Distr.
GENERAL

CES/AC.49/1999/9
5 August 1999

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

INTERNATIONAL LABOUR ORGANIZATION (ILO)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/ILO Meeting on Consumer Price Indices
(Geneva, 3-5 November 1999)

ITEM SAMPLING IN THE CPI: THE SUCCESS OF CUT-OFF SELECTION METHODS

Invited paper submitted by Statistics Netherlands*

Summary

Most statistical offices select the samples of commodities (items) of which prices are collected for their Consumer Price Indexes judgementally. In the Netherlands, judgmental sampling resembles cut-off selection where the items with the lowest expenditure are deliberately left unobserved. Cut-off item selection yields biased price index estimates. This paper addresses the question whether probability sampling leads to better results in terms of the mean square error. Monte Carlo simulations using scanner data indicate that cut-off selection generally is a powerful strategy for item sampling in the CPI, mainly because of the skewness of the distribution of item expenditures.

Introduction

1. In most countries, including the Netherlands, the Consumer Price Index (CPI) is essentially a Laspeyres-type index. This index weights the partial price indexes of the various commodities by expenditure shares that are fixed

---

* Prepared by Mr. Jan de Haan, Department of Consumer Prices, Statistics Netherlands, P.O. Box 4000, 2270 JM Voorburg (Netherlands) – E-mail-jhln@cbs.nl. This paper is based on De Haan et al. (1999). Mr. A.C. Nielsen (Netherlands) kindly provided scanner data at marginal costs.

at base period levels. Sampling methods are needed to estimate the population value. Ideally, the mean square error of the estimator would be minimised. 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 stylised manner, the estimation procedure involves drawing three kinds of samples in consecutive steps. First, a sample of households taking part in a budget survey is drawn to estimate the commodity group weights. Second, from each commodity group a sample of commodities (referred to as items) is selected. Third, the prices of these items are collected in a sample of outlets.

2. This paper focuses on the second step. 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 rely on expert judgements to determine which items should represent the item group. Probability selection methods do not necessarily outperform non-probabilistic techniques methods. It will in fact be argued that cut-off item sampling in the CPI, which mimics judgmental practices in the Netherlands, is a successful strategy, particularly in case of a skewed distribution of item expenditures. Empirical evidence using scanner data on coffee, baby's napkins and toilet paper illustrates the argument.[1]

3. Section 2 describes cut-off sampling and three probability techniques: simple random sampling, stratified sampling and sampling proportional to size. Section 3 gives an overview of the scanner data sets used for empirical estimation. Section 4 presents the results of Monte Carlo experiments that were carried out to determine the mean square error (the standard measure of statistical accuracy) of the estimated commodity group price indexes under the various sampling designs. Section 5 concludes.

Estimating Laspeyres-type Price Indexes

4. Suppose that item group *A* consists of *N* items; $g \in A$ means that item *g* belongs to group *A*. Group *A* is assumed to be fixed during time. Although in reality some products may disappear from the market and new products may enter, the constant item group assumption enables us to concentrate on the sampling aspect. Our data set will be adjusted accordingly. The Laspeyres-type (fixed-weight) price index of item group *A* in period *t* is defined as

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

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

<u>Simple Random Sampling</u>

5. Probability sampling refers to situations in which all possible samples have a known (and positive) probability of selection. Under s*imple random sampling* (without replacement) all possible samples have equal selection probabilities. Hence, 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 of all possible samples under a given sampling design, in this particular case simple random sampling. Despite its unbiasedness, $\hat{P}_A^t$ will never be used because of two undesirable properties. If the price index numbers of all sampled items happen to be equal, the estimated group price index differs from that value unless the population mean and the sample mean of expenditures coincide. Price index statisticians probably dislike this feature. 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 the 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 linearisation (Särndal et al. 1992, pp. 172-176), the variance of $\hat{P}_B^t$ can be estimated from the sample data. However, Taylor linearisation tends to lead to underestimated variances for small samples. The CPI item samples are indeed typically small. For some item groups there may even be only one or two representative items. Thus besides being unstable (that is, having a large variance itself), the variance is probably also underestimated when based on Taylor linearisation techniques.

