

Distr.  
GÉNÉRALE

CES/AC.49/1999/9  
5 août 1999

FRANÇAIS  
Original : ANGLAIS

COMMISSION DE STATISTIQUE et  
COMMISSION ÉCONOMIQUE POUR L'EUROPE

ORGANISATION INTERNATIONALE  
DU TRAVAIL (OIT)

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Réunion commune CEE/OIT sur les indices  
des prix à la consommation  
(Genève, 3-5 novembre 1999)

**ÉCHANTILLONNAGE DES ARTICLES AUX FINS DE L'ÉTABLISSEMENT DE L'IPC :  
LE SUCCÈS DES MÉTHODES DE PRÉLÈVEMENT FONDÉES  
SUR UN SEUIL D'INCLUSION**

Communication sollicitée de Statistics Netherlands \*

Résumé

La plupart des services de statistique constituent les échantillons de produits (articles) dont ils vont relever les prix aux fins du calcul de leurs indices des prix à la consommation de manière subjective. Aux Pays-Bas, cet échantillonnage discrétionnaire s'apparente à une méthode de prélèvement fondée sur un seuil d'inclusion (cut-off selection method), en ce sens que les articles auxquels correspondent les dépenses les moins importantes sont délibérément exclus du champ de l'observation. Ce mode de prélèvement de l'échantillon conduit à des estimations biaisées de l'indice des prix. Le présent document cherche à déterminer si un échantillonnage aléatoire donnerait de meilleurs résultats du point de vue de l'erreur quadratique moyenne. Des simulations selon la méthode de Monte-Carlo, au moyen de données saisies par lecture optique, révèlent que la sélection fondée sur un seuil d'inclusion constitue généralement une procédure efficace d'échantillonnage des articles dans le contexte de l'IPC, essentiellement en raison de l'asymétrie de la distribution des dépenses.

---

\*/ Élaborée par M. Jan de Haan, Département des prix à la consommation, Statistics Netherlands, B.P. 4000, 2270 JM Voorburg (Pays-Bas) - courrier électronique : jhln@cbs.nl. Le présent document est fondé sur De Haan *et. al.* (1999). M. A.C. Nielsen (Pays-Bas) a aimablement fourni des données saisies par lecture optique, pour une somme modique.

## Introduction

1. Dans la plupart des pays, y compris les Pays-Bas, l'indice des prix à la consommation (IPC) est essentiellement un indice du type Laspeyres, selon lequel les indices des prix élémentaires sont pondérés par les parts respectives de la consommation en valeur des divers produits, fixées aux niveaux de la période de base. Des méthodes d'échantillonnage sont nécessaires pour estimer la valeur de l'ensemble. Idéalement, l'erreur quadratique moyenne de l'estimateur devrait être minimale. Même si la formule de Laspeyres est extrêmement simple, les méthodes d'estimation appliquées à l'IPC en font une statistique assez complexe. Schématiquement, la méthode d'estimation consiste à procéder à un triple tirage d'échantillons, en trois étapes consécutives. Premièrement, un échantillon de ménages participant à une enquête sur le budget est constitué en vue d'estimer les pondérations des groupes de produits. Deuxièmement, pour chaque groupe de produits, un échantillon de produits (appelés "articles") est sélectionné. Troisièmement, les prix de ces articles sont relevés dans un échantillon de points de vente.

2. Le présent document porte principalement sur la deuxième étape. Seuls quelques organismes statistiques, par exemple le Bureau of Labor Statistics des États-Unis, choisissent par échantillonnage aléatoire les articles dont les prix seront pris en compte dans l'indice. La plupart des autres organismes s'appuient sur l'avis d'experts en vue de déterminer quels sont les produits qui devraient représenter le groupe d'articles. Les méthodes de prélèvement aléatoire ne donnent pas forcément de meilleurs résultats que les méthodes non aléatoires. En fait, le présent document fait valoir qu'un échantillonnage basé sur un seuil d'inclusion aux fins du calcul de l'IPC, sur le modèle de l'échantillonnage au jugé pratiqué aux Pays-Bas, constitue une méthode efficace, notamment en cas de distribution asymétrique des dépenses. Les résultats d'expériences réalisées au moyen de données saisies par lecture optique sur le café, les couches pour bébés et le papier toilette illustrent cette argumentation <sup>1</sup>.

