5 %) quelle que soit la valeur de k. Supposer que la réponse est uniforme dans toute la population est une hypothèse forte, si bien que l'inférence sera valide quel que soit le modèle utilisé pour créer les cellules. Pour les mécanismes de réponse *négligeable1* et *négligeable2*, nous constatons que, à mesure que la rigueur des contraintes augmente (c.-à-d. que le nombre minimal de donneurs dans chaque cellule augmente), le biais relatif de l'estimateur augmente; p. ex., pour $k = 1$ et le mécanisme *négligeable1*, le biais relatif de l'estimateur est égal à 0,6 % tandis que, pour $k = 8$, il est égal à 9,02 %. En cas de mécanisme de réponse non négligeable, l'estimateur ponctuel est biaisé même si $k = 1$ (4,52 %), comme nous l'avions prévu, mais la valeur du biais augmente parallèlement à celle de k (11,46 % pour $k = 8$). Cette situation pourrait tenir au fait que, à mesure qu'augmente la rigueur des contraintes, il devient de plus en plus difficile d'avoir des enregistrements donneurs dans une cellule, si bien que davantage de variables seront éliminées de la liste. Conséquemment, certains prédicteurs importants pourraient être omis, ce qui risque de produire des estimateurs biaisés.

Pour conclure, il semble que, pour toutes les catégories de mécanismes de réponse (sauf la réponse uniforme), la qualité de l'estimateur diminue à mesure que les contraintes deviennent plus sévères. Il est intéressant de noter que la qualité de l'estimateur pourrait également être sensible à l'ordre des variables. En effet, si l'on ne classifie pas ces dernières convenablement, certains prédicteurs importants risquent d'être omis de la liste à une étape très précoce du processus, ce qui pourrait causer un biais considérable. En outre, la qualité de l'estimateur pourrait être sensible au taux de réponse, car, à mesure que ce taux diminue, la deuxième contrainte pourrait être plus difficile à satisfaire, ce qui, de nouveau donnera lieu à un biais important. Enfin, on devrait également examiner la taille de l'échantillon par rapport au nombre de cellules créées par la combinaison des variables sélectionnées. Si la taille de l'échantillon est considérablement plus petite que le nombre de cellules, l'effet des contraintes sur la qualité de l'estimation ponctuelle pourrait être plus important.

Tableau 3 Biais relatif de l'estimateur pour $k = 3$

Mécanisme	**Estimation**	**Biais relatif (%)**	**IC à 95 % pour le biais relatif**
Uniforme	552,64	-0,42	(-0,52; -0,32)
Négligeable 1	591,04	6,50	(6,10; 6,89)
Négligeable 2	561,28	1,14	(1,02; 1,25)
Non négligeable	600,04	8,12	(7,72; 8,52)

Tableau 4 Biais relatif de l'estimateur pour $k = 8$

Mécanisme	Estimation	Biais relatif (%)	IC à 95 % pour le biais relatif
Uniforme	553,58	-0,25	(-0,35;-0,15)
Négligeable 1	604,98	9,02	(8,70;9,31)
Négligeable 2	564,42	1,70	(1,55;1,85)
Non négligeable	618,57	11,46	(11,05;11,87)

Tableau 5 Biais relatif de l'estimateur pour $k = 1$

Mécanisme	Estimation	Biais relatif (%)	IC à 95 % pour le biais relatif
Uniforme	552,79	-0,39	(-0,49; -0,30)
Négligeable 1	558,27	0,60	(0,52; 0,67)
Négligeable 2	558,93	0,71	(0,64; 0,78)
Non négligeable	580,03	4,52	(4,45; 4,58)

BIBLIOGRAPHIE

Chow, O. S. Y., (2001), "Model Building for construction of cellules d'imputation in the Labour Force Survey", rapport interne, Ottawa, Canada: Statistique Canada.

Eltinge, J. L., et Yansaneh, I. S. (1997), "Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey", Techniques d'enquête, 23, pp. 37-45.

Kalton, G., et Kasprzyk, D. (1986), "Le traitement des données d'enquête manquantes", Techniques d'enquête, 12, pp. 1-17.

Kass, G. V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29, No 2, pp. 119-127.

Lorenz, P. (1996), "Head Office Hot Deck Imputation System Specifications", rapport interne, Ottawa, Canada: Statistique Canada.

Smith, P. J., Hoaglin, D. C., Battaglia, M. P., Rao, J. N. K., et Daniels, D. (2001), "Evaluation of Adjustment for Partial Nonresponse Bias, Applied to Provider nonresponse in the National Immunization Survey", présenté au Annual Meeting of the Statistical Society of Canada, Ottawa, Canada.

Steinberg, D., et Colla, P. (1995), "Tree-Structured Non-Parametric Data Analysis", San Diego, CA: Salford Systems.

ANNEXE

Suit la description des mécanismes de réponse.

Uniforme	p_i est constante, c.-à-d. $\forall i, p_i = p$
Négligeable1	p_i est une fonction de $\hat{y}_i$ telle que $p_i = \alpha + (1-\alpha)\exp(1-\lambda\hat{y}_i)$, $0 \le \alpha \le 1$
Négligeable2	p_i est une fonction de x_i (Âge) telle que $p_i = \alpha + (1-\alpha)\exp(1-\lambda x_i)$, $0 \le \alpha \le 1$
Non négligeable	p_i est une fonction de y_i (*rémunération hebdomadaire*), telle que $p_i = \alpha + (1-\alpha)\exp(1-\lambda y_i)$, $0 \le \alpha \le 1$

La valeur des constantes α et λ est fixée de façon à produire le taux de réponse désiré.