Techniques d'enquête, décembre 1992
Vol. 18, n° 2, pp. 295–309
Statistique Canada

Une méthode multivariée pour l'estimation composite des dépenses de consommation en vue du calcul des indices des prix à la consommation aux États-Unis

P. LAHIRI et WENYU WANG[1]

RÉSUMÉ

Les auteurs se penchent sur l'estimation du "poids en valeur" et de l'"importance relative" de diverses strates de produits et services pour les régions de consommation. L'estimation de ces paramètres est une opération indispensable pour la construction des indices des prix à la consommation aux É.-U. Dans cet article, on se sert de modèles à plusieurs variables pour construire des estimateurs composites qui intègrent de l'information provenant de sources pertinentes. L'erreur quadratique moyenne (EQM) des estimateurs proposés et des estimateurs existants est estimée au moyen des demi-échantillons répétés tirés de l'enquête. D'après les résultats obtenus, les estimateurs proposés semblent être supérieurs aux autres estimateurs.

MOTS CLÉS: Dépenses de consommation; estimation composite; indice des prix à la consommation; poids en valeur; enquête journal; demi-échantillon; indice de Laspeyres; erreur quadratique moyenne; estimation synthétique.

1. INTRODUCTION

L'indice des prix à la consommation (IPC) des É.-U. est un indicateur des variations de prix touchant un ensemble de biens et services dont le volume et la qualité sont constants pendant une période quelconque. À chaque mois, le Bureau of Labor Statistics (BLS) des É.-U. calcule un certain nombre d'indices des prix à la consommation pour diverses régions géographiques, unités de consommation et classes de produits (voir le BLS Handbook of Methods, 1988).

La plus petite unité de classification des produits pour laquelle le BLS calcule des IPC est la "strate de produits et services". Il s'agit d'un ensemble préétabli de biens ou services de consommation (par ex.: lait entier frais) qui peuvent être achetés au détail durant une "période de référence" par un ensemble déterminé d'unités de consommation. Une unité de consommation peut être formée de tous les membres d'un ménage qui sont unis entre eux par les liens du sang, du mariage ou de l'adoption ou par toute autre forme de contrat. Un groupe de strates de produits et services forme ce que l'on appelle une catégorie de dépenses (par ex.: produits laitiers).

À des fins d'échantillonnage, on a divisé les É.-U. en huit régions dites "principales". Une région principale peut être "autoreprésentative" ou "non autoreprésentative" et fait partie de l'une ou l'autre des quatre régions géographiques suivantes: Nord-Est, Midwest, Sud et Ouest. Une région autoreprésentative se compose de toutes les grandes villes d'une région. Une région non autoreprésentative correspond généralement à un comté ou à un groupe de comtés contigus. À des fins de publication, les régions principales sont subdivisées en "régions de consommation" ou "régions de publication".

Nous décrivons ci-dessous la formule de Laspeyres qu'utilise le BLS pour calculer l'IPC pour une région donnée et la catégorie de dépenses E par exemple. Soit

[1] P. Lahiri, Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, NE 68588 0323, USA. Wenyu Wang, SUNY Health Science Center at Brooklyn, Box 1203, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.

P_{it} = le prix moyen de tous les produits ou services de la strate i à la période t ($t = 0,T$);

Q_{i0} = la quantité totale de biens ou de services de la strate i achetés à la période $t = 0$ (période de référence).

Alors, l'indice de Laspeyres au temps $t = T$ est défini

$$I_T = \sum_{i \in E} Q_{i0} P_{iT} \Big/ \sum_{i \in E} Q_{i0} P_{i0}$$

$$= \frac{\sum_{i \in E} C_i (P_{iT}/P_{i0})}{\sum_{i \in E} C_i}$$

$$= \sum\nolimits_{i \in E} R_i (P_{iT}/P_{i0}),$$

où

$C_i = Q_{i0} P_{i0}$ = dépenses totales pour les produits ou services de la strate i à $t = 0$,

$R_i = C_i / \sum_{i \in E} C_i$ = proportion de la dépense totale consacrée aux produits ou services de la strate i à $t = 0$.

Les quantités C_i et R_i représentent respectivement le "poids en valeur" et l' "importance relative" de la strate de produits et services i dans la catégorie de dépenses E.

Le Bureau of Labor Statistics se sert des données de la Consumer Expenditure Survey (CES) des É.-U. pour calculer les indices des prix à la consommation. La CES est constituée de deux volets: une enquête journal et une enquête par interview, chacune ayant ses plans d'échantillonnage et ses questionnaires propres. Dans cet article, nous ne considérons que des données de l'enquête journal. Selon le plan de sondage, toutes les unités primaires d'échantillonnage (UPÉ) d'une région autoreprésentative particulière sont tirées avec une probabilité égale à un, mais dans le cas des régions non autoreprésentatives, on ne tire qu'un échantillon probabiliste d'UPÉ. Dans un deuxième temps, on prélève un échantillon d'unités de consommation (UC) dans chacune des UPÉ échantillonnées en utilisant une fois de plus un plan probabiliste. Pendant deux périodes consécutives d'une semaine, chaque répondant consigne dans un journal les dépenses qu'il fait pour divers produits. Pour en savoir plus sur l'IPC et la CES, le lecteur est prié de consulter le **Handbook of Methods** du BLS (1988).