3. La section 2 du document soutient une description de l'échantillonnage fondé sur un seuil d'inclusion et de trois méthodes d'échantillonnage probabiliste : l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage proportionnel à la taille. La section 3 donne une vue d'ensemble des séries de données saisies par lecture optique utilisées pour les estimations empiriques. La section 4 présente les résultats d'expériences effectuées selon la méthode de Monte-Carlo pour déterminer l'erreur quadratique moyenne (mesure type de la précision statistique) des indices des prix de groupes de produits estimés en fonction des divers plans d'échantillonnage. La section 5 fait état des conclusions.

## Estimation des indices des prix du type Laspeyres

4. Supposons que le groupe d'articles A comprend  $N$  articles;  $g \in A$  signifie que l'article  $g$  appartient au groupe A. Le groupe A est censé ne pas varier dans le temps. Même s'il arrive en réalité que certains groupes de produits disparaissent du marché et que de nouveaux produits fassent leur apparition, l'hypothèse selon laquelle le groupe d'articles reste constant nous permet de nous concentrer sur l'aspect échantillonnage. Notre ensemble de données sera

ajusté en conséquence. L'indice des prix du type Laspeyres (à pondération fixe) du groupe d'articles A pour la période t est défini comme suit :

$$P^t = \frac{\sum_{g \in A} e_g^0 P_g^t}{\sum_{g \in A} e_g^0} = \sum_{g \in A} w_g^0 P_g^t, \quad (1)$$

où  $P_g^t$  désigne l'indice des prix de l'article g (qui est censé être connu),  $e_g^0$  la dépense consacrée à g durant la période de base 0 et  $w_g^0$  la part des dépenses que représente g au sein du groupe A. Durant la période de base, un échantillon  $\hat{A}$  de taille fixe n est prélevé dans le groupe A en vue d'estimer  $P^t$ . Puisque l'on suppose que A ne varie pas dans le temps, il semble logique que  $\hat{A}$  reste fixe également.

#### Échantillonnage aléatoire simple

5. On parle d'échantillonnage probabiliste (ou aléatoire) dans les cas où tous les échantillons possibles ont une probabilité connue (et positive) d'être sélectionnés. Dans l'échantillonnage aléatoire simple (sans remplacement), tous les échantillons possibles ont une probabilité égale de sélection. Par conséquent, l'estimateur d'Horvitz-Thompson  $\hat{P}_A^t = (N/n) \sum_{g \in \hat{A}} w_g^0 P_g^t$  est sans biais pour  $P^t$ , c'est-à-dire que  $E(\hat{P}_A^t) = P^t$  où l'espérance mathématique  $E(.)$  désigne la moyenne de tous les échantillons qui peuvent être constitués avec un plan d'échantillonnage donné, en l'occurrence un échantillonnage aléatoire simple. Malgré son absence de biais,  $\hat{P}_A^t$  ne sera jamais utilisé parce qu'il présente deux particularités peu souhaitables. S'il se trouvait que les indices des prix de tous les articles sélectionnés soient égaux, la valeur estimée de l'indice des prix du groupe d'articles différerait de cette valeur, à moins que la moyenne de la population et la moyenne des dépenses de l'échantillon ne coïncident. Il est probable que cette particularité déplaît aux statisticiens chargés d'établir l'indice des prix. D'autre part - et cet inconvénient est plus important encore -  $\hat{P}_A^t$  présente nécessairement une variance d'échantillonnage extraordinairement grande. Pour surmonter ces deux difficultés, l'estimation de  $P^t$  est établie en utilisant des estimateurs sans biais du numérateur et du dénominateur :

$$\hat{P}_B^t = \frac{(N/n) \sum_{g \in \hat{A}} e_g^0 P_g^t}{(N/n) \sum_{g \in \hat{A}} e_g^0} = \sum_{g \in \hat{A}} \hat{w}_g^0 P_g^t, \quad (2)$$

où  $\hat{w}_g^0$  représente la part des dépenses correspondant à l'article g dans l'échantillon. Avec une linéarisation de Taylor du premier ordre (Särndal et al., 1992, p. 172 à 176), la variance de  $\hat{P}_B^t$  peut être estimée à partir des données de l'échantillon. Cependant, la linéarisation de Taylor conduit généralement à une sous-estimation de la variance pour les petits échantillons. Or, les échantillons d'articles utilisés pour le calcul de l'IPC sont le plus souvent restreints. Pour certains groupes de produits, ils se composent même parfois d'un seul ou de deux articles représentatifs. Ainsi, outre qu'elle est instable (c'est-à-dire qu'elle présente elle-même une grande variance), la variance est sans doute aussi sous-estimée lorsqu'elle repose sur une linéarisation de Taylor.

