Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Item selection in the Consumer Price Index: Cut-off versus probability sampling

by Jan de Haan, Eddy Opperdoes and Cecile M. Schut

June 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 31-41
Statistics Canada, Catalogue No. 12-001

Item selection in the Consumer Price Index: Cut-off versus probability sampling

Jan de Haan, Eddy Opperdoes and Cecile M. Schut [1]

Abstract

Most statistical offices select the sample of commodities of which prices are collected for their Consumer Price Indexes with non-probability techniques. In the Netherlands, and in many other countries as well, those judgemental sampling methods come close to some kind of cut-off selection, in which a large part of the population (usually the items with the lowest expenditures) is deliberately left unobserved. This method obviously yields biased price index numbers. The question arises whether probability sampling would lead to better results in terms of the mean square error. We have considered simple random sampling, stratified sampling and systematic sampling proportional to expenditure. Monte Carlo simulations using scanner data on coffee, baby's napkins and toilet paper were carried out to assess the performance of the four sampling designs. Surprisingly perhaps, cut-off selection is shown to be a succesful strategy for item sampling in the consumer price index.

Key Words: Laspeyres price index; Monte Carlo simulation; Sampling; Scanner data; Substitution bias.

1. Introduction

Outsiders may think that measuring inflation is an easy job: just visit shops, collect a lot of prices and average them. However, statisticians engaged in the compilation of the Consumer Price Index (CPI), which is the most widely used measure of inflation, face many theoretical and practical problems. In most countries the CPI is essentially a Laspeyres price index. This index weights the partial price indexes of the various commodities by expenditure shares that are fixed at base period levels. Sampling procedures are needed to estimate the population value. Ideally, the mean square error of the estimator would be minimized. Even though the Laspeyres index formula is extremely simple, the estimation procedures applied to the CPI make it a rather complex statistic. Described in a stylized way, the estimation involves three different kinds of samples. A sample of households taking part in an expenditure survey is used to estimate the commodity group expenditure weights. From each commodity group a sample of commodities (items for short) is selected. The prices of these items are collected in a sample of outlets.

In this paper we focus on the sampling of items. Only a few statistical agencies, *e.g.*, the U.S. Bureau of Labor Statistics, use probability sampling to select items to be priced. Most others, for instance Statistics Netherlands, rely on the judgements of experts working at the central office for determining which items should represent the commodity group. In the past this method could be defended by referring to the lack of appropriate sampling frames. Due to the rapidly increasing automation of the retail industry, registers of consumer goods become more and more available, and probability sampling of items comes in sight. Before changing over to a new sampling strategy, however, it seems worthwile to experiment with alternative strategies in order to assess their impact on the accuracy of the estimated price index numbers. The question to be answered is whether current non-probabilistic selection practices perform worse, in terms of the mean square error, than probability techniques. This is the main topic of the present paper. Simulation studies were carried out for three commodity groups, *i.e.*, coffee, disposable baby's napkins and toilet paper.

Not so long ago, empirical price index number research was hampered by the fact that highly disaggregated expenditure and quantity information at the individual outlet level was lacking or at best available for small samples. Nowadays, some market research firms have managed to set up vast micro data bases on sales of consumer goods, especially in the field of fast moving consumer goods. These are derived from electronic scanning by bar-code reader or the associated bar-code typed in at the cashier's desk. Bradley, Cook, Leaver and Moulton (1997) give an overview of potential uses of scanner data in CPI construction. Processing large scanner data bases is a rather time-consuming task. For CPI compilation as such, this could prevent an extensive use in the near future. But scanner data certainly provide a rich source of information for empirical analysis. In addition to studies into sampling, scanner data also enable us to calculate price index numbers according to different index formulas. The (fixed weight) Laspeyres price index does not take the households' reactions to relative price changes into account. We therefore examined to what extent the Laspeyres population price indexes of the three commodity groups are biased with respect to the Fisher price index, an index that does account for commodity substitution.

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

Section 2 gives an overview of the scanner data that we used. Section 3 addresses four different commodity sampling designs. Three of these (*i.e.*, simple random sampling, stratified sampling, and sampling proportional to size) are probability techniques, whereas the fourth (cut-off sampling) is a judgemental one that mimics official practices in the Netherlands. Section 4 describes Monte Carlo experiments we performed to determine the accuracy of the estimated commodity group price indexes under the various sampling designs mentioned. Section 5 deals with the use of Fisher indexes at the item level and the item group level, respectively. The within-group substitution bias of the Laspeyres commodity group price indexes is shown. Section 6 summarizes and discusses the findings.

2. Bar-code scanning data

2.1 An overview

Scannable products are defined in Europe by the European Article Number (EAN). Manufacturers should assign one and only one EAN to every variety, size, type of packaging, *etc.* of a product. This has two implications. In the first place, EANs sometimes change very rapidly, for instance because of a new packaging. Clearly, this makes it difficult to follow a specific item over time. Secondly, some EANs have negligible expenditures. It seems that the system of classification is too detailed; what is really one item has been classified as a multitude of items. In a test study using scanner data on coffee, Reinsdorf (1995) also found that "items that are, for all practical purposes, the same may occasionally have different UPC's" (the US Universal Product Code). Some aggregation over EANs is required. Fortunately, several product characteristics such as brandname and subname are included in the scanner data sets. We will treat EANs having the same product characteristics as identical items. If the number of characteristics is insufficient, there will of course be a danger of over-aggregation, that is of putting heterogeneous items together.

From A.C. Nielsen (Nederland) B.V. we received scanner data sets containing weekly supermarket sales on coffee, disposable baby's napkins and toilet paper. The initial data sets contained 320, 569 and 294 different EANs, respectively. For each EAN, the number of packages sold and the corresponding value is included, together with several product and outlet characteristics. Prices are not included; average prices (unit values) must be calculated from the values and quantities. The coffee data relate to sales over a period of two and a half years, beginning with week 1 of 1994 and ending in week 24 of 1996, in a sample of 20 supermarkets located in a Dutch urban area unknown to us. The data on the other two item groups refer to a sample of 149 shops spread over the whole country, and cover a period of two years, beginning with week 1 of 1995 and ending with week 52 of 1996.

