

Article

Le choix des produits pour l'indice des prix à la consommation : le seuil d'inclusion par opposition au sondage probabiliste

par Jan de Haan, Eddy Opperdoes et Cecile M. Schut

Juin 1999



Le choix des produits pour l'indice des prix à la consommation : le seuil d'inclusion par opposition au sondage probabiliste

Jan de Haan, Eddy Opperdoes et Cecile M. Schut¹

Résumé

La plupart des bureaux de la statistique utilisent des techniques non probabilistes pour choisir l'échantillon de produits dont les prix permettent de calculer les indices des prix à la consommation. Aux Pays-Bas, comme dans beaucoup d'autres pays, ce genre de sondage raisonné se rapproche en quelque sorte de la sélection par seuil d'inclusion, une bonne partie de la population (normalement les produits suscitant le moins de dépenses) étant délibérément exclue des observations. Bien sûr, cette méthode donne lieu à des chiffres biaisés pour l'indice des prix. On peut se demander si un échantillonnage probabiliste donnerait de meilleurs résultats quant à l'erreur quadratique moyenne. Les auteurs ont considéré l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage systématique proportionnel aux dépenses. Ils ont mené des simulations de Monte Carlo à l'aide de données de lecture optique pour le café, les couches de bébés et le papier hygiénique afin d'évaluer le rendement des quatre plans d'échantillonnage. Il est assez surprenant de constater que la sélection par seuil d'inclusion est une bonne stratégie d'échantillonnage des produits pour l'indice des prix à la consommation.

Mots clés : Indice de prix de Laspeyres; simulation de Monte Carlo; échantillonnage; données de lecture optique; biais de substitution.

1. Introduction

Les non-initiés pourraient croire que la mesure de l'inflation est une tâche facile : il suffit de visiter des magasins, de recueillir une foule de prix et d'en établir la moyenne. Toutefois, les statisticiens chargés de calculer l'indice des prix à la consommation (IPC), la mesure la plus répandue de l'inflation, doivent surmonter de nombreux problèmes théoriques et pratiques. Dans la plupart des pays, l'IPC est essentiellement un indice de prix de Laspeyres. Cet indice établit le poids des indices de prix partiels des divers produits selon des proportions de dépenses fixes pour une période de référence. Il faut des procédures d'échantillonnage pour estimer la valeur de la population. Idéalement, l'erreur quadratique moyenne de l'estimateur est minimisée. La formule de l'indice de Laspeyres est extrêmement simple, mais les procédures d'estimation appliquées à l'IPC en font une statistique relativement complexe. Décrite de façon stylisée, l'estimation comporte trois types d'échantillons. Un échantillon des ménages participant à l'enquête sur les dépenses permet d'estimer le poids des dépenses par groupe de produits. Pour chaque groupe de produits, on choisit un échantillon de marchandises (produits). On recueille le prix de ces articles pour un échantillon de points de vente.

Les auteurs du présent article se penchent sur l'échantillonnage des produits. Il existe seulement quelques bureaux de la statistique, par exemple le Bureau of Labor Statistics des États-Unis, qui utilisent un échantillonnage probabiliste pour choisir les produits dont le prix sera calculé. La plupart des autres bureaux, y compris Statistique Pays-Bas, font

appel au raisonnement de spécialistes travaillant au bureau central afin de déterminer quels produits devraient représenter le groupe de marchandises. Dans le passé, on a pu justifier cette méthode en soulignant l'absence de bases de sondage appropriées. Compte tenu cependant de l'automatisation de plus en plus poussée du commerce de détail, des registres de biens de consommation sont de plus en plus accessibles, de sorte qu'un échantillonnage probabiliste est possible. Toutefois, avant de passer à une nouvelle stratégie d'échantillonnage, il est bon d'examiner d'autres possibilités afin d'évaluer leur effet sur l'exactitude des indices de prix estimatifs. Il s'agit de déterminer si les méthodes actuelles de sélection non probabiliste donnent des résultats qui, relativement à l'erreur quadratique moyenne, sont inférieurs à ceux des techniques probabilistes. C'est là le thème principal du présent article. Des études de simulation ont été menées pour trois groupes de produits, c'est-à-dire le café, les couches jetables de bébés et le papier hygiénique.

Il n'y a pas si longtemps, le calcul empirique des indices de prix était d'autant plus difficile que les données hautement désagrégées sur les dépenses et les quantités au niveau des points de vente étaient inexistantes ou tout au plus accessibles pour de petits échantillons. De nos jours, des compagnies d'études de marché ont réussi à établir de vastes bases de microdonnées sur la vente de biens de consommation, en particulier dans le secteur des produits de grande consommation. Il s'agit de données obtenues par balayage électronique (lecture optique d'un code à barres ou inscription d'un code à barres à la caisse). Bradley, Cook, Leaver et Moulton (1997) présentent un aperçu des applications possibles des données de lecture optique pour le calcul de l'IPC. Le traitement de grandes bases de

1. Jan de Haan, Eddy Opperdoes et Cecile M. Schut, Division Socio-economic, Statistics, Statistics Netherlands, P.O. Box 4000, 2270 JM, Voorburg, The Netherlands. Courriel : jhnh@cbs.nl.

données de lecture optique est une tâche assez fastidieuse. Pour ce qui est du calcul de l'IPC, c'est peut-être un obstacle à son application pour le moment. Toutefois, les données de lecture optique représentent certainement une riche source de renseignements pour les analyses empiriques. Les données de lecture optique permettent non seulement d'examiner l'échantillonnage, mais aussi de calculer les indices de prix en fonction de différentes formules. L'indice de prix de Laspeyres (poids fixe) ne tient pas compte des réactions des ménages aux changements de prix. Les auteurs ont donc examiné le degré de biais des indices de prix de Laspeyres (population) pour les trois groupes de produits relativement à l'indice de prix de Fisher, qui lui tient compte de la substitution des produits.

On trouvera à la section 2 un aperçu des données de lecture optique utilisées par les auteurs. À la section 3, il est question de quatre plans d'échantillonnage de produits. Trois d'entre eux (l'échantillonnage aléatoire simple, l'échantillonnage stratifié et l'échantillonnage proportionnel à la taille) relèvent de techniques probabilistes, tandis que le quatrième (échantillonnage par seuil d'inclusion) relève d'une méthode raisonnée qui reflète la stratégie officielle des Pays-Bas. La section 4 décrit les expériences de Monte Carlo utilisées par les auteurs pour déterminer l'exactitude des indices de prix des groupes de produits estimés à l'aide des plans d'échantillonnage mentionnés. La section 5 traite de l'utilisation des indices de Fisher pour des produits et des groupes de produits. Il y est question du biais de substitution intra-groupe des indices de prix de Laspeyres (groupes de produits). À la section 6, enfin, les résultats sont résumés et discutés.

2. Données de lecture optique (code à barres)

2.1 Aperçu

En Europe, les produits lisibles optiquement sont définis en fonction du numéro européen des articles (NEA). Les fabricants devraient attribuer un seul NEA à chaque variété, taille et type d'emballage d'un produit. Il en résulte deux phénomènes. Premièrement, les NEA changent parfois très rapidement, par exemple à cause d'un nouvel emballage. Bien sûr, il est alors difficile de suivre l'évolution d'un article. Deuxièmement, certains NEA comportent des données négligeables. Il semble que le système de classification soit trop détaillé; ce qui représente en réalité un article donne lieu à une multitude d'articles classés. Dans une étude fondée sur des données de lecture optique pour le café, Reinsdorf (1995) a trouvé que des articles pratiquement identiques portaient parfois un CUP (code universel des produits des États-Unis) différent. Une certaine agrégation relativement aux NEA est nécessaire. Heureusement, les séries de données de lecture optique englobent plusieurs caractéristiques de produits comme le nom commercial et le sous-nom. Les NEA comportant des

caractéristiques de produits identiques sont traités ici comme un seul article. Si le nombre de caractéristiques est insuffisant, il existe bien sûr un risque de suragrégation, c'est-à-dire de regroupement d'articles hétérogènes.