6. L'estimateur  $\hat{P}_B^t$  est entaché d'un petit biais d'échantillonnage avoisinant  $o(1/n)$ . Si l'échantillon d'articles est retreint et la variabilité des dépenses de la période de base importante, le biais  $\hat{P}_B^t$  peut se révéler non négligeable par rapport à son erreur type. Il ne faut pas pour autant en conclure que l'IPC pour l'ensemble des articles sera considérablement biaisé, puisque le biais est une moyenne (pondérée) des biais positifs et négatifs des indices des différents groupes de produits.

#### Échantillonnage proportionnel à la taille

7. L'échantillonnage proportionnel à la taille offre l'avantage suivant : les articles les plus importants ont la plus forte chance d'être inclus dans l'échantillon. En cas d'échantillonnage proportionnel à la taille, on choisira très vraisemblablement un plan d'échantillonnage fondé sur un échantillon d'effectif fixe sans remplacement, que l'on combinera avec un estimateur d'Horvitz-Thompson. On parle alors parfois d'échantillonnage *pps*. Les dépenses de la période de base font office de mesure de la taille, et la probabilité d'inclusion du premier ordre requise pour l'article  $g$  est  $ne_g^0 / \sum_{g \in A} e_g^0 = nw_g^0$ . Ainsi,  $\sum_{g \in \hat{A}} P_g^t / n$  est un estimateur sans biais de  $P^t$ .

8. Comme la plupart des systèmes d'échantillonnage proportionnel à la taille fondés sur un échantillon d'effectif fixe sont assez compliqués, on utilise plutôt une version systématique. Dans le cas des groupes d'articles, pour lesquels on constate une importante variabilité des dépenses de la période de base, il ne s'avère pas toujours possible de constituer un échantillon strictement proportionnel à l'ampleur des dépenses. Pour remédier à cette difficulté, on procède comme suit <sup>2</sup>. Premièrement, un sous-groupe  $A_H$  d'effectif  $N_H (< n)$  correspondant aux dépenses les plus importantes de la période de base est constitué à partir du groupe  $A$  avec certitude. Ensuite, un échantillon  $\hat{A}_L$  d'effectif  $n_L (= n - N_H)$  est prélevé, en veillant à ce qu'il soit exactement proportionnel à la dépense, dans le sous-groupe restant  $A_L$  correspondant aux dépenses de faible ampleur. L'estimateur sans biais qui en résulte représente une moyenne pondérée en fonction des dépenses de  $P^t(H) = \sum_{g \in A_H} e_g^0 P_g^t / \sum_{g \in A_H} e_g^0$ , le véritable indice des prix de  $A_H$ , et  $\sum_{g \in \hat{A}_L} P_g^t / n_L$ , la valeur estimée de l'indice des prix de  $A_L$ .

#### Échantillonnage stratifié

9. L'échantillonnage aléatoire simple présente un avantage évident par rapport à l'échantillonnage proportionnel à la dépense, à savoir qu'en dehors d'un registre d'articles faisant office de base d'échantillonnage, il n'exige pas d'autres données. Voir également Balk (1994). Lorsque la distribution des dépenses par article est assez inégale, il existe une forte probabilité que les produits les plus populaires soient exclus de l'échantillon ce qui, intuitivement, ne paraît pas souhaitable. Une réduction de la variance pourrait être obtenue si nous parvenions à stratifier le groupe d'articles en sous-groupes homogènes en fonction de leurs variations de prix. Toutefois, on ne sait pas a priori comment les prix vont varier. Une autre solution consiste à réduire la variance en stratifiant le groupe de produits en deux sous-groupes, l'un ( $A_H$ ) correspondant aux dépenses élevées de la période de base que l'on observe intégralement, et l'autre ( $A_L$ ) correspondant aux faibles dépenses, dans lequel on prélève un échantillon aléatoire  $\hat{A}_L$ . Le nouvel estimateur de l'indice des prix du groupe de produits est une moyenne pondérée

