Comparing scanner data and survey data for measuring price change of drugstore articles

Antonio G. Chessa[1]

Abstract. Scanner data of turnover and quantities sold per EAN were introduced in the Dutch CPI in 2002. Today, the CPI makes use of scanner data from supermarkets, DIY stores, travel agencies and of fuel prices. After ten years, scanner data cover almost a quarter of the total weight in the CPI and possible uses for other types of stores are being studied.

This paper compares price indices based on survey data with price indices calculated from recently obtained scanner data on sales of drugstore articles. In theory, scanner data offer optimal coverage of articles and brands sold. However, such data sets also expose users to pitfalls, which are partly linked to the rules for assigning EAN numbers to articles. In this study, particular attention is paid to the so-called phenomenon of "relaunches". This term is used in situations where articles are replaced by items that mainly differ from their predecessors by external appearance of their packaging rather than content or ingredients. The follow-up items are assigned a new EAN and are usually introduced at a higher price. Price increases with respect to the preceding items are therefore missed when calculating price changes per EAN.

Price indices are calculated and compared for drugstore articles at different levels of product aggregation (for article groups and at Coicop level). The price index calculated from survey data underestimates the price increase based on scanner data by about six percentage points at Coicop level (taken over 2011 and 2012). The survey-based approach gives a good approximation for article groups when the survey contains articles that underwent a relaunch. Relaunches are the most important driver behind price increase for various toiletry groups, while shifts in sales among articles of the same brand and consumer target group dominate the price increase for beauty products.

This study shows that scanner data should be preferred over traditional surveys, when data quality allows using scanner data. The results also show that price indices are sensitive to different choices regarding product characterisation, which should therefore be treated with great care. It is clear from this study that EANs should not be used to differentiate products.

Keywords: Consumer price index (CPI), scanner data, EAN, surveys, drugstore articles.

1. Introduction

The first use of scanner data for CPI calculations at Statistics Netherlands goes back to 2002. Since then, the use of such data has been extended through a number of phases and possibilities for widening its range of application are still being investigated. Scanner data have clear advantages over traditional survey data collection, notably because such data sets offer a better coverage of articles sold, sales data offer complete information (prices and quantities), while the data collection process can be carried out at a much lower cost compared to traditional surveys. However, scanner data have drawbacks as well, and their use therefore proceeds in a stepwise manner. In spite of their potential, scanner data are still used by a limited number of statistical agencies (e.g., see Rodriguez and Haraldsen (2006)).

In this paper, by scanner data we mean transaction data specifying turnover and numbers of articles sold by EAN (barcode). Scanner data were introduced in the CPI in 2002, which then involved two supermarket chains. In January 2010, the data were extended to six supermarket chains, as part of an ongoing re-design of the CPI (de Haan, 2006; van der Grient and de Haan, 2010; de Haan and van

[1] Statistics Netherlands (CBS), Prices Department, P.O. Box 24500, 2492 HA The Hague, The Netherlands.
E-mail: ag.chessa@cbs.nl The views expressed in this paper are those of the author and do not necessarily reflect the views of Statistics Netherlands.

der Grient, 2011). At present, scanner data of nine supermarket chains are used, while surveys are not carried out anymore for supermarkets since January 2013. Scanner data are also used for DIY stores (do it yourself stores), travel agencies and for fuel prices. At present, scanner data are used for more than 22% of the Dutch CPI (measured as a share of the sum of the Coicop weights, see Table 1).[2]

Other types of electronic price collection, such as "web scraping", are being studied but are not used at the moment in CPI calculations. Traditional survey methods are used for the other parts of the Dutch CPI, thus covering almost 78% in terms of Coicop weights. Traditional price collection methods include field surveys, surveys by telephone and mail and (non-automatised) internet data collection.

The use of scanner data has thus grown considerably in ten years and possibilities of using scanner data for other parts of the CPI are being investigated. One of the data sets that are currently under study is a scanner data set containing weekly sales information of drugstore articles sold in the Netherlands. The data set has been analysed to find out whether it is suitable for future use in the Dutch CPI. The data set offers an excellent opportunity for addressing different research questions concerning price measurement and index number calculation.

The objective of this paper is to calculate and compare price indices of drugstore articles based on scanner data and on survey data. Particular attention is given to the phenomenon of so-called "relaunches" when using scanner data. The term "relaunch" is coined for situations where a manufacturer decides to "refresh" its assortment. Articles are replaced by 'new' items, which return to the stores with a modified appearance of their packaging. Beside this, the follow-up items may also undergo a marginal change of the content of their package. But, more importantly, the follow-up items are assigned a new EAN and usually have a higher price than their predecessors. Price index methods should therefore link or group the old and new EANs in some way in order not to miss price increases after relaunches.

A part of this paper is thus dedicated to the questions whether it is possible to measure the contribution of relaunches to price change and to what extent survey based price indexes capture this 'price effect' for drugstore articles. Beside relaunches there could be other factors that contribute to price change, and this paper also attempts to quantify the size of additional factors. The study of the factors that may contribute to price change is closely linked to the degree of detail according to which products are characterised or differentiated. In this respect, this study also offers insight into the sensitivity of price change to variations in the level of detail of product characterisation.

This paper is organised as follows. In Section 2, the pros and cons of survey data and scanner data are listed, which are illustrated for drugstore articles. Section 3 briefly discusses the methods and results for price indices based on survey and scanner data. Differences between the two approaches are analysed in Section 4. The contributions of relaunches and other factors to price change are quantified for several article groups. The most important findings from this study are summarised in Section 5.

2. Pros and cons of survey and scanner data

Price indices for drugstore articles are calculated in the Dutch CPI by making use of prices collected from surveys, but also from scanner data of articles sold in supermarkets. We exclude the supermarket scanner data from further consideration in this study and focus on the survey part. The survey contains 24 drugstore articles, for which prices are collected exclusively at drugstores, so not in supermarkets.

