Statistics Netherlands

Division of Macro-economic Statistics and Dissemination
Development and support department

P.O.Box 4000
2270 JM Voorburg
The Netherlands

The use of scanner data in the CPI: a curse in disguise?

Peter Hein van Mulligen and May Hua Oei

Remarks:
The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Project number: *MOO-*
BPA number: *-MOO*
Date: *13 May 2004*

THE USE OF SCANNER DATA IN THE CPI: A CURSE IN DISGUISE?

Abstract:

Despite a well-developed theory of price indices, which goes back to the 19th century, statistical agencies have always been plagued by practical difficulties which usually forced them to use index numbers that were only second best. The rise in the availability of so-called scanner data has made it possible, at least for certain types of goods, to construct superlative chained indices. However, scanner data also pose several problems for statistical agencies, both conceptual and practical (Triplett, 2003).

The conceptual issues this paper focuses on are those of volatile sales and seasonality. Both loom large in the case of chained indices, which are traditionally preferred to fixed base indices. High frequency substitution in scanner data results in biased chained indices, as acquisition behaviour can result in extreme quantities, for example due to periodic (promotional) sales (Feenstra and Shapiro, 2003). Seasonality especially poses problems in chained indices, because the quantity sold of seasonal products is by definition zero in some or many periods (see Chapter 22 in the forthcoming CPI manual; ILO et al., 2004).

This paper uses scanner data, which Statistics Netherlands obtains from a major supermarket chain on a weekly basis and uses for the official CPI. Several types of index number formulas are investigated, to determine the best solutions to deal with the problems of volatile sales and seasonality. We argue that despite several drawbacks, scanner data provide an excellent source to confront these problems.

Keywords: Consumer price index, scanner data, volatile sales, seasonality

1. Introduction

The theory of price index numbers goes back a long time. Already in the 1920s, Irving Fisher (1922) listed several hundreds of different index numbers. So, when constructing price indices, statistical agencies have a substantial theoretic underpinning at their disposal in the choice of which index number formulas to use.

Unfortunately, theoretical possibilities are curtailed by practical issues. Many of these issues relate to the observation of prices and quantities. Traditionally, statistical agencies construct a sample out of all possible goods and services, and send interviewers to collect these prices. As for quantities, these are generally not used at all. Rather, statistical agencies compile weights, which are associated with distinct (types of) goods and services.

These weights are generally not associated with actual transactions, and they are held fixed for relatively long periods, often years. In the Netherlands, and many other countries as well, weights for the CPI are based on national account data and budget interviews among consumers.

The increasing availability of scanner data provides the opportunity to use data on actual transactions in the compilation of the CPI. Although scanner data are themselves associated with several conceptual problems (see Triplett, 2003), they have the distinct advantage of offering observations of both prices and quantities with a regularity and reliability that traditional methods of data collecting cannot match. For this and other reasons, Statistics Netherlands decided to use scanner data from several supermarket chains in the CPI in June 2002.

This decision was preceded by intensive research.[1] Some issues this research focused on were which part of the scanner data to implement, how to use weights and the choice of an index number formula. The current paper is a follow-up to this research. It addresses two questions. First, how can we use scanner data to adequately incorporate seasonal products? Currently, such products are excluded from the scanner data that are used for the Dutch CPI. Second, what are the possible effects of products that are frequently on offer? Scanner data may provide a solution to deal with such products, but it may also raise unexpected problems. If not used correctly, scanner data may actually increase potential biases in the CPI. As such, it may prove to be a curse in disguise.

This paper has the following set-up. Section 2 gives a summary of the way in which scanner data are currently employed in the Dutch CPI. Section 3 explores different index number formulas first for a set of fruit products with both strong and weak

[1] See Schut (2003) for a review.

seasonal behaviour, and second for children's napkins, an article where promotional sales are very frequent. Section 4 concludes.

2. The use of scanner data in the Dutch CPI

In the 1990s, most Dutch retailers introduced bar code scanners in their stores. This led to an increase in efficiency for those retailers, including a complete overview of their turnover on a daily basis for each store. Such complete transaction information is not only useful for companies, but also for the compilation of price index numbers. Scanner data offer the possibility of observing transactions with actual prices paid and quantities purchased. This contrasts with the traditional method of sending interviewers, who only observe list prices and no quantities at all.

Statistics Netherlands recognised the potential of scanner data to improve both the collection of data and the quality of the data themselves and in 2000, it settled an agreement with several supermarket chains for the supply of scanner data. After thorough research, scanner data were introduced in the CPI in June 2002.

All retailers deliver their data on a weekly basis. Because of timeliness, only data from the first two weeks of each month can be used for the compilation of the monthly CPI. Therefore, the assumption is made that prices and sales in the first two weeks properly represent the entire month.

Scanner data are coded according to the European Article Number (EAN) system. Each article gets a unique EAN. For the scanner data part of the CPI, these EANs are used to match articles within the same retailer. EANs are grouped according to COICOP (Classification of Individual Consumption by Purpose). For each COICOP group containing articles sold in supermarkets, several index numbers are calculated: one each for the supermarket chains using scanner data, and one using data collected at other points of sale.[2]

For the scanner data indices, fixed baskets of several thousand EANs are determined each year, one each for all retailers. Each EAN in this basket gets a weight, which is based on its expenditure share in that year. Price indices are then constructed, where the price of an EAN in the current month is matched with its price in the previous month. This price ratio is chained with preceding price ratios, resulting in the relative of the current price and the base year price.[3] The scanner data indices are therefore annually chained Lowe indices.

---

[2] These indices are weighted with the number of price quotes that were collected at each retailer before the implementation of scanner data. Starting from next year, they will be based on actual expenditure shares.

[3] The base year of the CPI is shifted every five years, both for scanner data and non-scanner data. Currently, the base year is 2000. Note that for scanner data, the base year which determines the selection of EANs and their weights in the scanner data index is shifted every year, so that at this moment the base year is 2003.