For reasons of convenience we deleted the minor brands. In the case of coffee, only the 15 brands with the highest turnover during the entire observation period were selected from the 55 brands actually sold. After aggregating over EANs with identical product characteristics, we further limited the population to those items that were sold in the base year 1994 and every month thereafter in order to have a complete data set for each month. We ended up with a total of 68 items (excluding coffee beans), among which 40 items of ground coffee (including decaffeinated coffee) and 28 items of instant coffee. These account for 94.5% of total base year coffee expenditure in the initial data set. For napkins and toilet paper (leaving out moist toilet paper), the brands with a turnover share of less than 1% were removed. Next, only those items were selected that were sold in the base year 1995 and at least eight months thereafter. This resulted in 58 napkins items and 70 toilet paper items, accounting for 90% and 86% of total 1995 expenditure in the initial data sets.

2.2 Descriptive statistics

The most striking feature of the item expenditures is the skewness of the distribution. Figure 1 shows the inequality of the base period expenditures in our adjusted data sets by means of so-called Lorenz curves. The vertical axis depicts the cumulative expenditure total, the horizontal axis the cumulative number of items, both expressed as percentages. The items are sorted in increasing order of expenditure. In case of equal expenditures, the Lorenz curve would lie on the diagonal. The more unequal the distribution becomes, the lower its position will be. Coffee item expenditures are distributed extremely unequal, with the three largest items accounting for over half of total base year (1994) coffee expenditure. For baby's napkins and toilet paper the largest six and eight items, respectively, account for nearly half of total base year (1995) expenditure.

Figure 2 shows unit value index numbers, that is the change in the value per package, irrespective of quantity, brandname, type *etc.*, taken over all outlets. This gives a first impression of the change in "prices" during the period under study. For coffee, there was a remarkable decrease in the second half of 1995 following large price rises in 1994 due to bad harvests in Brazil. Coffee prices are largely determined by world market prices for coffee beans. We did not find evidence of significant differences in price changes between outlets. Baby's napkins differ in this respect. A heavy competition was going on between the various producers (which may have caused the decline of the unit values during 1996), while discounts and other kinds of special actions were offered frequently. Hence, the unit value taken over all items and outlets gives an inaccurate picture of the aggregate price change of baby's napkins.

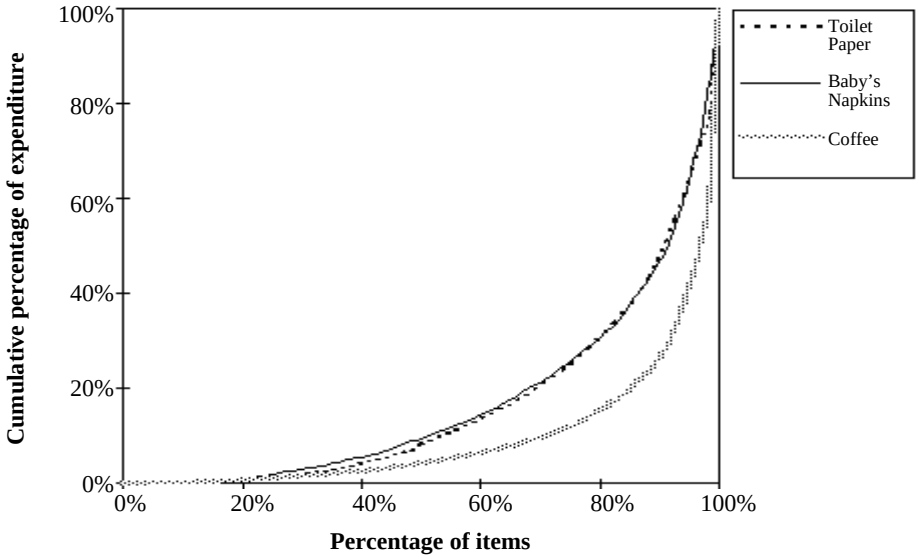

Figure 1 Distribution of base period item expenditures

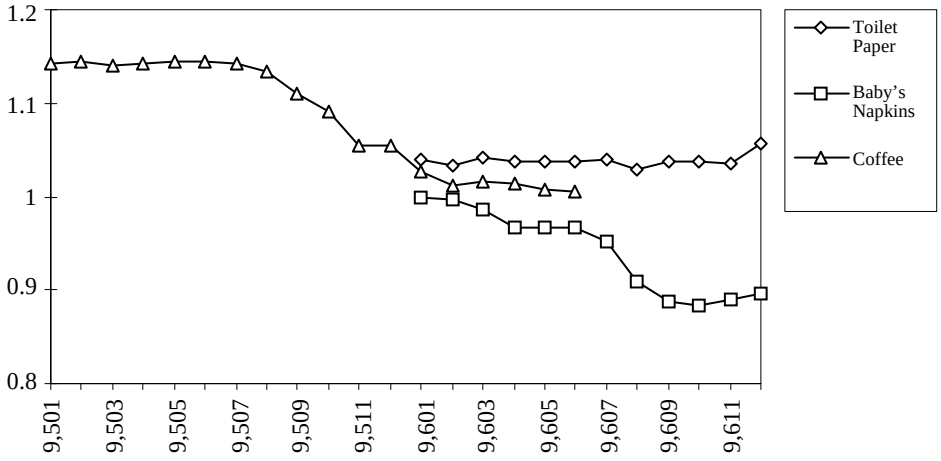

Figure 2 Unit value index numbers (1994 = 100 for coffee, 1995 = 100 for baby's napkins and toilet paper)

3. Estimating Laspeyres-type price indexes

We start this section by introducing some notation. Let commodity group A consist of a finite number, say N, of commodities (items); $g \in A$ means that item g belongs to group A. We assume that A is fixed during time. In real life this is not true: some products disappear from the market, while new products enter. In the short run, however, the constant item group assumption seems reasonable. Note that we adjusted our initial data set accordingly. The reason behind this is that we want to concentrate solely on the sampling aspect. The Laspeyres (fixed weight) price index of commodity group A in period t is

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

where P_g^t denotes the price index of item g, e_g^0 the expenditure on g during base period 0 and w_g^0 the corresponding expenditure share of g within item group A. In the base period a sample $\hat{A}$ with fixed size n is taken from A. Because A is supposed to be fixed during time, it seems natural to keep $\hat{A}$ fixed as well.

