

**SUPPORTING PAPER No 1 (12)*
19 October 2001**

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**INTERNATIONAL LABOUR
ORGANISATION (ILO)**

**CONFERENCE OF EUROPEAN
STATISTICIANS**

**Joint ECE/ILO Meeting
on Consumer Price Indices
(Geneva, 1-2 November 2001)**

CLASSIFICATION STRATEGIES FOR EXCESSIVELY DETAILED DATA

Supporting paper submitted by Statistics Finland**

* Paper posted on Internet as submitted by the country.

** Prepared by Mr. Timo Koskimäki, Statistics Finland.

1. Summary¹

During the past few years, in combination with increasing computerisation of the retail outlets, more and more “electronic” data on retail prices – and, indeed, also data on quantities sold – have become available. This type of data is referred to as EPOS-data (“**E**lectronic **P**oint of **S**ale”) or scanner data (referring to the scanning equipment used in outlets).

Data derived directly from the cash-registers is typically extremely detailed. For example, the scanner data set used in this study contains some 1 200 distinct products. The number of distinct products collected for the same product domain in the Finnish CPI is around 50. Scanner data also includes information of the volumes sold and a possibility to use rather detailed geographical distinctions. As compared to traditional price data used in CPI's, the scanner data can be characterised as excessively detailed.

At the first sight, the availability of scanner data may seem to solve some or all of the major problems that price index compilers have been facing. The availability of simultaneous price- and volume information provides the possibility to use superlative index formulas and hence can be seen as a solution to the substitution bias caused by the use of Laspeyres-type index formula. The exhaustive coverage in the product dimension may seem to solve of the problems of coverage on the elementary aggregate level.

The attempts to use scanner data in CPI compilation, seem, however, to bring up a set of new problems. When using scanner data, the volume of price observations to be dealt with in the index calculation increases by a factor of 20 as compared with the traditional CPI practice. The question posed in this study is how, in practice and in theory, should such a vast amount of extremely detailed data be dealt with in the CPI compilation

The paper presents two possible strategies that may be applied when using scanner-type data in CPI compilation. We denote the two strategies as the **matching approach** and as the **classification approach**. Both of the strategies will be illuminated by empirical analysis of a geographically representative scanner data set covering some 350 outlets and 8 product categories. Advantages and disadvantages of the approaches will be discussed.

The paper is divided into three sections: In chapter 2 the two strategies are briefly presented and their advantages and disadvantages are discussed in a general level. Chapter 3 describes the design and empirical results of the study. Chapter 4 summarises the conclusions.

On the basis of the analysis carried out, the use of traditional matching approach does not seem to be the optimal solution when using scanner data in index compilation. The more promising approach seems to be to calculate unit values using a less detailed classification scheme. The use of scanner data as such does not resolve the classical index problems of substitution and new and disappearing products. Construction and use of proper classifications may give an additional tool to solve these issues.

¹ This research paper has been produced as a part of research projet being carried out at Statistics Finland. The other members of the research team, professor Yrjö Vartia, Ms. Mari Ylä-Jarkko, Mr. Juha Mylläri and Mr. Anssi Vuorio have actively participated in the formulation of the research plan, creation of the classifications and data sets used for this study. I wish to express my sincere thanks to all of them.

2. *The two approaches*

2.1. *The Matching Approach - an extremely detailed classification*

The most obvious strategy to deal with the excessively detailed scanner data is to treat the data in a similar manner as the data is treated in current CPI compilation, i.e. one-to-one matching of the observations at the most detailed level, within outlets, across time. Indeed, most of the current studies utilising scanner data seem to take one-to-one matching as a self-evident starting point without reflecting on other possibilities.

The number of products to be included in the "scanner-CPI" would be essentially larger than in the traditional CPI. However, the products would still be matched in a very detailed level and the non-matches would be omitted from the compilation. In practice, this means matching of the products on the basis of the most detailed product codes available (in Europe the so called EAN-codes) on the level of single outlets.

From a more general perspective the matched models approach can be characterised as an extremely detailed classification system that has been formed by combining product- and outlet typologies. On the most disaggregated level of index compilation the product classification to be used is often very narrow. Products at the elementary aggregate level are typically classified by package size, manufacturer of the product, flavour or colour, specific brand name et c.

2.2. *Releasing the details - the classification approach*

Once we have realised that the traditional one-to-one matching is a form of classification we may also consider alternative ways of classifying the data. In essence, we may choose - at our will - the level the detail to be used in the formation of elementary aggregates for the index compilation. The choice of the level of classification should, of course, be justified by some form of rational reasoning.