It appears form the scanner data that the turnover rate of all EANs is very high. EANs constantly enter and exit the market, even on a weekly basis.[4] Only EANs which were sold in at least 48 of the 52 weeks of the base year, were selected for the basket.[5] Each year, it occurs fairly often that an EAN disappears, which would result in a missing observation. This problem is solved in a fairly traditional way: when the turnover share of this EAN in its CBL-group[6] is small, the class mean method is used. In other cases, a replacement EAN is found. When the old EAN and the replacement EAN are deemed too different, a quality adjustment factor is applied. The part of the Dutch CPI that is based on scanner data is therefore a hybrid of traditional matching procedures combines with a new way of collecting data and determining expenditure shares at the lowest level of aggregation.

Summarising, we obtain the following formula for the scanner data price index P_{Ak}^{rt} of product group *A* of retailer *k*, where *r* is the base year and *t* the current month (Schut, 2003):

$$P_{Ak}^{rt} = \sum_{i \in A} w_{ik}^{r} \left(\frac{\overline{p}_{ik}^{t}}{\overline{p}_{ik}^{r}} \right) = \sum_{i \in A} w_{ik}^{r} \prod_{\tau=r+1}^{t} \left(\frac{\overline{p}_{ik}^{\tau}}{\overline{p}_{ik}^{\tau-1}} \right) \qquad (1),$$

where *i* denotes an EAN and w_{ik}^{r} is the weight of EAN *i* for retailer *k* in base year *r*. The price $\overline{p}_{ik}$ is the average price (a unit value) of EAN *i* across all stores of retailer *k* in the respective period.

3. Scanner data issues: frequent sales and seasonality

Scanner data provide the obvious advantage of containing transaction data with actual prices and quantities, on a very regular basis. Does this mean that statistical agencies should abandon their traditional methods of collecting prices, and should make integral use of scanner data? Unfortunately, scanner date have some associated conceptual and practical problems. Triplett (2003) points at some issues which are often overlooked or ignored in the use of scanner data. His main point is that scanner data measures acquisition rather than consumption. This implies that search, shopping and inventory behaviour of consumers are incorporated in them. However, standard price index theory, among which the cost-of-living approach, does not provide adequate tools to deal with these issues. This is especially apparent when articles are on sale. Consumers tend to hoard articles when they are on offer,

[4] This does not necessarily mean that also different products constantly enter and exit. Even small changes in the package design or the fact that an article is on sale may lead to a different EAN for this article.

[5] Seasonal products are therefore excluded.

[6] A CBL-group is the level of aggregation below a COICOP group. This aggregation is only made for convenience; price indices are not published at this level.

consuming from inventory when they are not. In a chained index formula, frequent sales can lead to severe biases. Feenstra and Shapiro (2003) illustrate this effect with a well-behaved scanner data set on tuna sales. In this section, we will explore the effect of frequent sales in the scanner data set used for the Dutch CPI. We will focus on children's napkins, an article group where promotional sales are frequent.

Although scanner data and seasonality in price index numbers are not often treated simultaneously, scanner data provide a good opportunity to investigate seasonal effects and how to deal with them. Fruit and vegetables are obvious examples of seasonal products which are largely purchased in supermarkets. For several reasons outlined below, seasonal products are excluded from the supermarket scanner data used at Statistics Netherlands. Chapter 22 in the ILO CPI manual (ILO *et al*., 2004) on seasonal products uses an artificial data set to investigate several index number formulas to deal with seasonal effects. This is, however, a very neat data set where the seasonal pattern in all articles is very regular. Actual data on the purchase of fruits and vegetables rarely show such a regular pattern, largely because of differences in the weather pattern. A second focus of this paper therefore is how scanner data can be used to deal with irregular seasonal patterns.

3.1 Seasonal effects in scanner data: the case of fruit

Chapter 22 of the ILO CPI manual (ILO *et al*., 2004) discusses the treatment of seasonal products in the CPI. Seasonal products pose price statisticians for severe difficulties, as the expenditure on these products varies greatly, so that keeping weights fixed is even more problematic than normally. This is especially the case for products where expenditures have a so-called strongly seasonal pattern, i.e. products that are not sold at all during certain periods of the year.

There are many kinds of seasonal products, of which fresh fruits are a prime example. Although many more kinds of fruit are available throughout the year now than in the past, some fruits are still only available during a certain period of the year. Moreover, many perennial fruits also witness a seasonal fluctuation in prices and quantities purchased. Such products are also referred to as weakly seasonal products.

Strongly seasonal products have the obvious disadvantage that a month-to-month prices index can only be calculated during the short period of time when they are available for consecutive months. Any price changes at the moments of entry and exit are not registered.

Furthermore, a problem with traditional data collection is that no quantities are observed, so that seasonal fluctuations in quantities are not registered for weakly seasonal products, whereas they are in the case of prices. When a strongly seasonal product is not available, its price is simply not registered, so that the seasonal pattern is witnessed to some extent. However, interviewers are usually instructed to only collect prices of strongly seasonal products in pre-defined periods. Such products may therefore be available, while their prices are not collected.

An additional disadvantage of using fixed weights for seasonal products is that the seasonal pattern may not be identical in each subsequent year. The availability of fresh fruits and vegetables is especially subject to climatologic factors, which may result in a difference of up to several months for the period when a specific item is available each year. For example, the scanner data used in this paper show that in 2000, strawberries were available from May through September, but in 2003 from March through June.

Because they contain transactions, scanner data are potentially much better suited to deal with seasonal products than traditional price collection methods. However, seasonal products, like fresh fruit, were excluded from the scanner data that were used for the Dutch CPI. The first reason for this is that only EANs that were available in at least 48 weeks in the base year are included, so that only articles that are purchased very frequently turn up in the basket. Strongly seasonal articles are thus eliminated automatically.