3.1 Simple random sampling

Probability sampling refers to situations in which all possible samples have a known probability of selection. Under simple random sampling (without replacement), all possible samples have equal selection probabilities. The Horvitz-Thompson estimator $\hat{P}_A^t = (N/n)\sum_{g \in \hat{A}} w_g^0 P_g^t$ is unbiased for P^t, that is $E(\hat{P}_A^t) = P^t$ where the expectation $E(.)$ denotes the mean over all possible samples under a given sampling design, in this particular case simple random sampling. Despite its unbiasedness, $\hat{P}_A^t$ will not be used in practice because of two undesirable properties.

Firstly, if the price indexes of all sampled items are equal, the estimated item group index differs from that value, unless the population and sample means of expenditures coincide. Price index makers probably dislike this feature. Secondly, and more importantly, $\hat{P}_A^t$ is bound to exhibit extraordinary large sampling variance. To overcome both difficulties, P^t is estimated by taking unbiased estimators of the numerator and denominator:

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

where $\hat{w}_g^0$ is the expenditure share of item g in the sample. Using a first-order Taylor linearization (Särndal, Swensson and Wretman 1992, pages 172-176), the variance of $\hat{P}_B^t$ can be written as

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

where σ^0 denotes the coefficient of variation or relative standard error of the sample mean of base period expenditures, μ^t is the ratio of the relative standard errors of the average base period expenditures expressed in prices of period t and 0, and ρ^t is the correlation coefficient between the average base period expenditures in prices of t and 0 (which is expected to have a positive sign). The choice for $\hat{P}_B^t$ instead of $\hat{P}_A^t$ can thus be elucidated by the fact that the former exploits the panel character of the sample; with $\rho^t > 1/(2\mu^t)$, a substantial reduction in variance is expected. An alternative expression for the variance of $\hat{P}_B^t$ is:

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

which can be estimated using sample data provided that the sampling fraction $f = n/N$ is known. This formula, earlier mentioned by Balk (1989), shows that the variance depends on the within-group dispersion of the item price indexes. Hence, the variance could be lowered either by constructing item groups made up of items having similar price changes or by enlarging the sample. Särndal *et al.*, (1992, page 176) caution that "the Taylor linearization method has a tendency to lead to underestimated variances in not so large samples". The CPI item samples are generally quite small. For some item groups there may even be only one or two representative items. Thus, besides being unstable (having a large variance itself), the variance will probably also be underestimated when based on (3).

We note that estimator $\hat{P}_B^t$, being a ratio, suffers from small sample bias of approximately $o(1/n)$. It can easily be verified that its absolute value $|B(\hat{P}_B^t)| \le \sigma^0\sqrt{V(\hat{P}_B^t)}$. If σ^0 is small, say less than 0.1, the bias of $\hat{P}_B^t$ may safely be regarded as negligible in relation to its standard error. However, with a small item sample and a large variability of base period expenditures, σ^0 could easily exceed 0.1 by far. We add that the all items CPI is unlikely to be biased to a large extent on this account, since the bias is a (weighted) average of positive and negative biases of the various item group indexes.

3.2 Sampling proportional to size

Sampling proportional to size has the advantage that the most important items have a big chance of being sampled. We will restrict ourselves to fixed size sampling without replacement, since this seems most likely to be chosen in case of item sampling proportional to size (see for example the Swedish case described by Dalén and Ohlsson 1995). Base period expenditure acts as our measure of size, and the required first-order inclusion probability for item g is $\pi_g = ne_g^0/e^0 = nw_g^0$, where $e^0 = \sum_{g\in A} e_g^0$. It follows that $\sum_{g\in\hat{A}} P_g^t/r$ is an unbiased estimator of P^t.

Sampling proportional to size without replacement, combined with the Horvitz-Thompson or π estimator, is sometimes called πps sampling. Most existing schemes for fixed-size πps sampling are draw-sequential and rather complicated. We will therefore use systematic πps selection instead. This scheme can be described by imagining the expenditures $e_g^0 (g \in A)$ as cumulatively laid out on a horizontal axis, starting at the origin and ending at e^0. A real number is randomly chosen in the interval $(0, e^0/n]$, and we proceed systematically by taking the items g identified by points at the constant distance e^0/n apart. This method yields exactly the desired sample size. For commodity groups with large variation in base period expenditures, it may not always be possible to select an item sample strictly proportional to expenditure. Obviously, $\pi_g \le 1$ must be satisfied for all g. If $n > 1$ and some e_g^0 values are extremely large, it may be true for some items that $ne_g^0/e^0 > 1$, contradicting the requirement $\pi_g \le 1$. The conflict will be dealt with as follows. The N items are ordered according to descending expenditures. First, if $e_1^0 > e^0/n$, we set $\pi_1 = 1$. Next, if $e_2^0 > (e^0 - e_1^0)/(n-1)$, we also set $\pi_2 = 1$. The procedure is repeated until the requirement for sampling proportional to base period expenditure is met for all remaining items. Our recursive approach differs somewhat from the method proposed by Särndal *et al.*, (1992, page 90). They suggest to set $\pi_g = 1$ for all g with $e_g^0 > e^0/n$. In our data sets this would lead to unnecessary large numbers of items with $\pi_g = 1$. The subgroup A_H of items with the highest base period expenditures which is selected with certainty will be called the self-selecting part of the sample. From the remaining low-expenditure subgroup A_L a sample $\hat{A}_L$ with size n_L is drawn strictly proportional to expenditure. The resulting unbiased estimator is an expenditure weighted average of $P^t(H)$, the population Laspeyres price index of A_H, and $\sum_{g\in\hat{A}_L} P_g^t/n_L$, the estimated price index of A_L.

