Distr.
GENERAL

CES/AC.49/2001/5
4 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)

USING SCANNER DATA TO COMPILE PRICE INDICES: PRACTICAL PROBLEMS

Invited Paper submitted by Statistics Netherlands*

Keywords: Consumer price index, Electronic data interchange, scanner data,

* Prepared by Ms. Cecile Schut, Division of Macro-Economic Statistics and Dissemination, Department of prices, short-term indicators and programming, Statistics Netherlands. The author thanks Tjalling Gelsema and Gwen Krul for their useful comments.
The views expressed in this paper do not necessarily reflect the policies of Statistics Netherlands. Email csht@cbs.nl

I. INTRODUCTION

1. For the composing of the Consumer Price Index (CPI), Statistics Netherlands uses different kinds of surveys at this moment. Monthly, more than 200 interviewers visit shops all over the country collecting prices. Furthermore, written surveys are sent out, prices are collected from price lists and from the Internet In total, Statistics Netherlands monthly collects about 90,000 prices.

2. In the past decade, a lot of Dutch retailers introduced the use of scanners in their shops. Articles are coded according to the European Article Number (EAN) system. For retailers, this development has led to more efficiency, because it offers a fully automized administration of both turnover and supplies. Statistics Netherlands recognized the potential of this development for improving both the collection of data and the overall quality of the CPI. Last year, Statistics Netherlands reached an agreement on the delivery of scanner data on a regular basis with two large supermarket chains in The Netherlands. This paper describes the nature of the data and the way in which these data can be used in de CPI. A lot has been written already about the use of scanner data in the calculation of price index numbers. Most of these papers contain research results on the use of various formulas, sampling designs, or methods for quality adjustment (e.g. Dalén, 1997, Hawkes, 1997, Reinsdorf, 1995, Silver, 1995, Silver and Heravi, 1999). The aim of this paper is not to add any new theoretical insights to all these papers, but it will sketch the difficulties of using scanner data of supermarkets on a large scale at the calculation of the CPI. The difficulties we met were partly of practical nature and had partly to do with the chained Fisher method we chose for calculating index numbers from of the data.

3. In section 2, we describe the nature of the data and the way in which the data are obtained. In section 3, we outline the method for calculating index numbers from scannerdata. In section 4, we describe some issues about implementing the method in a production system. In section 5, we give some empirical results.

II. SCANNERDATA FROM SUPERMARKETS

4. This section gives a description of the way in which the scanner data are obtained. Subsection 2.1 describes some history of the procurement process. In subsection 2.2 a brief description of the nature of the data is given.

Obtaining scannerdata from supermarkets

5. In the past years, Statistics Netherlands followed the implementation of EAN-systems in different branches. In particular, large supermarketchains introduced the use of scannerdata on a large scale. Both supermarkets and producers of commodities experience big advantages from the fully automated supply- and demand systems. On a continuous basis, the number of sold commodities can be controlled, so the provisioning can be done just in time. Moreover, scanner data provide lots of interesting marketing information. Companies like GfK and A.C.Nielsen collect these information from different companies, and use them for market research. In the past years, Statistics Netherlands did some research on small sets of scannerdata which were bought from market research firms (De Haan and Opperdoes (1997), De Haan and Boon (1998) and De Haan, Opperdoes and Schut (1999)). The studies concerned specific topics like sampling issues and the possible range of biases. In the meanwhile, we developed ideas for using scannerdata in the actual compilation of the CPI. The

advantages of using large sets of scannerdata were clear: compared to the current limited sample that is used for the CPI, all transactions and their prices would be measured, prices would be real transaction prices and the turnover shares for each commodity would be known.

6. Therefore, Statistics Netherlands asked a large supermarketchain in The Netherlands for admission to their scannerdata for compilation of the CPI. At first, the company had some reservations, mainly because of the danger of disclosure of the information to competitors. In the end, Statistics Netherlands succeeded in getting an agreement with two different supermarketchains for the supply of scannerdata on a regular basis.

