Journal of Official Statistics, Vol. 13, No. 2, 1997, pp. 143–158

Estimating Consumer Price Indices for Small Reference Populations

Martin Boon[1] *and Jan de Haan*[2]

The weights of a consumer price index are usually estimated from sample means or totals of a single household expenditure survey. This article discusses alternative estimation methods for (small) domains. Applications are given for employees' families in the Netherlands. Because of nonsampling errors, the weights could be biased. Linking the estimates to national accounts consumption data is used to assess the bias.

Key words: Expenditure weights; small-domain estimators; sampling variance; micro-macro link.

1. Introduction

In many countries the Consumer Price Index (CPI) is used as an indicator of changes in the cost of living. The cost of living is determined by purchasing habits, which can differ greatly between households. This means there is a case for constructing CPIs for various reference groups consisting of households with more or less similar expenditure patterns. Statistics Netherlands publishes not only a CPI for all households, but also (monthly) CPIs for low income and high income employees' families and (yearly) CPIs for a number of even smaller household groups. These are all fixed-weight (i.e., Laspeyres-type) indices, with weights reflecting base year expenditure shares of the various goods and services. To date the weights are estimated directly from sample means or totals of a Household Expenditure Survey (HES) held in the base year. For small reference groups – also called domains – this leads to large sampling errors. Therefore, research was undertaken to provide alternative expenditure share estimators for domains by pooling several years or by econometric modelling.

During the last decade, the accuracy of CPIs received an increasing interest. Studies into the sampling error are e.g., Balk and Kersten (1986), Biggeri and Giommi (1987), Leaver et al. (1991), Leaver and Swanson (1992), Leaver and Valliant (1995) and Dalén and Ohlsson (1995). The total sampling error of a CPI results from errors in the partial price index numbers and errors in the weights. At least in the Netherlands, with its relatively small overall HES sample size, the latter cannot be neglected. Boon (1991) concluded for

[1] Statistics Netherlands, Division of Research and Development, P.O. Box 4000, 2270 JM VOORBURG, The Netherlands.
[2] Statistics Netherlands, Division of Socio-Economic Statistics.
Acknowledgments: We are grateful to Bert Balk and Hans van Driel as well as to three anonymous referees for stimulating comments and suggestions and to Frank Linder for research assistance. The views expressed in this article are solely those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

the CPIs of employees' families that the contribution of the weights to the total variance is about the same as the contribution of the sampling of outlets where prices are measured.

It is well-known that choosing a fixed weight formula to measure the true cost-of-living index gives rise to several kinds of biases, such as substitution bias with respect to goods and outlets. Wynne and Sigalla (1994) review the literature on this subject. We shall not address this issue, but simply take the Laspeyres CPI as the object of estimation. Another type of bias stems from nonsampling errors, such as underreporting of expenditures in the HES. National accounts data will be used to make an inference about the bias in the CPI this causes.

The remainder of this article is organized as follows. Section 2 gives a formal description of the ordinary direct estimator and two small-domain estimators. Section 3 goes into variance estimation. Section 4 discusses the data used for the empirical application. Section 5 describes the model specification applied for model-based estimation. In Section 6 we present standard errors for employees' families in the Netherlands. Section 7 deals with linking the direct estimates to national accounts data. Section 8 concludes.

2. Estimation of the Laspeyres Consumer Price Index

The Laspeyres CPI of period t with respect to the base year 0 is a weighted average of partial price indices:

$$P^t = \Sigma_g w_g^0 \pi_g^t$$

where π_g^t denotes the price index of commodity group g, the summation being over all $g = 1, \ldots, G$ commodity groups that fall within the scope of the CPI. The weight w_g^0 equals the share of the expenditure on g (e_g^0) in the total base year expenditure on all commodities ($e^0 = \Sigma_g e_g^0$) for a certain group of households. Let $\hat{e}_g^0$ be an unbiased estimator of the expenditure on g, $\hat{e}^0 = \Sigma_g \hat{e}_g^0$ and $\hat{w}_g^0 = \hat{e}_g^0/\hat{e}^0$. The weight estimators are ratios and therefore biased. However, this bias tends to zero with increasing sample size and may in practice be neglected, even with very moderate sample sizes (see Särndal et al. 1992, p. 177). Treating the partial price indices as if they were known with certainty, the estimator of the Laspeyres CPI becomes

$$\hat{P}^t = \Sigma_g \hat{w}_g^0 \pi_g^t = \Sigma_g \hat{e}_g^0 \pi_g^t / \hat{e}^0$$

The CPI is mainly used as a short-term indicator. It is not so much the CPI itself that matters for most ends but rather its relative change over time, estimated by

$$(\hat{P}^t/\hat{P}^s) - 1 = (\Sigma_g \hat{w}_g^0 \pi_g^t / \Sigma_g \hat{w}_g^0 \pi_g^s) - 1 = (\Sigma_g \hat{e}_g^0 \pi_g^t / \Sigma_g \hat{e}_g^0 \pi_g^s) - 1 \quad (t > s) \tag{1}$$

