Estimation of the Coffee Price Index Using Scanner Data: Simulation of Official Practices

Jan de Haan and Eddy Opperdoes[1]

Abstract: With the use of Nielsen scanner data on coffee sales Statistics Netherlands is undertaking empirical research into the effects of alternative micro indexes and different commodity sampling designs on the (Laspeyres-type) consumer price subindex of coffee. In this paper we show how scanner data may be used to approximate official concepts in constructing such an index.

1. Introduction

In compiling a consumer price index (CPI) various decisions are taken which can have a substantial influence on the outcome. These decisions relate for example to the sampling of items and outlets, the micro index formula and the aggregation of the micro indexes. Statistics Netherlands is undertaking empirical research in this area using scanner data from A.C. Nielsen (Nederland) B.V. with weekly coffee sales in 20 supermarkets during 1994, 1995 and the first half of 1996. Our main purpose is to gain insight into the effects of alternative micro index formulas and commodity sampling designs on the price index of coffee, and empirically check some of the assumptions made. Hopefully the findings can, at least to some extent, be generalised to other product groups.

This paper reports on the first stage of the research project in which we have simulated current CPI practices using the Nielsen data. It shows how these data may be used to construct (or rather approximate) a price index based on official concepts, while at the same time serving as a benchmark against which alternative methods can be evaluated. Section 2 of the paper goes into the Laspeyres-type commodity group price index and the estimator chosen by Statistics Netherlands. Section 3 describes the sampling design with respect to commodities and outlets that is currently being followed. Section 4 gives a brief summary of the scanner data and how these are adapted in order to simulate the official coffee price index. Section 5 presents the empirical results.

2. Estimation of the Laspeyres commodity group price index

A certain commodity group *A*, for example coffee, consists of a finite number of commodities (also called items); $g \in A$ means that item *g* belongs to group *A*. We assume that *A* is fixed during time. The Laspeyres-type (fixed weight) price index of commodity group *A* in period *t* with respect to base period *0* is the weighted average

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

where P_g^t denotes the price index of item *g* and w_g^0 the corresponding weight, that is the base period expenditure share of *g* within group *A*. The entire set of outlets in which item *g* can be bought may be

[1] We thank Bert M. Balk and Leendert Hoven for comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

subdivided into strata, typically according to type and size-class. The outlet strata, numbered *i=1,...,I*, are assumed to be fixed during time. P_g^t can then be written as

$$(2) \qquad P_g^t = \sum_{i=1}^{I} w_{gi}^0 P_{gi}^t ,$$

where P_{gi}^t is the price index for item *g* in stratum *i* and w_{gi}^0 the corresponding weight, which reflects relative base period expenditures.

The estimation of the commodity group price index involves a two-stage sampling procedure. In the first stage a fixed number of items is selected and in the second stage for each selected item a sample of outlets is taken. The actual sampling design is described in section 3. Let $\hat{A}$ denote the sample of commodities drawn from *A*. The *first stage* yields

$$(3) \qquad \hat{P}^t = \sum_{g \in \hat{A}} \hat{w}_g^0 P_g^t$$

as the estimator of P^t, with

$$(4) \qquad \hat{w}_g^0 = \frac{z_g^0}{\sum_{g' \in \hat{A}} z_{g'}^0} \quad (g \in \hat{A}),$$

where z_g^0 denotes base period expenditures on item *g*. Expression (4) describes the base period expenditure share of the sampled item *g* with respect to the entire sample. Proxies for the item weights $\hat{w}_g^0$ can be derived from retail trade turnover statistics or from market information.

In the *second stage* the item price indexes P_g^t are estimated for $g \in \hat{A}$. The stratum weights w_{gi}^0 in expression (2) are again estimated from turnover statistics or from other sources. Let $\hat{B}_{gi}$ be a sample of outlets from outlet stratum *i* with sample size n_{gi}. The sample is fixed during time, so that we have a panel of outlets that remains unchanged. Since outlet-specific quantity and/or expenditure data can hardly be observed, P_{gi}^t has to be estimated from price observations only. Statistics Netherlands has chosen the ratio of average prices

$$(5) \qquad \underline{P}_{gi}^t = \frac{\sum_{b \in \hat{B}_{gi}} p_{gb}^t / n_{gi}}{\sum_{b \in \hat{B}_{gi}} p_{gb}^0 / n_{gi}} = \frac{\sum_{b \in \hat{B}_{gi}} p_{gb}^t}{\sum_{b \in \hat{B}_{gi}} p_{gb}^0} ,$$

where p_{gb}^s is the price of item *g* bought at outlet *b* in period *s* (s=0,*t*). Because $\underline{P}_{gi}^t$ constitutes the lowest level of aggregation in the estimation procedure it is called a *micro index* or sometimes an *elementary aggregate index.*