7. For the continuation of the monthly CPI production, it is important that the timely delivery of the scannerdata is secured. For this reason, we formulated a written agreement with both chains in which the delivery of the data was settled precisely.

Description of the data

8. Statistics Netherlands receives from both supermarkets scannerdata of about 70 different outlets on a weekly basis. For each outlet, the total amount of scanned units in a week and the corresponding turnover is registered per EAN-code. Besides, a short product description of each EAN is provided. For one of the supermarketchains, data were delivered from July 1999 onwards, the other supermarketchain provides data from April 2000.

9. During the period from July 1999 until today, there were about 29,000 different EAN codes present in the database. Per month, we have the disposal of about 2,000,000 records. In the current CPI-sample, we have about 450 representative items that are priced in supermarkets. Monthly, we collect about 14,500 prices for these items in outlets of the two supermarketchains.

III. FROM DATA TO PRICE INDEXES

10. De Haan, Van Poppel and Spaanderman (2000), gives a comprehensive description of the way in which Statistics Netherlands would use the supermarket scanner data in practice. This paper was presented in 1999 by Jan de Haan at the Joint ECE/ILO meeting on Consumer Price Indices. In this section we describe the proposed method in general terms.

Calculation method

11. The basic principle of the proposed method is to divide the total sample of outlets into different strata: strata B_I and B_{II} that contain the data of the supermarkets I and II for which we have scanner data at our proposal and a stratum B_{III} which contains all other outlets. For each stratum, commodity group indexes are calculated separately. These index numbers are weighed together by using the market shares. Note that the market shares can be quite different per commodity group.

12. For stratum B_{III}, we can calculate commodity group index numbers in the traditional way, i.e., by calculating the ratio of unweighted arithmetic average prices in the observation month and the base period 0 for all goods that are representative items in a commodity group and weighing these indexes together to a commodity group index number.

13. To compute the commodity group index numbers for the strata B_I and B_{II}, we start with computing transaction prices (unit values) for all EAN codes, by aggregating the quantities and corresponding turnover over all outlets. In principle, the sample of outlets in the received data is fixed, but in practice it occurs that outlets close, or that there is no information from some outlets in a certain month. Therefore, we choose to base the computation of the transaction prices on monthly matched outlets. The second step is to calculate month-to-month unit value indexes for all EANs which are present in both months. There were three main reasons for using the unit value index for individual goods. First, in our view, a good bought in a certain outlet of supermarket X does not differ essentially from the same good bought in another outlet of the same supermarket X. For the supermarkets of which we receive scanner data, this assumption is a quite reasonable one. Supermarkets use nationwide advertising campains and the same standard service levels and quality programs for their outlets. They also use the same company logo all over the country. Besides, from the scannerdata we see that most goods have the same price in each outlet. It is clear that if a good that is sold in different outlets cannot be seen as identical, another index formula should be chosen. The other reasons for chosing the unit value index formula at this level are of more practical nature: the formula is easy to compute and there is no need for imputations.

14. The last step in the calculation procedure is to aggregate the month-to-month unit value indexes per EAN to commodity group price index numbers. For this step, we decided to calculate chained Fisher indexes. The use of the (chained) Fisher price index can also be justified with reference to micro-economic index number theory. The Fisher index belongs to the class of what Diewert (1976) calls superlative indexes. To account for substitution, the choice of the Fisher index (or any other superlative formula) is a good one. Furthermore, month-to-month chaining is necessary to incorporate (almost) immediate introduction of new goods in the calculation of the commodity group price index numbers.

Classification of goods

15. The goods in the Dutch CPI are grouped according to the Classification Of Individual Consumption by Purpose (COICOP). Thus, it is necessary to group the EAN codes accordingly. Classifying all 29,000 EAN by hand is a seemingly endless task and initially, we felt this was a big obstacle in implementing the use of scannerdata on a large scale. Fortunately, the Dutch Centraal Bureau Levensmiddeldnhandel (CBL) – a branche organisation of supermarkets – classifies EAN codes according to its own CBL classification system. This classification is more refined than COICOP. Statistics Netherlands reached an agreement with CBL about receiving the CBL-classification data. The classification method we use consists of two stages. In the first stage, Statistics Netherlands sends all EAN codes concerned electronically to the CBL. The CBL sends the file back, enlarged with the CBL classification numbers for each EAN. In the second stage, we linked the CBL categories to the COICOP groups.