The selection of articles in Table 2 shows that an 'article group' contains two brands at most, and one article per brand. Surveys are thus very restrictive in this sense, as price collection is a time consuming process. Scanner data have the advantage of electronic delivery and full coverage of brands and articles sold. In addition to this, scanner data contain complete sales information. That is, both prices and quantities are available, while surveys only collect prices and therefore have no information on turnover. This implies a narrower basis for setting up weights for the articles in surveys, which are used to combine price changes of articles to different levels of product aggregation. The weights of the drugstore articles in the survey are given in Table 2.

[2] Tables and figures are included at the end of this paper.

The advantage of complete information on articles sold, and of both prices and quantities in scanner data, also allows the calculation of accurate price indices. The accuracy may be affected when using survey information. As survey prices constitute a sample, the uncertainty in price indices will be larger than in price indices calculated with complete information offered by scanner data. For example, effects of discounts may be missed in surveys, which, in principle, is not the case when using scanner data. As was already mentioned in the preceding paragraph, also the reliability of the weights may affect the accuracy of price indices based on survey data.

Beside the major advantages mentioned, scanner data sets may also have weak points. If rules for coding product information are changed over time, then adjustments need to be made to index methods. For instance, if products are characterised in terms of the brand of an article, and by other characteristics in order to differentiate products, then index methods may not work properly when brand names are abbreviated or when other characteristics are excluded by the data supplier at some point. Table 3 gives an overview of advantages and disadvantages for survey and scanner data.

3. Price changes for the traditional approach and scanner data

In order to compare price changes calculated from survey and scanner data, two essential problems need to be dealt with: (1) how to characterise products, and (2) which index method to select for calculating price changes. The choices will be described below for the approaches using survey and scanner data.

Method based on survey data

Product characterisation

A selection of articles is made that is believed to be representative of the drugstore sector (Table 2). These are the separate products according to the survey.

Index method

Prices of the drugstore articles in the survey are collected in the first three weeks of a month. The number of prices collected per drugstore article per month ranges from 12 to 40. Articles with a higher weight tend to have a larger sample. The collected prices are used to calculate arithmetically averaged monthly prices for each article.

In this study, the month to month price changes of different articles are combined according to a weighted geometric mean (a weighted version of the Jevons index).[3] The weights of the articles are determined by combining information on expenditures from the national accounts and budget surveys. The values of the weights in Table 2 apply to 2012 but are reconsidered, and possibly adjusted, every year in the Dutch CPI. Combining the weights of the articles with their price changes results in price changes for three so-called L-Coicops[4] of drugstore articles: toiletries, beauty articles and other articles for personal care.

Method based on scanner data

Summary of the data

Scanner data of drugstore articles containing sales information on a weekly basis are supplied to Statistics Netherlands by a market leader in the Benelux. The data contain sales information of two

[3] In fact, the method used here is simplified compared to the method used for the Dutch CPI, which uses a Jevons index for an intermediate level of aggregation and a Laspeyres index for subsequently aggregating the price indices to Coicop level.
[4] The term "L-Coicop" denotes an article classification one level below Coicop level in the Dutch CPI. L-Coicops specify a more refined level of article classification (i.e., a Coicop may consist of one or more L-Coicops).

major drugstore chains in the Netherlands. The first week with sales data is week 3 in 2011. The data set contains the following information for every article sold:

- EAN number;
- Description of the article, which includes brand name;
- Week number and year, in which sales have been realised;
- Content of package and unit of measurement (millilitre, gram, units);
- VAT percentage;
- Turnover;
- Quantities sold;
- Drugstore chain;
- A classification system for grouping articles, which contains six levels of detail.

Only a few inconsistencies in the data were found.[5] Their effect on price change can be ignored. The scanner data were therefore judged to be suited for price index calculations.

The data contain all necessary information for the Dutch method that is used for the CPI. Beside turnover, quantities sold and EANs, the data contain an article classification, which is helpful for linking large numbers of EANs to different levels of aggregation like L-Coicops. According to the scanner data in this study, more than 15,000 EANs are sold every week.

Product characterisation

The first step in the development of an index method should be to differentiate products, for which prices are computed and compared between successive periods. In fact, this step consists of identifying a number of characteristics that distinguish products from others. Ideally, the products created should be 'homogeneous': products can be understood as sets of articles that can be considered as substitutes of each other.

What characteristics should be chosen in order to guarantee product homogeneity? The scanner data offer us a range of possible attributes to choose from, such as: EAN, brand name, content of a package and article classification. What level of detail should be chosen: should we distinguish products by EAN or should we select a level of less detail? Products differentiated by EANs are always homogeneous, since EANs are assigned to only one article. But different EANs may be substitutes of each other. The problem therefore is to find a level of detail at which products are still considered to be homogeneous, such that homogeneity is lost at levels of less detail.

We are unaware of any attempt to formalise and solve the problem of product homogeneity, and it is beyond the scope of this paper to contribute to this. Being a comparative study, the choices on product characterisation were guided by choices made in setting up the survey article list. From Table 2 we can recognise brand name as a product characteristic (e.g., Prodent for toothpaste). A second characteristic could be referred to as "consumer target group" (e.g., adults for toothbrush, and ladies and men for eau de toilette). We therefore decided to characterise products by:

- Brand name;
- Consumer target group.

Both attributes are available in the scanner data. Brand name is always mentioned at the beginning of an article description and target group can be selected from one of the levels of the article classification. For several article groups, such as eyeshadow, no target group is specified in the data, in which case brand name is used as the only product characteristic.

An additional characteristic for differentiating products could be package content. Reasons can be found for selecting content as well in theory, but it is omitted in this study.[6] The Dutch survey

[5] The numbers of EANs with different product descriptions and with changes in content specification over time were very small (several tens of cases on more than 10,000 EANs per week).

allows to switch between articles with different content for some article groups. For example, for hairspray an article of 400 ml is currently priced, but a switch to an article with a content of 300 ml is allowed in the future. This implies that survey-based price indices may also capture price changes due to shifts between articles that differ in content. In order to make a fair comparison between the price indices based on scanner data and survey data, we decided to leave out content as an additional factor for differentiating products when using scanner data to calculate price indices.