3.3 Stratified sampling

The obvious advantage of simple random sampling as opposed to sampling proportional to expenditure is that,

apart from a register of items serving as a sampling frame, no other data are required. See also Balk (1994). With very unequally distributed item expenditures chances are big that the market leaders fall outside the sample, a situation that seems intuitively unappealing. We will argue that it would indeed be preferable if they were selected. Recall that the variance of the item group price index under simple random sampling depends on the within-group dispersion of the item price indexes. A variance reduction could be achieved if it were possible to stratify the item group into homogeneous subgroups according to their price changes. However, *a priori* knowledge of item price changes is not available. Another way to lower the variance might be to stratify the item group into two subgroups, one (A_H) with high base period expenditures which is observed entirely and the other one (A_L) with low expenditures from which a random sample $\hat{A}_L$ is taken. The new item group price index estimator is an expenditure weighted average of $\hat{P}_B^t(L)$, the Laspeyres index of the low-expenditure subgroup, estimated in accordance with (2), and $P^t(H)$. Its sampling variance is $(1-\tau_H)^2 V[\hat{P}_B^t(L)]$, where τ_H is the expenditure share of A_H within A. This method does not necessarily reduce the variance of the estimated price index, but it is likely to do so under certain conditions. The variance of the new estimator will be smaller than the variance of $\hat{P}_B^t$ when

$$1-\tau_H < \frac{\mathrm{se}(\hat{P}_B^t)}{\mathrm{se}[\hat{P}_B^t(L)]}, \qquad (4)$$

where se(.) denotes standard error. Inequality (4) is expected to hold if the item expenditures are distributed extremely unequal, since $1-\tau_H$ will then become much smaller than 1. Stratification may be especially productive as the overall sample size n increases.

The choice of τ_H and thus of the size N_H of the "take-all" stratum A_H is a bit of a problem. Preferably we would have some optimality criterion in order to minimize the variance. But since *a priori* knowledge of item price changes is lacking and past trends do not forecast future price changes very accurately, the optimal size of A_H can hardly be computed in practice. In the empirical analysis we will try two different relative sample sizes $\lambda_H = N_H/n$ of A_H, namely $\lambda_H = 1/3$ and $\lambda_H = 2/3$. These values suffice to give a clear indication of the performance.

3.4 Cut-off sampling

When the sample size is very small it seems rather likely that stratification with $\lambda_H = 2/3$ leads to a larger standard error of the estimated price index than with $\lambda_H = 1/3$. But what happens if A_L is not observed at all, so that $\lambda_H = 1$ and thus $n = N_H$? We would then be using (a special type of) cut-off sampling. The item group price index is estimated simply by $\hat{P}_C^t = P^t(H)$. All $g \in A_H$ now have an inclusion probability of 1, whereas all $g \in A_L$ have zero inclusion probability (Särndal *et al.*, 1992, pages 531-533). Since we know exactly which items will be selected there is no randomness involved and the sampling variance of the $\hat{P}_C^t$ is zero by definition. The bias equals the actual error, *i.e.*, the difference between the estimated value and the true population index

$$\hat{P}_C^t - P^t = (1-\tau_H)[P^t(H) - P^t(L)]. \qquad (5)$$

With an extremely unequal distribution of item expenditures, even a small sample size would cause a large value for τ_H. In that case cut-off estimation may outperform stratification, in terms of the mean square error. We may either fix the cut-off rate τ_H, so that the sample size n is determined by τ_H, or fix the sample size, in which case τ_H depends on the choice of n. The latter option was chosen by us since fixed size sampling designs are common practice in selecting CPI items, and because this allows a suitable comparison with other fixed size designs.

The use of cut-off procedures can be justified on the grounds that i) the costs prohibit the construction of a reliable sampling frame for the whole population, and ii) the bias is deemed negligible. Assumption ii) cannot be verified in general, of course. The deliberate exclusion of part of the target population from sample selection may nevertheless give satisfactory results when appropriate corrections are made. Statistics Netherlands makes use of cut-off sampling in various other business surveys, for instance in production and foreign trade statistics where very small enterprises are left unobserved. In the Dutch National Accounts, that use production and foreign trade data as important inputs, explicit estimates are being made for small firms. The cut-off method for CPI item selection, on the other hand, does not correct for the excluded items. In addition to cost-considerations, this method is sometimes defended by the belief that, at least in the longer run, the price changes of less important items will not differ much from those of the market leaders within the same product category because of similar production cost structures.

4. Empirical estimation

4.1 Monte Carlo simulation

With the exception of cut-off selection it is difficult to find reliable measures of the sampling distributions based on a single sample. Under simple random sampling the estimator $\hat{P}_C^t$ has an unknown bias whereas variance estimation based on Taylor linearization techniques gives inaccurate results because CPI item samples are generally very small. Systematic πps sampling raises the question of how to estimate the variance of the estimator since the second-order inclusion probabilities are unknown. To obtain the exact sampling distribution we would have to consider all samples $\hat{A}$ that are possible under a certain sampling design. For every $\hat{A}$ the probability of drawing $\hat{A}$ and the estimated value of the commodity group price index must be known in order to calculate the exact values of the expected value, the bias and the variance of the estimator.

This is virtually impossible because of the extremely large number of possible samples. To describe the sampling distribution, we will therefore carry out Monte Carlo simulations. A large number of samples, say K, is drawn from the (same) population A according to the given design and for each sample the estimate is calculated. If K is large enough, the distribution of the K estimates will closely approximate the exact sampling distribution. Let $\hat{P}_k^t$ denote the result for the k^{th} sample under a certain sampling design. Then

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

is an unbiased estimate of the expected value $E(\hat{P}^t)$. We will calculate

$$\bar{\hat{P}}^t - P^t,$$

which is an unbiased estimate of the bias $B(\hat{P}^t)$;

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

which is an unbiased estimate of the variance $V(\hat{P}^t)$; and

$$\sqrt{(\bar{\hat{P}}^t - P^t)^2 + S_P^2},$$

which is an approximately unbiased estimate of the root mean square error (rmse) of $\hat{P}^t$. Särndal *et al.* (1992, page 280), remark that "the imperfection caused by the finite number of repetitions is more keenly felt in the case of a variance measure... than in the case of measures calculated as means".