6. Estimator $\hat{P}_B^t$ suffers from small sample bias of approximately o(1/*n*). With a small item sample and a large variability of base period expenditures, the bias of $\hat{P}_B^t$ may be non-negligible in relation to its standard error. The all-items CPI will probably not 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.

Sampling Proportional to Size

7. *Sampling proportional to size* has the advantage that the most important items have the biggest chance of being sampled. In case of item sampling proportional to size a fixed size design without replacement in combination with the Horvitz-Thompson estimator is most likely to be chosen. Such a design is sometimes called pps-sampling. Base period expenditure acts as the measure of size, and the required first-order inclusion probability for item g is $ne_g^0 / \sum_{g \in A} e_g^0 = nw_g^0$ Thus, $\sum_{g \in \hat{A}} P_g^t / n$ is an unbiased estimator of P^t.

8. As most existing schemes for fixed-size sampling proportional to size are rather complicated, a systematic version is used instead. For item groups with a large variability in base period expenditures, it may not always be possible to select a sample strictly proportional to expenditure. The conflict is solved as follows.[2] First, a subgroup A_H of size $N_H (< n)$ with the highest base period expenditures is selected from A with certainty. Next, a sample $\hat{A}_L$ with size $n_L (= n - N_H)$ is drawn strictly proportional to expenditure from the remaining low-expenditure subgroup A_L. The resulting unbiased estimator is an expenditure weighted average of $P^t(H) = \sum_{g \in A_H} e_g^0 P_g^t / \sum_{g \in A_H} e_g^0$, the true price index of A_H, and $\sum_{g \in \hat{A}_L} P_g^t / n_L$, the estimated price index of A_L.

Stratified Sampling

9. 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 there is a large probability that the market leaders fall outside the sample, which seems intuitively unappealing. A variance reduction could be achieved if we were able to stratify the item group into homogeneous subgroups according to their price changes. A priori knowledge of item price changes is not available, however. Alternatively, the variance might be reduced by stratifying 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 estimator of the item group price index is an expenditure weighted average of $\hat{P}_B^t(L)$, that is the Laspeyres-type price index of the low-expenditure subgroup estimated in accordance with expression (2), and $P^t(H)$. Its sampling variance is $(1 - t_H)^2 Var[\hat{P}_B^t(L)]$, where t_H denotes the expenditure share of subgroup A_H within group A. While this method does not necessarily lower the variance of the estimated price index, it is likely to do so as the overall sample size n increases.

10. The choice of t_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 minimise 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 two different relative sample sizes $l_H = N_H / n$ of A_H will be tried, namely $l_H = 1/3$ and $l_H = 2/3$. These values should suffice to give a clear indication of the performance.

Cut-off Sampling

11. When the sample size is very small it seems rather likely that stratification with $l_H = 2/3$ leads to a larger standard error of the estimated price index than with $l_H = 1/3$. But what happens if A_L is not observed at all, so that $l_H = 1$ 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, pp. 531-533). Since we know exactly which items will be selected there is no randomness involved and the sampling variance of $\hat{P}_C^t$ is zero by definition. The bias equals the actual error, being the difference between the estimated value and the population value: $\hat{P}_C^t - P^t = (1 - t_H)[P^t(H) - P^t(L)]$. With a very unequal distribution of item expenditures, even a small sample size would cause a large value for t_H. In that case cut-off estimation may outperform stratification, in terms of the mean square error. One can either fix the cut-off rate t_H, so that the sample size *n* is determined by t_H, or fix the sample size, in which case t_H depends on the choice of *n*. I have chosen the latter option since fixed size sampling designs are common practice in selecting CPI items, and because this allows a suitable comparison with other fixed size designs.

12. 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. However, the cut-off procedure for CPI item selection 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 the less

important items will not differ much from those of the market leaders within the same product group because of similar production cost structures.