2.1. *Direct estimation*

In the Netherlands' HES a substantial fraction of the sampled households is obtained by stratified random sampling, while the remaining part (mainly in low response strata) is selected by several non-probability techniques. Furthermore, the sample is poststratified. In view of the complexity of the actual sample design and the weighting of the sample, we assume stratified random sampling with fixed stratum sizes. Let N_k and n_k denote the sizes of the kth population stratum and sample stratum, respectively, and let e_{ig}^0 denote the base

year expenditure of household i on commodity group g. The expenditure on g for a domain that comprises the first K strata can be estimated by the stratified Horvitz-Thompson estimator

$$\hat{e}_g^0(D) = \sum_{k=1}^{K}(N_k/n_k)\sum_{i=1}^{n_k} e_{ig}^0 \tag{2}$$

We will call $\hat{e}_g^0(D)$ the direct (D) estimator of e_g^0. If the household group is partly made up of incomplete strata, the expenditure of households not belonging to this group will be put to zero, following Särndal et al. (1992, p. 390).

For domains with a small number of sampled households, direct estimation according to (2) leads to CPI changes with high standard errors. One way to reduce standard errors is to construct CPI weights simply by averaging the directly estimated expenditure shares $\hat{w}_g^t(D) = \hat{e}_g^t(D)/\Sigma_g\hat{e}_g^t(D)$ from successive years t. We shall take into consideration the base year 0 and the years before (-1) and after (1). The three-year moving average

$$\hat{w}_g^0(C) = [\hat{w}_g^{-1}(D) + \hat{w}_g^0(D) + \hat{w}_g^1(D)]/3 \tag{3}$$

will be referred to as the combined direct (C) estimator of w_g^0. It is generally believed that expenditure patterns are stable in the short run, or at least that expenditure shares change in a fairly smooth way. If expenditure shares are linear functions of time and the direct estimators $\hat{w}_g^t(D)$ are (approximately) unbiased, then $\hat{w}_g^0(C)$ is an (approximately) unbiased estimator of w_g^0 too.

2.2. *Model-based estimation*

Another way to increase precision is model-based estimation, which uses data (including auxiliary information) from the domain itself as well as from outside the domain. Suppose that the following linear regression model for the expenditure share of g in year t holds for every household in the entire population:

$$w_{ig}^t = \Sigma_m\beta_{gm}x_{mi}^t + \varepsilon_{ig}^t \tag{4}$$

with (time-independent) parameters β_{gm}, $x_{1i}^t = 1$ for all i and explanatory variables x_{mi}^t ($m = 2, \ldots, M$); ε_{ig}^t is a stochastic disturbance term with expected value $E(\varepsilon_{ig}^t) = 0$, $var(\varepsilon_{ig}^t) = \sigma_g^2$, and $cov(\varepsilon_{ig}^t, \varepsilon_{jg}^t) = 0$ ($i \neq j$). The parameters will be estimated with *ordinary least squares* regression (OLS) on HES microdata. Under the assumptions mentioned, the OLS estimates $\hat{\beta}_{gm}$ are unbiased, conditional on the x-values in the sample. Since model-based estimation is not design-unbiased, model misspecification causes a bias component that cannot be estimated. In Section 5 the specification applied for empirical estimation is given, along with some regression diagnostics in order to check the appropriateness of the model.

The base year expenditure share of household i is predicted by $\hat{w}_{ig}^0 = \Sigma_m\hat{\beta}_{gm}x_{mi}^0$. Multiplying by total expenditure e_i^0 yields the expenditure predictor $\hat{e}_{ig}^0$. If the x-values and total expenditure were known for the entire population with size N, we would be able to estimate e_g^0 as

$$\hat{e}_g^0(R) = \sum_{i=1}^{N}\hat{e}_{ig}^0 = \sum_m \hat{\beta}_{gm}z_m^0$$

in which $z_m^0 = \Sigma_{i=1}^N z_{mi}^0$ with $z_{mi}^0 = e_i^0 x_{mi}^0$, and where R stands for 'regression'. This approach strongly resembles regression estimation in the usual sense (see e.g., the small-domain estimators in Särndal et al. 1992). However, since expenditure data are available only for the sampled households, the z-variables cannot be calculated for the entire population. In addition, some of the x-values may be unknown due to the absence of a census or any other useful register, as is the case in the Netherlands. We therefore replace z_m^0 by the stratified Horvitz-Thompson estimator $\hat{z}_m^0 = \Sigma_{k=1}^K (N_k/n_k)\Sigma_{i=1}^{n_k} z_{mi}^0$, so that our model-based (M) estimator becomes

$$\hat{e}_g^0(M) = \sum_m \hat{\beta}_{gm}\hat{z}_m^0 = \sum_{k=1}^K (N_k/n_k)\Sigma_{i=1}^{n_k}\hat{e}_{ig}^0 \tag{5}$$