De la société A.C. Nielsen (Nederland), nous avons reçu des séries de données de lecture optique sur les chiffres de vente hebdomadaires du café, des couches jetables de bébés et du papier hygiénique dans des supermarchés. Les séries de données initiales comportaient 320, 569 et 294 NEA, respectivement. Pour chaque NEA, nous avons le nombre d'emballages vendus et la valeur correspondante, de même que plusieurs caractéristiques des produits et des points de vente. Les prix n'étaient pas indiqués; les prix moyens (valeurs unitaires) ont dû être calculés à partir des valeurs et des quantités. Les données sur la vente de café couvraient une période de deux ans et demi, depuis la première semaine de 1994 jusqu'à la vingt-quatrième semaine de 1996, pour un échantillon de 20 supermarchés situés dans une région urbaine néerlandaise non connue des auteurs. Les données sur la vente des deux autres produits se rapportaient à un échantillon de 149 magasins répartis dans l'ensemble du pays et couvraient une période de deux ans, depuis la première semaine de 1995 jusqu'à la cinquante-deuxième semaine de 1996 inclusivement.

Par souci de simplification, nous avons supprimé les marques moins importantes. Dans le cas du café, seules les 15 marques se vendant le mieux durant toute la période d'observation ont été retenues parmi les 55 marques vendues. Après l'étape d'agrégation des NEA comportant des caractéristiques de produits identiques, nous avons limité la population aux articles vendus au cours de l'année de référence 1994 et durant chaque mois par après de façon à constituer une série de données complètes pour chaque mois. Nous avons ainsi obtenu 68 produits (à l'exclusion des graines de café), dont 40 produits de café moulu (y compris le café décaféiné) et 28 articles de café instantané, représentant 94,5 % des dépenses totales en café pour l'année de référence dans le fichier initial. Pour ce qui est des couches et du papier hygiénique (à l'exclusion du papier hygiénique humide), nous avons éliminé les marques dont le chiffre des ventes représentait moins de 1 % du total. De plus, nous avons retenu uniquement les articles vendus pendant l'année de référence 1995 et au moins huit mois par après. Nous avons ainsi obtenu 58 produits de couches et 70 produits de papier hygiénique, représentant 90 % et 86 % respectivement du total des dépenses de 1995 dans les fichiers initiaux.

2.2 Statistiques descriptives

L'aspect le plus surprenant des articles de dépense est l'asymétrie de la distribution. La figure 1 indique l'inégalité des dépenses de la période de référence dans nos fichiers de données rajustés à l'aide de courbes dites de Lorenz. Le total cumulatif des dépenses figure en ordonnée, tandis que le nombre cumulatif de produits figure en abscisse, en pourcentage dans les deux cas. Les produits

sont triés dans l'ordre croissant des dépenses. En cas de dépenses égales, les courbes de Lorenz décrivent une diagonale. Plus la répartition est inégale et plus sa position est inférieure. La distribution des articles de dépense pour le café est très inégale, les trois articles les plus importants représentant plus de la moitié du total des dépenses en café pour l'année de référence (1994). Pour ce qui est des couches de bébés et du papier hygiénique, les six et huit articles les plus importants, respectivement, représentent près de la moitié du total des dépenses pour l'année de référence (1995).

La figure 2 fournit des chiffres d'indice de valeurs unitaires, c'est-à-dire le changement de valeur par emballage, peu importe la quantité, le nom commercial, le type, *etc.* pour tous les points de vente. On y trouve une première indication du changement de « prix » au cours de la période à l'étude. Pour le café, une baisse remarquable a été observée au cours du deuxième semestre de 1995 à la suite d'une forte hausse des prix survenue en 1994 à cause de mauvaises récoltes au Brésil. Le prix du café est largement déterminé par le prix des graines de café sur le marché mondial. Nous n'avons trouvé aucune différence appréciable de changement de prix d'un point de vente à l'autre. Les couches de bébés font exception à cet égard. Une forte concurrence sévissait entre les différents producteurs (ce qui explique peut-être le déclin des valeurs unitaires en 1996), tandis que les rabais et autres mesures spéciales étaient fréquents. Par conséquent, la valeur unitaire pour tous les produits et points de vente donne un aperçu inexact du changement de prix agrégé des couches de bébés.

3. Estimation des indices de prix de type Laspeyres

Nous présentons d'abord des éléments de notation. Soit A , le groupe de marchandises constitué d'un nombre fini N de marchandises (produits); $g \in A$ signifie que le produit g appartient au groupe A . A est supposé fixe dans le temps. En réalité, ce n'est pas vrai : certains produits disparaissent du marché, tandis que d'autres y apparaissent. À court terme, toutefois, l'hypothèse de groupes de produits constants semble raisonnable. À noter que notre fichier de données initial a été rajusté en conséquence, car nous voulions considérer uniquement l'aspect de l'échantillonnage. L'indice de prix de Laspeyres (poids fixe) pour le groupe de marchandises A au cours de la période t est

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

où P_g^t désigne l'indice de prix du produit g , e_g^0 les dépenses pour g au cours de la période de référence 0 et w_g^0 la part de dépenses correspondante de g au sein du groupe de produits A . Au cours de la période de référence, un échantillon \hat{A} de taille fixe n est tiré de A . Puisque A est supposé fixe dans le temps, il semble naturel que \hat{A} soit fixe également.

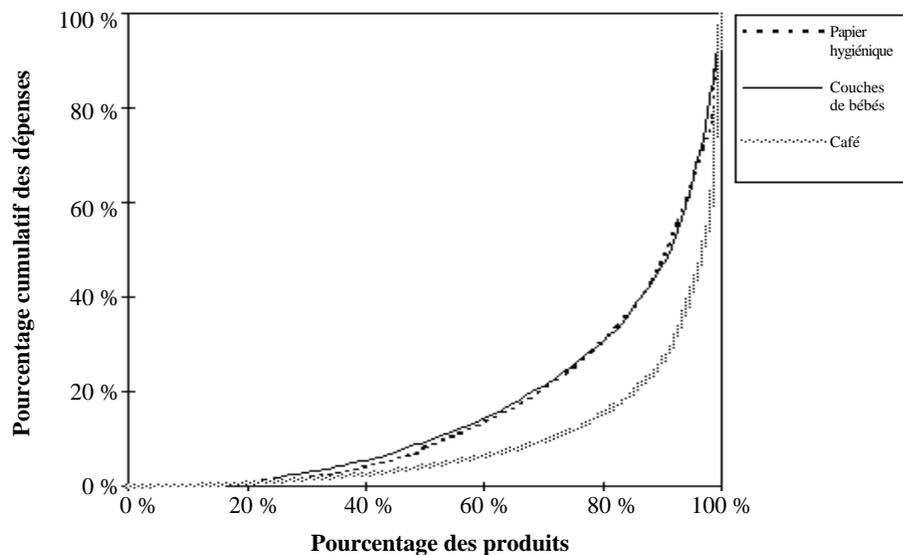


Figure 1 Distribution des articles de dépense pour l'année de référence

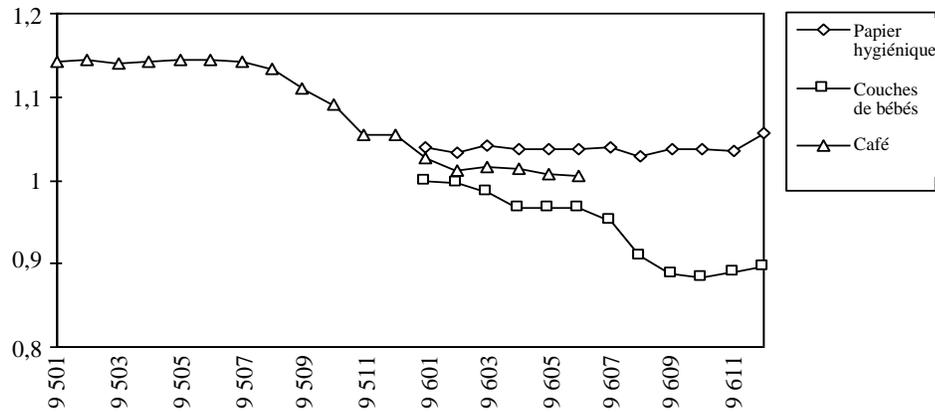


Figure 2 Chiffres d'indice de valeurs unitaires (1994 = 100 pour le café, 1995 = 100 pour les couches de bébés et le papier hygiénique)