Products were defined for almost every article in the survey, which gave rise to the article groups in Table 4. The survey list of Table 2 was extended with article groups that make a substantial contribution to turnover (electronic toothbrushes, hairstyling, mascara and sun protection/aftersun). Baby wipes were excluded for technical reasons. The article groups cover almost 70% of total turnover in 2012. The index number calculations were done in Excel; considering all articles would have slowed down the computations considerably. The limitation on the number of article groups was therefore merely made for computational reasons rather than sampling considerations or other reasons.

Table 4 shows that the two product characteristics brand name and consumer target group lead to fairly large numbers of products for most article groups. Of the 24 article groups, 19 groups have an average number of EANs per product smaller than 10, while 10 article groups have an average number of EANs per product of 5 at most. Although this observation cannot be used to make hard statements on product homogeneity, it could serve as a rough indication. We emphasise again that the objective of this study is to compare price indices based on survey and scanner data, which is realised in part by identifying brand name and consumer target group as product characteristics in both approaches.

Index method

As with product characterisation, we decided to follow the choices made for the survey-based index method. This means that we used a weighted version of the Jevons index. The weights of the survey articles are reconsidered every year, so for the scanner data we decided to base the weights on the yearly turnover shares of the products (the shares are given in Table 4 for the article groups).

Arithmetic average prices per product are calculated for every week. Week to week price changes are calculated for every product, which are combined in a weighted geometric mean to obtain price changes at different levels of aggregation (article groups, L-Coicops and Coicop). Price changes are computed only for products that generated turnover in two successive weeks. We make the implicit assumption that the price change for the set of products that do not generate turnover in both weeks is equal to the price change computed for the products with turnover in both weeks.

Results

The price indices calculated with survey and scanner data are shown in Figure 1 for the three L-Coicops and at Coicop level. The two scatter plots in each of the four graphs represent a monthly chained price index for survey data and a weekly chained price index for scanner data. The results for scanner data should have been converted to month to month price changes in order to make a one-to-one comparison with the price indices for survey data. This was not done in this study, as there are different ways to think about aggregating price changes over time scales. The differences stated in this paper between the results for survey and scanner data should therefore be considered as indicative.

We looked at the price change in week 52 of 2012 with respect to week 4 of 2011 for the price index calculated with scanner data. We compared these results with the price change for the survey measured in December 2012 with respect to January 2011. The results for the three L-Coicops and at Coicop level are as follows:

[6] Content should be used as a product characteristic for food and beverages because they have to be consumed within a limited period of time. But there are reasons for selecting content also for goods with no limits on preservation. For example, should we treat a package with two bottles of shampoo as the same product as two bottles of shampoo that are sold separately, or should we treat these two cases as different products? Buying a single bottle of shampoo gives the consumer freedom of buying a different type of shampoo after consuming the first bottle. The consumer may therefore not want to spend additional money buying a duo-pack for this reason, and possibly also for other reasons. Content should then be selected as an additional characteristic for differentiating drugstore products as well.

- For toiletries, the price change for scanner data is 12.2% and 2.0% for the survey;
- For beauty articles, the price change for scanner data is 5.0% and 0.9% for the survey;
- For other articles for personal care, the price changes are 6.7% for scanner data and 13.0% for the survey;
- At Coicop level, this gives price changes of 7.8% for scanner data and 1.9% for the survey.

The results show a large difference in price change for toiletries. The difference for beauty articles is much smaller, but is still quite large. However, the difference for beauty articles is strongly influenced by the weeks and months selected for measuring price changes. For instance, Figure 1 shows that the difference for beauty articles gets substantially larger in December 2012. The survey data give a price decrease with respect to November 2012, while the scanner data lead to a price increase. But looking at the whole period in Figure 1, the price indices that follow from scanner data are at higher levels than for survey data in almost every month, for both toilet and beauty articles.

The picture for the third L-Coicop (other articles for personal care) is different. The differences between the two price indices are rather small for 2011, but the price index for the survey data ends up at a much higher level in December 2012 than for scanner data. This L-Coicop has the smallest share in turnover (14.8% in 2012). Consequently, the price index for scanner data at Coicop level results in higher levels than for the survey. In the next section, we try to gain more insight into the possible causes behind some of the differences described above.

4. Analyses of the results

One of the questions that will be addressed in this section is why the survey data give rise to smaller price changes compared to the scanner data. This is presumably due to the limitations that follow from sampling brands and articles. We will investigate whether evidence can be found for this expectation from scanner data, which contains information of all articles sold.

The drugstore scanner data contain article information at different levels of detail. The most detailed level is the level of EANs. One level above EAN level we find article description. Next, articles are grouped into six levels of article classification. This makes the data set an ideal case for analysing contributions to price change. We will do this first, with the aim of setting up a context for explaining the differences between the price indices for survey and scanner data.

As a first step, we calculated price changes also by differentiating products at the lowest level of detail in the scanner data, that is, at EAN level. Subsequently, we decompose the differences between the price changes for the ‘EAN-based’ index and the ‘product-based’ index shown in Figure 1. We used the same index method for combining price changes of EANs at L-Coicop and Coicop level. The results are shown in Figure 2. We note the following characteristics:

- The differences between the product-based indices of Figure 1 and the EAN-based indices are large, except for other articles for personal care. The difference for toiletries is 15 percentage points, measured in week 52 of 2012;
- The price index level based on survey data, in December 2012, lies between the two price indices calculated from scanner data, except for other articles for personal care;
- The survey-based price index is closer to the product-based price index for beauty articles, but is closer to the EAN-based price index for toiletries.

As the differences between the product- and EAN-based price indices are small for other articles for personal care, we further analyse the differences for toiletry and beauty article groups with the highest turnover. The differences are decomposed into three components:

- Price change caused by relaunches;
- Price change due to shifts in sales between articles with the same brand and consumer target group, but with different content;

- A residual effect, which refers to price changes that result from shifts in sales between articles with the same brand, target group and also with the same content.

