Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

The QU-method: A new methodology for processing scanner data

Antonio G. Chessa[1]

Abstract

This paper presents a new price index method for processing electronic transaction (scanner) data. Price indices are calculated as a ratio of a turnover index and a weighted quantity index. Product weights of quantities sold are computed from the deflated prices of each month in the current publication year. New products can be timely incorporated without price imputations, so that all transactions can be processed. Product weights are monthly updated and are used to calculate direct indices with respect to a fixed base month. Price indices are free of chain drift by this construction. The results are robust under departures from the methodological choices. The method is part of the Dutch CPI since January 2016, when it was first applied to mobile phones.

Key Words: Scanner data, GTIN, relaunch, product homogeneity, price index, chain drift.

1. Introduction

Scanner data sets contain both turnover and numbers of items sold for every individual item, which is uniquely identified by its barcode (GTIN, Global Trade Item Number). Besides sales information, scanner data should also contain descriptive information for characterising items. Countries that use scanner data in their CPI usually request retailers to specify transactions by week.

Scanner data have important advantages compared to data collected from traditional surveys. Average transaction prices per GTIN can be calculated, while turnover data can be used to compute expenditure shares at GTIN level. Survey data merely contain shelf prices, which are not necessarily equal to the prices eventually paid by consumers. Information on numbers of items sold and turnover is not collected in traditional surveys. Scanner data offer possibilities of calculating more accurate indices, with expenditure based weighting at GTIN level. Another obvious advantage of scanner data is that the transactions of all items sold are available. As a consequence, scanner data make it possible to replace sample-based methods by methods that allow integral data processing. In spite of their potential, scanner data are still used by a small number of statistical agencies in their CPI, but the number is likely to increase in the coming years[2].

At the time of introduction into the Dutch CPI in 2002, scanner data involved two supermarket chains. At present, scanner data of 10 supermarket chains are used and surveys are not carried out anymore for supermarkets since January 2013. Beside supermarkets, scanner data and other electronic transaction data are used in the CPI, amongst others for do-it-yourself stores, drugstores and mobile phones. More than 20% of the Dutch CPI is now based on electronic transaction data (in terms of Coicop weights of 2015).

The index methods that make use of scanner data in the Dutch CPI are different across retailers and consumer goods. For instance, the method for supermarkets is a monthly chained Jevons index for elementary aggregates. The methods for other retailers use Laspeyres type indices for samples of items that cover a certain turnover share. The methods used for different retailers have evolved from different historical perspectives.

[1]Statistics Netherlands, CPI; Henri Faasdreef 312, 2492 JP The Hague, the Netherlands (ag.chessa@cbs.nl). The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

[2] In Europe, six countries are using scanner data since January 2016 (Belgium, Denmark, the Netherlands, Norway, Sweden and Switzerland). Iceland will start using scanner data in the CPI in April 2016. The scanner data workshops in Vienna (2014) and Rome (2015) evidenced that several countries are expecting their first data, while other countries made concrete steps towards acquiring their first scanner data.

The current methods have a number of issues. The method for supermarkets makes use of equal weights at GTIN level (a turnover threshold is used in order to exclude items with turnover shares below the threshold). Follow-up items are not matched to their predecessor when the GTIN changes. This may occur with "relaunches", a term used for items that return to the stores after undergoing modifications, which often concern the packaging. Package content (volume) and ingredients often remain the same. If the repositioned, follow-up item is accompanied by a higher price, the price increase will be missed when the old and new GTIN are not matched. Dump prices of outgoing items are removed in the method for supermarkets in order to reduce a possible downward bias of the price index (see de Haan and van der Grient (2011) for details). Methods for other retailers in the Dutch CPI do match GTINs under relaunches, but these methods make use of samples of items. Some of these methods perform manual replacements of items, which may be a time consuming activity.

The above-mentioned issues motivated a search towards a generic methodology that can be applied to electronic transaction data of different retailers and consumer goods. Section 2 draws the attention on one of the essential problems in price index calculation: the problem of defining homogeneous products. The index method is described in Section 3. Section 4 summarises some results of a comparative study, which quantifies the impact of departures from the choices made in the index method on price indices. Section 5 concludes with a summary of first experiences in the CPI and future plans with the new methodology.