3.1 Échantillonnage aléatoire simple

L'échantillonnage probabiliste renvoie à des situations comportant une probabilité connue de sélection de tous les échantillons éventuels. Pour l'échantillonnage aléatoire simple (sans remise), il existe une probabilité égale de sélection de tous les échantillons éventuels. L'estimateur de Horvitz-Thompson $\hat{P}_A^t = (N/n) \sum_{g \in \hat{A}} w_g^0 P_g^t$ n'est pas biaisé pour P^t , c'est-à-dire que $E(\hat{P}_A^t) = P^t$ où l'espérance $E(\cdot)$ désigne la moyenne de tous les échantillons éventuels pour un plan d'échantillon donné, dans le cas présent un échantillonnage aléatoire simple. Malgré son caractère non biaisé, \hat{P}_A^t ne sera pas utilisé dans la pratique à cause de deux propriétés indésirables. Premièrement, si les indices de prix de tous les produits échantillonnés sont égaux, l'indice du groupe de produits estimatif diffère de cette valeur-là, à moins que la population et les moyennes empiriques des dépenses ne coïncident. Les personnes qui calculent les indices de prix n'apprécient probablement pas cet aspect. Deuxièmement, et surtout, \hat{P}_A^t manifeste inévitablement une variance d'échantillonnage extraordinairement grande. Afin de surmonter ces deux difficultés, on prépare une estimation de P^t en établissant des estimateurs non biaisés 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 est la part de dépenses du produit g de l'échantillon. À l'aide d'une linéarisation de Taylor de premier ordre (Särndal, Swensson et Wretman 1992, pages 172-176), on peut écrire la variance de \hat{P}_B^t sous la forme

$$V(\hat{P}_B^t) \approx V(\hat{P}_A^t) - (P^t)^2 (2\mu^t \rho^t - 1) (\sigma^0)^2,$$

où σ^0 désigne le coefficient de variation ou l'erreur type relative de la moyenne empirique des dépenses pour l'année de référence, μ^t est le rapport des erreurs types relatives des dépenses moyennes pour la période de référence exprimé

sous forme de prix pour la période t et 0, et ρ^t est le coefficient de corrélation entre les dépenses moyennes pour la période de référence en prix de t et 0 (présupposé de signe positif). Le choix de \hat{P}_B^t au lieu de \hat{P}_A^t s'explique donc par le fait que le premier bénéficie du caractère de l'échantillon; pour $\rho^t > 1/(2\mu^t)$, une réduction appréciable de la variance est attendue. La variance de \hat{P}_B^t peut également s'exprimer comme suit :

$$V(\hat{P}_B^t) \approx \frac{1-f}{n} \frac{N^2}{N-1} \sum_{g \in \hat{A}} (w_g^0)^2 (P_g^t - P^t)^2, \quad (3)$$

que l'on peut estimer à l'aide de données d'échantillon à la condition que la fraction de sondage $f = n/N$ soit connue. Cette formule, déjà mentionnée par Balk (1989), montre que la variance dépend de la dispersion intra-groupe des indices de prix des produits. Par conséquent, on pourrait réduire la variance soit en établissant des groupes de produits comportant des changements de prix semblables, soit en élargissant l'échantillon. Särndal et coll. (1992, page 176) font remarquer toutefois que la méthode de linéarisation de Taylor risque d'entraîner une sous-estimation de la variance pour des échantillons qui ne sont pas très grands. Les échantillons de produits de l'IPC sont généralement assez petits. Pour certains groupes de produits, il n'y a parfois qu'un ou deux produits représentatifs. Ainsi, en plus d'être instable, comportant elle-même une grande variance, la variance sera probablement sous-estimée si elle se fonde sur (3).

À noter que l'estimateur \hat{P}_B^t , du fait qu'il est un rapport, subit pour de petits échantillons un biais à peu près égal à $o(1/n)$. On peut facilement vérifier que sa valeur absolue $|B(\hat{P}_B^t)| \leq \sigma^0 \sqrt{V(\hat{P}_B^t)}$. Si σ^0 est petit, par exemple inférieur à 0,1, le biais de \hat{P}_B^t peut sans danger être considéré comme négligeable relativement à son erreur type. Par contre, pour un petit échantillon de produits et une grande variabilité des dépenses pour la période de référence, σ^0 peut facilement dépasser 0,1 de beaucoup. À noter également qu'il est peu probable que l'IPC pour tous les produits soit biaisé dans

une large mesure à cause de cela, puisque le biais est une moyenne (pondérée) de biais positifs et négatifs de différents indices de groupes de produits.

3.2 Échantillonnage proportionnel à la taille

L'échantillonnage proportionnel à la taille a ceci d'avantageux que les produits les plus importants ont une bonne chance d'être sélectionnés. Nous nous limitons ici à l'échantillonnage de taille fixe sans remise, car ce semble être la méthode la plus probable en cas d'échantillonnage proportionnel à la taille du produit (voir par exemple le cas suédois décrit par Dalén et Ohlsson 1995). Les dépenses pour la période de référence nous servent de mesure de la taille; la probabilité d'inclusion de premier ordre requise pour le produit g est $\pi_g = ne_g^0/e^0 = nw_g^0$, où $e^0 = \sum_{g \in A} e_g^0$. Il s'ensuit que $\sum_{g \in \hat{A}} P_g^t/r$ est un estimateur non biaisé de P^t .

L'échantillonnage proportionnel à la taille sans remise, combiné à l'estimateur de Horvitz-Thompson ou π , s'appelle parfois l'échantillonnage $\pi\pi$. La plupart des schémas existants pour l'échantillonnage $\pi\pi$ de taille fixe sont de type tirage-progressif et passablement compliqués. Nous allons donc nous en tenir plutôt à une sélection $\pi\pi$ systématique. On peut décrire ce schéma en imaginant que les dépenses e_g^0 ($g \in A$) sont étalées de façon cumulative et horizontalement, depuis l'origine jusqu'à e^0 . Un nombre réel est choisi de façon aléatoire dans l'intervalle $(0, e^0/n]$, et nous continuons systématiquement en prenant les produits g identifiés par des points séparés par la distance constante e^0/n . Cette méthode fournit exactement la taille d'échantillon souhaitée. Pour des groupes de marchandises comportant une grande variation des dépenses pour la période de référence, il n'est pas toujours possible de tirer un échantillon de produits strictement proportionnel aux dépenses. Bien sûr, $\pi_g \leq 1$ doit être respecté pour tous les g . Si $n > 1$ et si certaines valeurs e_g^0 sont extrêmement grandes, il est possible que pour certains produits $ne_g^0/e^0 > 1$, ce qui va à l'encontre de l'exigence $\pi_g \leq 1$. La difficulté est surmontée comme suit. Les N produits sont rangés dans l'ordre descendant des dépenses. D'abord, si $e_1^0 > e^0/n$, nous prenons $\pi_1 = 1$. Ensuite, si $e_2^0 > (e^0 - e_1^0)/(n - 1)$, nous prenons également $\pi_2 = 1$. La procédure est répétée jusqu'à ce que l'exigence d'un échantillonnage proportionnel aux dépenses pour la période de référence soit respectée pour tous les produits qui restent. Notre stratégie récursive diffère quelque peu de la méthode proposé par Särndal et coll. (1992, page 90), qui suggèrent de prendre $\pi_g = 1$ pour tous les g comportant $e_g^0 > e^0/n$. Dans nos fichiers de données, cela entraînerait des nombres inutilement grands des produits comportant $\pi_g = 1$. Le sous-groupe A_H de produits comportant les dépenses les plus élevées pour la période de référence, qui est choisi avec certitude, s'appelle la partie auto-sélection de l'échantillon. À partir du sous-groupe A_L à faibles dépenses qui reste, on tire un échantillon \hat{A}_L de taille n_L strictement proportionnel

aux dépenses. L'estimateur non biaisé qui en résulte est une moyenne pondérée en fonction des dépenses de $P^t(H)$, l'indice de prix de Laspeyres (population) de A_H , et de $\sum_{g \in \hat{A}_L} P_g^t/n_L$, l'indice de prix estimatif de A_L .