4.2 Results

Monte Carlo simulations were carried out with three different sample sizes: $n = 3$, $n = 6$ and $n = 12$. The number of repetitions (K) per experiment was set to 500,000. Table 1 shows the results for coffee in January 1995 (1994 = 100), tables 2 and 3 those for baby's napkins and toilet paper, respectively, in January 1996 (1995 = 100). The choice of the formula with which individual price observations are aggregated into a single item price index is discussed in the Appendix. Throughout this section all item price indexes are calculated as unit value indexes over all outlets. Simple random sampling performs particularly bad. For example, with $n = 3$ the true (Laspeyres) coffee price increase of 17.2% is understated by 1.4%-points. Together with a standard error of 5.1%-points, the rmse amounts to 5.3%-points, that is almost one third of the true price increase. Even with $n = 12$, so that the sampling fraction is 0.18 (which would be unusually large in practical situations), the rmse still remains considerably high. Notice that, as expected, the small sample bias is halved when the sample size is doubled. Stratification works reasonably well with larger sample sizes but gives disappointing results with $n = 3$. In the latter case, stratification increases the rmse compared to simple random sampling for baby's napkins and toilet paper when $N_H = 2$ (that is, when $\lambda_H = 2/3$). Our favourite probabilistic design would be systematic sampling proportional to expenditure because the estimates are unbiased and their standard errors relatively low. The most surprising finding perhaps is the good performance of cut-off selection. Except for $n = 3$ and $n = 6$ in case of baby's napkins, this method produces the best results.

Table 1
Estimated Laspeyres price index numbers for coffee (1994 = 100), January 1995 ($N = 68$)

Sampling	$n = 3$				$n = 6$				$n = 12$			
scheme	exp. value	se	bias	rmse	exp. value	se	bias	rmse	exp. value	se	bias	rmse
S.R. *)	115.7	5.1	-1.4	5.3	116.4	3.4	-0.7	3.5	116.7	2.3	-0.4	2.3
πps	117.2	2.2	0	2.2	117.2	1.3	0	1.3	117.2	0.7	0	0.7
Stratified												
$\lambda_H = 1/3$	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
Cut-off	117.0	0	-0.2	0.2	117.2	0	0.0	0.0	117.5	0	0.3	0.3

*) Simple random

Table 2
Estimated Laspeyres price index numbers for baby's napkins (1995 = 100), January 1996 ($N = 58$)

Sampling	$n = 3$				$n = 6$				$n = 12$			
scheme	exp. value	se	bias	rmse	exp. value	se	bias	rmse	exp. value	se	bias	rmse
S.R.	99.4	5.0	2.3	5.5	98.7	3.9	1.5	4.2	97.9	2.9	0.8	3.0
πps	97.2	2.8	0	2.8	97.2	1.6	0	1.6	97.2	1.5	0	1.5
Stratified												
$\lambda_H = 1/3$	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
Cut-off	92.0	0	-5.1	5.1	93.4	0	-3.8	3.8	95.5	0	-1.6	1.6

Table 3
Estimated Laspeyres price index numbers for toilet paper (1995 = 100), January 1996 ($N = 70$)

Sampling scheme	$n = 3$				$n = 6$				$n = 12$			
	exp. value	se	bias	rmse	exp. value	se	bias	rmse	exp. value	se	bias	rmse
S.R.	103.9	4.5	0.1	4.5	103.9	3.5	-0.1	3.5	103.9	2.6	0.1	2.6
πps	103.9	3.4	0	3.4	103.9	1.8	0	1.8	103.9	1.2	0	1.2
Stratified												
$\lambda_H = 1/3$	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
Cut-off	105.0	0	1.1	1.1	104.0	0	0.1	0.1	104.0	0	0.1	0.1

For coffee we also tried another form of stratified sampling. The entire population of items was subdivided into ground coffee and instant coffee, and we took random samples from each stratum. Although the price changes of instant coffee are smoothed and lag behind as compared to ground coffee, Monte Carlo results using stratified sampling were similar to those using unstratified sampling for all four sampling methods. This contradicts earlier findings (see De Haan and Opperdoes 1997a). The reason is that we deleted some instant coffee items for this study to have a complete data set for each month, and ended up with a minor fraction (8%) of instant coffee in total base year coffee expenditures.

It would be hazardous to draw conclusions about the performance of the various sampling designs based on simulations for one particular month since it is likely that the outcomes depend on the frequency distribution of the item price indexes. Figure 3 shows these distributions for coffee and baby's napkins in two months. Both distributions move to the left, indicating that the unweighted mean has declined. Apart from that, the frequency distribution for coffee remains quite stable. The shape of the curve for napkins, on the other hand, changes dramatically; the variance of the item indexes has grown.

Monte Carlo experiments were run for each month of the period under study. Figure 4 shows the rmse with $n = 3$. The pattern that emerges for coffee and toilet paper is surprisingly robust: cut-off selection always comes out as best. Apparently, if sample sizes are small, the exclusion of the smaller items does not seem to matter much. This is what many statistical offices have been appreciating for a long time, without being able to test it empirically before. The reason why cut-off selection performs better than sampling proportional to expenditure is in case of toilet paper partly caused by the fact that there is no self-selecting part under the latter sampling scheme. With larger samples the results under cut-off selection and sampling proportional to size are very much alike. For baby's napkins the outcomes differ somewhat. Because of the high volatility of the item indexes, the rmse under cut-off selection varies considerably; it seems to meander around the rmse resulting from systematic sampling proportional to expenditure. The high variability of the error can be considered a drawback of cut-off selection.