16. In practice, CBL is not able to classify all EAN codes at once. About 20% of the EAN codes was not classified. The most important EANs (in terms of turnover share) are still coded by hand at Statistics Netherlands.

IV. Implementation of the method

17. In the summer of 2000, Statistics Netherlands started to build a computer system that could process the scannerdata into price indexes on commodity group level as described earlier. The aim was to build a system which calculated monthly indices on the basis of scannerdata that could be used in the regular CPI calculation. One of the main conditions for implementation of the use of scannerdata in the regular CPI process was that the current publication dates had to be maintained. At present, the Dutch CPI of month t is published in the first or second week of month t+1. To ensure having enough time for quality checks and preparing the dissemination, this means that the final calculation for month t has to be ready on the last day of month t at the latest.

18. The scannerdata of a certain week are sent to our bureau within 3 or 4 days after the end of the week. This means that it is impossible to process the dataset which contains the information about the last (part of the) week of the month in time. Another restriction comes from HICP regulation 2601/2000. Article 2 of this regulation states that *"Prices for goods shall be entered into the HICP for the month in which they are observed."*.

19. In practice, these restrictions imply that per calendar-month, only data of 2 or sometimes 3 weeks can be used. Because quantity data are used in the chained Fisher index, there will be some problems with using quantities that cover 2 weeks in one month and 3 weeks in the next month. So, we finally decided to use only the first two full weeks per month. This means that more than half of the data is not used in practice. Currently, we have not investigated the consequences of this restriction thoroughly, but some quick analysis learnt that the effect of using all available data leads to almost the same results[1].

V. EMPIRICAL RESULTS AND DISCUSSION

20. After the main part of the system was built, it was possible to calculate the first results based on all available data. The results contained index numbers for over 50 relevant commodity groups for the period of May 2000 until July 2001 for both supermarketchains[2]. For this period, we also recalculated all CPI commodity group index numbers, without using the prices collected by the interviewers in the two supermarkets for which scannerdata were available. Then, it was possible to weight the scannerdata results together with the results based on prices collected in other stores in the way described in section 3[3]. The weight of the supermarketscannerdata in the total CPI was about 4%. Because of the limited amount of weight, we expected the influence of introducing the scannerdata on the total CPI to be rather small. However, in practice we experienced that this was not the case. Soon, we recognized that the month-to-month chaining gave problems with seasonal goods. In section 5.1. We give some empirical results of commodity groups with a lot of seasonal goods. Nevertheless, there were more "strange" developments in the scannerdata indexes in comparison with the traditional CPI Laspeyres index. From the data, it was not directly clear what caused this big difference. We could think of the following three possible causes:

(1) The effect of using transaction prices. In particular, the discounts given to customer card holders, which are incorporated in the measured turnover figures in the scannerdata, could possibly have large effects.

(2) The effect of extending the amount of representative items in supermarkets from about 450 in the current CPI to about 20,000 different EAN codes.

(3) The effect of using a chained Fisher index formula.

21. The effect of using transaction prices was large in some cases. It turned out that discounts of up to 30% were offered and almost all customers make use of a customer card. Because transaction prices are real prices that are paid by customers, measuring transaction prices gives a better CPI estimation.

22. The second effect is also a major improvement of the quality of the index. Unfortunately, we found some problems in the use of the chained Fisher index formula. In the next section, we show different problems that arise when using a chained Fisher index. In section 5.1 we discuss the problem of seasonal goods and in section 5.2 we show some problems which have to do with enormous fluctuations in the data.