One may wonder why the variance of this estimator should be lower than the variance of the direct estimator. After all, we are still raising the same number of sampled households up to the entire subpopulation. The variance reduction comes from the smoothing of individual expenditure shares induced by regression, which reduces the (unknown) population variance S^2 of expenditures. The reduced variance is offset by a model component. This does not completely balance out the variance reduction, however, because the $\hat{\beta}$'s borrow strength from a larger sample. In that respect we note that it is not necessary to restrict the sample from which the parameters are estimated to the base year HES. The regression may be performed on pooled cross section data of consecutive years to increase precision even further, provided the model is suitably specified.

3. Variance Estimation

Linearizing the last expression in (1) by a first order Taylor series approximation leads to the following general variance formula (Särndal et al. 1992, pp. 172–176):

$$var[(\hat{P}^t/\hat{P}^s) - 1] = var(\hat{P}^t/\hat{P}^s) \simeq (P^t/P^s)^2 \Sigma_g \Sigma_h B_{gh} cov(\hat{e}_g^0, \hat{e}_h^0)/(e^0)^2 \tag{6}$$

where $B_{gh} = [(\pi_g^t/P^t) - (\pi_g^s/P^s)][(\pi_h^t/P^t) - (\pi_h^s/P^s)]$.
The variance of the CPI change, conditional on the partial price indices, is expressed in (6) as a weighted sum of the variances and covariances of expenditures on all commodity groups. It is a function of time, although CPI weights remain unchanged. If prices change with equal rates ($\pi_g^t = \pi_h^t = \pi^t$ and $\pi_g^s = \pi_h^s = \pi^s$ for all g, h) then $var[(\hat{P}^t/\hat{P}^s) - 1] = 0$, for in that case weighting of partial price indexes does not affect the CPI change. Notice that (6) simplifies to a first order approximation of $var(\hat{P}^t - 1) = var(\hat{P}^t)$ if $\pi_g^s = \pi_h^s = 1$ for all g, h.

For each of the estimation methods mentioned in Section 2 we will point out how the $G \times G$ (co)variance matrix of base year expenditures may be estimated. In case of the direct method the (co)variances are given by

$$cov(\hat{e}_g^0(D), \hat{e}_h^0(D)) = \sum_{k=1}^K \{(N_k)^2(1 - f_k)/n_k\} S_{gh_k} \tag{7}$$

where S_{gh_k} denotes the population (co)variance of base year expenditures on g and h within the kth stratum. Estimation of the stratum (co)variances is straightforward by using the conventional estimator. A similar 'classical' approach will be used for the combined

direct estimator. We shall act as if the three HES samples were independently drawn. In reality a minor part of the sampled households, especially in low response strata, prolong their co-operation for another year. This panel character is likely to increase the variance of the CPI-change to some extent. As a result, we will be estimating a lower bound of the variance.

The stochastic properties of the model-based estimator depend on both the sampling design of the HES and the model that is assumed to have generated individual expenditure shares. Pannekoek and Zeelenberg (1995) decompose the (co)variance of model-based expenditure estimators into

$$cov(\hat{e}_g^0(M),\ \hat{e}_h^0(M)) = \Sigma_m\Sigma_n E_d[\hat{z}_m^0\hat{z}_n^0 cov_\xi(\hat{\beta}_{gm}, \hat{\beta}_{hn})] + \Sigma_m\Sigma_n\beta_{gm}\beta_{hn}cov_d(\hat{z}_m^0, \hat{z}_n^0) \qquad (8)$$

where $cov_\xi(.)$ and $cov_d(.)$ denote the covariance with respect to the model and the sampling design, respectively. An unbiased (pseudo) estimator of the model component, the first term on the right hand side of (8), is $\Sigma_m\Sigma_n\hat{z}_m^0\hat{z}_n^0 cov_\xi(\hat{\beta}_{gm}, \hat{\beta}_{hn})$. A consistent estimator of the covariance matrix of regression coefficients can easily be derived from the covariance matrix of regression residuals. The formula for $cov_d(\hat{z}_m^0, \hat{z}_n^0)$ in the design component, the second term on the right of (8), is analogous to (7).

4. About the Data

Statistics Netherlands currently computes monthly CPIs with base year 1990 for three population groups:

- all private households (in 1990 a number of 6.1 million);
- low income employees' families (having an income below the median of the base year income distribution of employees' families; the corresponding CPI will be denoted CPI-L);
- high income employees' families (with an income above the median: CPI-H).

Employees' families are private households consisting of a married couple without children or with non-earning children of which the reference person (i.e., the head of the household) is a full-time employee. According to this definition, employees' families form a subset of all employees' households; for instance, single living employees and non-married couples are excluded. Income is measured as total gross household income, including income transfers.