3.3 Échantillonnage stratifié

L'échantillonnage aléatoire simple offre un avantage évident relativement à l'échantillonnage proportionnel aux dépenses: en effet, à part l'utilisation d'un registre de produits comme base de sondage, aucune autre donnée n'est nécessaire (voir aussi Balk 1994). En présence d'articles de dépense à distribution très inégale, il y a de fortes chances que les chefs de file du marché se trouvent hors de l'échantillon, situation qui est intuitivement jugée indésirable. Nous estimons qu'il serait effectivement préférable qu'ils soient sélectionnés. Rappelons que la variance de l'indice de prix des groupes de produits dans le cadre de l'échantillonnage aléatoire simple dépend de la dispersion intra-groupe des indices de prix des produits. La variance pourrait être réduite s'il était possible de stratifier le groupe de produits en sous-groupes homogènes en fonction des changements de prix. Toutefois, les changements de prix des produits ne sont pas connus d'avance. Une autre façon de réduire la variance serait de stratifier le groupe de produits en deux sous-groupes, le premier (A_H) comportant de fortes dépenses pour la période de référence, que l'on observe intégralement, l'autre (A_L) comportant de faibles dépenses, dont on tire un échantillon aléatoire \hat{A}_L . Le nouvel estimateur de l'indice de prix des groupes de produits est une moyenne pondérée en fonction des dépenses de $\hat{P}_B^t(L)$, l'indice de Laspeyres du sous-groupe à faibles dépenses, estimé en conformité avec (2), et de $P^t(H)$. Sa variance d'échantillonnage est de $(1 - \tau_H)^2 V[\hat{P}_B^t(L)]$, où τ_H est la partie dépenses de A_H dans A . Cette méthode ne réduit pas nécessairement la variance de l'indice de prix estimatif, mais elle la réduit probablement dans certaines circonstances. La variance du nouvel estimateur sera plus faible que la variance de \hat{P}_B^t lorsque

$$1 - \tau_H < \frac{\text{e.t.}(\hat{P}_B^t)}{\text{e.t.}[\hat{P}_B^t(L)]}, \quad (4)$$

où e.t.(.) désigne l'erreur-type. L'inégalité (4) est présumée intacte si les articles de dépense sont distribués de façon extrêmement inégale, puisque $1 - \tau_H$ devient alors nettement inférieur à 1. Il est possible que la stratification soit particulièrement productive à mesure que la taille globale de l'échantillon n augmente.

Le choix de τ_H et donc de la taille N_H de la strate « à tirage complet » A_H est un peu difficile. Il serait préférable d'avoir un certain critère d'optimalité afin de minimiser la variance. Toutefois, puisque les changements de prix des produits ne sont pas connus d'avance et que les tendances antérieures ne permettent pas une prévision très exacte des futurs changements de prix, il n'est guère possible de

calculer la taille optimale de A_H . Pour l'analyse empirique, nous ferons l'essai de deux tailles d'échantillon relatives $\lambda_H = N_H/n$ de A_H , c'est-à-dire $\lambda_H = 1/3$ et $\lambda_H = 2/3$. Ces valeurs permettent d'indiquer clairement le rendement.

3.4 Échantillonnage par seuil d'inclusion

Lorsque la taille de l'échantillon est très petite, il semble assez probable qu'une stratification fondée sur $\lambda_H = 2/3$ donnera une erreur type plus grande de l'indice de prix estimatif qu'une stratification fondée sur $\lambda_H = 1/3$. Qu'arrive-t-il si A_L n'est pas observé du tout, de sorte que $\lambda_H = 1$ et donc $n = N_H$? Nous aurions alors recours à un échantillonnage par seuil d'inclusion (de type spécial). L'estimation de l'indice de prix des groupes de produits se fonde simplement sur $\hat{P}'_C = P'(H)$. Dès lors tous les $g \in A_H$ ont une probabilité d'inclusion de 1, tandis que tous les $g \in A_L$ ont une probabilité d'inclusion nulle (Särndal et coll. 1992, pages 531-533). Puisque nous savons exactement quels produits seront choisis, il n'y a rien d'aléatoire et la variance d'échantillonnage du \hat{P}'_C est nulle par définition. Le biais correspond à l'erreur réelle, c'est-à-dire à la différence entre la valeur estimée et l'indice de population véritable

$$\hat{P}'_C - P' = (1 - \tau_H)[P'(H) - P'(L)]. \quad (5)$$

Lorsque la distribution des articles de dépense est extrêmement inégale, même une petite taille d'échantillon peut entraîner une grande valeur de τ_H . Dans un tel cas, une estimation par seuil d'inclusion peut donner de meilleurs résultats que la stratification, relativement à l'erreur quadratique moyenne. Nous avons le choix de fixer le taux d'inclusion τ_H , de façon que la taille de l'échantillon n soit déterminée par τ_H , ou de fixer la taille de l'échantillon, et alors τ_H dépend du choix de n . Nous avons retenu cette dernière possibilité puisque les plans d'échantillonnage à taille fixe sont un outil répandu pour la sélection des produits de l'IPC, et parce que cela permet une comparaison appropriée à d'autres plans à taille fixe.

Le recours à des procédures de seuil d'inclusion est justifié du fait que i) les frais d'établissement d'une base de sondage fiable pour l'ensemble de la population sont trop élevés et que ii) le biais est jugé négligeable. Bien sûr, l'hypothèse ii) ne peut pas être vérifiée en général. Néanmoins, l'exclusion délibérée d'une partie de la population cible du tirage de l'échantillon peut donner des résultats satisfaisants sous réserve de corrections appropriées. Statistique Pays-Bas fait appel à un échantillonnage par seuil d'inclusion pour diverses autres enquêtes-entreprises, par exemple pour la statistique de la production et du commerce étranger, de très petites entreprises restant inobservées. Quant aux comptes nationaux néerlandais, pour lesquels les données sur la production et le commerce étranger sont un intrant important, on établit des estimations explicites pour les petites entreprises. La méthode du seuil d'inclusion pour la sélection des produits de l'IPC, par contre, n'apporte pas

de correction pour les produits exclus. La question du coût mise à part, on défend parfois cette méthode du fait que, au moins à long terme, les changements de prix des produits moins importants ne diffèrent pas tellement de ceux des chefs de file du marché pour la même catégorie de produits à cause de la structure analogue des coûts de production.

4. Estimation empirique

4.1 Simulation de Monte Carlo

Exception faite pour la sélection par seuil d'inclusion, il est difficile de trouver des mesures fiables des distributions d'échantillonnage fondées sur un seul échantillon. Pour ce qui est de l'échantillonnage aléatoire simple, l'estimateur \hat{P}'_C comporte un biais inconnu, tandis que l'estimation de la variance fondée sur les techniques de linéarisation de Taylor donne des résultats inexacts parce que les échantillons de produits de l'IPC sont généralement très petits. L'échantillonnage $pt\pi$ systématique pose la difficulté du choix d'une méthode d'estimation de la variance de l'estimateur puisque les probabilités d'inclusion de second ordre sont inconnues. Afin d'obtenir la distribution d'échantillonnage exacte, il faudrait considérer tous les échantillons \hat{A} qui sont possibles pour un certain plan d'échantillonnage. Pour chaque \hat{A} la probabilité de sélection de \hat{A} et la valeur estimative de l'indice de prix des groupes de marchandises doivent être connues si l'on veut calculer la valeur attendue, le biais et la variance de l'estimateur de façon exacte. Or cela est pratiquement impossible à cause du nombre extrêmement grand d'échantillons possibles. Afin de décrire la distribution d'échantillonnage, nous menons donc des simulations de Monte Carlo. Un grand nombre d'échantillons K est tiré de la (même) population A en fonction du plan choisi et, pour chaque échantillon, on calcule l'estimation. Si K est assez grand, la distribution des K estimations se rapproche étroitement de la distribution d'échantillonnage exacte. Soit \hat{P}'_k , le résultat du $k^{\text{ième}}$ échantillon pour un certain plan d'échantillonnage. Dès lors

$$\bar{P}' = \frac{1}{K} \sum_{k=1}^K \hat{P}'_k$$

est une estimation non biaisée de la valeur attendue $E(\hat{P}')$. Nous calculons alors

$$\bar{P}' - P',$$

qui est une estimation non biaisée du biais $B(\hat{P}')$;

$$S_p^2 = \frac{1}{K-1} \sum_{k=1}^K (\hat{P}'_k - \bar{P}')^2,$$

qui est une estimation non biaisée de la variance $V(\hat{P}')$; et

$$\sqrt{(\bar{P}' - P')^2 + S_p^2},$$

qui représente une estimation à peu près non biaisée de la racine de l'erreur quadratique moyenne (reqm) de \hat{P}' . Särndal et coll. (1992, page 280) font remarquer que l'imperfection résultant du nombre fini de répétitions est plus sérieuse dans le cas d'une mesure de la variance... que dans le cas de mesures calculées comme moyennes.