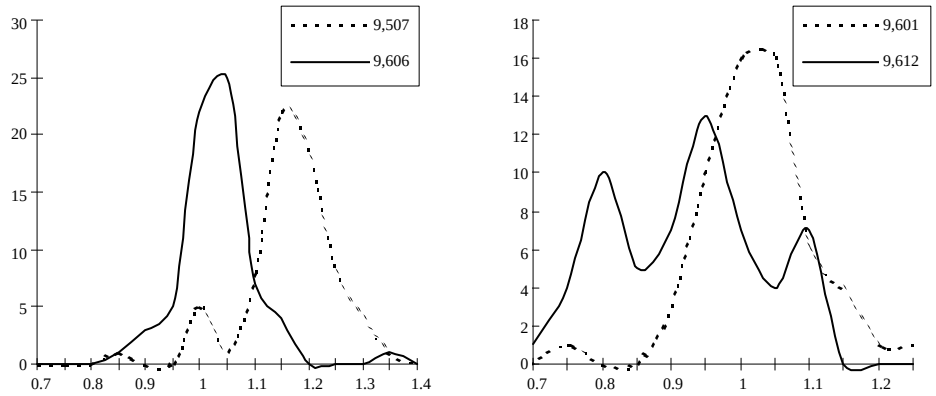

Figure 3 Frequency distribution of item price index numbers

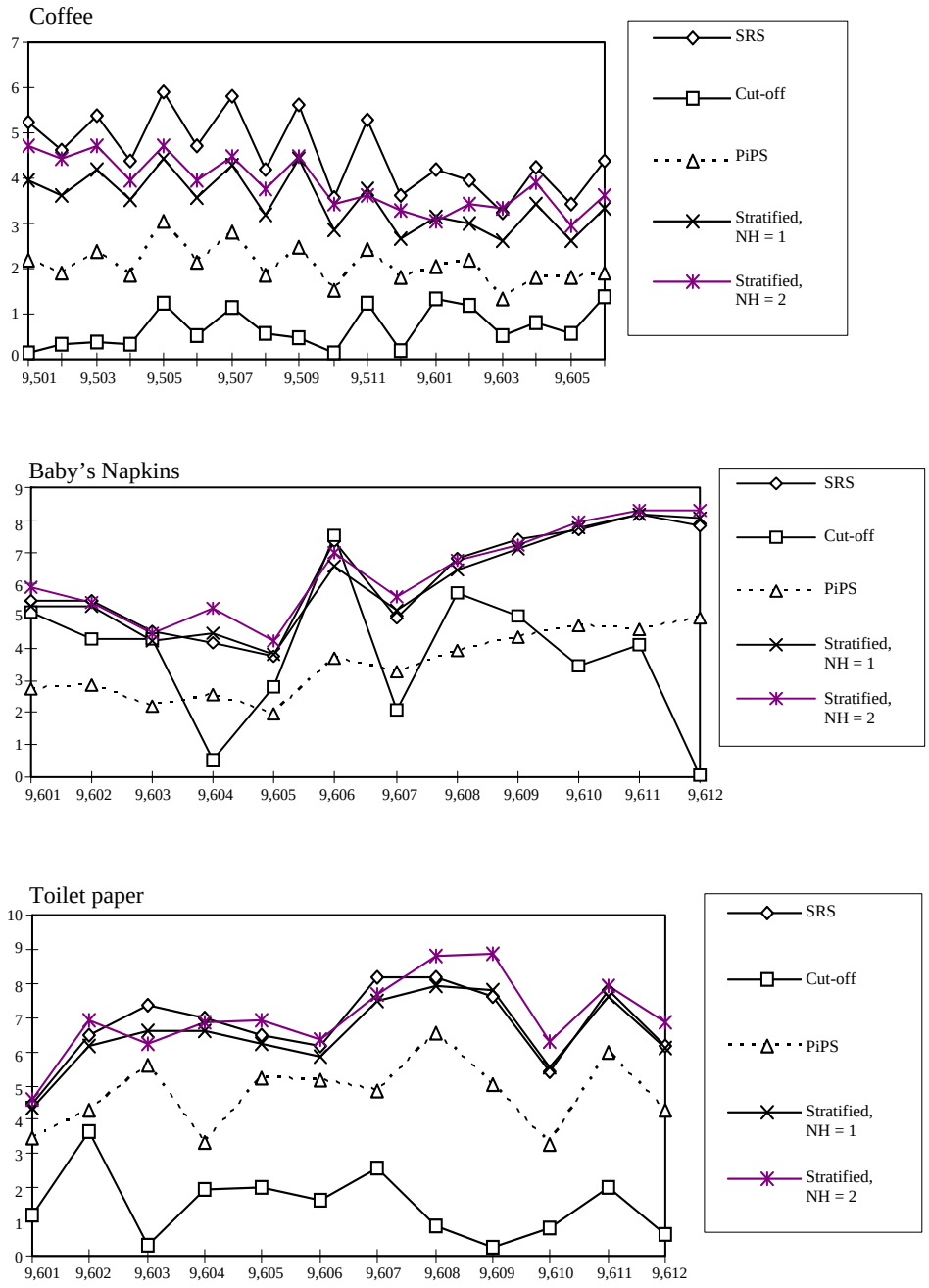

Figure 4 Rmse of estimated Laspeyres price indexes (n = 3)

5. The use of Fisher indexes

5.1 Unit value versus Fisher item indexes

In section 4 the item price indexes were calculated as unit value indexes over all outlets. To assess the impact of the choice of the item index formula on the outcomes of the simulation study, Table 4 compares Monte Carlo results with $n = 3$ based on unit value item index numbers (as in tables 1-3) and Fisher item price index numbers; see the Appendix for details. For coffee, we notice hardly any differences. For napkins and toilet paper, on the other hand, the rmse decreases when Fisher index numbers are used instead, especially in case of simple random sampling. This is caused by the fact that unit value indexes tend to show a more erratic pattern. If physically identical types of napkins or toilet paper are deemed heterogeneous across outlets, so that the Fisher formula would be more appropriate, the use of unit value indexes overstates the price variability of particularly small items and exaggerates the poor performance of simple random sampling. Nevertheless, we would still have to conclude that simple random sampling does not work very well.