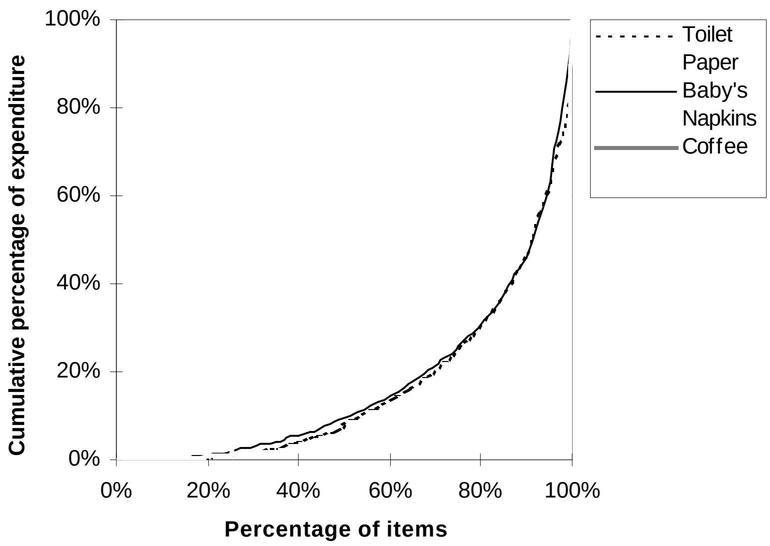
Bar-code Scanning Data

13. In Europe scannable products are defined by the European Article Number (EAN). Manufacturers assign a different EAN to every variety, size, type of packaging, etc. of a product. This has two important implications. Firstly, rapidly changing EANs – which occur very frequently – make it difficult to follow a specific item through time. Secondly, some EANs appear to have negligible expenditures. It seems that the classification system is too detailed: what is really one item has been classified as a multitude of items, and some aggregation over EANs is required for practical purpose.[3] Fortunately, several product characteristics such as brand name are included in scanner data sets. EANs with the same product characteristics will be treated as identical items. Of course, if the number of characteristics available is insufficient, there can be a danger of over-aggregation, *i.e.* of putting heterogeneous items together.

14. The scanner data sets exploited here contain weekly supermarket sales on coffee, disposable baby's napkins and toilet paper, and originally had 320, 569 and 294 different EANs, respectively. They include for each EAN the number of packages sold and the corresponding value. Prices are not included explicitly. Average transaction prices (unit values) are calculated from 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. The data on the other two item groups refer to a sample of 149 supermarkets and cover a period of two years, beginning with week 1 of 1995 and ending with week 52 of 1996.

15. For reasons of convenience the minor brands were deleted. In the case of coffee, the 15 brands with the highest expenditure during the entire period studied were chosen 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 beans), among which 40 items of ground 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 the brands with a turnover share of less than 1% were removed. Next, only those items were selected that had been sold in 1995 and at least eight months thereafter.[4] 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.

Figure 1: Distribution of base period item expenditures



16. A striking feature of the item expenditures is the skewness of their distribution. Figure 1 shows the unevenness of the base period expenditures in the adjusted data sets by means of Lorenz curves. The vertical axis depicts the cumulative expenditure total, the horizontal axis shows the cumulative number of items, both expressed as percentages. The items have been sorted in increasing order of expenditure. With evenly distributed expenditures, the Lorenz curve would lie on the diagonal. The more unequal the distribution becomes, the lower its position will be. Coffee expenditures appear to be distributed extremely unequal. The three largest items account 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.

Monte Carlo Results

17. 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 (2) has an unknown bias whereas variance estimation based on Taylor linearisation techniques gives inaccurate results because of the small CPI item samples. Systematic sampling proportional to size raises the question of how to estimate the variance since the second-order inclusion probabilities are unknown. Monte Carlo simulations have been carried out to describe the sampling distribution. Half a million samples were drawn from the item group in question according to the given design, and for each sample the price index was estimated. The distribution of the 500,000 estimates will closely approximate the exact sampling distribution.[5]

18. Three different sample sizes were used: *n*=3, *n*=6 and *n*=12. Note that all item price indexes were calculated as unit value indexes over all outlets in the sample. Part A of Table 1 shows the Monte Carlo results for coffee in January 1995 (1994=100), parts B and C those for napkins and toilet paper, respectively, in January 1996 (1995=100). 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 square root of the mean square error (*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), the *rmse* still remains considerably high. Notice that, as expected, the small sample bias is halved when the sample size is doubled.