4.2 Résultats

Des simulations de Monte Carlo ont été effectuées avec trois tailles d'échantillon : $n = 3$, $n = 6$ et $n = 12$. Le nombre de répétitions (K) par expérience a été fixé à 500 000. Le tableau 1 fournit les résultats pour le café en janvier 1995 (1994 = 100), tandis que les tableaux 2 et 3 fournissent les résultats pour les couches de bébés et le papier hygiénique, respectivement, en janvier 1996 (1995 = 100). Le choix de la formule qui permet d'agréger les observations de prix individuelles en un même indice de prix des produits est décrit à l'annexe. Pour ce qui est de la présente section, tous les indices de prix des produits sont calculés sous forme d'indices de valeurs unitaires pour tous les points de vente. L'échantillonnage aléatoire simple donne des résultats particulièrement médiocres. Ainsi, pour $n = 3$, la véritable augmentation du prix du café (Laspeyres)

de 17,2 % est sous-estimée de 1,4 point. Jumelée à une erreur type de 5,1 points, la reqm représente 5,3 points, c'est-à-dire près du tiers de l'augmentation véritable des prix. Même pour $n = 12$, qui donne une fraction d'échantillonnage de 0,18 (valeur inhabituellement élevée dans la pratique), la reqm demeure néanmoins considérablement élevée. À noter que, comme prévu, le biais pour les petits échantillons est réduit de moitié lorsque la taille de l'échantillon est doublée. La stratification fonctionne assez bien pour de grandes tailles d'échantillon, mais donne de piètres résultats pour $n = 3$. Dans ce dernier cas, la stratification fait augmenter la reqm comparativement à l'échantillonnage aléatoire simple pour les couches de bébés et le papier hygiénique lorsque $N_H = 2$ (c'est-à-dire lorsque $\lambda_H = 2/3$). Notre plan probabiliste préféré serait un échantillonnage systématique proportionnel aux dépenses parce que les estimations ne sont pas biaisées et que les erreurs types sont relativement faibles. Le résultat le plus surprenant est sans doute le bon rendement de la sélection par seuil d'inclusion. Sauf pour $n = 3$ et $n = 6$ dans le cas des couches de bébés, cette méthode donne les meilleurs résultats.

Tableau 1
Chiffres estimatifs de l'indice de prix de Laspeyres pour le café (1994 = 100), Janvier 1995 ($N = 68$)

Plan d'échantillonnage	$n = 3$				$n = 6$				$n = 12$			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S. *)	115,7	5,1	-1,4	5,3	116,4	3,4	-0,7	3,5	116,7	2,3	-0,4	2,3
π pt	117,2	2,2	0	2,2	117,2	1,3	0	1,3	117,2	0,7	0	0,7
Stratifié												
$\lambda_H = 1/3$	116,4	3,9	-0,7	4,0	116,6	2,3	-0,5	2,3	117,0	1,2	-0,1	1,2
$\lambda_H = 2/3$	115,6	4,5	-1,5	4,7	116,4	2,5	-0,7	2,6	117,0	1,1	-0,2	1,1
Seuil d'inclusion	117,0	0	-0,2	0,2	117,2	0	0,0	0,0	117,5	0	0,3	0,3

*) Aléatoire simple

Tableau 2
Chiffres estimatifs de l'indice de prix de Laspeyres pour les couches de bébés (1995 = 100), Janvier 1996 ($N = 58$)

Plan d'échantillonnage	$n = 3$				$n = 6$				$n = 12$			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S.	99,4	5,0	2,3	5,5	98,7	3,9	1,5	4,2	97,9	2,9	0,8	3,0
π pt	97,2	2,8	0	2,8	97,2	1,6	0	1,6	97,2	1,5	0	1,5
Stratifié												
$\lambda_H = 1/3$	98,9	5,0	1,8	5,3	98,1	3,3	1,0	3,4	97,4	1,7	0,2	1,7
$\lambda_H = 2/3$	98,3	5,8	1,1	5,9	97,4	3,3	0,3	3,3	97,0	1,6	-0,2	1,6
Seuil d'inclusion	92,0	0	-5,1	5,1	93,4	0	-3,8	3,8	95,5	0	-1,6	1,6

Tableau 3

Chiffres estimatifs de l'indice de prix de Laspeyres pour le papier hygiénique (1995 = 100), Janvier 1996 ($N = 70$)

Plan d'échantillonnage	$n = 3$				$n = 6$				$n = 12$			
	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm	valeur attendue	e.t.	biais	reqm
A.S.	103,9	4,5	0,1	4,5	103,9	3,5	-0,1	3,5	103,9	2,6	0,1	2,6
π pt	103,9	3,4	0	3,4	103,9	1,8	0	1,8	103,9	1,2	0	1,2
Stratifié												
$\lambda_H = 1/3$	103,5	4,3	-0,3	4,3	103,7	3,2	-0,1	3,2	104,0	2,1	0,1	2,1
$\lambda_H = 2/3$	103,7	4,6	-0,2	4,6	104,2	3,4	0,4	3,4	103,9	1,6	0,0	1,6
Seuil d'inclusion	105,0	0	1,1	1,1	104,0	0	0,1	0,1	104,0	0	0,1	0,1

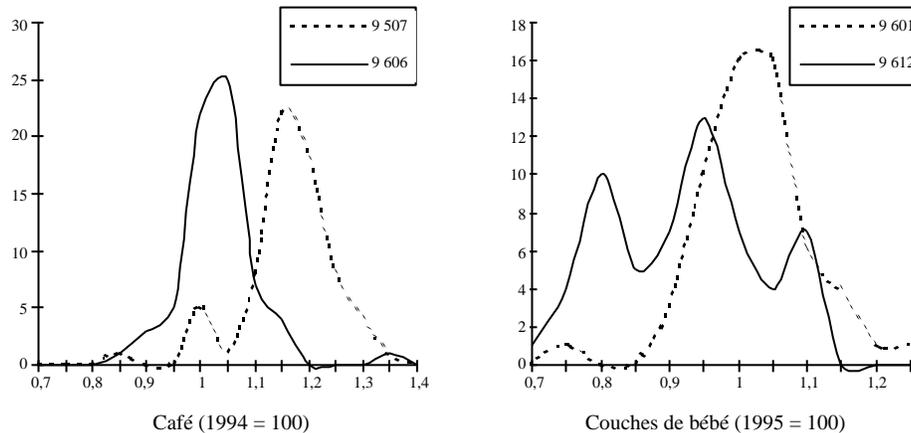


Figure 3 Distribution de fréquences des chiffres des indices de prix des produits

Pour le café, nous avons également fait l'essai d'un autre type d'échantillonnage stratifié. La population globale des produits a été divisée en café moulu et en café instantané et nous avons tiré de chaque strate des échantillons aléatoires. Même si les changements de prix du café instantané sont lissés et prennent du retard comparativement au café moulu, les résultats de Monte Carlo obtenus à l'aide de l'échantillonnage stratifié sont semblables à ceux d'un échantillonnage non stratifié pour l'ensemble des quatre méthodes d'échantillonnage. Cela va à l'encontre de résultats antérieurs (voir de Haan et Opperdoes 1997a). Il en est ainsi parce que nous avons supprimé certains produits de café instantané pour cette étude afin d'obtenir un fichier complet de données pour chaque mois; en fin de compte nous avons obtenu une fraction mineure (8 %) de café instantané dans les dépenses totales de café pour l'année de référence.