The Netherlands' HES is a continuing survey, held every year since 1980. Usually the total sample size amounts to some 2,000 households. In 1990 the sample size was enlarged after the usual sampling procedure had taken place. This second stage, referred to as over-sampling, was added to meet CPI demands: the initial number of sampled employees' families was thought to be too small to estimate CPI-L and CPI-H accurately. The over-sampling consisted of 620 households, of which 595 employees' families (mostly with lower income). Hence, the entire 1990 sample contains a substantially larger fraction of employees' families than the samples of 1989 and 1991, as can be seen from Table 1. To finance the additional costs, the 1991 sample was halved with respect to the normal size.

Our database contains detailed information about the yearly expenditures of all 5,780 sampled households in 1989–1991, together with a great number of household characteristics. It should be remarked that the households in the Netherlands' HES report their

Table 1. Sample frequency distribution of households by occupation of the reference person

	1989		1990				1991	
			(1)		(2)			
	abs.	%	abs.	%	abs.	%	abs.	%
Employees' households	1,128	58	1,835	66	1,227	57	555	52
of which								
low income families	388	20	1,076	39	583	27	261	25
high income families	466	24	514	19	412	19	168	16
other employees' households	274	14	245	9	232	11	126	12
Households of self-employed	172	9	202	7	198	9	132	12
Households of inactive	646	33	730	26	722	34	380	36
All households (total)	1,946	100	2,767	100	2,147	100	1,067	100

(1) = including oversampled households.
(2) = excluding oversampled households.

expenditures on many commodity groups only during a short period. The observed expenditures are raised to yearly figures. We do not take this stochastic element into account, however, and treat the household's yearly expenditures as if they were measured without error. Total expenditure is defined in accordance with the scope of the CPI. The scope includes all goods and services acquired out of net household income, with the exception of health insurance (Balk 1994). At the lowest level of aggregation we distinguish 186 commodity groups. The corresponding partial price index numbers for 1991–1994 are identical to those used for the official CPI-L and CPI-H.

The stratification scheme of the Netherlands' HES is very comprehensive. For example in 1990 there are over 300 sample strata, implying an average of less than 9 households per stratum. To obtain reasonably stable stratum (co)variance estimates, we collapsed the stratification scheme and recalculated raising factors. For 1990 this resulted in 111 strata with a minimum size of 10 households. For 1989 and 1991 similar procedures were adopted. The new raising factors will not only be used to estimate the variance of CPI (changes), but also to estimate the CPI itself. This may cause slight deviations from officially published figures.

5. Model Specification

In the appendix a linear regression model has been derived from econometrics. Apart from the year-dummies, the resulting model postulates expenditure shares as a function of total expenditure (income for short) and a set of household characteristics. The following household characteristics have been selected for empirical estimation: the equivalence factor (a measure of standardized household size; see Schiepers 1992), household composition, age, sex, and occupation of the reference person, housing situation, and region of residence. All explanatory variables are categorical and represented by dummies, except for total expenditure and the equivalence factor which are specified as logarithms. The 'income' elasticity implied by the model is income-dependent and the model admits luxury ($\beta_g > 0$) as well as necessary ($\beta_g < 0$) commodity groups.

The model parameters are estimated on pooled microdata of the entire 1989–1991 HES samples, excluding oversampled households. Table 2 shows the OLS results for the nine highest expenditure aggregates. Not surprisingly in working with (pooled) cross section data, the coefficient of determination R^2 is low. Except for non-insured medical care, the income coefficient always differs significantly from zero at the 95 per cent level. The negative signs for food, rents, and consumption-related taxes indicate that these are necessary expenditure groups. The year-dummies show little effect on the whole.

Table 2 also presents some regression diagnostics. The F test of the overall relation indicates that the included variables have a significant joint effect on the expenditure share of every commodity group. However, some values of the Ramsey Reset test for omitted variables are fairly high, which may be due to the omission of interaction effects of the explanatory variables. The Cook-Weisberg χ^2 test suggests non-constant residual variance. In that case the variance estimates of the regression coefficients will be biased upwards, which in turn means that our model-based variance estimates of the CPI changes are upward biased as well.

6. Standard Error Estimates

Employees' families will not be oversampled in the Netherlands' HES of future CPI base years. The alternative estimation methods described in Section 2 have been used to re-estimate CPI-L and CPI-H with base year 1990. The results should be helpful in deciding which method Statistics Netherlands must choose to calculate the most accurate CPIs. Therefore, the oversampled part of the 1990 sample is excluded from the empirical analysis as far as the alternative methods concern. Table 3 shows estimated annual CPI changes for 1991–1994. The direct estimates are also calculated including oversampled households. These figures serve as a benchmark for the alternative estimates and we discuss them first. Cost of living changes have been very modest in the nineties up till 1995, with consumer prices going up about three per cent annually. Low and high income employees' families experienced slightly different price increases: CPI-L increased 0.15 percentage points more than CPI-H in 1991 and 1994. The standard errors are low: some 0.02 per cent each year, implying a relative standard error between 0.7 and 1.0 per cent. Although the exclusion of oversampled employees' families hardly affects the magnitude of direct estimates, their standard errors increase of course.