19. Stratification works reasonably well with larger sample sizes but leads to disappointing results with *n*=3. In the latter case, stratification increases the *rmse* as compared to simple random sampling for baby’s napkins and toilet paper when $N_H = 2$ (*i.e.* when $l_H = 2/3$). Our favourite probabilistic design would clearly be sampling proportional to expenditure because the estimates are unbiased and their standard errors relatively low. But the most interesting finding 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.

20. It would be hazardous to draw conclusions about the performance of the various sampling designs based on results for a single month. Therefore, Monte Carlo experiments were performed for each month of the period under study. Figure 2 shows the *rmse* with *n*=3. The pattern that emerges for coffee and toilet paper is quite 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. Note that with larger sample sizes the results under cut-off selection and sampling proportional to size are much alike. For baby’s napkins the outcomes differ slightly. Due to the high volatility of the napkins item price indexes, the *rmse* under cut-off selection varies considerably; it meanders around the *rmse* under sampling proportional to expenditure.

Conclusion

21. Although scanner data may have some deficiencies, they provide an excellent opportunity to undertake empirical research into CPI sampling issues. Monte Carlo simulations show that, for coffee, disposable baby’s napkins and toilet paper at least, simple random sampling of items should be advised against. I believe that this recommendation can be extended to all item groups with a very skewed distribution of expenditures. Statistical offices that wish to apply probability sampling should consider using sampling proportional to size. However, the empirical evidence given in this paper supports the use of cut-off CPI item selection as a good or even better alternative.

Table 1: Monte Carlo estimates of Laspeyres price index numbers

A: Coffee (N= 68), January 1995 (1994=100)

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. *)	115.7	5.1	-1.4	5.3	116.4	3.4	-0.7	3.5	116.7	2.3	-0.4	2.3
pps	117.2	2.2	0	2.2	117.2	1.3	0	1.3	117.2	0.7	0	0.7
Stratified												
$l_H = 1/3$	116.4	3.9	-0.7	4.0	116.6	2.3	-0.5	2.3	117.0	1.2	-0.1	1.2
$l_H = 2/3$	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, estimator (2)

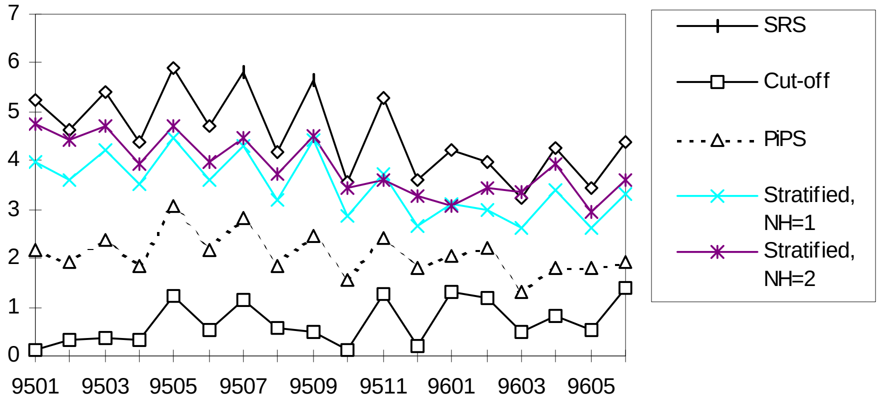
B: Baby's napkins (N= 58), January 1996 (1995=100)

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.	99.4	5.0	2.3	5.5	98.7	3.9	1.5	4.2	97.9	2.9	0.8	3.0
pps	97.2	2.8	0	2.8	97.2	1.6	0	1.6	97.2	1.5	0	1.5
Stratified												
$l_H = 1/3$	98.9	5.0	1.8	5.3	98.1	3.3	1.0	3.4	97.4	1.7	0.2	1.7
$l_H = 2/3$	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