en fonction des dépenses de  $\hat{P}'_B(L)$ , c'est-à-dire l'indice des prix du type Laspeyres pour le sous-groupe d'articles correspondant aux faibles dépenses estimé selon la formule (2) et  $P'(H)$ . Sa variance d'échantillonnage est  $(1 - \tau_H)^2 \text{Var}[\hat{P}'_B(L)]$ , où  $\tau_H$  représente la part des dépenses correspondant au sous-groupe  $A_H$  à l'intérieur du groupe A. Si cette méthode ne réduit pas nécessairement la variance de l'indice estimé des prix, elle permettra vraisemblablement d'obtenir un tel résultat à mesure que l'on augmentera la taille globale de l'échantillon  $n$ .

10. Le choix de  $\tau_H$  et, par conséquent, de la taille  $N_H$  de la strate "à caractère exhaustif"  $A_H$  est un peu problématique. De préférence, il nous faudrait avoir un critère quelconque d'optimalité en vue de réduire au minimum la variance. Toutefois, puisqu'on ignore à priori comment les prix des articles vont varier et puisque les tendances enregistrées dans le passé ne permettent pas de prévoir avec beaucoup de précision les variations futures des prix, la taille optimale de  $A_H$  peut difficilement être calculée dans la pratique. Dans le cadre de l'analyse empirique, deux tailles relatives différentes d'échantillons  $\lambda_H = N_H/n$  de  $A_H$  seront testées, à savoir  $\lambda_H = 1/3$  et  $\lambda_H = 2/3$ . Ces valeurs devraient suffire pour avoir une idée claire de l'efficacité de la méthode.

#### Échantillonnage basé sur un seuil d'inclusion

11. Lorsque l'échantillon est de très petite taille, il semble assez probable que la stratification fondée sur  $\lambda_H = 2/3$  conduise à une erreur type plus importante dans la valeur estimée de l'indice des prix qu'une stratification fondée sur  $\lambda_H = 1/3$ . Mais que se passe-t-il si le sous-groupe  $A_L$  n'est pas du tout observé, de sorte que  $\lambda_H = 1$  et que, par conséquent  $n = N_H$  ? Nous utiliserions alors (une forme particulière d') *échantillonnage basé sur un seuil d'inclusion*. L'indice des prix du groupe d'articles est estimé simplement par la formule  $\hat{P}'_C = P'(H)$ . Tous les articles  $g \in A_H$  ont maintenant une probabilité d'inclusion égale à 1, tandis que tous les articles  $g \in A_L$  ont une probabilité d'inclusion nulle (Särndal et al., 1992, p. 531 à 533). Puisque nous savons exactement quels seront les articles sélectionnés, le processus ne présente aucun caractère aléatoire et la variance d'échantillonnage de  $\hat{P}'_C$  est égale à zéro par définition. Le biais est égal à l'erreur réelle, étant la différence entre la valeur estimée et la valeur de la population :  $\hat{P}'_C = (1 - \tau_H)[P'(H) - P'(L)]$ . Avec une distribution très inégale des dépenses par article, on obtiendra une valeur importante pour  $\tau_H$  même si l'effectif de l'échantillon est restreint. Dans ce cas, un échantillonnage basé sur un seuil d'inclusion peut donner de meilleurs résultats qu'une stratification, du point de vue de l'erreur quadratique moyenne. On peut soit fixer le niveau d'inclusion  $\tau_H$ , qui détermine alors l'effectif de l'échantillon  $n$ , ou fixer l'effectif de l'échantillon, auquel cas  $\tau_H$  dépend du choix de  $n$ . J'ai opté pour cette dernière solution parce que l'emploi de plans de sondage fondés sur un échantillon à effectif fixe est pratique courante pour le choix des articles pris en compte dans l'IPC, et parce que cette formule permet de procéder à une comparaison valable avec d'autres plans reposant sur le même principe.