It frequently happens that the price in a certain outlet is temporarily not observable. Non-response of this kind (missing prices) could easily lead to adverse effects on the stratum price index, in particular if the stratum is not homogeneous. Therefore, an imputation method for handling missing prices seems useful. When p_{gb}^t is missing for a certain $b \in \hat{B}_{gi}$, the imputed price is calculated as p_{gb}^{t-1} times the ratio of

average prices in period *t* and *t*-1 for all (other) outlets in $\hat{B}_{gi}$ whose prices are observed in period *t*. Notice that the average price in period *t*-1 may partly be based on imputed prices, and that even p_{gb}^{t-1} may itself be an imputed price.

In real life the set of outlets is not fixed during time; some stores close down or stop to sell item g and new ones are established. Whenever an outlet in the panel disappears (or refuses to co-operate any longer), a new outlet is sampled. Something similar goes for the universe of items. When a sampled item ceases to be sold, another one will be substituted. Because this does not seem to be the case in the Nielsen database, we will not address the question of how substitution of commodities and the necessary adjustments for quality change are dealt with in practice.

3. The sampling design

In theory commodities may be sampled in various ways. A great advantage of probability sampling techniques is that standard errors can be calculated. Random sampling of course requires knowledge of *A*, the sampling frame. However, with a few minor exceptions, such registers of consumer products have not (yet) been compiled in the Netherlands. For that reason probability sampling was never seriously considered.

In practice the item selection is often a two-stage procedure. In the first stage a number of commodity subgroups are chosen using a cut-off method. Only the subgroups with the largest market shares are selected. In the second stage one or more specific items are chosen from each subgroup by means of judicious (purposive) sampling. The selection of these so-called representative items is based on the judgement of experts, who (think they) know the market in question. In order to measure pure price change these items should be described in great detail, to ensure that exactly the same items are observed in all outlets. But in some cases the quality description is widened for practical reasons. This is done whenever a more detailed description would lead to an unacceptable high nonresponse (missing prices). The selection of the specific item is then left to the price collectors in the field.

In the case of coffee we distinguish three subgroups: (roasted) coffee beans, grinded coffee and instant coffee. The last two subgroups were selected by cut-off sampling; the price of coffee beans is not observed at all. For grinded coffee the two most frequently bought brands (described in detail) were selected centrally at the office, while the price collectors had to make a choice out of ten outlet-specific 'house brands'. For instant coffee the field staff had to choose between the two most common brands. Note that there are four items selected: two specific brands of grinded coffee plus two composite items (house brand and instant coffee). Although this has some obvious disadvantages, a composite item is treated as if it were an identical item in all outlets.

Unlike the sample of commodities, the sample of outlets is to some extent based on probability techniques. Somewhere in the past a sample of outlets was drawn. The exact sampling design unfortunately cannot be traced. Starting with the most recent base year (1990), the sample is poststratified according to type of outlet (which is in fact an indicator of the type of goods sold) and size class. Whenever an outlet is removed from the sample it is replaced by an outlet in the same stratum, which is sampled at random from the so-called General Business Register. Unfortunately, this sampling frame has extensive overcoverage, so that the newly sampled outlet may turn out not to sell the sampled commodity in question, especially in case of a detailed quality description. In that case a new outlet is sampled.

4. Nielsen scanner data

4a. An overview

The data set from A.C. Nielsen (Nederland) B.V. covers the entire commodity group coffee: coffee beans, grinded coffee and instant coffee. It contains weekly sales over a period of 128 weeks, beginning with week 1 of 1994 and ending in week 24 of 1996, from 20 supermarkets in a Dutch urban area unknown to us. Variables included are the number of packages and value of coffee items sold, together with a number of product characteristics (brand name and subname, net weight) and a stratum indicator for large/small supermarkets. Prices are not included. The data relate to scanner (bar code) information. Scannable items are identified by European Article Number (EAN). Below we briefly summarize the most important features of the Nielsen data.