Les estimateurs classiques du poids en valeur et de l'importance relative des strates de produits et services sont généralement beaucoup moins efficaces au niveau de la région de publication qu'aux niveaux d'agrégation géographique supérieurs (par ex., régions principales). Cela s'explique par le fait que les régions de publication ne comptent chacune qu'un petit nombre d'unités de consommation. D'où la nécessité d'améliorer les estimateurs classiques en "empruntant" de l'information à des sources connexes. Marks (1978) et Cohen et Sommers (1984) ont étudié certains estimateurs composites qui groupent de l'information en provenance de régions connexes. Ghosh et Sohn (1990) ont obtenu des estimateurs composites du poids en valeur et de l'importance relative à l'aide d'une méthode empirique de Bayes.

La méthode qu'utilise actuellement le Bureau of Labor Statistics comporte plusieurs étapes. Premièrement, on calcule des estimateurs composites de l'importance relative à l'aide d'une méthode proposée par Cohen et Sommers (1984). On se sert ensuite d'une méthode itérative de "balayage" pour déduire les estimateurs du poids en valeur des estimateurs de l'importance

relative. Les estimations finales du poids en valeur pour la catégorie de dépenses et pour la région principale sont égales aux estimations provisoires correspondantes. Si nous prenons les moyens d'obtenir par un procédé de balayage des données qui concordent, c'est probablement parce que, aux niveaux d'agrégation supérieurs les estimateurs préliminaires ont en général un rendement plus satisfaisant qu'aux niveaux d'agrégation inférieurs. La dernière étape de la méthode du BLS consiste à déduire directement des estimateurs finals du poids en valeur les estimateurs finals de l'importance relative pour chaque division.

Contrairement à d'autres auteurs, nous tenons compte de la corrélation entre les strates de produits et services lorsque nous présentons dans la section 2 nos estimateurs composites. Le facteur de rétrécissement de l'estimateur composite qui présente l'erreur quadratique moyenne la moins élevée d'une catégorie appropriée d'estimateurs comprend des paramètres inconnus. Ceux-ci sont estimés au moyen des échantillons répétés compensés de l'enquête. L'estimateur que proposent Cohen et Sommers (1984) est en fait un cas particulier de notre estimateur dans l'hypothèse où il n'existe aucune corrélation entre les estimateurs préliminaires.

Dans la section 2, nous cherchons essentiellement à estimer le poids en valeur d'une strate de produits et services pour une région de publication. Il est toutefois possible d'estimer des poids en valeur pour des niveaux d'agrégation plus élevés (par ex.: catégorie de dépenses pour une région de publication, *etc.*) en effectuant les sommations voulues. D'après les résultats de notre étude, si l'on en juge par l'erreur quadratique moyenne, les estimateurs que nous proposons surclassent constamment les estimateurs préliminaires correspondants et, donc, les estimateurs du BLS (notons qu'à cause du procédé de balayage, les estimateurs du BLS sont identiques aux estimateurs préliminaires pour les niveaux d'agrégation supérieurs).

Dans la section 3, nous proposons un estimateur composite de l'importance relative d'une strate de produits et services pour les régions de publication. Au lieu d'utiliser les estimateurs préliminaires des poids en valeur, nous nous servons des estimateurs préliminaires de l'importance relative de chacune des strates de produits et services qui constituent la catégorie de dépenses à l'étude. Comme la somme de ces estimateurs est égale à un, la matrice de variances-covariances des estimateurs préliminaires est singulière, ce qui fait que l'estimation de l'importance relative est un problème différent de l'estimation des poids en valeur. La méthode que nous proposons permet de supprimer de façon optimale une strate de produits et services, ce qui élimine le problème de la singularité de la matrice de variances-covariances des estimateurs préliminaires. Les résultats numériques de notre analyse montrent que l'estimateur proposé surclasse invariablement tous les autres estimateurs étudiés en ce qui regarde l'erreur quadritique moyenne.

La section 4 renferme tous les résultats numériques. Nous avons évalué différents estimateurs du poids en valeur et de l'importance relative en nous fondant sur l'erreur quadritique moyenne estimée établie à l'aide des demi-échantillons compensés (voir McCarthy (1969), Ghosh et Sohn (1990)). D'après nos résultats, les estimateurs proposés semblent être supérieurs à tous les autres estimateurs étudiés.

2. ESTIMATION DU POIDS EN VALEUR

Soit X_{ijl} la moyenne des dépenses de deux semaines consécutives faites pour tous les produits ou services de la strate i par l'unité de consommation l dans la région de publication j d'une région principale particulière ($i = 1, \ldots, I; j = 1, \ldots, m; l = 1, \ldots, n_j$). Soit W_{jl} le poids d'échantillonnage rattaché à l'unité de consommation l dans la région de publication j ($j = 1, \ldots, m$; $l = 1, \ldots, n_j$). Ce poids représente un certain nombre d'unités de consommation dans la population et est déterminé par le Census Bureau selon une méthode complexe qui tient compte de divers facteurs comme les probabilités de sélection, la non-réponse, *etc.* Le but de cette section est d'estimer θ_{ij}, c.-à-d. les dépenses hebdomadaires moyennes réelles d'une unité de consommation pour les produits ou services de la strate i dans la région de publication j. Le poids en valeur est défini simplement comme $N_j\theta_{ij}$, N_j étant le

nombre total d'unités de consommation dans la région de publication j. L'estimateur préliminaire de θ_{ij} est défini par l'expression

$$Y_{ij} = \sum_{l=1}^{n_j} W_{jl}X_{ijl} \Big/ \sum_{l=1}^{n_j} W_{jl}, \ (i = 1, \ldots, I; \ \ j = 1, \ldots, m). \tag{2.1}$$