5.2 Within-group substitution bias

Many statistical agencies and users are of the opinion that the CPI should be an approximation to the true cost of living index. This theoretical concept is derived from micro-economics and measures the change in the minimum costs for a representative consumer, or household, necessary to retain the same standard of living or utility. Since utility cannot be measured, a feasible index formula should be chosen that closely approximates the concept. Diewert (1976) showed that (what he calls) superlative indexes provide second order approximations to the cost of living index. The most important feature of superlative price indexes is that they take account of consumers' substitution towards goods and services exhibiting relatively small price increases. These index formulas make use of expenditure data relating to both the base period 0 and the current period *t*. In practice it takes some time before expenditure data are known, so that superlative indexes cannot be compiled in real time. For the sake of timeliness most national statistical offices adopt the Laspeyres (fixed weight) formula for constructing their CPIs.

The Fisher index is one of the best-known superlative indexes. When applied to the item group level, the difference between the population price indexes calculated according to the Laspeyres and the Fisher formula can be interpreted as within-group item substitution bias (Figure 5). For coffee it is less than 1%-point per year. For toilet paper and particularly for napkins the biases are very large, about 1.5-3%-points per year. Within-group substitution bias is generally positive and increases over time. Notice, however, that for baby's napkins in a few months of the first half of 1996 the Laspeyres index numbers are lower than the Fisher index numbers. This unexpected effect, and possibly also the large magnitude of the positive bias in other months, may be due to a deficiency of the data set which only contains supermarkets. It is well-known that baby's napkins are bought in the Netherlands also in other kinds of shops such as drugstores that do not make use of bar-code scanning. Substitution between the included and excluded outlets in the data base may damage our population index numbers as accurate approximations of the true values. We are convinced though that it does not seriously affect the assessment of the sampling methods presented in section 4.

Table 4
Estimated Laspeyres price index numbers using alternative item indexes ($n = 3$)

Sampling scheme	Coffee, January 1995 (1994 = 100)				Napkins, January 1996 (1995 = 100)				Toilet paper, January 1996 (1995 = 100)			
	(1)		(2)		(1)		(2)		(1)		(2)	
	exp. value	rmse	exp. value	rmse	exp. value	rmse	exp. value	rmse	exp. value	rmse	exp. value	rmse
S.R.	115.7	5.3	115.8	5.3	99.4	5.5	100.4	4.2	103.9	4.5	104.0	3.5
πps	117.2	2.2	117.2	2.2	97.2	2.8	98.6	2.1	103.9	3.4	104.3	2.6
Stratified												
$\lambda_H = 1/3$	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
Cut-off	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) Based on unit value item index numbers.
(2) Based on Fisher item index numbers.

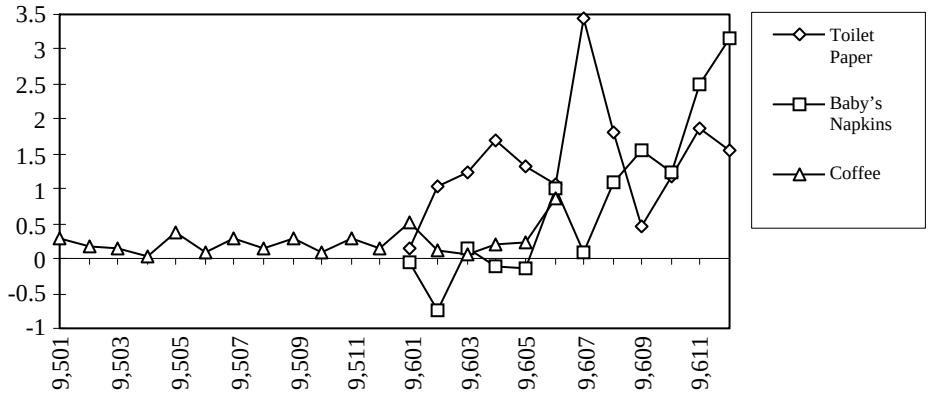

Figure 5 Difference between Laspeyres and Fisher population price index numbers

6. Discussion

Although bar-code scanning data have some deficiencies, they provide an excellent opportunity to undertake empirical research into various sampling issues concerning CPIs. Our simulations show that, for coffee, disposable baby's napkins and toilet paper at least, simple random sampling of items should be advised against. We believe that this recommendation can be extended to all item groups where the distribution of expenditures is very skewed. If statistical offices want to apply probability sampling, they would do a better job using sampling proportional to expenditure. However, cut-off selection might be a good or even better alternative for those item groups where the various item price changes are not too volatile. As a matter of fact, as far as we are aware this is the first study to supply empirical evidence in support of cut-off CPI item selection methods. Aggregated scanner data – that is, scanner data aggregated over outlets – should give a clear indication of the required cut-off rate. Statistics Netherlands already made use of aggregated Nielsen data on a range of commodity groups in the past in order to select items for the CPI sample.

Cut-off methods are applied extensively in the Netherlands and many other European countries (Boon 1997). In the Netherlands the actual item selection is a little more complex than the situation described above. First, a number of item subgroups instead of specific items are chosen using the cut-off method. Next, a number of specific items are selected from each subgroup through so-called judgemental sampling. The selection of these representative items is based on the judgement of experts working at the central office who should have a firm knowledge of the consumer market in question. Usually the most frequently bought items or those with the highest turnover will be selected, so that the entire sampling scheme is a two-stage cut-off procedure. It is unlikely that such a two-stage method would yield results much different from the single-stage procedure we have used in this paper.