Seasonal goods

23. For seasonal goods such as vegetables, fruit, potatoes but also sweets that are only available at Easter or at Christmas' time, month-to-month chaining leads to severe biases. In the proposed chained Fisher index, seasonal goods are treated as disappearing and (newly) appearing goods, but never as reappearing goods. However, this treatment is not quite legitimate: seasonal goods are goods that just disappear temporarily and come back again after some time. Empirical results showed that using the chained Fisher index leads to large biases of commodity group indices.

24. In Figure 1, the chained Fisher index based on scannerdata from one of the two supermarkets for the commodity group fresh fruit is shown together with the CPI index for the same commodity group. For the scannerdata, there was only information available from July 1997 onwards. Therefore, we put both indices on 100 for July 1999. From this picture, it is clear that the scannerdata index is upward biased. The range and the direction of the bias which is caused by seasonal goods, of course depends on the average price movements in the beginning an at the end of the period in which seasonal goods are available.

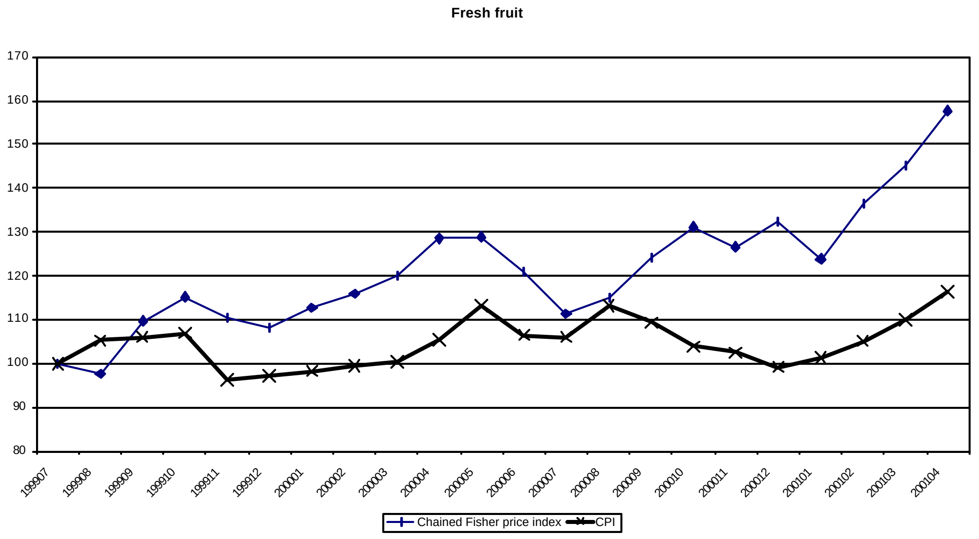


Figure 1: Chained Fisher index and the CPI for fresh fruit

25. To deal with this problem, we experimented with an imputation method. The idea behind the proposed method is that it is necessary that the price of a seasonal good in the last month it occurs, somehow has to be linked to its next occurrence, which is some months later.

26. To explain the method, let's suppose a good is last sold in month t. In the next 3 months, the good is not available, but in month t+4 it appears again in the shops. In month t+1, the good is just treated as a disappearing good in the chained Fisher index. In month t+4, when the good appears again, we used the last known price from period t as a "virtual price" in period t+3. Because the good was not available in period t+3, we supposed that there was sold only one piece of the good in period t+3. In this method, the price of the good that occurred in period t could be linked to the measured price in period t+4 by making an artificial imputation for both the price and the quantity of the good in period t+3. For fresh fruit, fresh vegetables and potatoes, we tested this method on the data. Unfortunately, the results were very similar to the normal chained Fisher. To the present day, we have not solved the problem of seasonal scannerdata.