De même, l'estimateur correspondant pour une région principale s'écrit

$$Y_{i.} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}X_{ijl} \Big/ \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}. \tag{2.2}$$

Comme la variabilité de $Y_{i.}$ est beaucoup moins grande que celle de Y_{ij}, il nous faut un estimateur composite de θ_{ij} qui accroisse le niveau de précision. Soit $Y_j = (Y_{1j}, \ldots, Y_{Ij})'$ et $\theta_j = (\theta_{ij}, \ldots, \theta_{Ij})', j = 1, \ldots, m$. Posons V_j comme la matrice de variances-covariances vraie de Y_j, $(j = 1, \ldots, m)$. Suivant une hypothèse artificielle, à savoir $\theta_j = \mu$, un vecteur colonne $I \times 1$, $(j = 1, \ldots, m)$, le meilleur estimateur de θ_j est défini

$$\tilde{\mu} = \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \sum_{j=1}^{m} V_j^{-1} Y_j, \tag{2.3}$$

expression que l'on obtient en minimisant $\sum_{j=1}^{m} (Y_j - \mu)' V_j^{-1} (Y_j - \mu)$ par rapport à μ. Or, l'hypothèse artificielle se vérifie difficilement. Par contre, lorsqu'il n'y a aucune espèce de similitude entre les θ_j, il convient d'utiliser Y_j comme estimateur de θ_j. Pour les cas intermédiaires, on peut se servir d'un estimateur composite comme celui-ci:

$$\hat{\theta}_{ij}(a_{ij}) = (1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu}, \tag{2.4}$$

où les a_{ij} sont des valeurs constantes $(0 \leq a_{ij} \leq 1)$, et e_i est un vecteur colonne $I \times 1$ dont l'élément i est égal à 1 et les autres éléments sont nuls.

On obtient a_{ij} en minimisant l'erreur quadratique moyenne

$$E[\{(1 - a_{ij}) Y_{ij} + a_{ij} e_i' \tilde{\mu} - \theta_{ij}\}^2 \mid \theta_{ij}] \tag{2.5}$$

par rapport à a_{ij}. La formule optimale est donc

$$\tilde{a}_{ij} = \frac{e_i' \left[V_j - \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \right] e_i}{E[(Y_{ij} - e_i' \tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]}. \tag{2.6}$$

Par conséquent, l'estimateur optimal de θ_{ij}, pour la classe d'estimateurs définis en (2.4), s'écrit

$$\tilde{\theta}_{ij} = (1 - \tilde{a}_{ij}) Y_{ij} + \tilde{a}_{ij} e_i' \tilde{\mu}. \tag{2.7}$$

Remarque 1: Dans le calcul de l'estimateur optimal $\tilde{\theta}_{ij}$, on suppose que les quantités V_j, $(j = 1, \ldots, m)$, et $E[(Y_{ij} - e_i'\tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m]$ sont fixes et connues.

Remarque 2: L'estimateur proposé par Cohen et Sommers (1984) est un cas particulier de $\tilde{\theta}_{ij}$ lorsque

$$V_j = \left(\sum_{l=1}^{n_j} W_{jl} \right)^{-1} \operatorname{Diag}(\sigma_1^2, \ldots, \sigma_I^2).$$

Notons que ces auteurs supposent qu'il n'existe pas de corrélation entre les strates de produits et services, ce qui semble être une hypothèse très restrictive du point de vue de notre étude.

Remarque 3: Il convient de souligner que par une simple inversion de matrice (voir Rao 1973),

$$V_j - \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} = V_j \left[V_j + \left(\sum_{s \neq j} V_s^{-1} \right)^{-1} \right]^{-1} V_j$$

qui est définie positive. De plus,

$$E[(Y_{ij} - e_i'\tilde{\mu})^2 \mid \theta_j, j = 1, \ldots, m] = e_i' V_j \left[V_j + \left(\sum_{s \neq j} V_s^{-1} \right)^{-1} \right]^{-1} V_j e_i$$

$$+ \left[\theta_{ij} - e_i' \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \left(\sum_{j=1}^{m} V_j^{-1} \theta_j \right) \right]^2.$$

Par ailleurs, lorsque $\theta_j = \mu$, $\tilde{a}_{ij} = 1$ et, par conséquent, $\tilde{\theta}_{ij} = e_i'\tilde{\mu}$. Autrement, la grandeur du facteur de rétrécissement dépend de la valeur de

$$\left[\theta_{ij} - e_i' \left(\sum_{j=1}^{m} V_j^{-1} \right)^{-1} \left(\sum_{j=1}^{m} V_j^{-1} \theta_j \right) \right]^2.$$

Plus l'écart entre θ_{ij} et $e_i'(\sum_{j=1}^{m} V_j^{-1})^{-1}(\sum_{j=1}^{m} V_j^{-1}\theta_j)$ est grand, plus $\tilde{a}_{ij}$ est faible. Autrement dit, si une région particulière se distingue nettement des autres par ses caractéristiques, notre méthode attribuera moins de poids à la portion synthétique de l'estimateur. C'est ce qui explique la forte variabilité des facteurs de rétrécissement du tableau 1.