In some other European countries, *e.g.*, the United Kingdom, cut-off selection does not take place at the central office but by field staff at the outlets where prices are measured. To illustrate this method, we choose one item per outlet, namely the item with the highest base period sales in the outlet. For coffee, baby's napkins and toilet paper this yields 2, 12 and 24 different items, respectively. The Laspeyres item group index is estimated in accordance with expression (2), where the item price indexes are calculated as outlet-specific unit value indexes and weighted by outlet-specific weights. Figure 6 shows the rmse resulting from this method. If we compare this with Figure 4 (cut-off selection done at the central office for $n = 3$), the accuracy of both cut-off selection methods seems "on average" to be of the same order of magnitude, although the pattern is slightly more erratic under selection at the outlets. But such a comparison is quite arbitrary. Why not compare cut-off selection at the outlets with cut-off selection at the central office for $n = 6$, or $n = 12$, or indeed for any other sample size? Another problem is that we treated the item price indexes as if they were known with certainty. In reality they will be based on a sample of outlets, so that our results are conditional on this sample. For a proper assessment of both cut-off selection procedures we need to take both the sampling of items and the sampling of outlets into account. However, that is beyond the scope of this paper.

Scanner data not only offer challenging perspectives for statistical research in the field of CPI sampling issues, they also enable us to compile all sorts of index numbers, including superlative indexes, using real and highly disaggregated data at the individual outlet level. We demonstrated that the Laspeyres item group price indexes used by statistical agencies can be biased by more than +1%-point on a yearly basis with respect to the (superlative) Fisher price index that accounts for item substitution. A related type of bias, caused by neglecting products that are introduced after the base period (see *e.g.*, Boskin, Dulberger, Gordon, Griliches and Jorgenson 1996), was not addressed by us. Scanner data do provide a good opportunity to investigate this new goods bias.

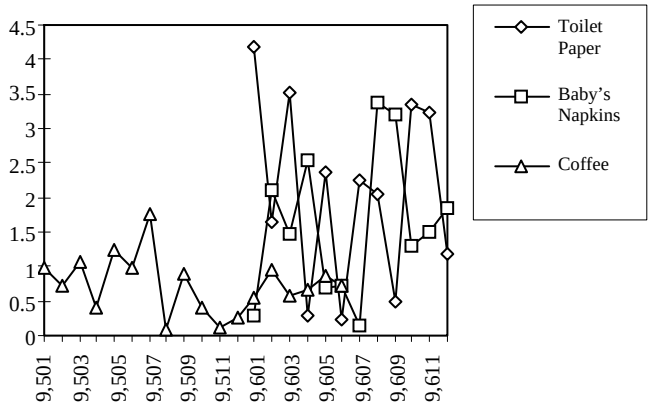

Figure 6 Rmse resulting from cut-off selection at the outlets

Acknowledgements

This research was partially supported by Eurostat (the Statistical Office of the European Communities) under SUP-COM 1996, Lot 1: Development of methodologies in consumer price indices and purchasing power parities. The authors are grateful to A.C. Nielsen (Nederland) B.V. for providing scanner data at marginal costs. They also wish to thank Bert M. Balk, Leendert Hoven and two anonymous referees for helpful comments on an earlier draft.

Appendix

The choice of the item index formula

To perform sampling simulations we need item index numbers. What index formula should be chosen? Statistical offices are generally forced to calculate indexes at the

lowest level of aggregation based on price data alone because quantity or expenditure data is lacking. See Szulc (1987), Dalén (1991), Balk (1994), and Diewert (1995) for a comprehensive treatment of this subject. With scanner microdata at hand, we are in the unique position to construct genuine price indexes (Silver 1995, Hawkes 1997). Consider a set of outlets B_g, assumed fixed during time, where item g can be bought; $b \in B_g$ means that g can be bought in outlet b. The price of g at outlet b in period $s (s = 0, t)$ and the corresponding quantity sold are denoted P_{gb}^s and x_{gb}^s, respectively. The item will be taken as the lowest aggregation level where price indexes are constructed. As a start we restrict ourselves to item indexes that can be written as ratios of weighted arithmetic mean prices in period t and period 0:

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

where $w_{gb}^z = x_{gb}^z / \sum_{b \in B_g} x_{gb}^z$ denotes the share of outlet b in the total quantity sold of item g in period $z (z = s, u)$. If $u = 0$ and $s = t$, the prices in period 0 and period t are weighted by the corresponding relative quantities. The average prices are called unit values, and P_g^t is a unit value index. De Haan and Opperdoes (1997b) and Balk (1998) discuss its merits. Adding up quantities makes sense only if item g can be conceived of as being homogeneous, that is identical across all $b \in B_g$. Unit values then yield the appropriate average transaction prices and the unit value index is the appropriate item price index.

The problem, of course, is to define homogeneity. It can be argued that physically identical products sold in different outlets are not identical items because of different services that accompany the transactions, so that homogeneity across outlets never occurs. Another index formula should then be chosen. If $u = s$ in expression (6), P_g^t can be called a fixed quantity price index with u acting as the quantity reference period. For $u = s = 0$, P_g^t turns into the Laspeyres price index, and for $u = s = t$, P_g^t is the Paasche price index. On theoretical grounds we cannot favour either one. For reasons of symmetry it seems natural to take the (unweighted) geometric average of the Paasche and the Laspeyres index, which is the Fisher (ideal) price index.

References

Balk, B.M. (1989). On Calculating the Precision of Consumer Price Indices. Report, 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*, Statistics 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. and Jorgenson, D. (1996). Toward a More Accurate Measure of the Cost of Living. Final report to the U.S. Senate Finance Committee, Washington, D.C.

Bradley, R., Cook, B., Leaver, S.G. and 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., and 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.

Haan, J. De, and Opperdoes, E. (1997a). Estimation of the Coffee Price Index Using Scanner Data: The Sampling of Commodities. Research Paper, Socio-economic Statistics Division, Statistics Netherlands, Voorburg.

Haan, J. De, and Opperdoes, 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. Presented at NBER Conference on Productivity, Cambridge, Mass., July 17, 1995.

Särndal, C.-E., Swensson, B. and 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*, Reprinted in Turvey (1989).

Turvey, R. (1989). Consumer Price Indices: An ILO Manual. Geneva: International Labour Office.