Note

« Un mode de pondération de données provenant d'enquêtes longitudinales »

Jacques A. Turcotte et Misa Gratton

Cahiers québécois de démographie, vol. 6, n° 2, 1977, p. 94-107.

Cahiers québécois de démographie
Vol. 6, no. 2, août 1977

TURCOTTE, Jacques A. et GRATTON, Misa: Un mode de pondération de données provenant d'enquêtes longitudinales

SOMMAIRE

Lors de l'étape du traitement de données provenant d'enquêtes longitudinales, le méthodologue désire vérifier si les distributions des estimations correspondent, soit dans le temps, soit à divers stades de la période d'enquête, à certaines distributions externes. A ce titre, les auteurs se réfèrent à des travaux méthodologiques et informatiques récents pour concevoir une nouvelle façon d'ajuster les estimations selon une séquence et un nombre d'itérations prescrits. Cette exploitation conjuguée de la technique de l'estimation par le quotient et du programme informatique THAID apporte, à toute fin, une possibilité inédite de valider le procédé de pondération.

Cahiers québécois de démographie
Vol. 6, no 2, août 1977.

UN MODE DE PONDÉRATION DE DONNÉES PROVENANT D'ENQUÊTES LONGITUDINALES

par

Jacques A. TURCOTTE* et Misa GRATTON**

Etape importante du traitement des données de toute enquête longitudinale, le procédé de pondération exige:

a) que l'on attache un poids à chaque unité d'échantillonnage au fichier, poids qui est déterminé en fonction de la nature de l'échantillon;

b) que l'on ajuste ces poids pour tenir compte des non-réponses lorsqu'il s'agit, notamment, d'enquête par questionnaire; et,

---

* Groupe des travaux de recherche, Emploi et Immigration, Canada, Ottawa, K1A 0J9.

** Division des systèmes généraux d'enquête, Statistique Canada, Ottawa, K1A 0T6.

Les auteurs remercient tous les collègues qui, à un moment ou l'autre, ont fait part de leurs commentaires. Les auteurs sont toutefois les seuls responsables de ce texte.

c) que l'on ajuste de nouveau les poids à partir d'information auxiliaire.

Puisque l'objectif des opérations de pondération est, en de tel cas, celui d'obtenir des estimations dont les distributions correspondent à certaines distributions externes, nous reverrons successivement le cas de l'ajustement des poids pour fin de non-réponse (t=0, dans la formule qui suit), et le cas de l'ajustement des poids en fonction d'information auxiliaire (t=4, dans cette même formule), selon une séquence et un nombre d'itérations prescrits.

UN APERÇU DU CALCUL DES ESTIMATIONS

Il est entendu que l'attribution du poids initial (item a, ci-haut) peut s'effectuer par le "Statistical Package for the Social Sciences" (Nie et al., 1975, 129-131) ou par un programme informatique répondant à cette fin, ex. le Système de prélèvement d'échantillon mis au point à Statistique Canada. Les opérations automatisées des items b et c ne peuvent toutefois être accomplies que par l'utilisation de programmes généraux, tel le Système d'estimation (Gratton, 1975; Satin, 1974).

Ce "système d'estimation" offre cinq options différentes. En plus des deux options dont nous faisons état, il y en a une pour le cas où l'on ne désire pas d'ajustement de poids, c'est-à-dire $A_1 \equiv 1$ pour tous les i, dans la formule qui suit; une deuxième option

pour le cas où l'on désire ajuster les poids en fonction des moyennes pondérées de tout un ensemble de variables; et, une troisième option pour le cas où l'on désire ajuster les poids en fonction d'une séquence donnée de variables. Dans tous les cas, l'estimation recherchée découle d'une généralisation de la formule:

$$\hat{Y}_p = \sum_{i \in p} (A_i^{(t)} W_i y_i)$$

où $\hat{Y}_p$: valeur estimée totale d'une caractéristique y pour une sous-population p;

$A_i^{(t)}$: facteur d'ajustement du poids pour l'unité d'échantillonnage i;

t : genre d'ajustement ou d'option (t=0,1,2,3,4);

W_i : le poids de l'unité d'échantillonnage i;

y_i : la valeur y rapportée pour l'unité d'échantillonnage i.