12. L'utilisation de méthodes d'échantillonnage basé sur un seuil d'inclusion peut être justifiée en arguant que i) des considérations de coût empêchent d'établir une base d'échantillonnage fiable pour l'ensemble de la

population, et ii) le biais est jugé négligeable. Évidemment, l'hypothèse ii) ne peut pas être vérifiée de manière générale. L'exclusion voulue d'une partie de la population cible lors du prélèvement des unités composant l'échantillon peut néanmoins donner des résultats satisfaisants lorsque les corrections appropriées sont effectuées. Toutefois, la méthode fondée sur un seuil d'inclusion qui est appliquée au choix des articles considérés dans l'IPC ne comporte pas d'ajustement tenant compte des articles exclus. Outre des considérations de coût, cette méthode est parfois préconisée en partant du principe qu'à long terme du moins, les variations des prix des articles les moins importants ne diffèrent guère de celles des produits les plus populaires dans le même groupe de produits parce que la structure de leurs coûts de production est semblable.

#### **Données saisies par lecture optique d'un code-barre**

13. En Europe, les produits pourvus d'une étiquette déchiffrable par lecture optique sont désignés en fonction du système de numérotation européenne des articles (NEA). Les fabricants attribuent un code NEA différent à chaque variété, format, type de conditionnement, etc., d'un produit. Cette politique a deux conséquences importantes. Premièrement, il est difficile en raison des modifications fréquentes des NEA de suivre un article donné dans le temps. Deuxièmement, on constate que certains NEA correspondent à des dépenses négligeables. Il semblerait que le système de classification soit trop détaillé : ce qui constitue en réalité un seul article est classé comme une multitude d'articles, et il est nécessaire, dans la pratique, de procéder à des agrégations de données en regroupant plusieurs NEA<sup>3</sup>. Heureusement, diverses caractéristiques des produits comme la marque figurent dans les données du NEA saisies par lecture optique. On considérera donc que les NEA indiquant les mêmes caractéristiques de produits désignent des articles identiques. Bien entendu, si le nombre de caractéristiques indiquées est insuffisant, il existe un risque de suragrégation, c'est-à-dire de regroupement d'articles hétérogènes.

14. Les séries de données recueillies par lecture optique qui ont été exploitées dans la présente étude portent sur les ventes hebdomadaires dans les supermarchés de café, de couches jetables pour bébés et de papier toilette, et, au départ, elles couvraient respectivement 320, 569 et 294 NEA différents. Elles comprennent, pour chaque NEA, le nombre de paquets vendus et la valeur correspondante. Les prix ne figurent pas expressément dans ces données. Les prix moyens par achat (valeurs unitaires) sont calculés à partir des valeurs et des quantités. Les données sur le café se rapportent aux ventes effectuées au cours d'une période de deux ans et demi, à partir de la première semaine de 1994 et jusqu'à la vingt-quatrième semaine de 1996, dans un échantillon de 20 supermarchés. Les données relatives aux deux autres groupes d'articles se rapportent à un échantillon de 149 supermarchés et s'étendent sur une période de deux ans, débutant à partir de la première semaine de 1995 et s'achevant à la fin de la cinquante-deuxième semaine de 1996.

15. Pour des raisons de commodité, les marques les moins importantes ont été éliminées. Dans le cas du café, les 15 marques qui représentaient les plus grosses ventes sur l'ensemble de la période considérée ont été choisies parmi les 55 marques effectivement vendues. Après une agrégation des données issues des NEA correspondant à des caractéristiques de produits identiques,

nous avons encore restreint la population cible aux articles vendus durant l'année de base 1994 et, tous les mois par la suite, afin d'obtenir une série complète de données pour chaque mois. Nous avons terminé avec un total de 68 articles (en excluant les cafés en grains), dont 40 articles se composent de café moulu et 28 articles de café instantané. Les articles en question représentent une part de 94,5 % des achats totaux de café durant l'année de base, dans la série initiale de données. Pour ce qui est des couches pour bébés et du papier toilette, les marques représentant moins de 1 % de la valeur des ventes ont été éliminées. Ensuite, on a sélectionné uniquement les articles qui avaient été vendus en 1995 et pendant au moins huit mois par la suite <sup>4</sup>. En définitive, on a retenu 58 articles dans le cas des couches pour bébés et 70 articles pour ce qui est du papier toilette, représentant respectivement 90 % et 86 % du total des dépenses pour 1995 dans les séries initiales de données.