Generally, the term used for this type of approach is "unit value index". The use of unit value indices as elementary aggregates may be perceived as somewhat peculiar (Balk 1998). Diewert (1995, ref Balk 1998), however, seem to give some support for the general idea. Unit values are also to the standard procedure proposed by some less recent textbooks (see e.g. Allen, 1975). As will be demonstrated, there always exists a multitude of alternatives to form "unit value indices". To define the borderline between "unit value index" and "price index" is always to certain extent an arbitrary choice.

2.3 *Pros and cons of the approaches*

"Price indexes, almost universally, have followed one fundamental methodological principle: The quality of the product sets to be compared is held constant by following the matched pairs strategy. The price index compiling agency chooses a sample of retail outlets or sellers and a sample of products. It collects an initial period, or base period, price for each of the products selected. It then collects at some later date the price for exactly the same product, from the same seller, that was selected in the initial period. The price index is computed by matching, observation by observation, the price at the later period with the initial price (Triplet, 2000)"

This kind of classification strategy, admittedly, creates very homogenous sets of observations. Triplet, again, summarises the advantages of one-to-one matching quite neatly:

"The great advantages of this matching methodology are sometimes not explicitly stated, and other times not fully appreciated. The "matched model" methodology

holds constant many price determining factors that are usually not directly observable. Examples are characteristics of the retailer, such as customer service, reputation of the manufacturer et.c. Matching the price quotes model by model (and outlet by outlet) is not just a methodology for holding quality change constant in the items selected for pricing. It is also a methodology for holding constant non-observable aspects of the transaction that might bias the measure of price change.

The disadvantages of the matching approach are quite obvious although seldom explicitly stated. The most general formulation of the drawbacks is that the matched models approach always causes quite a considerable loss of information. This loss of information stems from the simple fact that new and disappearing varieties can not be taken into account in the index calculation.

Quite a number of contemporary problems in the field of price indices are derivatives of this general feature of the matched models approach. The need for specific "quality adjustment" procedures in cases where a variety that has been followed in the CPI disappears from the markets and has to be replaced by another variety is one class of problems. The other class of problems stems from appearance of "new products" . As there is no natural pair for comparison, the new product is generally omitted from the calculation.

A third class of problems stem from the practicalities of the matched pairs - approach. It is often the case that price collectors are advised to follow the initially selected variety until it entirely disappears from the markets. This, in turn, leads rapidly to a situation where the CPI sample is no longer representative, i.e. gives a distorted picture of the markets (see Koskimäki and Vartia 2001, Silver 2001).

The main drawback of the classification strategy is often denoted as "the unit value bias". In essence this term refers to a situation where the elementary cells - or elementary classes created using pre-specified classification - are not homogenous enough. The idea of comparing "like with the like" may thus be violated and differences in quality might appear as differences in price.

3. Indices based on unit values and matched pairs

3.1. The data

The data used in this study was provided by ACNielsen Ltd. The "elementary transaction" in the data is sales of a specific product during one week in a single outlet. "Specific product" is identified as a distinct EAN code. Data relating to one elementary transaction are as follows:

- number of packages sold
- package size
- unit of the package (Kilogram, Litre, MI et.c.)
- total sales during the week
- Product identification code ("EAN")
- brief description of the product
- manufacturer
- ACN's own product class
- "Brand"
- outlet code
- region

For the purpose of index calculations a simple package size adjustment was performed. The price used in the examples is the "weekly mean price" of one unit sold

The products covered in this study are typical "fast moving consumer goods" where no rapid quality changes is expected to occur. The data covers the following product groups: Butter, Margarine and other vegetable fats, Vegetable oils, Soft drinks, Fruit juices and Detergents.

The weights for various index calculations were derived directly from the sales data.

The data covers all weeks in years 1998, 1999 and 2000. For the purpose of this study, two weekly data sets - last week of September 1998 and last week of September 2000. were extracted from the material

In addition, products were classified in the spirit of the standard classification of individual consumption by purpose (COICOP). The version of COICOP classification used is presented in annex 1. Other classifications used in the index calculations are presented in table 1 below:

Table 1: Classifications used in the study

(Number of distinct product codes (EAN) refers to 1998 data)

Regional dimension	Product dimension		
	Number of levels	Number of levels	
Whole country	1	Coicop 5-digit	6
Province	4	Coicop 7-digit	26
ACN region	15	ACNielsen brand	266
Outlet	338	ACNielsen EAN	1 028

3.2 The general design of the study

The regional and product dimensions were then combined to form a typology - or a matrix - to be used in the compilation of indices. For each cell in the matrix an index with differing weighting structure were calculated. The typology is shown in table 2.