The second reason is of a more practical nature: many fresh fruits are not sold in a fixed quantity. Instead, consumer choose the quantity they purchase themselves and use a balance to determine the price. In this procedure, individual stores themselves attach an EAN to the product, so-called 'in-store' EANs. Unfortunately, these in-store EANs are not the same between stores or in different months. Using EANs to match articles may therefore result in, for example, matching mangos with sliced bread, another type of product with in-store EANs.

Not every fruit and vegetable gets an in-store EAN, however. Articles with a set quantity, like a 500 grams box of strawberries, have their own standard bar code, and therefore a 'conventional' EAN, and can be matched without difficulties.

In this section, five kinds of fresh fruit were used from the scanner data: strawberries, white grapes, red grapefruits, mangos and golden delicious apples. All articles have a set quantity, so no problems with in-store EANs occur. To simplify calculations, only scanner data from one retailer were used. The data contains weekly observations from 2000 to 2003. Appendix table A.1 shows the monthly fluctuations in prices and quantities. Strawberries and grapes are strongly seasonal goods, the other fruits are weakly seasonal. Every index number formula without any form of seasonal adjustment will thus show strong fluctuations in the resulting index.

Chapter 22 of the CPI manual discusses the problems posed by seasonal products in the calculation of the CPI, and presents several alternative index number formulations. Some of these index number formulas measure aggregate price change when there are seasonal products, others eliminate seasonal patterns altogether. These indices are illustrated with an artificial data set consisting of some strongly and weakly seasonal kinds of fruit. An important difference between the artificial data set and real-world data used in CPIs, is that the former includes quantities in every month, whereas these are generally not available in a CPI. As seasonal behaviour is especially expressed in quantities, conventional CPI data is unsuitable to compare different methods of seasonal adjustment. On the other hand, artificial

data may be just too artificial, bearing too little similarity to reality. For example, it is much more well-behaved than the scanner data we use here. In the artificial data set, the strongly seasonal commodities are available in the same months every year. This is not the case in our data, so that not only month-to-month indices will yield difficulties, but also monthly year-to-year indices. Scanner data therefore present an ideal data set to compare methods of seasonal adjustment. Like the artificial data set, it contains quantities. Moreover, scanner data are 'real', and so reflect actual seasonal patterns, which are much less well-behaved than those in an artificial data set. Any recommendation to what type of index number formula to use in the face of seasonal products needs to be founded on actual rather than artificial data.

The CPI manual discusses several price indices, of which three types are considered here: monthly year over year indices, monthly rolling indices and the Rothwell index.

Year over year indices simply compare prices in the current month with prices in the same month in the base year. The base year can either be a fixed reference year, yielding a fixed index, or the previous year, yielding a chained index. Hence, each month only prices are compared of goods that are present in both the current month and the same month in the base year. For the *n* articles, the monthly year-to-year Laspeyres ($P_L^{t,m}$) and Paasche ($P_P^{t,m}$) indices in month *m* of year *t* compared with base year t_0 can be written as:

$$P_L^{t,m} = \frac{\sum_{i=1}^{n} p_i^{t,m} q_i^{t_0,m}}{\sum_{i=1}^{n} p_i^{t_0,m} q_i^{t_0,m}} \tag{2}$$

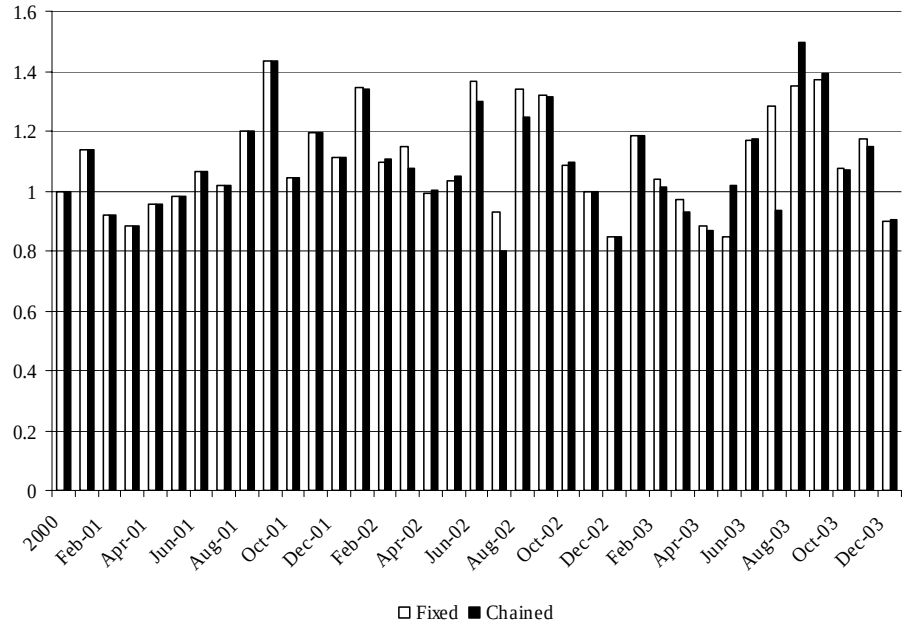
$$P_P^{t,m} = \frac{\sum_{i=1}^{n} p_i^{t,m} q_i^{t,m}}{\sum_{i=1}^{n} p_i^{t_0,m} q_i^{t,m}}, \tag{3}$$

If the seasonal pattern of strongly seasonal goods is the same every year, there will be no missing prices, like in month-to-month indices. In a month-to-month index, a strongly seasonal commodity is excluded from the index in the month it appears, and the month it disappears.