Figure 1 : Distribution des dépenses de la période de base

16. Ce qui frappe dans les dépenses par article, c'est le caractère asymétrique de leur distribution. La figure 1 montre l'inégalité de la distribution des dépenses durant la période de base, telle qu'elle ressort des séries de données ajustées, à l'aide de courbes de Lorenz. L'axe vertical indique le total cumulé des dépenses, l'axe horizontal le nombre cumulé des articles, tous deux exprimés en pourcentage. Les articles ont été classés par ordre croissant de dépenses. Si les dépenses étaient également distribuées, la courbe de Lorenz suivrait la diagonale. Plus la distribution est inégale, plus basse est la courbe. Les dépenses consacrées à l'achat de café semblent être distribuées de manière extrêmement asymétrique. Les trois principaux articles représentent plus de la moitié des dépenses totales en café au cours de l'année de base (1994). Dans le cas des couches pour bébés et du papier toilette, les sixième et huitième articles, respectivement, par ordre d'importance, représentent près de la moitié des dépenses totales de l'année de base (1995).

Résultats de l'application de la méthode de Monte Carlo

17. À l'exception du prélèvement basé sur un seuil d'inclusion, il est difficile de trouver une méthode de mesure fiable des distributions d'échantillonnage fondée sur un échantillon unique. Dans l'échantillonnage aléatoire simple, l'estimateur (2) est entaché d'un biais inconnu, tandis que l'estimation de la variance basée sur les techniques de linéarisation de Taylor donne des résultats inexacts à cause de l'effectif restreint des échantillons d'articles utilisés aux fins du calcul de l'IPC. Un échantillonnage systématique proportionnel à la taille soulève la question de savoir comment il convient d'estimer la variance puisqu'on ignore les probabilités d'inclusion du second ordre. Des simulations selon la méthode de Monte-Carlo ont été entreprises en vue de définir la distribution d'échantillonnage. Un demi-million d'échantillons ont été prélevés dans le groupe d'articles en question, selon le plan d'échantillonnage donné, et pour chaque échantillon on a procédé à l'estimation de l'indice des prix. La distribution des 500 000 estimations sera très proche de la distribution d'échantillonnage exacte <sup>5</sup>.

18. Des échantillons de trois tailles différentes ont été utilisées :  $n = 3$ ,  $n = 6$  et  $n = 12$ . Il convient de noter que les indices des prix de tous les articles ont été calculés en tant qu'indices de la valeur unitaire sur l'ensemble des points de vente considérés dans l'échantillon. La partie A du tableau 1 indique les résultats obtenus en appliquant la méthode de Monte-Carlo aux prix du café en janvier 1995 (1994 = 100) et les parties B et C font état des résultats concernant, respectivement, les prix des couches pour bébés et ceux du papier toilette, en janvier 1996 (1995 = 100). L'échantillonnage aléatoire simple donne des résultats particulièrement mauvais. Par exemple, pour  $n = 3$ , l'augmentation réelle du prix du café (selon la formule de Laspeyres) qui se chiffrait à 17,2 % est sous-évaluée de 1,4 point de pourcentage. En même temps qu'une erreur type de 5,1 points de pourcentage, on note que l'erreur quadratique moyenne s'élève à 5,3 points de pourcentage, c'est-à-dire presque un tiers de l'augmentation réelle des prix. Même en portant  $n$  à 12, de sorte que la fraction d'échantillonnage soit égale à 0,18 (ce qui serait anormalement élevé), l'erreur quadratique moyenne reste encore très substantielle. Veuillez noter que, comme on pouvait s'y attendre, le biais d'échantillonnage est réduit de moitié quand on double la taille de l'échantillon.

19. La stratification fonctionne assez bien avec les échantillons de plus grande taille mais donne des résultats décevants lorsque  $n = 3$ . Dans ce dernier cas, la stratification accentue l'erreur quadratique moyenne, par rapport à l'échantillonnage aléatoire simple, pour ce qui est des couches pour bébés et du papier toilette lorsque  $N_H = 2$  (c'est-à-dire lorsque  $\lambda_H = 2/3$ ). Notre plan d'échantillonnage aléatoire de prédilection serait sans aucun doute l'échantillonnage proportionnel à la dépense parce que les estimations obtenues ne sont pas biaisées et que l'erreur type est relativement faible. Mais la conclusion la plus intéressante concerne les bons résultats obtenus avec la méthode fondée sur un seuil d'inclusion. Sauf pour  $n = 3$  et  $n = 6$  dans le cas des couches pour bébés, cette méthode est celle qui donne les meilleurs résultats.