Tableau 1

Facteurs de rétrécissement $\hat{a}_{ij}$ pour la région non autoreprésentative de l'Ouest

i \ j	1	2	2
1	0.8479225	0.7057626	0.9214804
2	0.8434894	0.5692695	0.8092725
3	0.0969009	0.0786758	0.6953904
4	0.4446537	0.5444809	1
5	0.6999551	0.3460123	0.5487382
6	0.0318442	0.4981756	0.2598752

Nous allons estimer $\tilde{a}_{ij}$ à l'aide des 20 demi-échantillons compensés provenant de l'enquête. Posons $w_{jl}^{(k)}$ comme le poids attribué à l'unité de consommation l de la région j pour l'échantillon répété k ($j = 1, \ldots, m$; $l = 1, \ldots, n_j$; $k = 1, \ldots, 20$). Ces poids sont élaborés par le Census Bureau selon une procédure complexe. Pour un échantillon répété quelconque, environ la moitié des unités de consommation reçoivent un poids nul et le reste reçoivent un poids positif. Définissons

$$\hat{a}_{ij}^* = \frac{e_i'\left[\hat{V}_j - \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\right]e_i}{\frac{1}{20}\sum_{k=1}^{20} [Y_{ij}^{(k)} - e_i'\hat{\mu}^{(k)}]^2},$$

$$\hat{\mu} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1} Y_j\right],$$

$$\hat{\mu}^{(k)} = \left[\sum_{j=1}^{m} \hat{V}_j^{-1}\right]^{-1}\left[\sum_{j=1}^{m} \hat{V}_j^{-1} Y_j^{(k)}\right],$$

$$Y_{ij}^{(k)} = \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} \Big/ \sum_{l=1}^{n_j} W_{jl}^{(k)},$$

$$Y_j^{(k)} = [Y_{1j}^{(k)}, \ldots, Y_{Ij}^{(k)}]',$$

$$\hat{V}_j = 1/20 \sum_{k=1}^{20} [Y_j^{(k)} - Y_j][Y_j^{(k)} - Y_j]'.$$

Nous proposons alors l'estimateur de θ_{ij} suivant:

$$\hat{\theta}_{ij}^* = (1 - \hat{a}_{ij}^*) Y_{ij} + \hat{a}_{ij}^* e_i'\hat{\mu}. \tag{2.8}$$

Remarque 4: Sur la base de l'argument donné dans la remarque 3, $\hat{a}_{ij}^* \geq 0$. Mais il peut arriver que $\hat{a}_{ij}^*$ soit supérieur à un. Par conséquent, nous considérons l'estimateur suivant:

$$\hat{\theta}_{ij} = (1 - \hat{a}_{ij}) Y_{ij} + \hat{a}_{ij} e_i'\hat{\mu}, \tag{2.9}$$

où $\hat{a}_{ij} = \min[1, \hat{a}_{ij}^*]$.

Le tableau 1 donne les valeurs de $\hat{a}_{ij}$ pour la région non autoreprésentative de l'Ouest.

3. ESTIMATION DE L'IMPORTANCE RELATIVE

Soit $R_{ij} = Y_{ij} / \sum_{i=1}^{I} Y_{ij}$ l'estimateur préliminaire de l'importance relative $r_{ij} = \theta_{ij} / \sum_{i=1}^{I} \theta_{ij}$, ($i = 1, \ldots, I$; $j = 1, \ldots, m$). Posons $R_j = (R_{1j}, \ldots, R_{Ij})'$, ($j = 1, \ldots, m$). Comme

$\sum_{i=1}^{I} R_{ij} = 1$, $(j = 1, \ldots, m)$, la matrice de variances-covariances de R_j est singulière. On ne peut donc appliquer directement la méthode exposée dans la section précédente. Pour contourner le problème de la singularité, nous supprimons une strate de produits et services de la catégorie de dépenses étudiée. Sans aucune perte de généralité, supprimons la strate I. Dans un deuxième temps, appliquons la méthode décrite dans la section 2 afin d'obtenir l'estimateur de r_{ij} suivant, $(i = 1, \ldots, I - 1; j = 1, \ldots, m)$

$$\hat{r}_{ij}^{*} = (1 - \hat{d}_{ij})R_{ij} + \hat{d}_{ij} e_i' \hat{\xi}, \tag{3.1}$$

où

$$\hat{d}_{ij} = \min[1, \hat{d}_{ij}^{*}],$$

$$\hat{d}_{ij}^{*} = \frac{e_i' \left[\hat{D}_j - \left[\sum_{j=1}^{m} \hat{D}_j^{-1} \right]^{-1} \right] e_i}{\frac{1}{20} \sum_{k=1}^{20} [R_{ij}^{(k)} - e_i' \hat{\xi}^{(k)}]^2},$$

$$R_{ij}^{(k)} = Y_{ij}^{(k)} \Big/ \sum_{i=1}^{I} Y_{ij}^{(k)},$$

$$R_j^{(k)} = (R_{1j}^{(k)}, \ldots, R_{I-1j}^{(k)})',$$

$$\hat{D}_j = \frac{1}{20} \sum_{k=1}^{20} (R_j^{(k)} - R_j)(R_j^{(k)} - R_j)',$$

$$\hat{\xi}^{(k)} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1} \right]^{-1} \left[\sum_{j=1}^{m} \hat{D}_j^{-1} R_j^{(k)} \right],$$

$$\hat{\xi} = \left[\sum_{j=1}^{m} \hat{D}_j^{-1} \right]^{-1} \left[\sum_{j=1}^{m} \hat{D}_j^{-1} R_j \right].$$