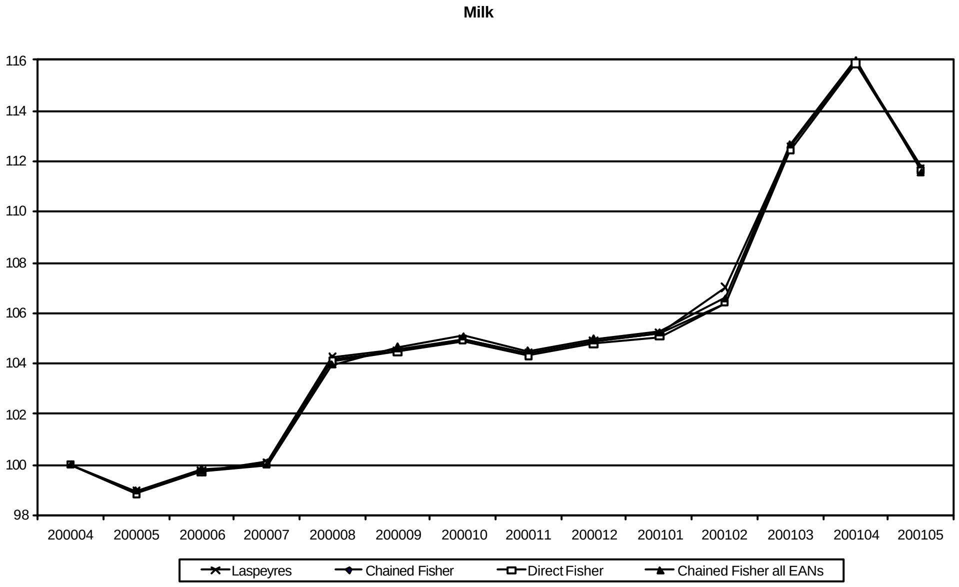
Behaviour of the chained Fisher index in special cases

27. To check the behaviour of the chained Fisher index, we decided to carry out some additional research. For certain commodity groups, we selected only those EANs which were present in all months form April 2000 until May 2001 for certain commodity groups. For those set of EANs, which contained no appearing nor disappearing EANs, we calculated a fixed base Laspeyres index, a direct Fisher index and a chained Fisher index. We expected that the direct Fisher index would be very near to the chained Fisher index. Furthermore, we expected (hoped) that the chained Fisher index based on all EANs of the corresponding commodity groups would give very similar results. In other words, we expected that the influence of new and disappearing goods would be small.

28. In figure 2, 3 and 4, results are shown for the commodity groups milk, coffee and sweets.

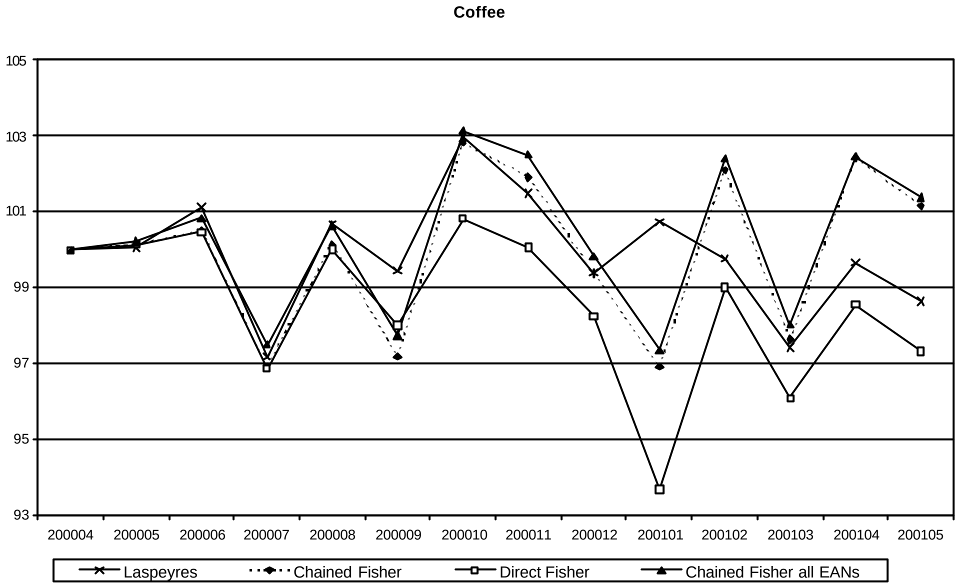
29. For milk, the different indexes are almost the same. This means that there is hardly any substitution effect (because the Laspeyres index is not very different from the direct Fisher index). Besides, when comparing the chained Fisher index - which is based on the EANs that were available during the whole considered period - with the all-EAN chained Fisher index, we see no influence of new and disappearing goods. These results depend on the nature of the commodity group; for milk, there were hardly any new or disappearing EANs and the monthly amount of sold items was quite stable. As can be seen in figures 3 and 4, for coffee and sweets this is totally different.