Seasonal effects will be eliminated in a monthly year-to-year index only if the seasonal pattern in both prices and quantities is exactly the same in each year. In our data, this regularity is not present, so that such an index will show strong fluctuations. This is illustrated in figure 1, which shows two monthly Fisher year-to-

year indices based on our scanner data: a fixed base index with 2000 as the base year, and a chained base index using *t-1* rather than t_0.[7]

Figure 1. Fixed and chained monthly year-to-year Fisher indices, fruit



Clearly, such year-to-year indices do not tell us anything about aggregate price changes on a monthly basis. In fact, figure 1 shows 24 price indices, two for each month in a year. Two types of aggregate index number formulas that also deal with seasonal products are considered here: rolling year indices and the Rothwell index. In essence, both types of indices use aggregations over an entire year. Seasonal fluctuations are thus smoothed out. A rolling (or moving) year index is actually an annual index, which is updated monthly. In a rolling year index, the prices in a period of twelve months are compared with the prices in the same months of the reference twelve-month period. In a fixed rolling year index, the reference period is simply January to December of the base year. In a chained or shifting base index, the reference is the same twelve-month period one year earlier. A rolling year index equals the average price change in the past twelve months. The Laspeyres ($P_{L,R,fixed}^{t,m}$) and Paasche ($P_{P,R,fixed}^{t,m}$) fixed base rolling year indices in month *m* of year *t* can then be written as:

[7] These indices and all indices shown in subsequent graphs and tables use weekly data aggregated over months. Since the Dutch CPI only uses the first two weeks of every month all (monthly) indices presented here also refer to the first two weeks each month.

$$P_{L,R,\,fixed}^{t,m} = \frac{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{t-1,j} q_i^{t_0,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{t,j} q_i^{t_0,j}}{\sum_{j=1}^{12}\sum_{i=1}^{n} p_i^{t_0,j} q_i^{t_0,j}} \tag{4}$$

$$P_{P,R,\,fixed}^{t,m} = \frac{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{t-1,j} q_i^{t-1,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{t,j} q_i^{t,j}}{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{t_0,j} q_i^{t-1,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{t_0,j} q_i^{t,j}} \tag{5}$$

where t_0 is the base year. Naturally, when t=1, t-1 en t_0 coincide.

Similarly, chained rolling year Laspeyres ($P_{L,R,chained}^{t,m}$) and Paasche ($P_{P,R,chained}^{t,m}$) indices in month m of year t (t_0 being the base year) are defined as:

$$P_{L,R,chained}^{t,m} = \prod_{\tau=t_0}^{t}\left[\frac{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{\tau-1,j} q_i^{\tau-2,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{\tau,j} q_i^{\tau-1,j}}{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{\tau-2,j} q_i^{\tau-2,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{\tau-1,j} q_i^{\tau-1,j}}\right] \tag{6}$$

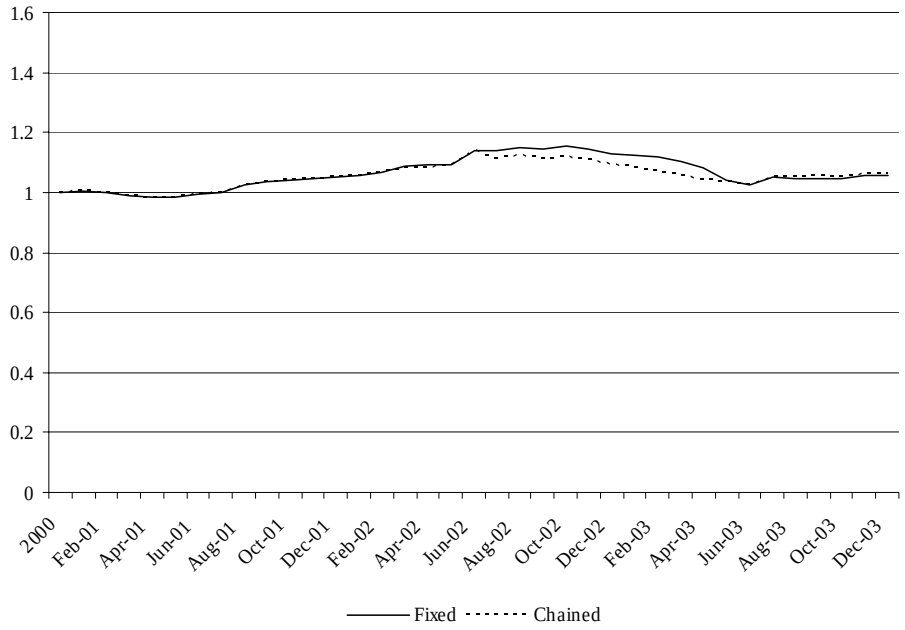
$$P_{P,R,chained}^{t,m} = \prod_{\tau=t_0}^{t}\left[\frac{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{\tau-1,j} q_i^{\tau-1,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{\tau,j} q_i^{\tau,j}}{\sum_{j=m+1}^{12}\sum_{i=1}^{n} p_i^{\tau-2,j} q_i^{\tau-1,j} + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i^{\tau-1,j} q_i^{\tau,j}}\right] \tag{7}$$

Again, note that, when t=1, t-1 and t-2 equal t_0 and when t=2, t-2 equals t_0.

As rolling year indices are annual indices rather than monthly ones, seasonal fluctuations are nearly entirely smoothed out. This is illustrated in figure 2, which shows the fixed base and chained Fisher rolling year indices.

Unfortunately, the rolling year index also comes with two drawbacks, that may be significant for statistical agencies. The CPI is essentially a short-term statistic. It measures inflation on a monthly basis, whereas a rolling year index measures annual inflation in a given month. In a rolling index, price changes in general are smoothed out, not only seasonal effects. These are therefore different concepts of inflation, which cannot easily be aligned. For central banks and other parties interested in annual inflation, a rolling year index is an adequate measure of price change, but for statistical agencies it is less suitable. Second, because a rolling year index is the average measure of price change over the past twelve months, it actually measures the average annual price change of six months ago, resulting in a lag of six months.