The data set contains 320 EANs. A large number of EANs have negligible expenditures. There are 55 different brands, some of which are subdivided and have subnames. Looking at brand names, the distribution of coffee sales is very skewed; it has an extremely long right tail. Taken over the entire period January 1994-June 1996, the most important brand alone accounts for 54% of the total expenditures on coffee while the 'top 15' of brands account for 97%. Average coffee sales per outlet amounted to 1.1 mln dfl or 8,433 dfl per week. Average turnover of the 13 larger supermarkets (in stratum 1) was twice as high as those of the 7 smaller supermarkets (stratum 2). Table 1 contains the unit value over all outlets - the total value of coffee sold per kilogramme, irrespective of brandname, type etc. - and unit value indexes over all outlets as well as per stratum. This gives a first impression of the change in coffee 'prices' during the period under study. There was a remarkable increase in 1994.

Manufacturers assign one and only one EAN to every variety, size and type of packaging of an item. Some EANs have very low expenditures because the system of classification is too fine; what is really one item has been classified as a multitude of items. In a similar study, Reinsdorf (1995) also found that "items that are, for all practical purposes, the same may occasionally have different UPC's" (the USA's universal product code). We will treat EANs that have the same product characteristics (see above) as identical items. This not only reflects standard practice best but is also needed in order to avoid large-scale imputations. Moreover, EANs that differ only in net weight are combined into a composite item.

4b. Adapting scanner data

A few questions need to be answered before we can simulate current official practices in the compilation of the price index of coffee using the Nielsen scanner data. Since prices are not available, the first and most important question is how to determine them. For the CPI, prices are measured once each month. To be specific: the price observation takes place on Thursday of the week in which the 15th falls, where a week is supposed to start with Sunday. We approximate this isolated price quotation by the unit value in the observation week:

$$(6) \qquad p_{gb}^{t} \approx \frac{v_{gb}^{t^*}}{x_{gb}^{t^*}}$$

where $v_{gb}^{t^*}$ and $x_{gb}^{t^*}$ denote the total value and total net weight, respectively, of item *g* sold in outlet *b* in the observation week * of month *t*. If the sampled item *g* is temporarily not sold in outlet *b* in that week, we treat this event as a missing price observation and impute a price according to the method mentioned in section 2.2. Imputation was only needed for instant coffee, after removing one outlet in the sample where no sales were recorded during many observation weeks.

The next question is how to compute elementary aggregate indexes and their corresponding weights. The official price index of coffee is estimated with elementary aggregate indexes of three size classes of supermarkets. Our data set distinguishes large and small supermarkets only. For these two strata ratios of average prices will be calculated. Following standard practice the base year price of an item is an unweighted twelve month average. We take 1994 as the base period and calculate elementary aggregate indexes from January 1994 to June 1996 (thus, including the base year months). The stratum weight for a certain item will be calculated as its relative base year expenditures in the Nielsen sample; that is, we will not be using additional information. The same holds for the aggregation of the item indexes in constructing the commodity group index.

The last question is whether we should use the entire Nielsen sample of outlets selling the selected commodities or draw a subsample thereof. We have chosen the former option, the overall sample size being small as it is already. The entire sample will be regarded as the total population of outlets in a certain area. As a consequence, outlet sampling errors do not come into play: only commodity sampling affects the accuracy of the estimated coffee price index, apart from any bias due to an inadequate micro index formula.

5. Empirical results

Table 2 compares the coffee price index (1994=100) calculated with the Nielsen data to the index that uses official data. The long run trend in both series is very similar, but the Nielsen data show a little bit more variability than the official series. Some months display rather large differences. In June 1996, for example, the official figures record a price decrease of 1.7% (with respect to 1994) while the Nielsen data show a decline of 3.8%. As should be clear from the previous sections there can be various reasons why the estimates differ. One obvious reason is the small size of the Nielsen sample; it contains only 20 outlets whereas the official sample is about four times as large. The second and third column of Table 2 present price index numbers for the larger (stratum 1) and smaller (stratum 2) supermarkets, respectively. There are no systematic differences between the coffee price changes in both strata.