These components are quantified by calculating two additional price indices:

- A variant of the product-based index, where products are characterised by brand name, target group, and also by content;
- A variant of the EAN-based method, where products are not uniquely described by EAN-codes, but according to their article description.

The latter method allows us to detect relaunches, if article descriptions remain the same when EANs are changed.[7]

The decomposition of the differences between the product- and EAN-based price indices into the three aforementioned components is shown in Figure 3 and Figure 4 for toiletry and beauty article groups. The decompositions are shown for one-year periods, for either 2011 or 2012, depending on which year shows the largest difference between the product- and EAN-based index. The contributions of the three components to the one-year price changes are quantified in Table 5. The following observations can be made:

- For toiletries, the largest contribution to price increase comes from relaunches, except for razor blades. Large contributions are also found for shifts in sales between articles with the same brand and target group, but with differences in content;
- For the beauty article groups analysed, it turns out that the contributions of relaunches to price change are small or almost equal to zero. The other two factors dominate, in particular shifts in sales between articles with the same brand, target group and content (see column Residual in Table 5).

These results shed a particular light on the price indices obtained from the survey data. Relaunches play a major role in price changes for toiletry groups. Surveys are restrictive in the selection of brands and articles for price collection, so that the probability of missing price changes due to relaunches could be high, especially when relaunches are limited to a few brands or articles per year. This may be a reason for the result, shown in Figure 2, that the price index based on survey data is closer to the EAN-based price index than to the product-based index.

In Figure 5, the survey-based price indices for shampoo and toothpaste in 2011 are compared with the product- and EAN-based price indices, including the price index that is based on price comparisons at article description level (that captures the effect of relaunches). The survey-based price index follows the product-based index quite well for shampoo, but for toothpaste the survey-based price index moves towards the EAN-based index.

Different shampoo articles of the brand L'Oréal Elvive underwent relaunches towards the end of the third quarter of 2011. This is one of the two shampoo brands contained in the survey (see Table 2), which is a major reason for the good approximation of the product-based price index by the survey-based index.

Figure 7 shows one of the shampoo articles of L'Oréal Elvive that underwent a relaunch in 2011. The 'new' shampoo (on the right) differs from the preceding article only in the shape and other appearance characteristics of the bottle. The content and substances that constitute the shampoo, including the enrichment by three multivitamins, are the same for both articles, but the EANs and the prices are different. The price increase that follows from Figure 7, together with the weight of L'Oréal Elvive in Table 2 compared to the other shampoo in the survey (Andrélon), allow us to explain the price increase for shampoo according to the survey in Figure 5.

We also note that the survey-based price index for shampoo lies between the product-based index and the EAN-based index that captures the effect of relaunches, in the last four months in Figure

[7] In practice, additional work was needed prior to calculating index numbers, as article descriptions do not always remain the same. For instance, words were abbreviated or the word order was changed in article descriptions after relaunches.

5. Two articles of different brands are followed in the survey, so that the survey should not give contributions to price change that arise from shifts in sales between articles with the same brand. We therefore would expect the survey-based price index to be close to the price index based on article description, capturing merely the effect of relaunches on price change. The survey probably attaches too large a weight to L'Oréal Elvive and to relaunches.

The survey-based price index for toothpaste moves towards the EAN-based index (without the effect of relaunches). The survey contains one article of the brand Prodent. The scanner data show that no relaunches took place for this brand in 2011, but the effect of relaunches on the price increase of toothpaste was substantial according to the scanner data (see Figure 3 and Table 5). In addition, although the survey data give a price decrease for Prodent, the scanner data evidence a price increase of about 10% for this brand in 2011. This can be largely explained by shifts in sales towards Prodent articles with a smaller content, which contributes to the price increase according to the product-based index method. Another factor that contributes to the large difference between the survey-based and the product-based index is that other brands underwent larger price increases (Elmex, Aquafresh). In summary, we can say that the survey-based price index misses all the components that contribute to price change for toothpaste as shown in Figure 3 and Table 5.

The above analyses seem to confirm the expectation stated at the beginning of this section. The probability that the survey misses the price effect due to relaunches may be large, because of its practical limitations on the selection of brands and articles. The price increase of shampoo in 2011 is partly captured by the survey, but is missed for toothpaste. The analyses on the four toiletry groups have shown that relaunches give the largest contribution to price increase (except for razor blades). This partly explains why the survey-based price index is closer to the EAN-based index for the L-Coicop toiletries.

The effect of relaunches on the price increase for beauty articles is small. The risk of missing a substantial price change in a survey due to relaunches is therefore small as well. Figure 2 shows that the survey-based price index for beauty articles tends more towards the product-based index. Also for this L-Coicop there are article groups for which the survey-based price index lies closer to the EAN-based index, while for other article groups the survey-based index tends towards the product-based index.

Figure 6 shows two examples with a different behaviour of the survey-based index. Apart from three months, the survey-based price index for day cream is closer to the EAN-based index. The survey contains one specific article (Nivea cream, 150 ml). The effect of relaunches is small, but the main components of price change are obviously missed. Shifts in sales between articles with different content is the largest of three components of price change (see Figure 4 and Table 5), which cannot be captured by the survey for day cream.

The survey-based price index for deodorant moves towards the product-based index in the last quarter of 2011. The survey is not restricted to one specific article, but a choice can be made from a set of variants of the brand Rexona. A survey aims to stick to one variant in successive periods. But if that variant becomes less popular, then another variant may be chosen. The survey for deodorant may therefore capture price changes caused by shifts in sales between articles with the same brand, target group and content. This is the component that explains the largest part of the price change for deodorant in 2011. A similar reasoning applies to other article groups (e.g., hair dye and hairspray).

For different beauty articles, the survey allows switches to other variants during the price collection. Because of this, the main component behind the price change of beauty articles may be captured to some extent, that is, shifts in sales between articles with the same brand and target group, and in some cases also with the same content.