Figure 2. Fixed and chained Fisher rolling year indices, fruit



For a short-term statistics like the CPI, an alternative method of seasonal adjustment seems necessary. Different statistical agencies use different methods to correct their price indices, as illustrated by Turvey (1979). One popular method of seasonal adjustment is the Rothwell index, of which several variants are in use.[8] In its basic form, the Rothwell index in month *m* of the current year *t* compares prices in this month with the annual average prices of the base year t_0. We refer to this index as the 'fixed base' Rothwell index, because the quantities used are those in the corresponding month *m* in the base year t_0.

$$P_{R,\,fixed}^{t,m} = \frac{\sum_{i=1}^{n} p_i^{t,m} q_i^{t_0,m}}{\sum_{i=1}^{n} p_i^{t_0} q_i^{t_0,m}}, \tag{8}$$

with

$$p_i^{t_0} = \frac{\sum_{m=1}^{12} p_i^{t_0,m} q_i^{t_0,m}}{\sum_{m=1}^{12} q_i^{t_0,m}} \tag{9}$$

The Rothwell is a short-term price index, showing monthly price change including seasonal fluctuations. Unfortunately, the Rothwell index still misses price changes if the seasonal pattern changes over the years, like in our data. A changing seasonal pattern means that for some months, $p_i^{t,m}$ is not observed, while $q_i^{t_0,m} > 0$, or vice

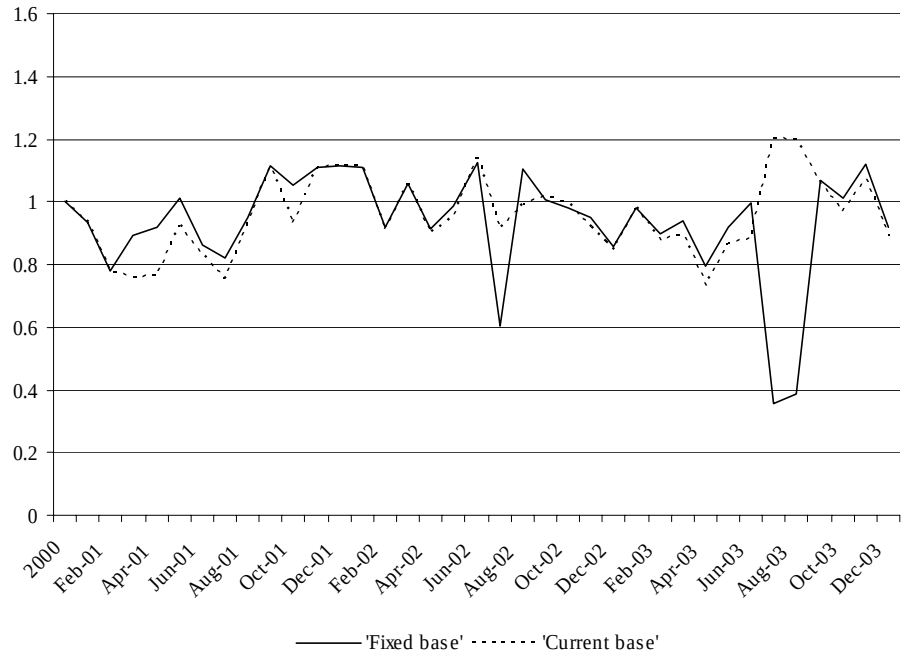
[8] Statistics Netherlands also uses a variant of the Rothwell index..

versa. To prevent this from happening, an alternative specification of the Rothwell index could include current quantities in (8) rather than quantities of the base year, yielding a 'current base' Rothwell index:

$$P_{R,current}^{t,m} = \frac{\sum_{i=1}^{n} p_i^{t,m} q_i^{t,m}}{\sum_{i=1}^{n} p_i^{t_0} q_i^{t,m}}, \qquad (8')$$

The average price of the base year $p_i^{t_0}$ would still use base quantities, so equation (9) remains unchanged.

Figure 3. Rothwell indices, 'fixed base' and 'current base'



In several months, the differences between both Rothwell indices are quite substantial. This is especially the case in July and August 2003. A closer inspection of the scanner data reveals that in these months, strawberries were not available anymore, while they were available in these months in each of the preceding years. This change in the seasonal pattern causes a difference between both index of more than 0.8 points (on a scale of 1)!

We prefer the 'current base' Rothwell index, as it gives the best reflection of current seasonal patterns. In this Rothwell index, no mismatch occurs between current prices and base year quantities, which is the case in the traditional 'fixed base' Rothwell index. In our view, Baldwin's (1990) recommendation of the Rothwell index needs some adjustment, in that the quantities in the index should be current quantities rather than base year quantities. Of course, in statistical practice this is usually not possible, because current quantities are generally not available. Scanner

data, however, do contain current quantities sold, making clear the distinct advantage of using scanner data for the treatment of seasonal products.

Summary

Scanner data expose the reality of actual transactions, and therefore lend themselves very well to investigate seasonal patterns in detail. Data rarely behave in ways that are pleasant to statisticians or researchers, and this is better reflected in scanner data than in price data collected with traditional methods. The traditional index to treat seasonal products, the Rothwell index, suffers from the fact that a change in the seasonal pattern over the years can create a serious mismatch between current prices and base year quantities. Scanner data offer the advantage of providing current quantities as well as prices, so that this mismatch can be eliminated in a alternative specification of the Rothwell index.

If the goal is to smooth out seasonal patterns altogether, a different approach is necessary. An annual index like the rolling year index succeeds very well in eliminating seasonal fluctuations in an aggregate price index, but comes at the price of not being a short term statistics and having a lag of six months. In some cases, annual indices may be preferred to monthly ones, but in the scope of most CPIs, our 'current base' Rothwell index seems the most suitable way to treat seasonal fluctuations in prices and quantities.

3.2 Frequent sales in scanner data: the case of children's napkins

A product group where promotional sales are fairly frequent is that of children's napkins. As any young parent can tell you, many consumers only buy napkins when they are on sale, hoarding them until the next or even beyond. This pattern of acquisitions is illustrated in appendix figure A.2, which shows the quantity purchased of one all napkins in our data set. The periods with a promotional sale can easily be distinguished.