Figure 2: Different indices for the commodity group milk



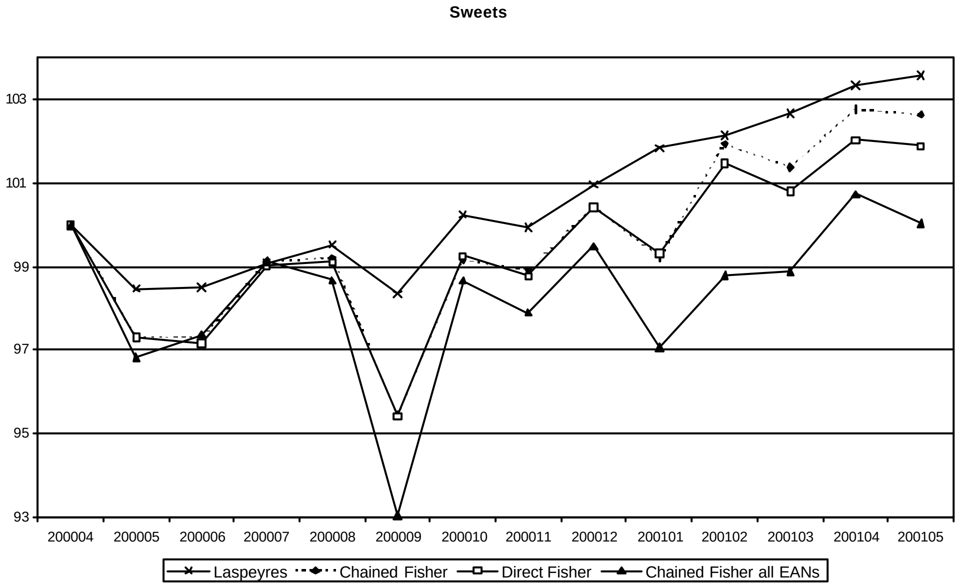
30. For coffee, we see large differences between the Laspeyres and the direct Fisher index. Another, more surprising result, is the discrepancy of the direct Fisher and the chained Fisher index, even in the case of no new or disappearing goods. The reason for these results is that there were some product offers for very popular types of coffee, which led to an increase of the turnover with 1000 or even 1500% in one month! Apparently, the chained Fisher index is not proof against this kind of Dutch panic-buying.

Figure 3: Different indices for the commodity group coffee



31. For sweets, just like for coffee, again there can be seen some substitution effects when comparing the Laspeyres with the direct Fisher indexes. Here, the direct Fisher and the chained Fisher index give very much the same results in the first months considered. However, the chained Fisher - which is calculated on all EANs - is quite different from the Fisher indexes based on the EANs that were present during the whole period considered. This effect is caused by the fact that there are quite a lot of new and disappearing EANs in this commodity group[4]. The question here is whether new EANs are really new goods or that new EANs refer to goods that already existed. In the latter case, the old and the new EAN-codes have to be linked together in such a way that the prices are compared in the index. In the next section, we describe this in more detail.

Figure 4: Different indices for the commodity group sweets



32. For Statistics Netherlands, the empirical results that are shown above, led to the decision of not using the chained Fisher index formula for scannerdata in the CPI at all. In section 6, some alternatives that we are currently investigating are described.

Linking disappearing goods to successors

33. In De Haan, Van Poppel and Spaanderman (2000), it was already mentioned that it is questionable whether or not each separate EAN code should be treated as a separate good. Different codes may refer to the same good, either physically or from a consumer's perspective. When changing EAN codes are accompanied by price changes, these price changes are missed in the proposed chained Fisher index. For these cases, adjustments should be made.