Pour $i = I$,

$$\hat{D}_{II}^{(j)} = \frac{1}{20} \sum_{k=1}^{20} (\hat{R}_{Ij}^{(k)} - R_{Ij})^2,$$

$$R_{I.} = \left[\sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1} R_{Ij} \right] \Big/ \sum_{j=1}^{m} (\hat{D}_{II}^{(j)})^{-1},$$

$$\hat{d}_{Ij} = \min[1, \hat{d}^*_{Ij}],$$

$$\hat{d}^*_{Ij} = \frac{\hat{D}^{(j)}_{II} - \left[\sum_{j=1}^{m} (\hat{D}^{(j)}_{II})^{-1}\right]^{-1}}{\frac{1}{20}\sum_{k=1}^{20} [R^{(k)}_{Ij} - R^{(k)}_{I\cdot}]^2},$$

$$R^{(k)}_{I\cdot} = \left[\sum_{j=1}^{m} (\hat{D}^{(j)}_{II})^{-1} R^{(k)}_{Ij}\right] \Big/ \sum_{j=1}^{m} (\hat{D}^{(j)}_{II})^{-1}.$$

Pour estimer r_{Ij}, nous nous servons d'une méthode univariée qui donne l'estimateur de r_{Ij}, suivant, $(j = 1, \ldots, m)$:

$$\hat{r}^*_{Ij} = (1 - \hat{d}_{Ij})R_{Ij} + \hat{d}_{Ij}R_I.$$

Comme estimateur final de r_j nous avons $\hat{r}_j = (\hat{r}_{1j}, \ldots, \hat{r}_{Ij})'$, où $\hat{r}_{ij} = \hat{r}^*_{ij}/\sum_{i=1}^{I}\hat{r}^*_{ij}$. Pour ce qui est de la suppression de la strate de produits et services, nous avons I choix possibles. Nous choisissons la combinaison pour laquelle l'erreur quadritique moyenne estimée (pour l'ensemble des strates) est la plus faible. On peut obtenir un autre estimateur de r_{Ij} en soustrayant $\sum_{i=1}^{I-1} r_{ij}$ de un. Cependant, selon la procédure, il y a des chances réelles que la valeur estimée de r_{Ij} soit négative.

4. RÉSULTATS NUMÉRIQUES

Dans cette section, nous évaluons divers estimateurs du poids en valeur et de l'importance relative en nous fondant sur l'erreur quadritique moyenne estimée. Nous comparons en fait quatre estimateurs: l'estimateur préliminaire, l'estimateur proposé par Cohen et Sommers (1984), l'estimateur qu'utilise actuellement le BLS et l'estimateur empirique de Bayes qu'ont étudié récemment Ghosh et Sohn (1990). L'estimateur de Cohen-Sommers du poids en valeur (avant balayage) est défini par l'expression

$$\begin{aligned}
\hat{\theta}^{\mathrm{CS}}_{ij} &= \hat{\theta}^{\mathrm{CS}*}_{ij} \quad \text{si} \quad |\hat{\theta}^{\mathrm{CS}*}_{ij} - Y_{ij}| < \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \\
&= Y_{ij} + \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \quad \text{si} \quad \hat{\theta}^{\mathrm{CS}*}_{ij} \geq Y_{ij} + \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \\
&= Y_{ij} - \mathrm{c}\cdot\mathrm{sd}(Y_{ij}) \quad \text{si} \quad \hat{\theta}^{\mathrm{CS}*}_{ij} \leq Y_{ij} - \mathrm{c}\cdot\mathrm{sd}(Y_{ij})
\end{aligned}$$

où

$$\hat{\theta}^{\mathrm{CS}*}_{ij} = (1 - \hat{a}^{\mathrm{CS}}_{ij})Y_{ij} + \hat{a}^{CS}_{ij}Y_{i\cdot},$$

$$\hat{a}^{\mathrm{CS}}_{ij} = \min\left[1, (1 - N_j/N)\left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y_{ij})^2\right] \Big/ \left[\frac{1}{20}\sum_{k=1}^{20}(Y^{(k)}_{ij} - Y^{(k)}_{i\cdot})^2\right]\right],$$

Tableau 2

Valeurs du facteur de sécurité (c) pour les régions principales

Région principale	NCNS	NCSR	NENS	NESR	SSNS	SSSR	WWNS	WWSR
	1	2	3	4	5	6	7	8
c	1.0	.5	1.0	.5	3.0	.25	1.0	.5

NCNS: Centre-Nord (Midwest) - non autoreprésentative.
NCSR: Centre-Nord - autoreprésentative.
NENS: Nord-Est - non autoreprésentative.
NESR: Nord-Est - autoreprésentative.
SSNS: Sud - non autoreprésentative.
SSSR: Sud - autoreprésentative.
WWNS: Ouest - non autoreprésentative.
WWSR: Ouest - autoreprésentative.

$$Y_{i\cdot}^{(k)} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} \Big/ \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl}^{(k)},$$

$N_j =$ nombre total d'unités de consommation dans la région de publication j,

$$N = \sum_{j=1}^{m} N_j,$$

$$\mathrm{sd}(Y_{ij}) = \sqrt{\left\{\frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2\right\}},$$

$c =$ facteur de sécurité déterminé par le BLS (voir tableau 2).