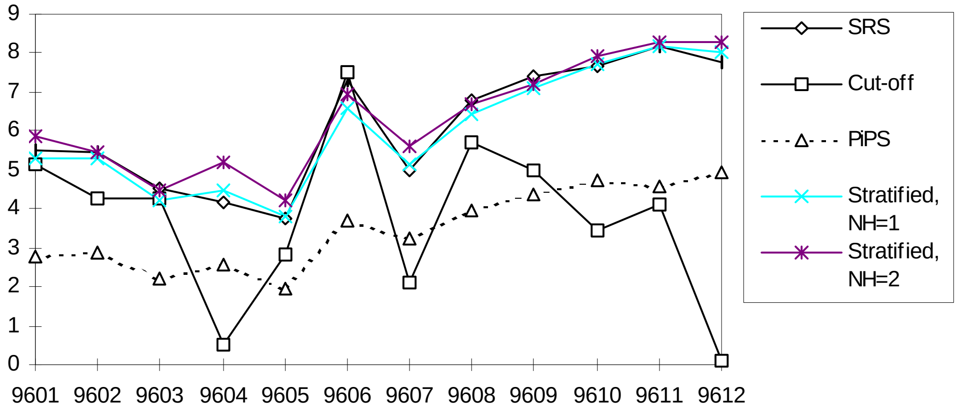
C: Toilet paper (N= 70), January 1996 (1995=100)

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
pps	103.9	3.4	0	3.4	103.9	1.8	0	1.8	103.9	1.2	0	1.2
Stratified												
$l_H = 1/3$	103.5	4.3	-0.3	4.3	103.7	3.2	-0.1	3.2	104.0	2.1	0.1	2.1
$l_H = 2/3$	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

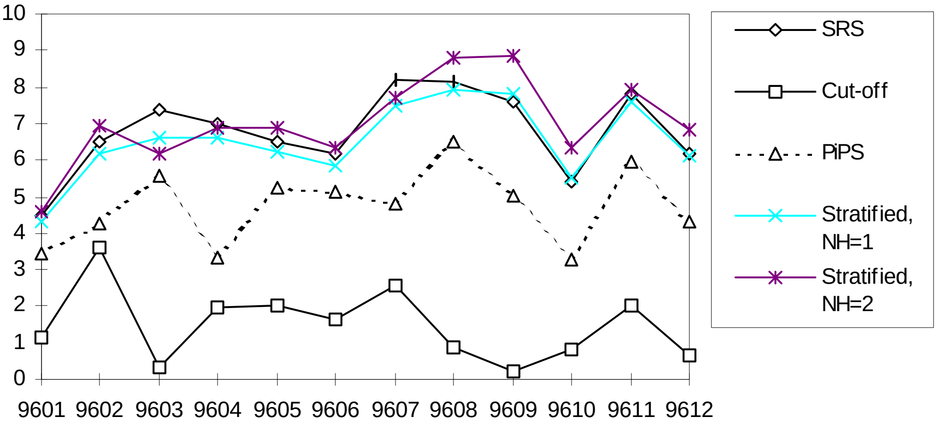
Figure 2: Rmse of estimated Laspeyres price indexes (n=3)



A: Coffee



B: Baby's napkins



C: Toilet paper

22. In this paper the items were tightly described to attain a high degree of ‘homogeneity’, following Statistics Netherlands’ practice. Some countries, on the other hand, use loose item descriptions instead. In addition, the item price indexes were treated as if they were known with certainty whereas in reality they are estimated from price observations in a sample of outlets. Opperdoes (1999) addresses the choice between tight and loose item descriptions, and takes into account both the sampling of items and the (random) sampling of outlets. The results of this preliminary study are still somewhat puzzling and inconclusive. More research into this important area would be welcome.[6]

References

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).

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

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

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

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

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

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

END NOTES

[1] Scanner data are derived from electronic scanning by bar-code reader. Bradley et al. (1997) give an overview of potential uses of scanner data in CPI construction.

[2] For the exact procedure followed, see De Haan et al. (1999).

[3] 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).

[4] For a limited number of (small) items imputation was needed in some months.

[5] For details, the reader is again referred to De Haan et al. (1999).

[6] The degree of ‘tightness’ of the item descriptions and the sampling design of outlets both have implications for the formula with which prices are aggregated into price indexes at the lowest level of commodity aggregation. See *e.g.* De Haan and Boon (1998).