5. Conclusions

In theory, scanner data have clear advantages over surveys for price index calculation, as was argued in Section 2. A major advantage is the availability of full sales information, both with regard to prices and quantities sold, and with respect to the completeness of brands and articles sold. This gives the possibility of calculating price indices more accurately, to study the factors that determine price

change and to compare the results with survey-based price indices. The drugstore scanner data of two Dutch chains have proven to be an excellent test case for addressing these issues.

Price indices were calculated for toiletries, beauty articles and other articles for personal care for 2011 and 2012. The survey-based price changes for toiletries and beauty articles, and at Coicop level, are larger than the EAN-based price changes but smaller than the product-based price changes, with products being characterised by brand and consumer target group. A decomposition of the product-based price changes showed that relaunches are the main driver behind price change for most toiletries analysed, while shifts in sales between articles with the same brand, content and consumer target group determine almost the entire price change for most beauty article groups. As surveys are restrictive in the selection of brands and articles, the components determining price change are partly captured, as was shown in the previous section. This offers an explanation for the result that the survey-based price change over the two-year period lies between the product-based and the EAN-based price changes.

The Dutch survey is not able to capture the price change due to relaunches for all articles. Relaunches play a significant role for toiletries. Although opinions about the level of detail used in product characterisation may differ, there should be no doubt about the requirement that any price index method should take into account the contribution of relaunches to price change. For this reason alone, scanner data should be preferred over surveys, at least when possible data inconsistencies can be ignored.

The objective of this study was to compare price indices calculated with scanner data and survey data. But the study also gives useful information about the choices that have to be made when it is decided to develop an index method for drugstore articles based on scanner data. Decisions need to be made first regarding product characterisation. This study shows that price indices may be very sensitive to the level of detail used in characterising products, which therefore should contribute to the awareness about the importance and sensitivity of product characterisation.

For instance, the decision whether to select package content as a characteristic for discriminating between products or not has a significant impact on the results. As the focus of this study was on comparing price indices for scanner data and survey data, it was decided to omit content as product characteristic for reasons linked to the construction of the survey article list. However, this does not mean that content, and possibly other characteristics as well, should also be excluded when an index method based on drugstore scanner data will be developed for the CPI. As was hinted in footnote 6, content should be considered as an additional product characteristic. How to characterise products and to choose the eventual index formula are topics of further research.

Acknowledgements

The author expresses his gratitude to the supplier of the scanner data, which is not mentioned here for reasons of confidentiality. The author also wants to thank various colleagues at the prices department of Statistics Netherlands for making available the survey data and for helpful comments on the paper.

References

van der Grient, H.A., de Haan, J., 2010. The use of supermarket scanner data in the Dutch CPI. Paper presented at the Joint ECE/ILO Workshop on Scanner Data. Geneva, 10 May 2010. Also available on: www.cbs.nl

de Haan, J., 2006. The re-design of the Dutch CPI. Statistical Journal of the United Nations Economic Commission for Europe 23, 101-118.

de Haan, J., van der Grient, H.A., 2011. Eliminating chain drift in price indexes based on scanner data. Journal of Econometrics 161, 36-46.

Rodriguez, J., Haraldsen, F., 2006. The use of scanner data in the Norwegian CPI: the new index for food and non-alcoholic beverages. Economic Survey 4, 21-28.

Table captions

Table 1. Use of scanner data versus traditional surveys in the Dutch CPI for various store types by 2013, expressed as percentages of the total weight in the CPI.

Table 2. Drugstore articles in the survey for the Dutch CPI and their relative weights for 2012 (sum to 100).

Table 3. Some advantages and disadvantages of survey and scanner data.

Table 4. Article groups selected from the scanner data and their turnover in 2012. The numbers of products and EANs, and the average number of EANs per product, are also given for every article group.

Table 5. Decomposition of the differences between the price changes for the product- and EAN-based index methods into three components for toiletry and beauty article groups (in percentage points).

Tables

Table 1

	Scanner data	Surveys
Supermarkets	13.5	0
DIY stores	0.5	0.9
Travel agencies	1.9	0
Fuel	6.1	0
Other	0	77.0
Total	**22.1**	**77.9**

Table 2

L-Coicop	Article name	Weight (%)
Toiletries	Toothpaste - Prodent	4.8
	Toothbrush - Adults	4.6
	Shampoo - L'Oréal Elvive	2.1
	Shampoo - Andrélon	3.9
	Liquid soap - Palmolive	1.7
	Shower gel - Sanex	1.9
	Shower cream - Dove	3.2
	Razor blades	4.0
Beauty articles	Deodorant - Rexona	5.9
	Bodymilk, self-tanning - Dove	4.2
	Eau de toilette - Ladies	5.3
	Eau de toilette - Men	5.3
	Hair dye - Permanent colouring	4.4
	Lipstick - Maybelline	4.5
	Night cream	8.4
	Aftershave - Nivea	8.3
	Day cream - Nivea	4.3
	Eyeshadow	7.7
	Hairspray - Wella	5.3
Other articles for personal care	Nappies - Pampers	3.3
	Nappies - Own brand	1.9
	Sanitary protection towels	3.3
	Paper handkerchiefs	0.8
	Baby wipes	1.0

Table 3

Survey data		Scanner data	
Pros	*Cons*	*Pros*	*Cons*
Flexibility with assortment changes	Restrictions on number of brands and articles per brand (sample)	Optimal coverage of brands and articles sold	Detection of assortment changes (relaunches) may be time consuming
Direct contact with real world during data collection	No data on quantities sold and on turnover	Full information on sales (prices and quantities)	Continuity of index methods depends on stability of data formats
	Narrow basis for setting weights of articles in price index	Data are available for setting weights of articles in price index	
	Data collection is time consuming	Data are received electronically	
		Contain information for product characterisation	
		Allow sensitivity analyses	