1. L'ajustement pour fin de non-réponse

Soit X une variable pour laquelle on désire un ajustement pour les non-réponses et soit p le nombre de catégories de cette variable, le facteur de pondération est alors:

$$A_i^{(0)} = \Sigma_p (X^{(p)} / \hat{X}^{(p)}) \, \delta_i^{(p)}$$

où $X^{(p)}$: total de X pour la catégorie p de l'univers;

$\hat{X}^{(p)}$: valeur estimée totale de X, c'est-à-dire le nombre de réponses à la catégorie p de l'échantillon;

$\delta_i^{(p)}$: $\begin{cases} 1 \text{ si le ième bloc d'information appartient à la catégorie p,} \\ 0 \text{ autrement.} \end{cases}$

Par exemple, assumons un ajustement pour les non-réponses à la variable âge. Pour chaque groupe d'âges défini, le nombre de réponses attendu divisé par le nombre effectif de réponses donne le facteur de pondération désiré.

2. L'ajustement itératif-séquentiel basé sur l'information auxiliaire

Parfois, il arrive aussi que le chercheur ait accès à des données auxiliaires, par exemple des données administratives sur l'immigration, qui sont, en fait, les paramètres correspondants de plusieurs estimations. Le système d'estimation utilise non seulement cette information de l'univers pour pondérer celle obtenue par l'échantillon mais ajuste également les estimations selon une séquence prescrite de variables. A ce titre, l'utilisateur peut avoir recours à un programme informatique complémentaire, tel THAID [("Theta Automatic Interaction Detector") Morgan et al., 1973].

Avant d'aborder les avantages qui découlent de l'utilisation conjointe de ces logiciels, revoyons d'abord les fondements méthodologiques qui ont guidé la fabrication de THAID et ce que ce programme accomplit.

Conçu spécifiquement pour l'analyse de variables dépendantes nominales ou ordinales, THAID permet à l'utilisateur d'identifier, d'après la valeur du coefficient partiel des variables en cause, les principales variables indépendantes et de déceler s'il y a interaction entre certaines de leurs catégories et non d'autres. Comme nous ne cherchons qu'à connaître, soit dans le temps, soit à divers stades de la période d'enquête, l'apport statistique d'un certain nombre de données multidimensionnelles, THAID se révèle un outil méthodologique approprié (voir Morgan et al., ibid., 2). Le programme peut aussi être employé pour identifier un tout autre ensemble de variables candidates à la conceptualisation de l'adaptation des immigrants ou à la simulation d'autres critères de sélection. Outre ces particularités, cette approche est aussi exempte de certaines hypothèses statistiques restrictives telles la linéarité et l'additivité (Morgan et al., ibid.).

L'utilisateur obtient simplement un modèle prédictif optimal résultant du partage de la distribution de chaque variable indépendante en deux sous-groupes. Le partage s'effectue selon l'une ou l'autre des deux statistiques, Theta ou Delta, au choix de l'utilisateur. Theta est employé lorsque le chercheur veut maximiser, en tenant compte de chaque cas, la proportion de l'échantillon classée correctement. Sous cette première statistique, la prédiction optimale s'obtient du mode de la distribution des fréquences marginales de la variable indépendante en cause, c'est-à-dire:

$$\Theta_{y/x} = \sum_{i=1}^{c} \left(\frac{F_i}{N} \right) \left(\frac{M_i}{F_i} \right) = \frac{1}{N} \sum_{i=1}^{c} M_i$$

où i : le nombre de catégories pour x (i = 1, 2, ..., c);

F_i : fréquence marginale de x pour la catégorie i;

N : la taille de l'échantillon;

M_i : le mode de F_i (Messenger et al., 1972, 768).