L'estimateur de Cohen-Sommers de l'importance relative est défini par l'expression

$$\begin{aligned} \hat{r}_{ij}^{\mathrm{CS}} &= \hat{r}_{ij}^{\mathrm{CS}*} \quad \text{si} \quad |\hat{r}_{ij}^{\mathrm{CS}*} - R_{ij}| \le \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \\ &= R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \quad \text{si} \quad \hat{r}_{ij}^{\mathrm{CS}*} \ge R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \\ &= R_{ij} + \mathrm{c}\cdot\mathrm{sd}(R_{ij}) \quad \text{si} \quad \hat{r}_{ij}^{\mathrm{CS}*} \le R_{ij} - \mathrm{c}\cdot\mathrm{sd}(R_{ij}), \end{aligned}$$

où

$$\hat{r}_{ij}^{\mathrm{CS}*} = (1 - \hat{d}_{ij}^{\mathrm{CS}}) R_{ij} + \hat{d}_{ij}^{\mathrm{CS}} R_{i\cdot}^{\mathrm{CS}},$$

$$R_{i\cdot}^{\mathrm{CS}} = \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl} \Big/ \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{l=1}^{n_j} W_{jl} X_{ijl},$$

$$\hat{d}_{ij}^{\mathrm{CS}} = \hat{d}_{ij}^{\mathrm{CS}*} \quad \text{si} \quad 0 < \hat{d}_{ij}^{\mathrm{CS}*} < 1,$$

$$= 0 \quad \text{si} \quad \hat{d}_{ij}^{\mathrm{CS}*} \leq 0,$$

$$= 1 \quad \text{si} \quad \hat{d}_{ij}^{\mathrm{CS}*} \geq 1,$$

$$\hat{d}_{ij}^{\mathrm{CS}*} = \frac{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})^2 - \frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})(R_{i\cdot}^{\mathrm{CS}(k)} - R_{i\cdot}^{\mathrm{CS}})}{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{i\cdot}^{\mathrm{CS}(k)})^2},$$

$$R_{i\cdot}^{\mathrm{CS}(k)} = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl}^{(k)} X_{ijl},$$

$$\mathrm{sd}(R_{ij}) = \sqrt{\frac{1}{20}\sum_{k=1}^{20}(R_{ij}^{(k)} - R_{ij})^2}.$$

Puisque $\sum_{i=1}^{I} \hat{r}_{ij}^{\mathrm{CS}} \neq 1$, nous avons divisé $\hat{r}_{ij}^{\mathrm{CS}}$ par $\sum_{i=1}^{I} \hat{r}_{ij}^{CS}$ pour les besoins de la comparaison.

La méthode qu'utilise actuellement le Bureau of Labor Statistics (voir United States Department of Labor 1988) comporte un certain nombre d'étapes.

Étape 1: Déterminer un estimateur du poids en valeur comme ci-dessous:

$$\hat{\theta}_{ij}^{\mathrm{CS}(1)} = \hat{r}_{ij}^{CS} \sum_{i=1}^{I} Y_{ij}.$$

Étape 2: Déduire l'estimateur final de θ_{ij} de $\hat{\theta}_{ij}^{\mathrm{CS}(1)}$ par un procédé de "balayage". L'estimateur final, désigné par $\hat{\theta}_{ij}^{\mathrm{BLS}}$, satisfait aux deux conditions suivantes:

$$\sum_{i=1}^{I} \hat{\theta}_{ij}^{\mathrm{BLS}} = \sum_{i=1}^{I} Y_{ij},$$

$$\sum_{j=1}^{m} N_j \hat{\theta}_{ij}^{\mathrm{BLS}} = \sum_{j=1}^{m} N_j Y_{ij}.$$

Étape 3: Enfin, déterminer un estimateur de l'importance relative comme ci-dessous:

$$\hat{r}_{ij}^{\mathrm{BLS}} = \hat{\theta}_{ij}^{\mathrm{BLS}} \Big/ \sum_{i=1}^{I} \hat{\theta}_{ij}^{\mathrm{BLS}}.$$

Dans nos calculs, nous avons estimé N_j par $\sum_{l=1}^{n_j} W_{jl}$.

L'erreur quadratique moyenne d'un estimateur e_{ij} de θ_{ij} est définie par l'expression

$$\begin{aligned} \text{EQM} &= E(e_{ij} - \theta_{ij})^2 \\ &= E(e_{ij} - Y_{ij})^2 - V(Y_{ij}) + 2\,\text{Cov}(e_{ij}, Y_{ij}), \end{aligned}$$

où l'on suppose que $E(Y_{ij} \mid \theta_{ij}) = \theta_{ij}$. La formule ci-dessus est définie dans Cohen et Sommers (1984). Comme dans Ghosh et Sohn (1990), nous estimons les trois termes de la formule au moyen des demi-échantillons compensés de l'enquête. Par exemple,

$$E(e_{ij} - Y_{ij})^2 \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - Y_{ij}^{(k)})^2,$$

$$V(Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (Y_{ij}^{(k)} - Y_{ij})^2,$$

$$\text{Cov}(e_{ij}, Y_{ij}) \doteq \frac{1}{20} \sum_{k=1}^{20} (e_{ij}^{(k)} - e_{ij})(Y_{ij}^{(k)} - Y_{ij}).$$

Dans l'expression ci-dessus, $e_{ij}^{(k)}$ désigne l'estimateur e_{ij} fondé sur le k-ième demi-échantillon ($k = 1, \ldots, 20$). Par exemple,

$$\hat{\theta}_{ij}^{\text{CS}(k)} = (1 - \hat{a}_{ij}^{CS}) Y_{ij}^{(k)} + \hat{a}_{ij}^{\text{CS}} Y_{i\cdot}^{(k)},$$

$$\hat{\theta}_{ij}^{(k)} = (1 - \hat{a}_{ij}) Y_{ij}^{(k)} + \hat{a}_{ij} e_i' \hat{\mu}^{(k)}.$$

Nous calculons $\hat{\theta}_{ij}^{\text{BLS}(k)}$ à l'aide de la méthode multi-étape utilisée pour calculer $\hat{\theta}_{ij}^{\text{BLS}}$, sauf que nous remplaçons Y_{ij}, R_{ij} et $\hat{r}_{ij}^{\text{CS}}$ par $Y_{ij}^{(k)}$, $R_{ij}^{(k)}$ et $\hat{r}_{ij}^{\text{CS}(k)}$ respectivement. Il convient de souligner que la procédure ci-dessus ne tient pas compte de la variation due à l'estimation des coefficients (c.-à-d. des a_{ij}) des estimateurs composites. Cohen et Sommers (1984) ont recommandé l'utilisation de demi-échantillons des demi-échantillons, ou quarts d'échantillons, pour tenir compte de cette variation. Nous n'avons pas pu appliquer cette méthode puisque notre ensemble de données ne renfermait pas de données relatives à des quarts d'échantillons.