34. In the data we tried to find some cases in which different EANs referred to the same or almost the same goods. As described in section 2.2, for each EAN code, a short product description is available. Finding the same or nearly the same goods with the help of only this short product description turned out to be very difficult in a lot of cases. In practice, it will be necessary to have some specialists available that follow new product introductions carefully. From our experiences with the data, we saw that in some specific commodity groups, like sweets, detergents and washing powder, there are more new or renewed products than in other commodity groups. Furthermore, there are some "special cases". First, when the excise duties on cigarettes are changed in the Netherlands, this automatically means that EAN codes are also change. Second, there were two well-known brands (crisps and detergent) that changed their name in the Netherlands recently because producers wanted to harmonise brand names internationally. For all product varieties from these brands, the EAN codes were kept the same. This kind of general information about the use of scannerdata by producers and supermarkets can help in the process of finding EANs that need to be linked together. Until now, we do not have real experience with this linking in practice.

VI. NEW RESEARCH

35. Statistics Netherlands is willing to use the available scannerdata for supermarkets in the (hopefully near) future. We adjusted our ambitions of using an index formula that takes the substitution effect into account. At this moment, we focus on a fixed base Laspeyres index in which EAN codes are represented as much as possible. To incorporate new goods, we propose a yearly change of the weights. Research topics include: how to construct a weighing scheme, which EAN codes are selected, what to do with EAN codes that disappear and how to deal with seasonal goods? From our experiences described in this paper, it will be clear that before implementing any new ideas, we have to be sure that the method leads to reliable results. The advantage of using a simple Laspeyres index is the relatively simple way in which results can be interpreted and calculated. Nevertheless, one can ask why not use a direct Fisher index instead?

NOTES

[1] During the research on which the choice of the method was based, only data of one week per month were used (De Haan, Van Poppel and Spaanderman, 2000).

[2] For one of the two supermarkets involved in this project, there were also data available from July 1999 until May 2000. The index numbers for this supermarket were also used for research, but the total CPI was not recalculated for this period.

[3] To link the partial commodity group scannerdata indexes into the CPI in May 2000, we set them in April 2000 to the level of the corresponding CPI commodity group indexes of April 2000.

[4] De Haan, Van Poppel and Spaanderman (2000) showed that “the turnover share of new EAN codes is typically low”. In practice, it turned out that the amount of new EAN codes differs from commodity group to commodity group.

REFERENCES

Dalén, J. (1997): “Experiments with Swedish scanner data”. In: B.M. Balk (ed.), Proceedings of the Third international conference on price indices. Voorburg: Statistics Netherlands.

Haan, J. de and E. Opperdoes (1997): “Estimation of the Coffee Price Index using scanner data: the choice of the micro index”. In: B.M. Balk (ed.), *Proceedings of the third international conference on price indices*. Voorburg: Statistics Netherlands.

Haan, J. de and M. Boon (1998): “Elementary index bias in the consumer price index”, *Research paper* (Statistics Nethrelands, Voorburg).

Haan, J. de, E. Opperdoes and C. Schut (1999): ”Item selection in the consumer price index: cut-off versus probability sampling”, *Survey Methodology*, 25, 31-41.

Haan, J.de, P. van Poppel and C. Spaanderman (2000): “Using scanner data in the compilation of the CPI: A Dutch pilot study”, *Research paper* (Statistics Netherlands, Voorburg).

Hawkes, W.J. (1997): “Reconciliation of consumer price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanning data”. In B.M. Balk (ed.), *Proceedings of the Third international conference on price indices*. Voorburg: 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.

Silver, M. (1995): “elementary aggregates, micro-indices and scanner data: some issues in the compilation of consumer price indices”, *Review of income and wealth*, 41, 427-438.

Silver, M. and S. Heravi (1999): “The measurement of quality-adjusted price changes”, *paper presented at the fifth international conference on price indices*, Reykjavik, August 25-27.