In the measurement of price changes, acquisition is generally assumed to equal consumption. This assumption may be rather plausible for non-durables, if not for durables. However, as shown in figure A.2 and in Feenstra and Shapiro (2003), this assumption does not even hold for (admittedly storable) non-durables like canned tuna and napkins.

As pointed out by Feenstra and Shapiro (2003), chained indices of articles with frequent promotional sales can suffer from severe biases. In the case of a Laspeyres index, the preferred index of many statistical agencies, this bias is upward because the price decline in the period when it is on offer gets a much smaller weight than the price increase when the price returns to its pre-sale level. They also find an upward bias in the chained Törnqvist index, which they attribute to the fact that lower prices only attract high purchases when they are accompanied by advertising, which usually takes place in the final weeks of a sale.

When Statistics Netherlands first acquired scanner data from two major retailers the first aim was to construct chained Fisher indices (Schut, 2001). However, the chained indices appeared to contain substantial biases. One of the causes of these biases was that some articles were on sale very frequently. Eventually, the decision was made to use a fixed base Laspeyres index, where the base year would be shifted every year. The corresponding formula was shown in (1).

Using current price index theory, we investigate several index number formulas to treat scanner data for articles that are frequently on sale, like napkins. As in the previous chapter, all indices are based on the first two weeks of sales per month only. This includes base period prices and quantities.

Because we have no data on the way articles on sale are promoted and when they are promoted during the sale, we cannot a priori expect to find an upward bias in a Törnqvist index, like Feenstra and Shapiro (2003) do. The chained Törnqvist index (which compares prices in each month with those in the previous month) in month *m* is given by:

$$P_T = \exp\left[\sum_{i=1}^{n} \frac{1}{2}\left(w_i^{m-1} + w_i^{m}\right)\ln\left(\frac{p_i^{m}}{p_i^{m-1}}\right)\right] \qquad (11)$$

We chose 2000 as our base period. This means that for January 2001, p_i^{m-1} is the average price of article *i* in month *m* in 2000, and w_i^{m-1} is the corresponding expenditure weight.

The resulting chained Törnqvist index for napkins is shown in figure 4. An upward bias is not visible from this figure, although the index looks rather volatile: in some months the difference with the previous month is more than ten percent. The chained Laspeyres is not shown here, but it appears to have a strong upward bias, as expected. In December 2003, it has a value of 4.1.

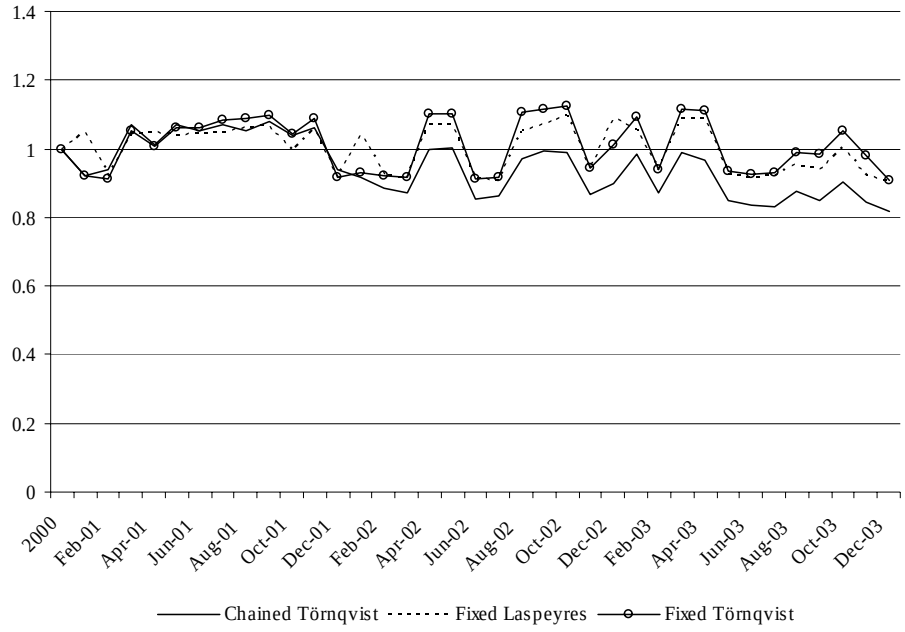
Fixed base indices are not as likely to contain substantial biases as chained ones. However, the strong monthly price fluctuations caused by periods of promotional sales will still be reflected in a fixed base index. The fixed base Laspeyres and Törnqvist indices are also shown in figure 4.

All indices in figure 4 show more or less the same volatile pattern, caused by periodic sales. Rather than an upward bias, the chained Törnqvist has a downward trend vis-à-vis the fixed base indices. When taking a closer look at the data, this is not that surprising. It appears that after a period of sale, in most cases the quantity sold of an article that was on sale dips somewhat below its pre-sale quantity. When this is the case, the price increase after a sale has a larger weight than the price decrease during a sale.

But the pattern of purchases of napkins that are regularly on sale tells another story as well. The quantity purchased of napkins that are on sale may be somewhat smaller just after than just before a sale, both quantities are dwarfed by the quantity sold during a period of sale, as shown in figure A.2. This suggests that there are

many consumers who only buy napkins when they are on sale: the 'inventory shoppers' described by Triplett (2003). For such consumers, only price changes from one period of sale to the next are relevant, rather than monthly price changes measured in traditional indices.