Il serait dangereux de tirer des conclusions au sujet du rendement des différents plans d'échantillonnage à partir de simulations pour un mois particulier car il est probable que les résultats dépendent de la distribution de fréquences des indices de prix des produits. La figure 3 montre ces distributions pour le café et les couches de bébés pendant deux mois. Les deux distributions se déplacent vers la gauche, indiquant une baisse de la moyenne non pondérée. Autrement, la distribution de fréquences pour le café demeure assez stable. Le profil de la courbe pour les couches, par

contre, change considérablement; il y a eu accroissement de la variance des indices de produits. Des expériences de Monte Carlo ont été menées pour chaque mois de la période à l'étude. La figure 4 montre la reqm pour $n = 3$. Le profil qui se dégage pour le café et le papier hygiénique est remarquablement robuste : la sélection par seuil d'inclusion est toujours la meilleure. Il semble que, pour de faibles tailles d'échantillon, l'exclusion des produits moins importants n'a guère d'importance. C'est ce que de nombreux bureaux de la statistique ont compris depuis longtemps, sans avoir pu le vérifier de façon empirique jusqu'à présent. Le meilleur rendement de la sélection par seuil d'inclusion relativement à l'échantillonnage proportionnel aux dépenses s'explique, dans le cas du papier hygiénique, en partie par le fait qu'il n'existe pas d'auto-sélection pour ce dernier schéma d'échantillonnage. Pour de grands échantillons, les résultats de la sélection par seuil d'inclusion et de l'échantillonnage proportionnel à la taille sont très semblables. Pour les couches de bébés, les résultats sont quelque peu différents. À cause de la grande volatilité des indices du produit, l'reqm varie considérablement pour ce qui est de la sélection par seuil d'inclusion; elle semble se promener autour de la reqm provenant d'un échantillonnage systématique proportionnel aux dépenses. La forte variabilité de l'erreur peut être considérée comme un désavantage de la sélection par seuil d'inclusion.

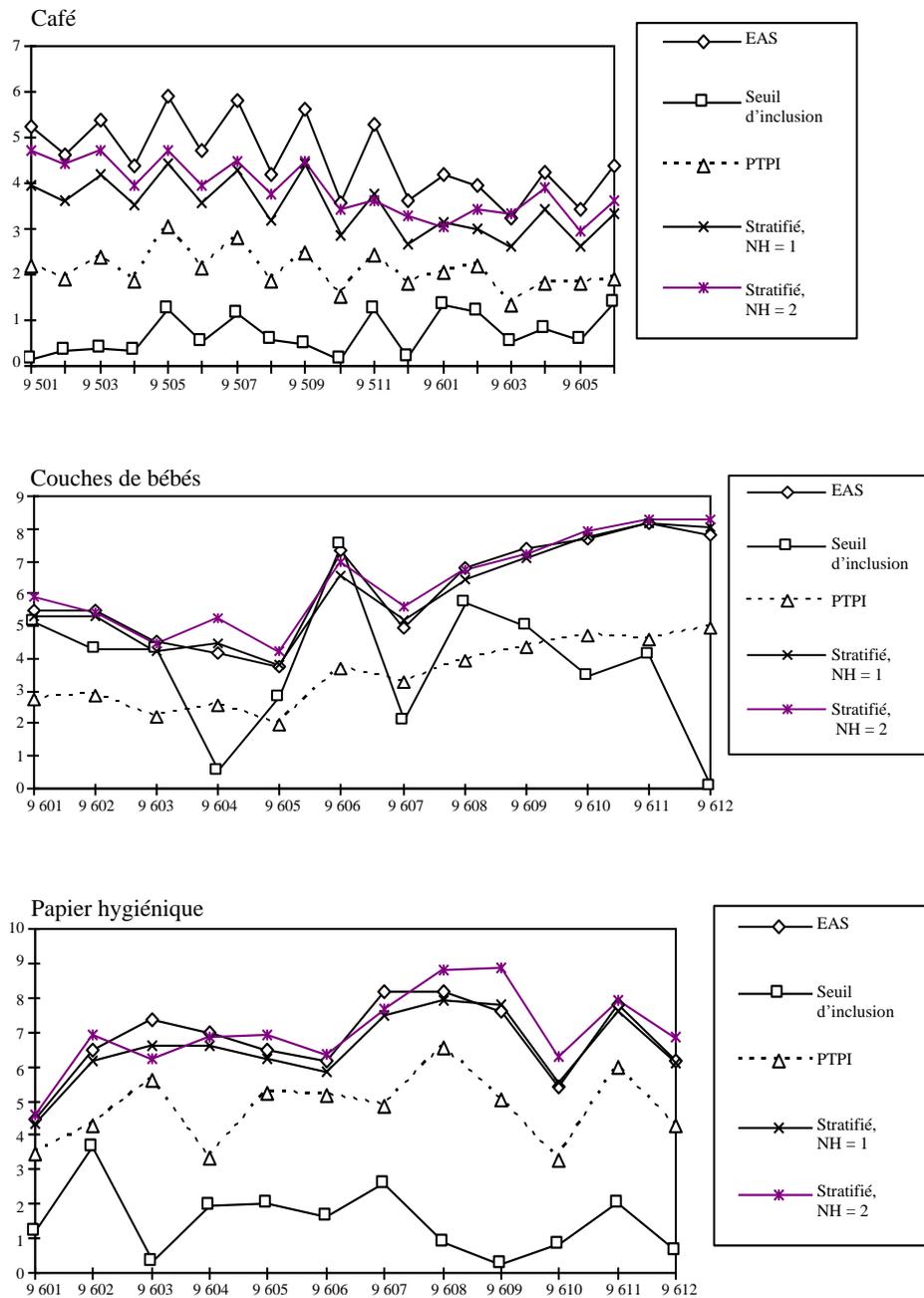


Figure 4 Racine de l'erreur quadratique moyenne (reqm) des indices de prix de Laspeyres estimatifs

5. Utilisation des indices de Fisher

5.1 Indices de produits de valeurs unitaires/de Fisher

À la section 4, les indices de prix des produits ont été calculés comme indices de valeurs unitaires pour tous les points de vente. Comme moyen d'évaluer l'effet du choix de la formule de l'indice de produits sur le résultat de l'étude de simulation, le tableau 4 compare les résultats de

Monte Carlo pour $n = 3$ d'après les chiffres de l'indice de produits de valeurs unitaires (comme dans les tableaux 1-3) et les chiffres de l'indice de prix de Fisher (on trouvera des détails à l'annexe). Pour le café, il n'y a guère de différence. Pour les couches et le papier hygiénique, par contre, la reqm diminue lorsqu'on fait appel aux chiffres de l'indice de Fisher, surtout pour l'échantillonnage aléatoire simple. Il en est ainsi parce que les indices de valeurs unitaires manifestent un comportement plus irrégulier. Si des types de couche ou de papier hygiénique physiquement

identiques sont jugés hétérogènes d'un point de vente à l'autre, de sorte que la formule de Fisher serait plus appropriée, le recours à des indices de valeurs unitaires entraîne une surestimation de la variabilité des prix des produits exceptionnellement petits et exagère le mauvais rendement de l'échantillonnage aléatoire simple. Il reste cependant que l'échantillonnage aléatoire simple ne fonctionne pas très bien.

5.2 Biais de substitution intra-groupe

De nombreux bureaux de la statistique et utilisateurs estiment que l'IPC devrait se rapprocher du véritable indice du coût de la vie. Cette notion théorique tirée de la micro-économie sert à mesurer le changement, pour un consommateur ou un ménage représentatif, des coûts minimaux nécessaires au maintien du même niveau de vie ou d'utilité. Puisque l'utilité ne se laisse pas mesurer, il convient de choisir une formule d'indice se rapprochant étroitement de la notion. Diewert (1976) a montré que (ce qu'il appelle) des indices superlatifs fournissent des approximations de second ordre de l'indice du coût de la vie. L'aspect le plus important des indices de prix superlatifs, c'est qu'ils tiennent compte de l'évolution des consommateurs vers des biens et services manifestant des augmentations de prix relativement faibles. Ces formules d'indice font appel à des données sur les dépenses tant pour la période de référence 0 que pour la période courante t . Concrètement, il faut un certain temps avant que les données sur les dépenses soient connues, de sorte que les indices superlatifs ne peuvent pas être compilés en temps réel. Par souci de rapidité de diffusion, la plupart des bureaux de la statistique nationaux adoptent la formule de Laspeyres (poids fixe) pour l'établissement de leur IPC.