Table 2: Typology of different classification strategies

Regional dimension	Product dimension			
	Coicop 5-digit	Coicop 7-digit	ACNielsen Brand	ACNielsen EAN
Whole country	1.	2.	3.	4.
Province	5.	6.	7.	8.
ACN region	9.	10.	11.	12.
Outlet	13.	14.	15.	16.

The idea of the approach is as follows: The intersection of row and column attributes describes the elementary aggregate level of the index to be calculated. Hence, in the index to be calculated for cell number one we only fix weights for the six COICOP 5-digit groups. In cell 5 we fix the weights also according to the the province (4 classes in regional dimension and six classes product dimension, i.e. 24 classes.).

In the lower right corner (cell 16) all available data has been fixed to form an entirely fixed weighting structure (338 outlets and 1028 distinct products).

Theoretically, if all outlets would sell all products - and stayed in the markets over the entire period of the study - this would mean 347 464 fixed cells.

Turning back to the two basic approaches, the rightmost cells in the matrix (4, 8, 12, 16) are variations of the matching approach. The rest of the cells are examples of different classification strategies.

The elementary aggregates below the given stratification levels have been calculated as weighted arithmetic means. The aim is to produce a meaningful mean price taking into account the fact that the market shares of the products under study vary from 0,5 per mill to some 10 ten per cent within a given product category.

In essence, elementary aggregates used here are analogous to the price concept inherent in the scanner data on the outlet level. The weights used in the construction of elementary aggregates are allowed to vary between the two periods concerned. This is consistent with the general idea that the elementary aggregates under study are considered to be homogenous, at least from a consumer's point of view.

3.3. *The results*

The results of the exercise are shown in tables 3 to 6. To summarise:

- Increasing the level of detail (of the weights) in the product dimension tends to lower the observed price increase. Using 6-class structure in the product dimension yields a price increase of 7,9 percent whereas fixing the weights on EAN-level indicates only 2,3 per cent price increase.
- Increasing the level of detail in regional dimension, especially if the calculus is fixed at the outlet level, tends to increase the observed price increase. The phenomenon, however, disappears if product dimension is tightly fixed.
- Classification which keeps the producer fixed (ACNielsen Brand - classification, 266 classes) tends to give higher price increase when compared to a reasonable consumer-oriented classification (COICOP 7-digit, 26 classes)
- Tight classification - both in product- and outlet- dimensions - tends to increase upper level substitution bias (measured as difference between Laspeyres and Fisher indices).
- The loss of information increases rapidly when classifications get more detailed. Keeping the outlet sample fixed between the two periods decreases the coverage by 10 per cent. The most detailed product classification enables only 80 per cent of the transactions to be included in the comparison. The joint effect - keeping both the products and outlets extremely fixed - excludes almost 40 per cent of the data from the index calculation.

3. Laspeyres price indices September 1998 - September 2000 (sales during one week)

Regional dimension	Product dimension			
	Coicop 5-digit	Coicop 7-digit	ACNielsen Brand	ACNielsen EAN
Whole country	107,9	103,1	104,6	102,3
Province	107,8	103,1	104,8	102,3
ACN region	107,8	103,1	105,1	102,5
Outlet	108,6	104,0	106,0	102,8

4. Fisher price indices September 1998 - September 2000 (sales during one week)

Regional dimension	Product dimension			
	Coicop 5-digit	Coicop 7-digit	ACNielsen Brand	ACNielsen EAN
Whole country	108,0	103,2	104,8	101,5
Province	107,9	103,1	104,8	101,4
ACN region	107,9	103,0	104,7	101,4
Outlet	108,9	103,4	104,9	101,1

5. Difference Laspeyres - Fisher

Regional dimension	Product dimension			
	Coicop 5-digit	Coicop 7-digit	ACNielsen Brand	ACNielsen EAN
Whole country	-0,1	-0,1	-0,2	0,8
Province	-0,1	0,0	0,0	0,9
ACN region	-0,1	0,1	0,4	1,1
Outlet	-0,3	0,6	1,1	1,7

6. Turnover covered by each calculation (Whole country, Coicop 5-digit = 100)

Regional dimension	Product dimension			
	Coicop 5-digit	Coicop 7-digit	ACNielsen Brand	ACNielsen EAN
Whole country	100,0	100,0	98,2	80,1
Province	100,0	100,0	97,5	77,4
ACN region	100,0	100,0	96,9	75,5
Outlet	90,4	90,4	84,6	61,7