20. Il serait risqué de tirer des conclusions concernant l'efficacité des diverses méthodes d'échantillonnage sur la base des résultats d'un seul mois. Par conséquent, des expériences selon la méthode de Monte-Carlo ont été effectuées pour chaque mois de la période considérée. La figure 2 indique l'erreur quadratique moyenne que l'on obtient pour  $n = 3$ . La tendance qui se dégage pour le café et le papier toilette est assez nette : la méthode fondée sur un seuil d'inclusion s'avère toujours la meilleure. Apparemment, si la taille des échantillons est restreinte, l'exclusion des articles les moins importants ne semble pas avoir beaucoup d'incidence. On notera qu'avec des échantillons de plus grande taille, les résultats obtenus par la méthode de tirage fondé sur un seuil d'inclusion et la méthode de l'échantillonnage proportionnel à la taille sont très semblables. Dans le cas des couches pour bébés, les résultats diffèrent légèrement. En raison de la forte instabilité des indices des prix de cet article, l'erreur quadratique moyenne pour la méthode de prélèvement fondé sur un seuil d'inclusion varie sensiblement; elle tourne autour de l'erreur quadratique moyenne obtenue avec la méthode de l'échantillonnage proportionnel à la dépense.

### **Conclusion**

21. Bien que les données saisies par lecture optique présentent parfois quelques insuffisances, elles offrent une excellente occasion d'entreprendre des études empiriques sur les questions relatives à l'échantillonnage dans le cadre de l'IPC. Les simulations à l'aide de la méthode de Monte Carlo montrent que pour le café, les couches jetables pour bébés et le papier toilette au moins, un échantillonnage aléatoire simple des articles devrait être déconseillé. Je pense que cette recommandation peut également s'appliquer à tous les groupes d'articles pour lesquels la distribution des dépenses est très asymétrique. Les services de statistique qui souhaitent recourir à un échantillonnage aléatoire devraient envisager d'utiliser un échantillonnage proportionnel à la taille. Cependant, les résultats des expériences décrites dans le présent document plaident en faveur de l'utilisation d'une méthode de tirage des articles sur la base d'un seuil d'inclusion car elle représente une solution de rechange satisfaisante, voire préférable, dans le cas du calcul de l'IPC.

Tableau 1 : Estimations selon la méthode de Monte Carlo d'indices des prix de Laspeyres  
A : Café (N = 68), janvier 1995 (1994 = 100)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne
EAS */	115,7	5,1	-1,4	5,3	116,4	3,4	0,7	3,5	116,7	2,3	0,4	2,3
$\pi$ ps	117,2	2,2	0	2,2	117,2	1,3	0	1,3	117,2	0,7	0	0,7
Échantillonnage stratifié												
$\lambda_{H1} = 1/3$	116,4	3,9	0,7	4,0	116,6	2,3	0,5	2,3	117,0	1,2	0,1	1,2
$\lambda_{H1} = 2/3$	115,6	4,5	-1,5	4,7	116,4	2,5	0,7	2,6	117,0	1,1	0,2	1,1
Tirage fondé sur un seuil d'inclusion	117,0	0	0,2	0,2	117,2	0	0,0	0,0	117,5	0	0,3	0,3

\*/ Estimateur aléatoire simple (formule 2).

B : Couches pour bébés (N = 58), janvier 1996 (1995 = 100)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne
EAS	99,4	5,0	2,3	5,5	98,7	3,9	1,5	4,2	97,9	2,9	0,8	3,0
$\pi$ ps	97,2	2,8	0	2,8	97,2	1,6	0	1,6	97,2	1,5	0	1,5
Échantillonnage stratifié												
$\lambda_{H1} = 1/3$	98,9	5,0	1,8	5,3	98,1	3,3	1,0	3,4	97,4	1,7	0,2	1,7
$\lambda_{H1} = 2/3$	98,3	5,8	1,1	5,9	97,4	3,3	0,3	3,3	97,0	1,6	0,2	1,6
Tirage fondé sur un seuil d'inclusion	92,0	0	5,1	5,1	93,4	0	3,8	3,8	95,5	0	-1,6	1,6

C : Papier toilette (N = 70), janvier 1996 (1995 = 100)