Figure 4. Chained base and fixed base indices, napkins



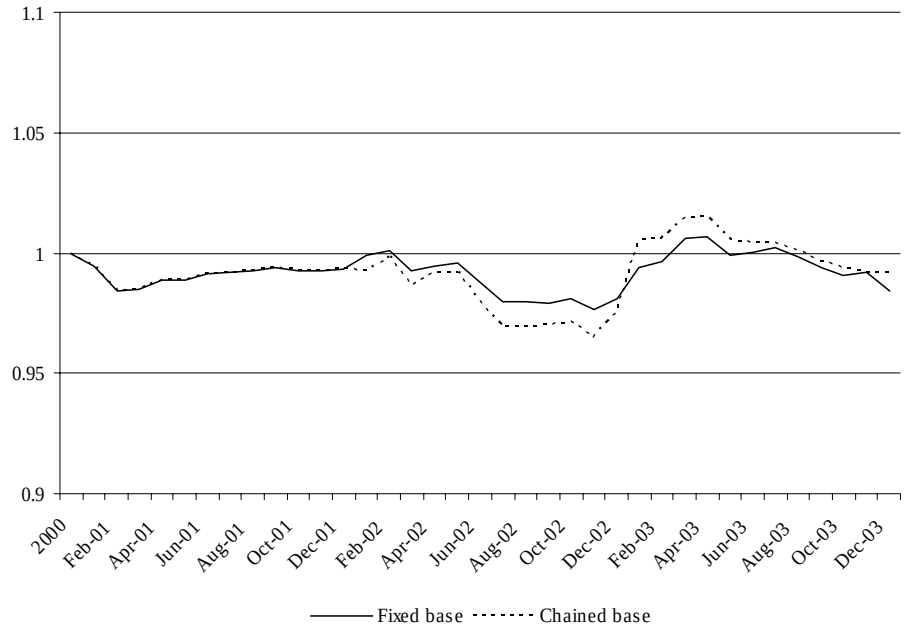
Clearly, consumption and acquisition of napkins do not coincide. A solution for this problem is to expand the time horizon in the index. In the case of napkins, a horizon of one year rather than one month looks reasonable. Within a given year, consumption and acquisition of napkins are more likely to coincide. Choosing such a long period, however, implies that the price index for napkins can only be computed once each year, instead of each month.

A solution may be found in the calculation of a rolling index, which can be updated monthly. As described in the previous section, a rolling index compares the prices in a period of twelve months (the 'rolling year') with those in the same months in the base twelve-month period. Fortunately, the periodicity of sales is quite regular in the case of napkins: they roughly occur in the same weeks every year. This is actually quite common for consumer goods, both durable and non-durable. An additional advantage of a rolling year index is therefore that price comparisons are made between different periods of sale, thus adequately capturing the shopping behaviour of 'inventory shoppers'.

Figure 5 shows the rolling year fixed base and chained base Fisher indices. The underlying Laspeyres and Paasche indices are based on equations (4) to (7). In these Fisher indices, the volatility of the month-to-month indices is smoothed out. Of course, the traditional disadvantages of rolling indices, mentioned in the previous section, also hold here. Moreover, the likelihood of 'missing prices' due to entries

and exits increases when prices are compared over one year instead of monthly. Nevertheless, a rolling year Fisher index seems ideally equipped to deal with 'inventory shoppers' and with the discrepancy between acquisition and consumption, two major problems of traditional monthly indices based on scanner data.

Figure 5. Rolling year fixed and chained base Fisher indices, napkins



4. Summary and conclusions

This paper uses scanner data to construct price indices for two problem areas in price index measurement: seasonal products and articles with frequent sales. Because they offer transaction data on a very frequent basis, scanner data reveal issues that are often overlooked with data that are collected using traditional methods.

As Triplett (2003) argues, these issues need to be confronted by developing adequate theoretic tools, rather than just adjusting the data to fit existing theory. He stresses the potential risks of scanner data, so that these data may appear to be a curse in disguise for price statisticians.

But without neglecting potential risks, the benefits of scanner data in price measurement are obvious, and should not be ignored: they provide statisticians with actual transaction prices, rather than list prices nobody may actually pay. Scanner data may measure acquisitions rather than consumption, traditional methods of price collection measure neither. Moreover, scanner data allow expenditure weights to be directly based on actual purchases and to adapt them frequently and up-to-date,

rather then using some indirect measure (like budget interviews or national accounts estimates), which usually can only be adjusted once every few years.

The part of the Dutch CPI where scanner data are currently employed (discussed in Section 2) appears to be a hybrid index: it uses fixed expenditure weights on the level of individual articles to avoid biases due to promotional sales. The consumption basket is adjusted each year, to allow both for substitution and entries and exits. To further mitigate the effect of entries and exits, replacements are found disappearing articles, using an explicit quality adjustment where necessary. This leads to an index that is arguably better than one based on traditional data.

As pointed out in this paper, fixed base price indices using scanner data still have problems when there are strong fluctuations in prices and quantities, like with seasonal products and articles that are on sale frequently. We argue that a Rothwell index using current quantities is the best method to adequately treat seasonal commodities. Scanner data are at the moment the only source for current quantities, which is an obvious advantage over traditionally collected data.

If the goal is to smooth out seasonal fluctuations, rolling year indices offer a possible way to reach such a goal. Rolling year indices also appear to provide an excellent way to remove the discrepancy between consumption and acquisition of a storable non-durable good like children's napkins.

Unfortunately, rolling year indices also have particular disadvantages for statistical agencies. First, they provide a measure of annual price change, rather than a short-term monthly index. This is not in line with a CPI, which is nearly always presented as a short-term statistic. Second, rolling indices have a lag of six months: they provide the average price change over the past twelve months, which equals the average annual price change of six months ago.

Statistical agencies need to decide whether these drawbacks of rolling year indices outweigh their advantages. One possible choice would be to use rolling year indices for partial indices of articles where they offer the best scope for improvement, thus yielding a hybrid CPI. Whatever the outcome of such decisions, it is clear that scanner data provide an excellent source of data for the index number formulas presented in this paper.

References

Baldwin, A. (1990), "Seasonal baskets in consumer price indexes", *Journal of Official Statistics*, vol. 6, no. 3, pp. 251-273

Feenstra, R.C. & M.D. Shapiro, "High-Frequency Substitution and the Measurement of Price Indexes", Scanner Data and Price Indexes (Chapter 5), Studies of Income and Wealth, volume 64, 2003.