Aggregates below classification level calculated as a ratio of (weighted) arithmetic mean prices (period-specific weights)

4. Conclusions

The above described results should not come as a great surprise to experienced price statisticians or economists. It is quite easy to figure out behavioural explanations for the above results: if we keep "the producer" (ACN brand-classification above) fixed, we seem to exclude from the index price effects caused by new producers or brands entering the markets. Also, if we keep our sample of outlets constant, we lose the price decrease brought up by new outlets competing on the markets. Similarly, if we keep our product set extremely tightly defined, we lose a considerable part of the transactions in the markets and hence may create a biased index.

The above mentioned substitution effects have traditionally been discussed in the context of different index formulas (see, for example, de Haan 2001). The exercise exposed here should show that the issue can - and should - be treated also as a problem of product classification.

The above results have, of course, been derived from an extremely exhaustive data. Data on volumes sold have been used extensively on the elementary aggregate level. For the third, the period for comparison is rather long, 24 months. None of these aspects is realised as such in the more traditional CPI compilation.

It is, however, possible to draw some conclusions also with regard to current index compilation. Notions on the effects of keeping outlets or brands fixed can be directly utilised also in the context of traditional CPI's. The same holds also for too narrow product definitions or practices where new varieties are incorporated into the index following some kind of "link to show no change" procedure. Loss of price information can easily be generated also in the traditional CPI environment as well.

References

- Allen, R.G.D: Index Numbers in Theory and Practice. The Macmillan Press Ltd. London and Basingstoke, 1975.
- Balk, B. M: On the Use of Unit Value Indices as Consumer Price Sub-indices. Paper presented at the 4th meeting of the Working Group on Price Indices, Washington D.C, April 22 - 24, 1998.
- Diewert, W. E: Axiomatic and Economic Approaches to Elementary price indexes. Discussion paper no 95-01, Department of Economics, University of British Columbia, Vancouver 1995. Referenced in Balk, 1998.
- De Haan, Jan: Generalized Fisher Price Indexes and the Use of Scanner Data in the CPI. Paper presented at the 6th meeting of the Working Group on Price Indices, Canberra, April 2001.
- Koskimäki, T and Vartia, Y: Beyond matched pairs and Griliches-type hedonic methods for controlling quality changes in CPI sub-indices. Mathematical considerations and empirical examples on the use of linear and non-linear hedonic models with changing quality parameters. Paper presented at the 6th meeting of the Working Group on Price Indices, Canberra, April 2001.
- Silver, Mick: Quality adjustment and price indices. Draft chapter for the forthcoming ILO manual on CPI:s ILO, 2001.
- Tripplert, Jack E: Handbook on quality adjustment of price indexes for information and communication technology products. Draft report prepared for Industry Committee, OECD Directorate for Science, Technology and Industry. October 1999.

Annex 1: COICOP classification used in the study

01.1.5.1 Butter

01.1.5.1.01 Dairy butter

01.1.5.1.02 Other butter

01.1.5.2 Margarine and other vegetable fats

01.1.5.2.01 Butter and vegetable oil mixture

01.1.5.2.02 Cooking margarine

01.1.5.2.03 Soft margarine

01.1.5.2.04 low fat margarine

01.1.5.2.05 other veg. fats

01.1.5.4 Vegetable oils

01.1.5.4.01 Rapeseed oil

01.1.5.4.02 Sunflower oil

01.1.5.4.03 Olive oil

01.1.5.4.04 Other Oil

01.2.2.2 Soft drinks

01.2.2.2.01 Veg. extract drinks (Coke)

01.2.2.2.02 Soda orange

01.2.2.2.03 Energy drinks

01.2.2.2.04 Soda soft drinks, other than orange

01.2.2.2.05 Other soft drinks

01.2.2.3 Fruit juices

01.2.2.3.01 Mixed fruit cordial

01.2.2.3.02 Orange juice, 100 per cent fruit

01.2.2.3.03 Other cordials

01.2.2.3.04 Other juice, 100 per cent fruit

01.2.2.3.05 Juice, less than 100 per cent fruit

01.2.2.3.06 Other juices and cordials

05.6.1 Non-durable household goods

05.6.1.1 Detergents

05.6.1.1.01 Dishwasher detergent

05.6.1.1.02 Synthetic detergent

05.6.1.1.03 Dish washing liquid

05.6.1.1.04 General purpose cleanser

05.6.1.1.05 Other detergents