Table 4

L-Coicop	Article group	Turnover 2012		Products	EANs	EANs per product
		In mln euro	*In %*			
Toiletries	Toothpaste	52.3	6.8	39	400	10.3
	Toothbrushes	11.7	1.5	33	90	2.7
	Electronic toothbrushes and top brushes	21.0	2.7	10	90	9.0
	Shampoo	51.3	6.7	72	400	5.6
	Hairstyling (gel, mousse, wax)	29.9	3.9	56	282	5.0
	Shower cream and gel	65.2	8.5	57	463	8.1
	Liquid soap	11.1	1.4	39	108	2.8
	Razor blades	49.9	6.5	33	197	6.0
Beauty articles	Deodorant	71.4	9.3	121	697	5.8
	Self-tanning products	1.0	0.1	5	26	5.2
	Eau de toilette - ladies	22.9	3.0	69	259	3.8
	Eau de toilette - men	15.7	2.0	47	202	4.3
	Hair dye	53.4	6.9	29	463	16.0
	Lipstick	15.7	2.0	19	400	21.1
	Night cream	15.2	2.0	30	129	4.3
	Aftershave	7.1	0.9	23	65	2.8
	Day cream	86.1	11.2	107	463	4.3
	Eyeshadow	6.2	0.8	14	581	41.5
	Mascara	42.5	5.5	36	462	12.8
	Hairspray	25.8	3.4	44	187	4.3
Other articles for personal care	Nappies	67.0	8.7	25	178	7.1
	Sanitary protection towels	13.3	1.7	19	118	6.2
	Paper handkerchiefs	3.6	0.5	9	32	3.6
	Sun protection and aftersun	29.9	3.9	49	273	5.6
Sum		**769.1**	**100**	**985**	**6,565**	
Total turnover of drugstore articles		**1,122.2**				
Turnover covered by selected article groups		**68.5 %**				

Table 5

L-Coicop	Article group	Year	Difference between product- and EAN-based index	Contributions to differences in price change		
				Relaunches	*Content*	*Residual*
Toiletries	Shampoo	2011	12.5	7.3	5.6	-0.4
	Toothpaste	2011	19.1	8.7	4.2	6.2
	Razor blades	2011	12.1	2.0	4.5	5.5
	Shower cream & gel	2012	12.5	4.8	4.8	2.9
Beauty articles	Day cream	2011	9.1	1.0	5.4	2.7
	Deodorant (men)	2011	8.2	1.0	1.0	6.1
	Hair dye	2012	8.1	0.0	0.0	8.0
	Mascara	2012	4.0	0.0	0.0	4.0

Figure captions

Figure 1. Price indices for the three L-Coicops of drugstore articles and for the Coicop between January 2011 and December 2012. Price indices are calculated with scanner data (week 4, 2011 = 100) and survey data (January 2011 = 100).

Figure 2. The price indices in Figure 1 are compared with price indices calculated from the scanner data, in which an EAN is assumed to uniquely characterise a product.

Figure 3. Four price indices for four toiletry groups calculated from the scanner data: the product- and EAN-based indices are shown, along with the indices based on article description ("EAN-based with relaunches") and with products characterised by brand, target group and content ("product-based incl. content").

Figure 4. The same price indices as in Figure 3 are shown, but here for four beauty article groups.

Figure 5. Product- and EAN-based price indices for shampoo and toothpaste in 2011 (week 4 = 100), compared with the survey-based price indices (January 2011 = 100).

Figure 6. Product- and EAN-based price indices for day cream and deodorant in 2011 (week 4 = 100), compared with the survey-based price indices (January 2011 = 100).

Figure 7. Example of a relaunch for a shampoo article of the brand L'Oréal Elvive in 2011.