Fisher, I. (1922), *The Making of Index Numbers*, Houghton Mifflin, Boston

International Labour Office, OECD, Eurostat, IMF, Worldbank, United Nations (2004), *Consumer Price Index Manual, Theory and Practice*, Geneva

Schut, C. (2001), *Using scanner data to compile price indices: experiences and practical problems*, Presented at the Joint ECE/ILO meeting on Consumer Price Indices, Geneva, November 1-2, 2001

Triplett, J.E., "Using Scanner Data in Consumer Price Indexes, Some Neglected Conceptual Considerations", Scanner Data and Price Indexes (Chapter 6), Studies of Income and Wealth, volume 64, 2003.

Turvey, R. (1979), "The treatment of seasonal items in consumer price indices", *Bulletin of Labour Statistics*, Fourth quarter, International Labour Office, Geneva, pp. 13-33

Appendix

Table A.1 Monthly prices and quantities sold of five kinds of fruit, 2000-2003

	Strawberries		White grapes		Red grapefruits		Mangos		Golden del. apples	
	price	quantity	price	quantity	price	quantity	price	quantity	price	quantity
Jan-00	0.00	0	0.00	0	1.81	10339	0.90	8909	1.56	5000
Feb-00	0.00	0	2.26	2535	1.81	10631	0.96	20388	1.59	4512
Mar-00	0.00	0	2.26	3125	1.81	10947	1.36	6145	1.59	4186
Apr-00	0.00	0	1.81	3026	1.81	11912	1.59	4928	1.59	3794
May-00	2.97	34836	1.81	3598	1.81	11865	1.01	17994	1.59	4610
Jun-00	1.95	66445	1.81	1	1.81	11706	0.96	21541	1.93	4439
Jul-00	1.91	43799	0.00	0	1.36	15571	1.23	5242	2.26	4138
Aug-00	1.99	40965	0.00	0	1.36	9779	0.81	17594	2.26	3760
Sep-00	2.72	346	0.00	0	1.44	10180	1.03	11764	1.70	4213
Oct-00	0.00	0	0.00	0	1.81	10965	1.36	5931	1.58	6004
Nov-00	0.00	0	0.00	0	1.81	10995	1.36	5946	1.69	4865
Dec-00	0.00	0	0.00	0	2.03	9400	1.36	5338	1.80	3593
Jan-01	2.72	1	0.00	0	2.04	10493	1.03	9881	1.80	3650
Feb-01	0.00	0	1.81	3091	1.55	19727	0.98	19786	1.58	3591
Mar-01	0.00	0	1.42	24572	2.04	12403	1.19	10906	1.58	3538
Apr-01	0.00	0	1.44	27452	2.04	13137	1.36	7570	1.58	4185
May-01	2.70	11478	1.49	9239	2.04	14537	1.24	8299	1.58	4050
Jun-01	1.97	83228	1.80	1611	2.23	12206	1.35	8055	2.01	4335
Jul-01	1.74	88415	1.81	2	2.26	10450	1.36	7153	2.25	3877
Aug-01	2.19	49337	0.00	0	2.26	9004	1.36	7538	2.26	4559
Sep-01	3.17	5	0.00	0	2.26	9216	1.36	6711	2.25	4102
Oct-01	0.00	0	0.00	0	2.26	9888	0.96	28565	2.25	3840
Nov-01	0.00	0	0.00	0	2.25	8760	1.36	8397	2.25	3839
Dec-01	0.00	0	0.00	0	2.26	9078	1.36	6604	2.25	3569
Jan-02	0.00	0	1.81	40	2.26	10767	1.36	6629	2.26	4166
Feb-02	0.00	0	1.99	3240	2.26	12192	0.87	24542	2.25	4006
Mar-02	0.00	0	1.98	4056	2.26	13386	1.29	7456	2.26	4612
Apr-02	1.99	3	1.98	4199	1.69	23761	1.29	7676	2.26	4754
May-02	2.49	2514	1.98	11211	2.26	13182	1.03	22790	2.26	4043
Jun-02	2.79	3184	0.00	0	2.26	12266	1.51	7866	2.32	4542
Jul-02	1.12	3458	0.00	0	1.80	10855	1.49	6730	2.38	4291
Aug-02	2.85	1188	0.00	0	1.59	11424	1.46	6694	2.49	5638
Sep-02	0.00	0	0.00	0	1.89	10959	1.37	7458	2.25	5596
Oct-02	0.00	0	0.00	0	1.89	9747	1.40	8315	1.99	4092
Nov-02	0.00	0	0.00	0	1.99	8180	0.99	10803	1.98	3675
Dec-02	0.00	0	0.00	0	1.67	11400	0.99	9723	1.99	3477
Jan-03	0.00	0	0.00	0	1.99	11149	1.22	7876	1.97	3857
Feb-03	0.00	0	1.95	2536	1.66	15150	1.11	6065	2.28	3495
Mar-03	0.00	0	1.75	2704	1.99	12206	0.99	24170	2.28	3776
Apr-03	1.87	24691	1.51	20987	1.42	19400	1.13	8063	2.28	4164
May-03	1.99	44375	1.95	5017	1.99	12777	1.59	7381	2.28	3642
Jun-03	2.49	21	1.77	21059	2.29	11635	1.05	28978	2.28	4372
Jul-03	0.00	0	2.29	6	2.48	9921	1.49	6464	2.28	3991
Aug-03	0.00	0	0.00	0	2.48	10674	1.49	8501	2.28	5483
Sep-03	0.00	0	0.00	0	2.48	9473	1.08	16481	2.28	4365
Oct-03	0.00	0	0.00	0	2.03	12748	1.04	13279	2.28	3355
Nov-03	0.00	0	0.00	0	2.48	8875	0.99	9271	2.27	3201
Dec-03	0.00	0	0.00	0	1.78	14151	0.99	10389	2.28	2801

Figure A.2 Quantity sold of baby's napkins, 2000-2003