It is important to check whether the alternative estimators produce CPI-L and CPI-H changes with at least the same precision as the direct estimation method including oversampling. The model-based estimator in particular meets this requirement. The standard errors of the model-based estimates coming from regression on pooled data are 0.5 to 0.8 times the standard errors of the direct estimates. We add that the design components appear to be rather small, on average a tenth of the model components. Limiting the regression to base year data increases standard errors slightly. Combining direct estimates gives rise to larger standard errors (which are likely to be underestimated) than model-based estimation on pooled data. This method therefore seems second-best.

Table 4 contains some estimated budget shares and their relative standard errors. We selected the four commodity groups with the largest difference between the relative standard errors of the direct estimates (excluding oversampling) and the model-based

Table 2. OLS estimates for aggregate commodity groups

Variable	Commodity group (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total expenditure (log)	−0.081*	0.008*	−0.136*	0.041*	0.000	0.114*	0.026*	0.043*	−0.015*
Equivalence factor (log)	0.119*	0.031*	−0.018	−0.044*	−0.001	−0.056*	−0.013	−0.051*	0.033*
Household composition									
married couple with children	0.014*	−0.001	0.009*	−0.001	0.000	−0.010	0.003	−0.010*	−0.004*
single parent	0.000	−0.019*	0.009	−0.013*	−0.001	−0.004	0.023*	0.001	0.004
single person and other	−0.025*	−0.013*	0.001	−0.017*	−0.001*	0.019*	0.023*	0.007*	0.005*
Age of reference person									
25–44	0.036*	−0.001	0.046*	0.012*	0.001	−0.035*	−0.012*	−0.016*	−0.031*
45–64	0.045*	−0.001	0.051*	0.008	0.003*	−0.033*	−0.019*	−0.025*	−0.028*
65 and above	0.043*	−0.002	0.068*	0.018*	0.004*	−0.038*	−0.034*	−0.022*	−0.037*
Sex of reference person									
female	−0.007	0.028*	0.009*	0.010*	0.001	−0.019*	−0.015*	−0.005	−0.002
Occupation of reference person									
self-employed	0.012*	0.004	0.021*	−0.001	−0.000	−0.026*	−0.014*	0.002	0.003*
inactive	0.010*	−0.007*	0.013*	−0.002	0.001*	0.001	−0.010*	−0.010*	0.003*
Housing situation									
owning a dwelling	−0.015*	−0.004*	0.077*	−0.010*	−0.000	−0.029*	−0.013*	−0.010*	0.005*
Region of residence									
outside big cities	−0.010*	0.002	0.027*	0.002	0.001	0.009*	−0.012*	−0.017*	−0.001
Year									
1989	−0.000	0.001	−0.003	0.004*	0.000	0.003	0.000	−0.003	−0.002*
1991	−0.002	−0.003	0.016*	−0.002	0.001*	−0.007	0.001	−0.004*	0.001
Intercept	0.971*	−0.022	1.587*	−0.341*	−0.001	−0.999*	0.131*	−0.280*	0.214*

Table 2. Continued

Variable	Commodity group (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
R^2	0.31	0.07	0.42	0.06	0.03	0.15	0.10	0.11	0.12
Overall relation test $F(15,5144)$	150.47	27.01	243.51	23.71	11.74	61.60	39.50	44.08	47.26
Omitted-variables test $F(3,5141)$	7.56	4.45	16.41	4.89	2.40	12.41	0.83	1.55	78.21
Heteroscedasticity test $\chi^2(1)$	112.15	25.62	201.62	286.42	909.27	302.78	192.98	175.90	695.80

* = significantly different from zero at the 95% level.
(1) = Food, beverages, tobacco.
(2) = Clothing, footwear.
(3) = Gross rents, energy.
(4) = Furniture, household equipment.
(5) = Medical care (non-insured).
(6) = Transport, communications.
(7) = Recreation, education, entertainment.
(8) = Miscellaneous goods and services.
(9) = Consumption-related taxes and government services.