The officially published index originally has 1990 instead of 1994 as base year. In Table 2 the official item price indexes are implicitly weighted by their ‘price updated’ expenditure shares of 1990. We have recalculated the official coffee price index numbers by using the 1994 Nielsen item expenditure shares (the last column). The outcomes change only slightly.

Diagram 1 shows the price changes of all four selected items using Nielsen data. The pattern for instant coffee differs somewhat from the grinded coffees (market leader, runner-up and house brands). As might be expected, price changes of instant coffee are smoothed and lag behind. We did not find evidence of significant differences in average prices or in price changes for any of the four items between large and small supermarkets.

Reference

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

Table 1. Unit values in the Nielsen coffee data base

Month	Unit value (dfl per kg)	Unit value index (1994=100)	Unit value index (1994=100) stratum 1	Unit value index (1994=100) stratum 2
9401	11.85	82.6	82.9	81.8
9402	12.38	86.3	86.3	86.5
9403	12.27	85.5	85.5	85.7
9404	12.35	86.1	86.2	86.0
9405	12.38	86.3	86.9	83.9
9406	12.84	89.5	89.3	90.4
9407	13.77	96.0	95.9	96.6
9408	15.29	106.6	106.2	108.3
9409	16.95	118.1	118.6	116.7
9410	17.94	125.1	125.6	123.3
9411	17.96	125.2	126.0	122.6
9412	17.90	124.8	125.0	124.1
9501	16.35	114.0	113.8	114.8
9502	16.49	115.0	114.7	115.9
9503	16.69	116.3	116.7	115.0
9504	16.34	113.9	115.1	109.8
9505	16.93	118.1	118.4	116.8
9506	16.52	115.2	115.5	113.9
9507	16.97	118.3	119.0	115.8
9508	16.17	112.8	112.5	113.7
9509	16.22	113.1	114.3	109.1
9510	15.39	107.3	107.1	108.0
9511	14.68	102.3	102.4	101.9
9512	15.27	106.5	107.0	104.5
9601	14.12	98.4	97.9	100.5
9602	14.42	100.6	100.8	99.5
9603	14.47	100.9	101.2	99.7
9604	14.24	99.3	99.5	98.5
9605	14.48	100.9	101.2	97.5
9606	13.65	95.1	94.5	98.0

Table 2: Price index numbers for coffee (1994=100)

Month	Nielsen data Total	Nielsen data stratum 1	Nielsen data stratum 2	Official CPI	Official (Nielsen weights)
9401	78.5	78.5	78.4	78.8	78.8
9402	82.5	82.4	82.9	83.0	83.0
9403	82.1	82.1	82.0	83.6	83.6
9404	82.8	82.8	83.0	83.6	83.6
9405	82.7	83.0	81.3	83.4	83.3
9406	86.5	86.3	87.3	88.3	88.1
9407	95.6	95.0	97.9	97.8	97.9
9408	116.3	116.4	115.7	113.9	114.2
9409	116.4	116.7	115.0	115.2	115.2
9410	124.2	124.1	124.4	123.1	123.1
9411	125.5	125.8	124.5	124.4	124.5
9412	124.4	124.3	124.8	124.9	124.8
9501	115.5	115.3	116.2	116.7	116.6
9502	114.3	113.8	116.1	116.5	116.2
9503	116.4	116.5	115.6	116.4	116.2
9504	114.8	115.6	112.0	115.6	115.3
9505	116.7	117.0	115.4	116.2	116.0
9506	114.5	114.8	113.5	115.6	115.5
9507	116.1	116.1	116.1	116.3	116.1
9508	112.7	113.3	110.5	113.6	113.4
9509	111.7	111.9	111.0	111.8	111.6
9510	110.7	111.5	108.1	107.3	107.1
9511	103.6	103.9	102.8	104.8	104.6
9512	100.8	101.2	99.3	104.4	104.3
9601	95.8	95.5	97.1	98.8	98.7
9602	96.5	96.7	96.2	98.4	98.3
9603	97.2	97.4	96.4	98.3	98.2
9604	97.0	97.2	96.4	98.1	98.0
9605	97.2	97.5	96.0	98.1	98.0
9606	96.2	96.2	96.4	98.3	98.1