Plan d'échantillonnage	n = 3				n = 6				n = 12			
	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne	valeur probable	erreur type	biais	erreur quadratique moyenne
EAS	103,9	4,5	0,1	4,5	103,9	3,5	0,1	3,5	103,9	2,6	0,1	2,6
$\pi$ ps	103,9	3,4	0	3,4	103,9	1,8	0	1,8	103,9	1,2	0	1,2
Échantillonnage stratifié												
$\lambda_{H1} = 1/3$	103,5	4,3	0,3	4,3	103,7	3,2	0,1	3,2	104,0	2,1	0,1	2,1
$\lambda_{H1} = 2/3$	103,7	4,6	0,2	4,6	104,2	3,4	0,4	3,4	103,9	1,6	0,0	1,6
Tirage fondé sur un seuil d'inclusion	105,0	0	1,1	1,1	104,0	0	0,1	0,1	104,0	0	0,1	0,1

Figure 2 : Erreur quadratique moyenne des indices des prix estimés selon la formule de Laspeyres (n = 3)

22. Dans le présent document, les articles ont été décrits avec précision pour obtenir un degré élevé d'"homogénéité", selon la pratique suivie par Statistics Netherlands, alors que certains pays utilisent plutôt des descriptions d'articles assez imprécises. En outre, on est parti de l'hypothèse que les indices des prix des articles étaient connus avec certitude, tandis qu'en réalité, ils sont calculés par estimation à partir des prix observés dans un échantillon de points de vente. Opperdoes (1999) traite du choix qui doit s'opérer entre une description précise et une description plus générale des articles et tient compte à la fois de l'échantillonnage des articles et de l'échantillonnage (aléatoire) des points de vente. Les résultats de cette étude préliminaire sont encore assez surprenants et peu concluants. Il serait souhaitable d'entreprendre d'autres recherches dans ce domaine important <sup>6</sup>.

### Références bibliographiques

Balk, B.M., 1994, On the First Step in the Calculation of a Consumer Price Index, *Proceedings of the Third International Conference on Price Indices* (Statistique Canada, Ottawa)

Bradley, R., B. Cook, S.G. Leaver et B.R. Moulton, 1997, An Overview of Research on Potential Uses of Scanner Data in the U.S. CPI. Dans : B.M. Balk (ed.), *Proceedings of the Third International Conference On Price Indices* (Statistics Netherlands, Voorburg)

Haan, J. de et M. Boon, 1998, Elementary Index Bias in the Consumer Price Index, *Research Paper* (Statistics Netherlands, Voorburg)

Haan, J. de, E. Opperdoes et C.M. Schut, 1999, Item Selection in the Consumer Price Index: Cut-off versus Probability Sampling, *Survey Methodology* (à paraître)

Opperdoes, E., 1999, Comparing Loose/tight Item descriptions in CPI Compilation Using Scanner Data, *Research Paper* (Statistics Netherlands)

Reinsdorf, M., 1995, Constructing Basic Component Indexes for the U.S. CPI from Scanner Data: A Test Using Data on Coffee, *Paper Presented at the NBER Conference on Productivity*, Cambridge, Massachusetts, 17 juillet 1995

Särndal, C.E., B. Swensson et J. Wretman, 1992, *Model Assisted Survey Sampling*. New York: Springer Verlag

Notes

1/ Les données en question proviennent d'un balayage électronique par lecteur de codes-barres. Bradley *et al.* (1997) donnent un aperçu général des utilisations potentielles des données obtenues par lecture optique dans le calcul de l'IPC.

2/ Pour une description précise de la méthode suivie, voir De Haan *et al.* (1999).

3/ Lors d'une étude expérimentale reposant sur des données recueillies par lecture optique sur le café, Reinsdorf (1995) a également constaté que "des articles qui, à toutes fins utiles, sont identiques peuvent occasionnellement être affectés de différents UPC" (UPC = code de produit universel employé par les États-Unis).

4/ Pour un nombre restreint de (petits) articles, il a fallu procéder à l'imputation des données manquantes pour certains mois.

5/ Pour plus de détails, le lecteur est de nouveau prié de se reporter à De Haan *et al.* (1999).

6/ Le degré de "précision" des descriptions d'articles et le plan d'échantillonnage des points de vente ont tous deux des incidences sur la formule selon laquelle les prix sont agrégés en indices des prix au niveau le plus détaillé d'agrégation des produits. Voir par exemple De Haan et Boon (1998).

-----