Table 3. Annual change of consumer price index numbers (base year 1990) for employees' families; standard errors in parentheses

	Low income				High income			
	1991	1992	1993	1994	1991	1992	1993	1994
	%							
Direct estimate[a]	3.10	3.14	2.52	2.75	2.95	3.14	2.52	2.60
	(0.018)	(0.022)	(0.019)	(0.019)	(0.023)	(0.031)	(0.022)	(0.025)
Direct estimate[b]	3.13	3.14	2.52	2.78	2.95	3.14	2.52	2.60
	(0.022)	(0.024)	(0.024)	(0.023)	(0.025)	(0.034)	(0.024)	(0.027)
Combined direct estimate	3.11	3.15	2.51	2.77	2.95	3.12	2.51	2.56
	(0.017)	(0.018)	(0.017)	(0.019)	(0.018)	(0.021)	(0.017)	(0.018)
Model-based estimate[c]	3.11	3.16	2.57	2.81	2.92	3.12	2.49	2.52
	(0.014)	(0.015)	(0.015)	(0.016)	(0.017)	(0.017)	(0.016)	(0.020)
Model-based estimate[d]	3.12	3.15	2.56	2.81	2.93	3.11	2.49	2.52
	(0.016)	(0.017)	(0.017)	(0.018)	(0.020)	(0.022)	(0.020)	(0.023)

[a] Including oversampled households.
[b] Excluding oversampled households.
[c] Pooled data.
[d] Base year data only.

Table 4. Budget shares (×10,000) of selected commodity groups for employees' families, 1990; relative standard errors (in %) in parentheses

	Low income				High income			
	liqueur	kitchen furniture	cutlery	motorcycles	liqueur	kitchen furniture	cutlery	motorcycles
Direct estimate[a]	2.01	21.08	3.19	10.97	2.59	29.27	11.86	12.39
	(24.38)	(22.87)	(20.69)	(59.62)	(36.29)	(35.53)	(56.16)	(41.40)
Direct estimate[b]	1.49	14.46	3.48	7.76	2.68	32.97	11.44	13.19
	(40.94)	(29.05)	(27.01)	(81.19)	(38.06)	(36.21)	(57.95)	(43.14)
Combined direct estimate	1.26	12.07	2.07	5.65	2.98	29.26	6.99	9.82
	(27.78)	(22.54)	(17.87)	(47.79)	(20.47)	(22.62)	(33.76)	(26.99)
Model-based estimate[c]	3.69	14.68	4.28	7.86	5.23	23.60	7.12	12.31
	(12.47)	(18.26)	(16.36)	(32.19)	(9.56)	(12.37)	(10.67)	(22.26)
Model-based estimate[d]	3.76	15.27	5.03	8.53	5.20	23.25	9.02	12.12
	(14.49)	(20.73)	(20.55)	(39.26)	(12.33)	(16.05)	(13.51)	(32.50)

[a] Including oversampled households.
[b] Excluding oversampled households.
[c] Pooled data.
[d] Base year data only.

estimates for high-income families. Most of the standard errors are extremely high, up to 81 per cent in case of the direct estimate (excluding oversampling) of motorcycles and mopeds for low-income employees' families. One should remember, however, that durable goods like motorcycles are bought occasionally, so that a great number of the sampled households report zero expenditures for these goods.

7. Micro-macro Link

The Fourteenth International Conference of Labour Statisticians 1987 recommended that "Before any of the survey results are used to provide weights for the index, it is necessary to examine them carefully, e.g., in the light of sampling and nonsampling errors, in order to judge whether the survey has provided reliable and representative information. Adjustments should be made, if necessary, using other available statistics." (see Turvey 1989, p. 126). National Accounts (NA) consumption data are constructed by confronting HES outcomes with various other sources, such as production and foreign trade data. As a result, they are probably less sensitive to nonsampling (as well as sampling) errors and may be regarded as the most accurate estimates of the population values (Dalén 1995). In this section, we will link the direct HES-expenditure estimates to the NA data. This gives some insight into the magnitude of the CPI bias caused by nonsampling errors. A two-step procedure is followed.

First, at the lowest level of aggregation (186 expenditure groups), the NA data are corrected for any differences in reference population between the CPI for all private households and the NA and for differences in the definition of total expenditure (see Linder 1996, for details). The most important differences are that the NA household sector includes people living in institutional households and private non-profit institutions and that total expenditure includes health care and excludes most consumption-related taxes and government services. Ignoring sampling errors, the relative difference between the direct HES-based estimate for all private households (including oversampled households) and the adjusted NA figure can be regarded as an, albeit rather crude, measure of the relative bias of the estimate. In 1990, the 'relative bias' ranges from -164 per cent (films and other photographic supplies) to 85 per cent (articles for the transport of children).

Second, for each commodity group the difference between the adjusted NA figure and the direct estimate for all private households is distributed among employees' families and other households proportional to the direct HES-based estimates. Assuming the relative bias is equal for all household groups, this leads to a bias correction of the initial estimates, at least to some extent. Table 5 illustrates the effect of this so-called micro-macro link on the expenditure pattern of employees' families on the highest aggregation level.