L'indice de Fisher est un des indices superlatifs les mieux connus. Lorsqu'on l'applique au niveau des groupes de produits, la différence entre les indices de prix (population) calculés en fonction des formules de Laspeyres et de Fisher peut être interprétée comme un biais de substitution de produits intra-groupe (figure 5). Pour le café, le biais est inférieur à 1 point par année. Pour le papier hygiénique et pour les couches en particulier, le biais est très considérable,

c'est-à-dire environ 1,5 à 3 points par année. Le biais de substitution intra-groupe est généralement positif et il augmente avec le temps. À noter, toutefois, que pour les couches de bébés durant quelques mois au cours du premier semestre de 1996, les chiffres de l'indice de Laspeyres sont inférieurs à ceux de l'indice de Fisher. Ce phénomène inattendu et peut-être aussi l'ampleur du biais positif au cours d'autres mois sont possiblement dus à un défaut du fichier de données qui ne comporte que des supermarchés. Il est bien connu que les couches de bébés aux Pays-Bas s'achètent aussi dans d'autres magasins comme les pharmacies, qui n'ont pas recours à la lecture optique de codes à barres. La substitution des points de vente inclus et exclus de la base de données pourrait porter atteinte à nos chiffres d'indice de la population comme approximation exacte des valeurs véritables. Nous sommes toutefois persuadés qu'elle ne porte pas sérieusement atteinte à l'évaluation des méthodes d'échantillonnage présentées à la section 4.

6. Discussion

Même si les données tirées de la lecture optique de codes à barres comportent certaines lacunes, elles sont un excellent outil de recherche empirique pour diverses questions d'échantillonnage se rapportant aux IPC. Nos simulations montrent que, du moins pour le café, les couches jetables de bébés et le papier hygiénique, l'échantillonnage aléatoire simple des produits n'est pas à conseiller. Nous estimons qu'une telle affirmation s'applique à tous les groupes de produits lorsque la distribution des dépenses est très asymétrique. Si les bureaux de la statistique veulent avoir recours à un échantillonnage probabiliste, ils auraient avantage à utiliser l'échantillonnage proportionnel aux dépenses. Toutefois, la sélection par seuil d'inclusion serait un choix convenable ou même meilleur lorsque les groupes de produits ne comportent pas de changements de prix trop volatils. En effet, autant que nous sachions, la présente étude est la première à fournir des arguments empiriques à l'appui des méthodes de sélection des produits de l'IPC par seuil d'inclusion. Les données de lecture optique agrégées relativement aux points

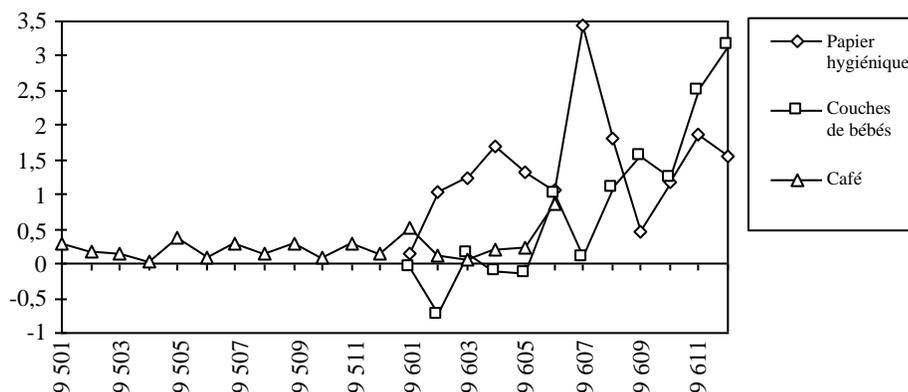


Figure 5 Différence entre les indices de prix (population) de Laspeyres et de Fisher

Tableau 4
Chiffres estimatifs de l'indice de prix de Laspeyres résultant d'autres indices de produits ($n = 3$)

Plan d'échantillonnage	café, janvier 1995 (1994 = 100)				couches, janvier 1996 (1995 = 100)				papier hygiénique, janvier 1996 (1995 = 100)			
	(1)		(2)		(1)		(2)		(1)		(2)	
	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm	valeur attendue	reqm
A.S.	115.7	5.3	115.8	5.3	99.4	5.5	100.4	4.2	103.9	4.5	104.0	3.5
π pt	117.2	2.2	117.2	2.2	97.2	2.8	98.6	2.1	103.9	3.4	104.3	2.6
Stratifié												
$\lambda_H = 1/3$	116.4	4.0	116.5	4.0	98.9	5.3	100.1	4.1	103.5	4.3	103.3	3.6
$\lambda_H = 2/3$	115.6	4.7	115.6	4.8	98.3	5.9	99.5	4.6	103.7	4.6	103.2	3.9
Seuil d'inclusion	117.0	0.2	117.0	0.2	92.0	5.1	94.8	3.8	105.0	1.1	104.7	1.0

(1) D'après les chiffres d'indice de produits de valeurs unitaires.

(2) D'après les chiffres d'indice de prix de Fisher.

de vente devraient fournir une indication claire du seuil d'inclusion requis. Statistique Pays-Bas a déjà fait appel à des données de Nielsen agrégées pour différents groupes de marchandises afin de choisir des produits pour l'échantillon de l'IPC. Les méthodes de seuil d'inclusion sont répandues aux Pays-Bas et dans d'autres pays européens (Boon 1997). Aux Pays-Bas, la sélection des produits comme telle est un peu plus complexe que la situation décrite ci-dessus. Tout d'abord, la méthode du seuil d'inclusion sert à choisir des sous-groupes de produits et non pas des produits particuliers. Ensuite, des produits particuliers sont choisis dans chaque sous-groupe à l'aide d'un échantillonnage dit raisonné. La sélection de ces produits représentatifs se fonde sur le raisonnement de spécialistes travaillant au bureau central et connaissant à fond le marché de consommation en question. Normalement on choisit les produits qui sont achetés le plus souvent ou qui comportent le taux de rotation le plus élevé, de sorte que le schéma d'échantillonnage est une procédure d'inclusion à deux degrés. Il est peu probable qu'une telle méthode à deux degrés donne des résultats bien différents de la procédure à un degré utilisée par les auteurs du présent exposé.

Dans d'autres pays européens, par exemple, au Royaume-Uni, la sélection par seuil d'inclusion ne se fait pas au bureau central, mais bien par le personnel régional là où les prix sont mesurés. Nous choisissons, par exemple, un produit par point de vente, c'est-à-dire le produit le plus vendu durant la période de référence pour le point de vente en question. Dans le cas du café, des couches de bébés et du papier hygiénique, cela donne 2, 12 et 24 produits différents, respectivement. L'estimation de l'indice de Laspeyres du groupe de produits se fait en conformité avec l'expression (2), où les indices de prix des produits sont calculés sous forme d'indices de valeurs unitaires particuliers aux points de vente et pondérés en fonction du point de vente. La figure 6 montre la reqm résultant de cette méthode. Si nous comparons cette figure à la figure 4 (sélection par seuil d'inclusion effectuée au bureau central pour $n = 3$), l'exactitude des deux méthodes de sélection par seuil d'inclusion semble être « en moyenne » du même ordre de

grandeur, même si le schéma est un peu plus irrégulier pour la sélection qui se fait aux points de vente. Une telle comparaison est cependant assez arbitraire. Pourquoi ne pas comparer la sélection par seuil d'inclusion qui se fait aux points de vente à la sélection par seuil d'inclusion qui se fait au bureau central pour $n = 6$, ou encore $n = 12$, ou même pour une autre taille d'échantillon quelconque ? Une autre difficulté, c'est que nous avons traité les indices de prix des produits comme s'ils étaient connus de façon certaine. En réalité, ils se fondent sur un échantillon de points de vente, de sorte que nos résultats dépendent de l'échantillon en question. Pour bien évaluer les deux procédures de sélection par seuil d'inclusion, il faudrait tenir compte à la fois de l'échantillonnage des produits et de l'échantillonnage des points de vente, ce qui toutefois dépasse le cadre du présent exposé.