Les données que nous analysons sont tirées de la Consumer Expenditure Survey de 1982-1983 (volet "enquête journal"). La catégorie de dépenses étudiée est celle des produits laitiers. Cette catégorie compte en tout six strates de produits et services: 1) lait entier frais, 2) autres sortes de lait frais et crème, 3) beurre, 4) fromage, 5) crème glacée et produits assimilés et 6) autres produits laitiers.

Nous estimons l'erreur quadratique moyenne de tous les estimateurs étudiés pour chaque région de publication et chaque strate de produits et services. Le tableau 3 donne l'erreur quadratique moyenne estimée (pour l'ensemble des strates et l'ensemble des régions de publication d'une région principale) des estimateurs de θ_{ij}. Notons que sur ce plan, tous les estimateurs composites sauf celui proposé par Ghosh et Sohn (1990) surclassent l'estimateur préliminaire pour toutes les régions principales. Les estimateurs θ_{ij}^{CS} et $\hat{\theta}_{ij}$ sont tous deux

Tableau 3

Erreur quadratique moyenne estimée de différents estimateurs de θ_{ij}

Région principale	Erreur quadratique moyenne de				
	Y_{ij}	$\hat{\theta}_{ij}^{GS}$	$\hat{\theta}_{ij}^{CS}$	$\hat{\theta}_{ij}^{BLS}$	$\hat{\theta}_{ij}$
NCNS	.020047	.011549 (22)	.009342 (53)	.014885 (25)	.009428 (52)
NCSR	.036620	.024783 (32)	.016017 (56)	.023627 (35)	.016155 (55)
NENS	.018162	.013299 (26)	.007327 (59)	.013046 (28)	.005504 (69)
NESR	.052883	.051100 (3)	.038911 (26)	.045610 (13)	.028958 (45)
SSNS	.021757	.013146 (39)	.009954 (54)	.014415 (33)	.006418 (70)
SSSR	.047500	.028984 (38)	.031743 (33)	.044238 (6)	.009270 (80)
WWNS	.052387	.029938 (42)	.017433 (66)	.030069 (42)	.010849 (79)
WWSR	.018223	.033529 (−83)	.009925 (45)	.014898 (18)	.005761 (68)

Nota: Les chiffres entre parenthèses représentent l'amélioration en pourcentage par rapport à l'estimateur préliminaire, Y_{ij}.

supérieurs à $\hat{\theta}_{ij}^{BLS}$. L'estimateur que nous proposons, $\hat{\theta}_{ij}$, surclasse $\hat{\theta}_{ij}^{CS}$ dans six des huit régions principales. Dans les deux autres régions (NCNS et NCSR), $\hat{\theta}_{ij}^{CS}$ est supérieur, mais de très peu, à $\hat{\theta}_{ij}$.

Dans les tableaux 4 et 5, nous tentons de démontrer que le procédé de "balayage" n'est peut-être pas nécessaire. Le tableau 4 a pour objet le paramètre $\sum_{i=1}^{I}\theta_{ij}$, c.-à-d. le poids en valeur vrai pour une catégorie de dépenses. À cause du procédé de "balayage", $\sum_{i=1}^{I}\hat{\theta}_{ij}^{BLS} = \sum_{i=1}^{I} Y_{ij}$. Voici que nous proposons un autre estimateur, $\sum_{i=1}^{I}\hat{\theta}_{ij}$, et que nous comparons son erreur quadritique moyenne estimée (pour l'ensemble des régions de publication d'une région principale) à celle de $\sum_{i=1}^{I} Y_{ij}$. Dans tous les cas, notre estimateur est notablement supérieur.

Le tableau 5 a pour objet le poids en valeur d'une strate de produits et services pour une région principale. L'estimateur préliminaire (qui est identique à l'estimateur BLS en raison du procédé de "balayage") s'écrit $(\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl} Y_{ij}) / (\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. La méthode que nous présentons permet aussi la construction d'estimateurs au niveau de la région principale. L'estimateur que nous proposons est $\hat{\theta}_{i.} = \sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl}\,\hat{\theta}_{ij} / (\sum_{j=1}^{m}\sum_{l=1}^{n_j} W_{jl})$. Le tableau 5 donne l'erreur quadratique moyenne estimée de ces deux estimateurs. On constate une fois de plus que notre estimateur surclasse l'estimateur préliminaire (BLS).

Les résultats des tableaux 4 et 5 donnent à penser que l'étape de l'harmonisation des données par un procédé de "balayage", comme le prévoit la méthode du BLS, n'est peut-être pas indispensable. En effet, on peut vraisemblablement améliorer les estimateurs classiques à des niveaux d'agrégation supérieurs.