Table 6 shows the annual CPI changes for 1994 before and after linking to the adjusted NA data. The differences between the initial estimates and the new ones are about ten times the standard error of the former. The initial estimates seem to overstate cost of living increases in 1994 by 0.2 per cent. However, we must be careful in drawing far-reaching conclusions. It is true that the NA consumption data are stable in time and may generally be regarded as the best figures available, but they remain estimates and are not completely without error. To give an example: in the Netherlands there is a suspicion that the NA figure on imputed rents for owner-occupied housing suffers from substantial downward bias.

Table 5. Aggregate budget shares for employees' families, 1990[a] (per cent)

	Low income			High income		
	(1)	(2)	(1)–(2)	(1)	(2)	(1)–(2)
Food, beverages, tobacco	19.98	20.68	−0.70	16.36	16.79	−0.43
Clothing, footwear	7.37	8.29	−0.92	8.06	8.98	−0.93
Gross rents, energy	26.19	22.48	3.71	23.27	19.51	3.76
Furniture, household equipment	7.52	7.73	−0.21	8.97	9.28	−0.31
Medical care (non-insured)	0.48	0.59	−0.11	0.46	0.56	−0.10
Transport, communications	13.88	14.64	−0.76	15.38	16.01	−0.63
Recreation, education, entertainment	10.62	10.76	−0.15	11.71	11.75	−0.04
Miscellaneous goods and services	10.00	11.74	−1.74	12.33	14.53	−2.20
Consumption-related taxes, etc.	3.98	3.09	0.89	3.46	2.59	0.87
Total (all items)	100	100	0	100	0	0

[a] Direct estimates, including oversampled households.
(1) = HES-based estimates.
(2) = NA-linked estimates.

8. Conclusions

The standard error of HES-based CPI changes can be lowered by combining directly estimated budget shares from successive HESs. Nevertheless, the use of this method must be advised against, because of the bias that arises when budget shares do not move exactly linearly in time. In addition, the combined direct estimator has the disadvantage that publication of the resulting CPI changes will be held up, since HES outcomes of year +1 are published one year after those of year 0. Application of the model-based estimator seems more promising. Not only does it induce a larger standard error reduction, it is also able to take account of price effects and other influences on consumer demand. The model-based estimator can be applied to HES data of year −2, −1 and 0 (instead of −1, 0 and +1, as we did) and will in that case cause no delay for publication.

The alternative estimators have been applied to domains of moderate size, with 400–600 sampled households in 1990. For smaller domains the gain in precision is expected to be larger. According to our results, there is no need to apply an alternative estimation method for the CPIs for employees' families. Without oversampling of employees' households the standard errors of the directly estimated (all items) CPI changes are still

Table 6. Annual change of consumer price index numbers for employees' families, 1994 (per cent)

	Low income			High income		
	(1)	(2)	(1)–(2)	(1)	(2)	(1)–(2)
Direct estimate[a]	2.75	2.52	0.23	2.60	2.38	0.22
Direct estimate[b]	2.78	2.54	0.24	2.60	2.37	0.23

[a] Including oversampled households.
[b] Excluding oversampled households.
(1) = HES-based estimates.
(2) = NA-linked estimates.

acceptable; detailed household group-specific subindices will not be published any longer, starting with the coming revision to base year 1995. For smaller household groups Statistics Netherlands intends to use the model-based estimation method in the near future.

In view of the large difference between CPIs calculated entirely with HES-based weights and those which are linked to NA consumption data, we advocate to undertake further research into this area. The difference between (direct) HES and NA outcomes for all private households should be analysed in greater detail and over a much longer period. Within the framework of the *Socio-economic accounts* (Statistics Netherlands 1994) this kind of analysis is available for a period of over ten years. Unfortunately, the classification of goods and services differs substantially from the CPI classification, making it less useful for our purpose.

9. Econometric Appendix

The Almost Ideal Demand (AID) System of Deaton and Muellbauer (1980, pp. 75–78), which is probably the most widely studied system of linear consumer demand equations, serves as our starting point for model-based estimation of expenditure shares. For individual household i it reads

$$w_{ig}^t = \alpha_{ig} + \beta_g \log(e_i^t/p_i^t) + \sum_{h=1}^{G} \gamma_{gh} log(p_{ih}^t) + \varepsilon_{ig}^t \ (g = 1, \ldots, G) \tag{A.1}$$

where p_{ih}^t denotes the 'price' paid by household i for commodity (group) h in year t and p_i^t the household-specific price level. Using cross-section data, we must take into account the fact that (unobservable) preferences diverge across households. To capture their effect on consumer demand, α_{ig} is replaced by a constant α_g plus a linear function $\Sigma_m \delta_{gm} x_{mi}^t$ of (mostly demographic) household characteristics. Under the assumption that p_{ih}^t is identical for all i, the yearly price effects $\gamma_{gh} log(p_{ih}^t)$ are also equal. The third term on the right-hand side of (A.1) represents the total price effect. Assuming that the price level is approximately equal for all i, we take this effect together with the effect $-\beta_g \log(p_i^t)$ and replace e_i^t/p_i^t by e_i^t. We may now estimate