Figure 1

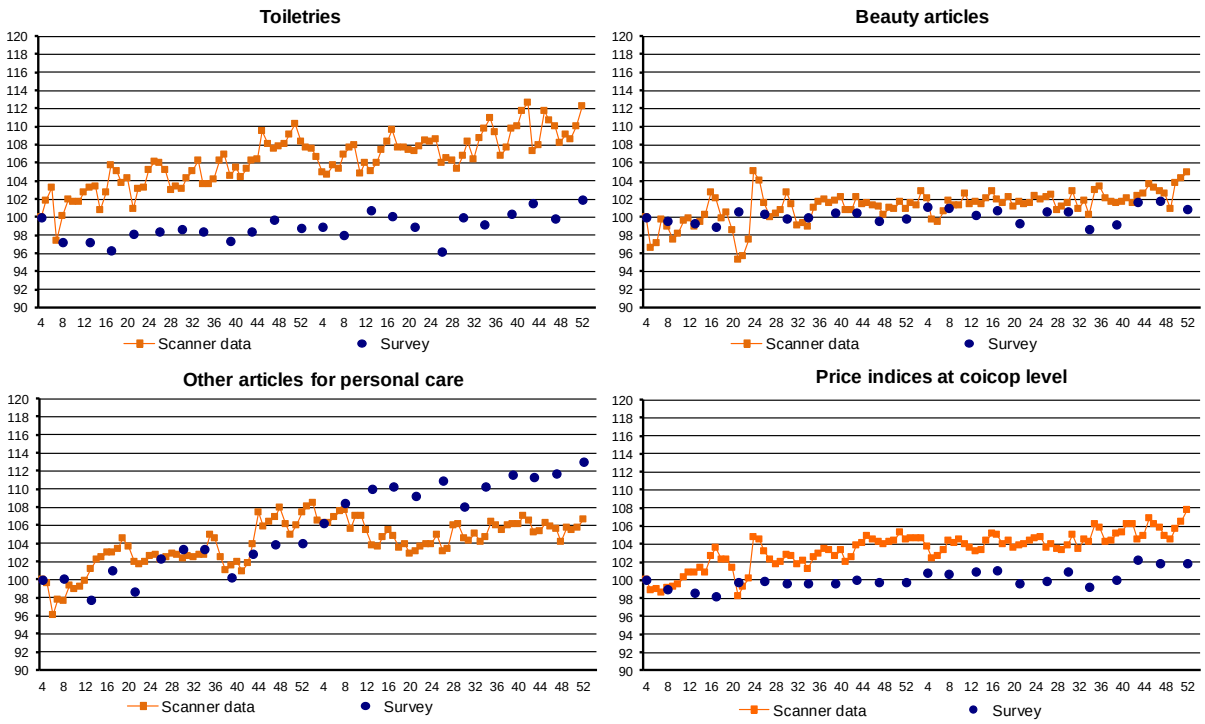

Figure 2

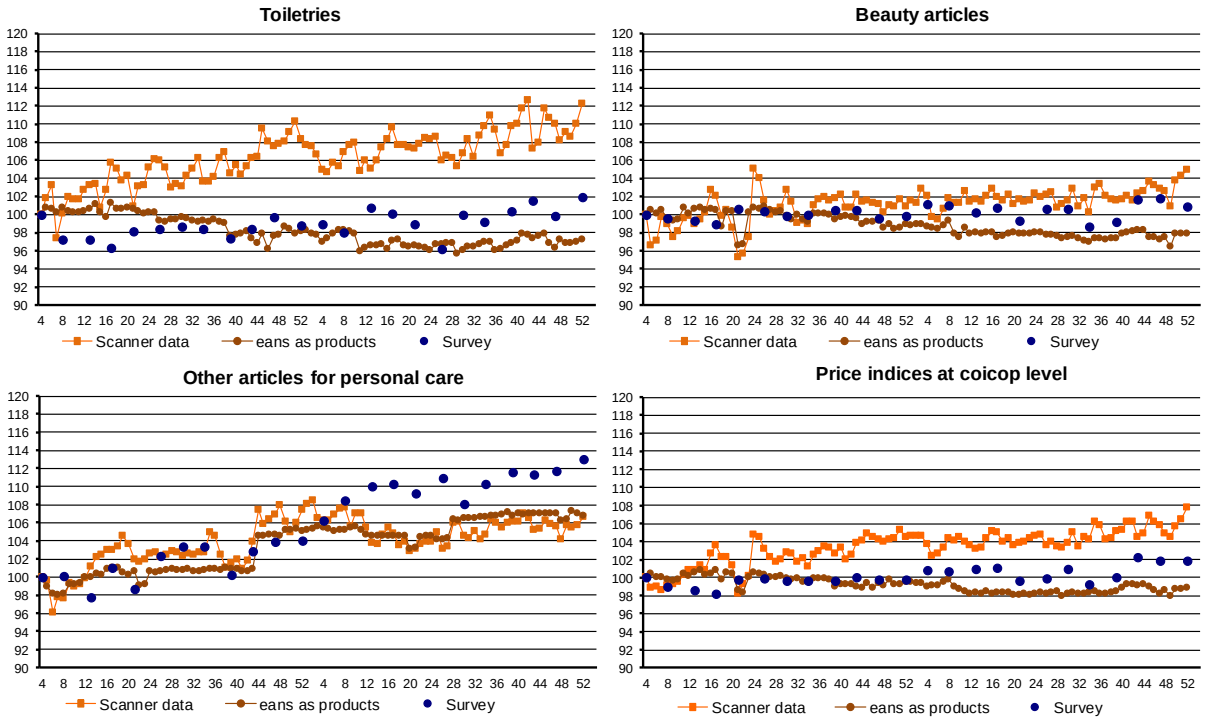

Figure 3

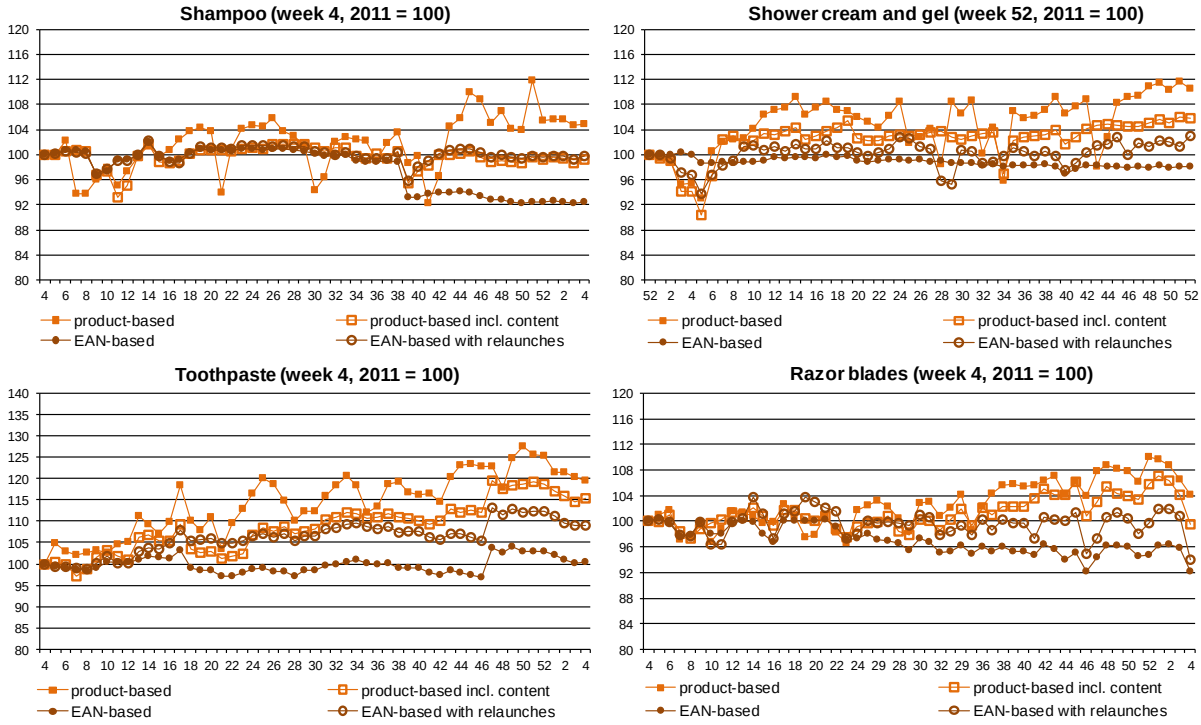

Figure 4

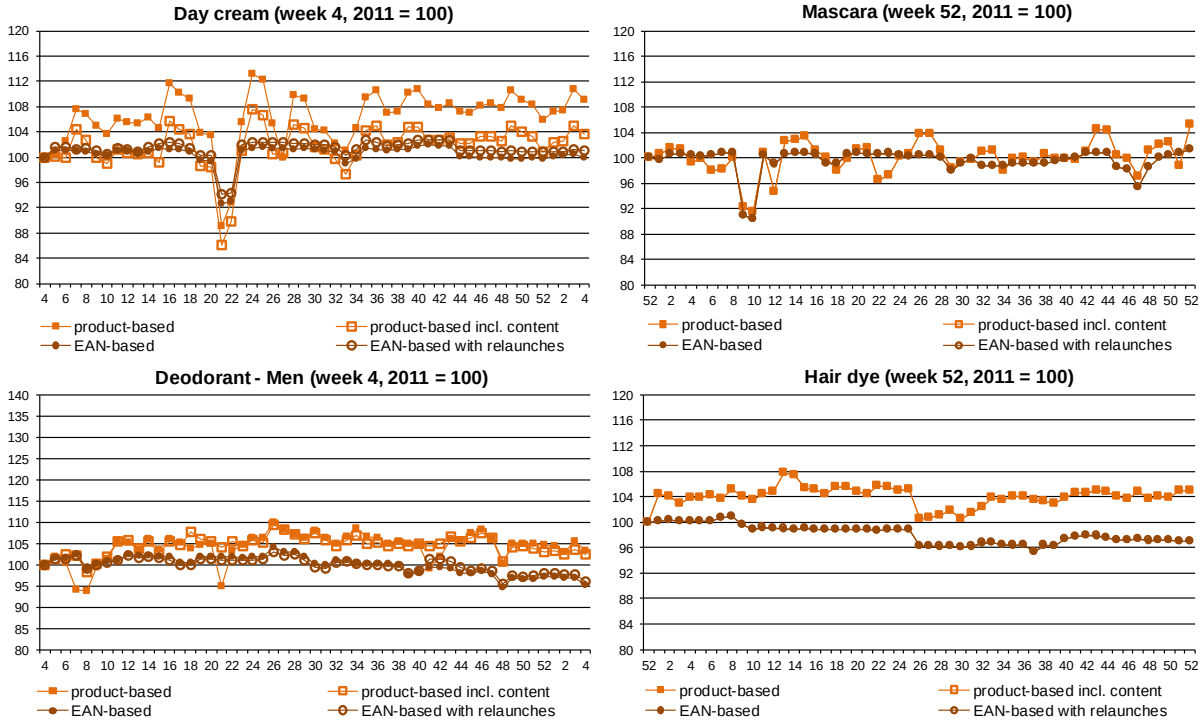

Figure 5

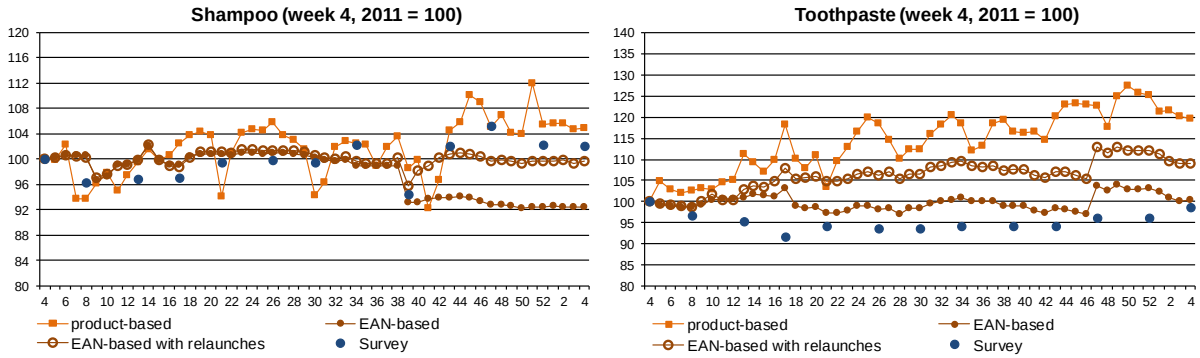

Figure 6

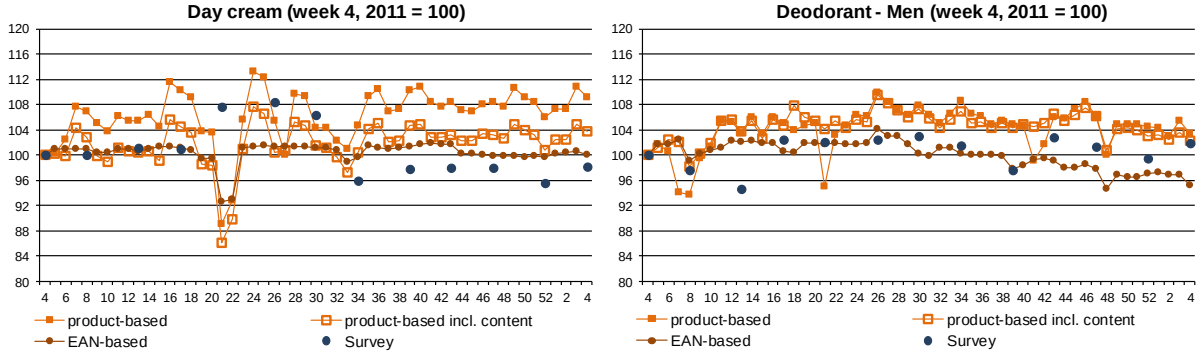

Figure 7

<u>EAN: 36-00521-74076-7</u>
Elvive shampoo 2-in-1 multivitamine
Content: 250 ML
Price week 38: **€ 3,18**
Price week 39: **€ 2,00**

<u>EAN: 36-00522-00499-8</u>
Elvive shampoo 2-in-1 multivitamine
Content: 250 ML
In week 39 sold for first time
Price week 39: **€ 3,98**