Ceci sous-entend toutefois qu'un changement de mode d'un groupe à l'autre est possible lors du partage d'une variable indépendante. Si la distribution de la variable dépendante est "extrêmement unimodale", par exemple, lorsque la fréquence de la catégorie modale par rapport à la fréquence de la catégorie résiduelle la plus élevée donne une proportion supérieure à 2:1, la statistique Theta ne peut pas être utilisée. Il en est ainsi soit parce que la fréquence d'un ou des deux groupes résultant du partage est inférieure à un minimum spécifié par l'utilisateur, soit parce que le partage ne maximise pas $\Theta_{y/x}$ suffisamment pour atteindre une valeur minimum pré-établie (ex. .005) [Morgan et al., op.cit., 33].

Puisque le choix de Theta n'est pas toujours possible, une statistique alternative Delta est offerte. Celle-ci maximise la somme pondérée des fréquences des différences absolues entre les proportions

originales (avant le partage) et celles des groupes scindés (Andrews et al., 1973, 49), c'est-à-dire:

$$\delta_{y/x} = \frac{N_1 \sum_{i=1}^{G} | P_j - p_{1j} | + N_2 \sum_{j=1}^{G} | P_j - p_{2j} |}{2\left(N - \frac{\Sigma N_j^2}{N}\right)}$$

où

j : indice des catégories de la variable dépendante;

P_j : proportion du groupe original dans la $j^{ième}$ catégorie de la variable dépendante;

p_{ij} : proportion de sujets du $j^{ième}$ sous-groupe correspondant à la $j^{ième}$ catégorie de la variable dépendante (voir University of Michigan, 1974, 200);

N : fréquence du groupe original;

N_j : fréquence de la $j^{ième}$ catégorie du groupe original (voir aussi Morgan et al., op.cit., 7, 15-18).

En d'autres mots, cette prédiction d'une variable dépendante nominale n'est plus sous la forme d'une catégorisation optimale de tous les cas - en utilisant le mode - mais bien sous la forme d'une prédiction optimale basée sur des sous-groupes qui ont une distribution différente. Bien que les auteurs du programme informatique préfèrent la statistique Delta, en outre, à cause de sa capacité de repérer une tendance d'association chez des catégories autres que celle du mode et que

Theta, d'ailleurs, omettrait (Morgan et al., ibid. 34), l'utilisateur choisira aussi cette statistique Delta puisque le programme imprime également les coefficients Theta dans le cas où l'option alternative est adoptée.

Un autre fondement méthodologique découle de l'analyse multivariée de données provenant d'enquêtes longitudinales. L'analyse comparée de telles données soulève, en effet, deux facettes de l'apport statistique d'un indice ou d'une variable: a) le niveau p au temps x - ex. un revenu d'emploi p après un an de résidence au Canada; et, b) le niveau p ou q au temps x + n, étant donné que certains autres attributs de l'individu ne sont plus les mêmes qu'au temps x. Dans de telles circonstances, le chercheur se trouve dans l'obligation de créer une nouvelle variable (ex. chef de ménage) ou d'en recombiner simplement d'autres [ex. un indice socio-économique combinant les niveaux de revenu et de scolarité pour une profession particulière (Blishen et al., 1976)], d'où la nécessité de se servir d'un programme informatique conçu pour divers genres de variables indépendantes.

Afin de souligner la complémentarité des programmes informatiques THAID et d'estimation, représentons-nous un cas hypothétique portant sur le taux de chômage des immigrants. Assumons que nous connaissons, d'après les études antérieures, certains attributs des immigrants qui influencent cette situatior d'emploi (ex. l'âge, le sexe, la catégorie d'immigrant, la connaissance des langues officielles, la scolarité, ...). Une première exécution du système d'estima-

tion permet d'ajuster les estimations brutes pour fin de non-réponse. THAID est alors exécuté afin de déterminer l'ordre d'importance statistique des variables de notre cas hypothétique. Tel qu'il a déjà été mentionné, les valeurs des coefficients Delta (ou Theta) et des coefficients partiels respectifs servent à établir le nombre de variables et l'ordre selon lequel elles seront pondérées (pour un exemple, voir Messenger et al., op.cit., 771). Lors d'un ajustement itératif-séquentiel, l'ordre des variables est important puisque les estimations de la dernière d'entre-elles correspondent toujours exactement à la distribution externe respective. De plus, le procédé d'ajustement séquentiel des poids sera répété de sorte que les estimations des autres variables puissent s'y apparenter autant que possible. Sous cette option itérative-séquentielle, les facteurs d'ajustement d'une itération donnée sont calculés comme suit:

$$A_i^{(4)} = \left[\frac{X_1^{(h)}}{\hat{X}_1^{(h)}} \cdots \frac{X_s^{(p)}}{\hat{X}_s^{(p)}}\right] \left[\frac{X_1^{(h)}}{\hat{X}_1^{2(h)}} \cdots \frac{X_s^{(p)}}{\hat{X}_s^{2(p)}}\right] \left[\frac{X_1^{(h)}}{\hat{X}_1^{t(h)}} \cdots \frac{X_s^{(p)}}{\hat{X}_s^{t(p)}}\right]$$

et où $X_k^{r(q)}$: total de la sous-population q de la variable X_k de l'univers, $r^{ième}$ itération;

$\hat{X}_k^{r(q)}$: valeur estimée de la variable X_k ajustée pour la variable X_{k-1}, X_{k-2}, ...X_1 de la $r^{ième}$ itération et pour toutes les variables (X_1, X_2, ..., X_s) des itérations (r-1) précédentes.

Le système d'estimation donne, en option, les tableaux recoupés résultant de ce dernier ajustement de poids, permettant ainsi de comparer ces nouvelles estimations aux estimations antérieures à cet ajustement.

Revoyons donc, en sommaire, les principales étapes de ce mode de pondération. Lorsqu'on a un échantillon aléatoire, la première étape consiste à rattacher à chaque unité d'échantillonnage au fichier un poids qui est tout simplement la fraction d'échantillonnage. L'ajustement pour fin de non-réponse suit et a pour effet de pondérer indépendamment le poids de chacune des variables d'intérêt. Sur la base des résultats de THAID, le système d'estimation assure, sous l'option itérative-séquentielle, une convergence rapide des valeurs des estimations vers celles des paramètres. Dès lors, le méthodologue a en mains les données qu'il lui faut pour juxtaposer (en utilisant THAID de nouveau) les résultats des estimations non-pondérées à celles qui le sont et d'obtenir, ainsi, une mesure de l'efficacité des opérations de pondération.

RÉFÉRENCES

Andrews, Frank M. et Robert C. Messenger, Multivariate Nominal Scale Analysis: a Report on a New Analysis Technique and a Computer Program, Ann Arbor: The University of Michigan, 1973, vi, 108 p.

Blishen, Bernard R. et Hugh A. McRoberts, "A Revised Socioeconomic Index for Occupations in Canada", La Revue canadienne de Sociologie et d'Anthropologie, 13, 1, février 1976, p. 71-79.

Coleman, James S., "Recent Developments in American Sociological Methods", The Polish Sociological Bulletin, n° 2 (30), 1974, p. 11-23.

Gratton, Misa, Estimation System User's Guide, Division des systèmes généraux d'enquête, Statistique Canada, Ottawa, octobre 1975, 77 p.

Messenger, Robert et Lewis Mandell, "A Modal Search Technique for Predictive Nominal Scale Multivariate Analysis", Journal of the American Statistical Association, 67, n° 340, décembre 1972, p. 768-772.

Morgan, James N. et Robert C. Messenger, THAID: a Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables, Ann Arbor: The University of Michigan, 1973, 92 p.

Namboodiri, N. Krishnan, Lewis F. Carter et Huber M. Blalock, jr, Applied Multivariate Analysis and Experimental Design, New York: McGraw-Hill, 1975, 600 p.

Nie, Norman H., C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner et Dale H. Bent, SPSS (Statistical Package for the Social Sciences), New York: McGraw-Hill, 2e édition, 1975, xx, 675 p.

Satin, A., Methodological Requirements for an Automated Estimation System, Elaboration d'enquêtes-ménages, Statistique Canada, Ottawa, juin 1974, 19 p.

University of Michigan (The), Institute for Social Research, OSIRIS III: an Integrated Collection of Computer Programs for the Management and Analysis of Social Science Data, Survey Research Center, volume 5, 1974, iv, 212 p.