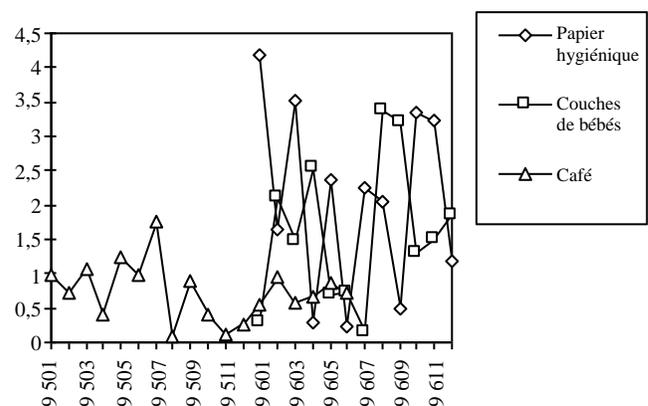


Figure 6 Racine de l'erreur quadratique moyenne (reqm) résultant de la sélection par seuil d'inclusion aux points de vente

Les données de lecture optique ouvrent des perspectives pour la recherche statistique dans le domaine de l'échantillonnage en vue de l'IPC, et servent aussi à compiler toutes sortes de chiffres d'indice, y compris les indices superlatifs, à l'aide de données réelles aussi bien que fortement désagrégées au niveau des points de vente particuliers. Nous avons montré que les indices de prix de

Laspeyres pour des groupes de produits utilisés par les bureaux de la statistique peuvent être biaisés de plus d'un point sur une base annuelle relativement à l'indice de prix de Fisher (superlatif) qui tient compte de la substitution des produits. Un type connexe de biais, résultant d'une considération insuffisante des produits introduits après la période de référence (voir par exemple Boskin, Dulberger, Gordon, Griliches et Jorgenson 1996), n'a pas été étudié dans le présent exposé. Or les données de lecture optique sont un bon moyen d'examiner ce type de biais lié aux nouveaux produits.

Remerciements

La présente recherche a été appuyée en partie par Eurostat (le bureau de la statistique des Communautés européennes) dans le cadre de SUP-COM 1996, Lot 1 : élaboration de méthodologies pour les indices de prix à la consommation et les parités de pouvoir d'achat. Les auteurs tiennent à remercier la société A.C. Nielsen (Nederland) d'avoir bien voulu fournir des données de lecture optique à un coût marginal. Ils tiennent également à remercier Bert M. Balk, Leendert Hoven et deux lecteurs anonymes de leurs précieuses remarques au sujet d'une version provisoire antérieure.

Annexe

Le choix d'une formule d'indice de produits

Les simulations d'échantillonnage supposent des chiffres d'indice de produits. Quelle formule d'indice choisir ? Les bureaux de la statistique doivent généralement calculer les indices au niveau d'agrégation le plus faible basé sur les seules données sur les prix à cause de l'absence de données sur les quantités ou les dépenses. On pourra consulter à ce sujet les travaux détaillés de Szulc (1987), de Dalén (1991), de Balk (1994) et de Diewert (1995). Grâce à la présence de microdonnées de lecture optique, nous nous trouvons dans une position unique de pouvoir établir des indices de prix authentiques (Silver 1995, Hawkes 1997). Soit B_g , une série de points de vente supposée fixe dans le temps, où g est un produit achetable, $b \in B_g$ signifie que g peut être acheté au point de vente b . Le prix de g au point de vente b pour la période s ($s = 0, t$) et le nombre correspondant d'articles vendus sont P_{gb}^s et x_{gb}^s , respectivement. Le produit est considéré comme le niveau d'agrégation le plus faible pour lequel des indices de prix sont établis. Pour commencer, nous nous limitons aux indices de produits que nous pouvons formuler comme des coefficients de prix moyens arithmétiques pondérés pour la période t et la période 0 :

$$P_g^t = \frac{\sum_{b \in B_g} w_{gb}^s P_{gb}^t}{\sum_{b \in B_g} w_{gb}^u P_{gb}^0}, \quad (6)$$

où $w_{gb}^z = x_{gb}^z / \sum_{b \in B_g} x_{gb}^z$ désigne la part du point de vente b dans la vente globale du produit g pour la période z ($z = s, u$). Si $u = 0$ et $s = t$, les prix au cours de la période 0 et de la période t sont pondérés à l'aide des quantités relatives correspondantes. Les prix moyens sont appelés des valeurs unitaires, et P_g^t est un indice de valeurs unitaires. Les mérites de cette méthode sont évalués par De Haan et Opperdoes (1997b) et par Balk (1998). L'addition des quantités n'a de sens que si le produit g peut être considéré comme homogène, c'est-à-dire identique pour l'ensemble des $b \in B_g$. Les valeurs unitaires donnent alors les prix de transaction moyens appropriés et l'indice de valeurs unitaires est l'indice de prix approprié des produits.

Bien sûr, le problème consiste à définir l'homogénéité. Il est possible de soutenir que des produits identiques physiquement mais vendus à différents endroits ne sont pas des produits identiques puisque différents services accompagnent les transactions, de sorte que l'homogénéité d'un point de vente à l'autre n'est jamais atteinte. Une autre formule d'indice devrait donc être choisie. Si $u = s$ dans l'expression (6), P_g^t peut constituer un indice de prix de quantité fixe avec u comme période de référence de la quantité. Pour $u = s = 0$, P_g^t devient l'indice de prix de Laspeyres et, pour $u = s = t$, P_g^t est l'indice de prix de Paasche. Théoriquement, nous ne pouvons préférer l'un ou l'autre. Pour des raisons de symétrie, il semble naturel de retenir la moyenne géométrique (non pondérée) des indices de Paasche et de Laspeyres, qui est l'indice de prix (idéal) de Fisher.

Bibliographie

- Balk, B.M. (1989). On Calculating the Precision of Consumer Price Indices. Rapport, Department of Price Statistics, Statistics Netherlands, Voorburg.
- Balk, B.M. (1994). On the first step in the calculation of a consumer price index. *Proceedings of the First International Conference on Price Indices*, Statistique Canada, Ottawa.
- Balk, B.M. (1998). On the use of unit value indices as consumer price subindices. *Proceedings of the Fourth International Conference on Price Indices*, U.S. Bureau of Labor Statistics, Washington, D.C.
- Boon, M. (1997). Sampling Designs in Constructing Consumer Price Indices: Current Practices at Statistical Offices. Research Paper no. 9717, Research and Development Division, Statistics Netherlands, Voorburg.

- Boskin, M.J., Dulberger, E.R., Gordon, R.J., Griliches, Z. et Jorgenson, D. (1996). Toward a More Accurate Measure of the Cost of Living. Rapport final à la U.S. Senate Finance Committee, Washington, D.C.
- Bradley, R., Cook, B., Leaver, S.G. et Moulton, B.R. (1997). An overview of research on potential uses of scanner data in the U.S. CPI. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.
- Dalén, J. (1991). Computing elementary aggregates in the Swedish consumer price index. *Journal of Official Statistics*, 8, 129-147.
- Dalén, J., et Ohlsson, E. (1995). Variance estimation in the Swedish consumer price index. *Journal of Business & Economic Statistics*, 13, 347-356.
- Diewert, W.E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4, 115-145.
- Diewert, W.E. (1995). Axiomatic and Economic Approaches to Elementary Price Indexes. Working Paper no. 5104, National Bureau of Economic Research, Cambridge.
- De Haan, J., et Oppendoes, E. (1997a). Estimation of the Coffee Price Index Using Scanner Data: The Sampling of Commodities. Rapport de recherche, Socio-economic Statistics Division, Statistics Netherlands, Voorburg.
- De Haan, J., et Oppendoes, E. (1997b). Estimation of the coffee price index using scanner data: The choice of the micro index. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.
- Hawkes, W.J. (1997). Reconciliation of consumer price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanning data. *Proceedings of the Third International Conference on Price Indices*, Statistics Netherlands, Voorburg.
- Reinsdorf, M. (1995). Constructing Basic Component Indexes For The U.s. Cpi From Scanner Data: A Test Using Data On Coffee. Présenté à NBER Conference on Productivity, Cambridge, Mass., 17 juillet, 1995.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York : Springer Verlag.
- Silver, M. (1995). Elementary aggregates, micro-indices and scanner data: Some issues in the compilation of consumer price indices. *Review of Income and Wealth*, 41, 427-438.
- Szulc, B.J. (1987). Price indices below the basic aggregation level. *ILO Bulletin of Labour Statistics*, Ré-imprimé dans Turvey (1989).
- Turvey, R. (1989). Consumer Price Indices: An ILO Manual. Geneva : International Labour Office.