$$w_{ig}^t = \alpha_g + \beta_g log(e_i^t) + \gamma_g^{-1} D_i^{-1} + \gamma_g^1 D_i^1 + \Sigma_m \delta_{gm} x_{mi}^t + \varepsilon_{ig}^t \ (g = 1, \ldots, G) \tag{A.2}$$

on the pooled microdata of years $t = -1, 0, 1$. Dummy D_i^t has the value 1 if household i belongs to the HES of year t (and 0 if not); the base year is chosen as reference year. Alternatively, (A.2) may be estimated on base-year data only, leaving out the year-dummies of course.

Because we are estimating a complete set of demand equations instead of the single equation (4) mentioned in Section 2.2, some additional assumptions about the covariance structure of disturbances are needed. We take $cov(\varepsilon_{ig}^t, \varepsilon_{ih}^t) = \sigma_{gh}$ and $cov(\varepsilon_{ig}^t, \varepsilon_{jh}^t) = 0$ $(i \neq j, g \neq h)$. Furthermore, since expenditure shares must sum to unity, we impose the so-called adding-up restrictions $\Sigma_g \alpha_g = 1$, $\Sigma_g \beta_g = \Sigma_g \gamma_g^{-1} = \Sigma_g \gamma_g^1 = \Sigma_g \delta_{gm} = 0$, and $\Sigma_g \varepsilon_{ig}^t = 0$. A set of equations like (A.2) should in general be estimated by *generalized least squares* (GLS). Zellner (1962) proposed a method, known as *seemingly unrelated*

regression equations. However, this method reduces to OLS if the explanatory variables are identical in all equations (Theil 1971, pp. 309–310). OLS satisfies the adding-up restrictions automatically (Cramer 1986, p. 112).

10. References

Balk, B.M. (1994). The New Consumer Price Indices of Statistics Netherlands. Statistical Journal of the United Nations, ECE, 11 (2), 119–123.

Balk, B.M. and Kersten, H.M.P. (1986). On the Precision of Consumer Price Indices Caused by the Sampling Variability of Budget Surveys. Journal of Economic and Social Measurement, 14, 19–35.

Biggeri, L. and Giommi, A. (1987). On the Accuracy and Precision of the Consumer Price Indices; Methods and Applications to Evaluate the Influence of the Sampling of Households. Bulletin of the International Statistical Institute 52, Book 3, 137–154.

Boon, M. (1991). The Variance of Consumer Price Indices: Some Estimation Results. Maandstatistiek van de Prijzen 16, no. 11, 8–14. [In Dutch].

Cramer, J.S. (1986). Econometric Applications of Maximum Likelihood Methods. Cambridge: Cambridge University Press.

Dalén, J. (1995). Quantifying Errors in the Swedish Consumer Price Index. Journal of Official Statistics, 11, 261–275.

Dalén, J. and Ohlsson, E. (1995). Variance Estimation in the Swedish Consumer Price Index. Journal of Business and Economic Statistics, 13, 347–356.

Deaton, A. and Muellbauer, J. (1980). Economics and Consumer Behavior. Cambridge: Cambridge University Press.

Leaver, S.G., Johnstone, J.E., and Archer, K.P. (1991). Estimating Unconditional Variances for the U.S. Consumer Price Index for 1978–1986. Proceedings of the Survey Research Methods Section, American Statistical Association, 614–619.

Leaver, S.G. and Swanson, D. (1992). Estimating Variances for the U.S. Consumer Price Index for 1987–1991. Proceedings of the Survey Research Methods Section, American Statistical Association, 740–745.

Leaver, S. and Valliant, R. (1995). Statistical Problems in Estimating the U.S. Consumer Price Index. In B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds.), Business Survey Methods. New York: John Wiley.

Linder, F. (1996). Reducing Bias in the Estimation of Consumer Price Indices by Using Integrated Data. Voorburg: Statistics Netherlands.

Pannekoek, J. and Zeelenberg, K. (1995). The Variance of Model-based Estimators. Voorburg: Statistics Netherlands. [In Dutch].

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

Schiepers, J. (1992). On the Choice of Equivalence Scales. Statistical Journal of the United Nations, ECE, 9 (2), 111–124.

Statistics Netherlands (1994). Socio-economic Accounts 1993. Voorburg.

Theil, H. (1971). Principles of Econometrics. New York: John Wiley.

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

Wynne, M.A. and Sigalla, F.D. (1994). The Consumer Price Index. Economic Review. Federal Reserve Bank of Dallas, Second Quarter, 1–22.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. Journal of the American Statistical Association, 57, 348–368.

Received January 1996
Revised November 1996