Tableau 4

Erreur quadratique moyenne estimée de deux estimateurs des dépenses de consommation moyennes pour une catégorie de dépenses

Région principale	Estimateur préliminaire	Estimateur proposé	Amélioration (en pourcentage)
NCNS	0.12384	0.07969	36
NCSR	0.29819	0.13040	56
NENS	0.21658	0.07602	65
NESR	0.67486	0.20119	70
SSNS	0.21506	0.08303	61
SSSR	0.68415	0.06462	90
WWNS	0.35446	0.05175	85
WWSR	0.19292	0.05524	71

Tableau 5

Erreur quadratique moyenne estimée de deux estimateurs des dépenses de consommation moyennes pour une région principale

Région principale	Estimateur préliminaire	Estimateur proposé	Amélioration (en pourcentage)
NCNS	0.008181	0.0045468	44
NCSR	0.003672	0.0031047	15
NENS	0.006174	0.0029128	53
NESR	0.011680	0.0056922	51
SSNS	0.007501	0.0036401	51
SSSR	0.004434	0.0013751	69
WWNS	0.008203	0.0022560	72
WWSR	0.002786	0.0007882	72

Le tableau 6 donne l'erreur quadritique moyenne estimée (pour l'ensemble des strates et l'ensemble des régions de publication d'une région principale) de divers estimateurs de l'importance relative. Notons que, comme dans le tableau 3, tous les estimateurs sauf $\hat{r}_{ij}^{GS}$ surclassent l'estimateur préliminaire $\hat{R}_{ij}$ pour toutes les régions principales. L'estimateur que nous proposons, $\hat{r}_{ij}$, est le plus efficace de tous les estimateurs étudiés.

Dernièrement, Swanson (1992) a comparé différentes méthodes d'estimation du poids en valeur pour 12 des quelque 70 catégories de dépenses de l'IPC. Son étude montre que dans l'ensemble, la méthode que nous proposons est supérieure à toutes les autres.

Tableau 6

Erreur quadratique moyenne estimé de différents estimateurs de l'importance relative

Région principale	Erreur quadratique moyenne estimé de				
	R_{ij}	$\hat{r}_{ij}^{GS}$	$\hat{r}_{ij}^{CS}$	$\hat{r}_{ij}^{BLS}$	$\hat{r}_{ij}$
NCNS	.0006342	.00046480 (27)	.00033143 (48)	.00042130 (34)	.00018592 (71)
NCSR	.0009125	.00071967 (21)	.00040226 (56)	.00044815 (51)	.00021309 (77)
NENS	.0003588	.00026894 (25)	.00014146 (61)	.0001620 (55)	.00011105 (69)
NESR	.0004264	.00072001 (−69)	.00028862 (32)	.00030555 (28)	.00016744 (61)
SSNS	.0005071	.00033736 (33)	.00019352 (62)	.00021385 (58)	.00011925 (76)
SSSR	.0006564	.00048569 (26)	.00053173 (19)	.00053603 (18)	.00030979 (53)
WWNS	.0013709	.00086849 (37)	.00051474 (62)	.00061901 (55)	.00028519 (79)
WWSR	.0003540	.00070770 (−100)	.00021384 (40)	.00023255 (34)	.00013750 (61)

Nota: Les chiffres entre parenthèses représentent l'amélioration en pourcentage par rapport à R_{ij}.

REMERCIEMENTS

Cet article est le résultat de recherches qui ont été rendues possibles grâce à une subvention de la National Science Foundation (NSF) (SES-9001399 – "On-Site Research to Improve the Quality of Labor Statistics.") Ces recherches ont été faites au Bureau of Labor Statistics des É.-U. pendant que les auteurs participaient au programme conjoint de recherches de l'American Statistical Association et du Bureau of Labor Statistics, financé par le BLS et la NSF. Les opinions, conclusions ou recommandations qui peuvent être exprimées dans cet article sont celles des auteurs et ne reflètent pas nécessairement la position de la National Science Foundation ni celle du Bureau of Labor Statistics. Nous tenons à remercier Sylvia Leaver, Stuart Scott, Malay Ghosh, Richard Valliant, Stephen Miller, So Young Sohn, Paul Hsen et Adriana Silberstein pour les nombreux et précieux échanges de vues que nous avons eus avec ces personnes. Nous voulons exprimer aussi notre reconnaissance aux deux arbitres et au rédacteur associé qui ont formulé des commentaires utiles sur une version antérieure de cet article.

BIBLIOGRAPHIE

COHEN, M.P., et SOMMERS, J.P. (1984). Evaluation of methods of composite estimation of cost weights for the CPI. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 466-471.

GHOSH, M. et SOHN, S.Y. (1990). An Empirical Bayes Approach Towards Composite Estimation of Consumer Expenditure. Rapport technique, U.S. Bureau of Labor Statistics.

MARKS , H. (1978). Composite estimation techniques used for the CPIR weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 311-315.

McCARTHY P.J. (1969). Pseudoreplication: half-samples. *Revue de l'Institut International de Statistique*, 37, 239-264.

SWANSON, D. (1992). An evaluation of 4 cost weight composite estimation methods for the CPI. Note de service pour Janet Williams, Chief, CPI Survey Research Branch, Statistical Methods Division, U.S. Bureau of Labor Statistics.

RAO, C.R. (1973). *Linear Statistical Inference and Its Applications*, (2ième Édition). New York: J. Wiley & Sons.

UNITED STATES DEPARTMENT OF LABOR (1988). *Handbook of Methods*. Bureau of Labour Statistics. Washington DC: U.S. Government Printing Office.