2. Product classifications and homogeneity

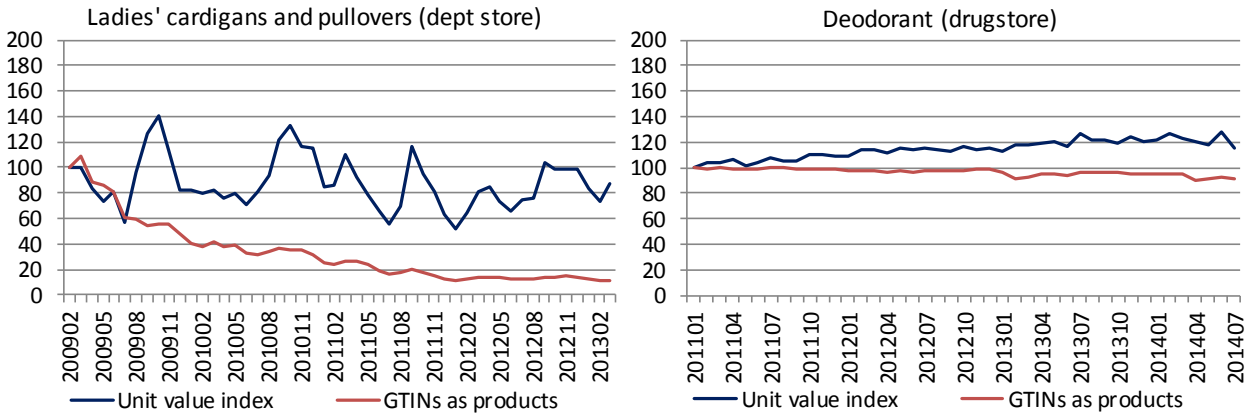
Price index calculations in the CPI follow a series of aggregation steps that start with the calculation of average prices at the most detailed product level as defined by statistical agencies. Product prices are subsequently used to calculate price indices at a first aggregate level, which are subsequently combined to price indices at higher levels of a nested classification of product groups. The highest level of this classification is the well-known Coicop system. The most detailed level of publication of CPI figures is referred to as "L-Coicop" in the Dutch CPI[3].

Scanner data contain transaction data specified by GTIN. Scanner data sets may contain very large numbers of GTINs, even more than 100,000 for a single retailer. Such large numbers require an efficient procedure for assigning GTINs to L-Coicops. This can be achieved by making use of the retailers' own classification of items (called "ESBAs" in the Dutch CPI). ESBAs should therefore be part of a scanner data set. Usually, we take the most detailed ESBA level for establishing the GTIN-Coicop links. However, the most detailed ESBAs may still cover more than one L-Coicop. A level below L-Coicop is therefore created, which we call "consumption segments". In the Dutch CPI, consumption segments usually represent classes of similar items, such as men's T-shirts, women's socks, mobile phones or chocolate.

Consumption segments may still be too heterogeneous in order to be considered as individual products for which meaningful prices can be calculated. A segment 'socks' may consist of different types of socks (walking, thermal, sport socks) and items may contain different numbers of pairs of socks. GTINs represent the most detailed level of item differentiation but may suffer from relaunches. The possible impact of relaunches is illustrated in Figure 2-1, especially in the leftmost graph. The price index calculated with GTINs as unique products drops to almost zero after three years. The assortment for cardigans and pullovers, in this case of a Dutch department store, is almost renewed each year. Items enter the store at high introduction prices, which subsequently decrease in the course of a year.

[3] L-Coicops are specified at the fifth digit level at most (depending on Coicop division).

Figure 2-1
Unit value indices and price indices with GTINs as products for two consumption segments. The x-axis represents year and month (yyyymm). Price indices (vertical axis) are set at 100 in the first month.



The result is a cascade of yearly declining indices, when old and new GTINs are not linked.

Figure 2-1 also shows considerable differences between the price indices based on GTINs as products and the unit value indices. Given that such large differences between the two levels of product differentiation may occur, an essential question in the whole exercise of compiling price indices is how to define products. How could an intermediate level be created between consumption segments and GTINs, such that GTINs combined at the intermediate level can be considered to be homogeneous? The GTIN groups at intermediate level will be referred to as homogeneous products or simply "products".

The relaunch problem plays a crucial part in selecting an approach for defining products. The following possibilities can be identified:

1. If relaunches do not occur, then GTINs would be a natural choice for homogeneous products.
2. If relaunches do occur, then a broader level of product differentiation is needed at which GTINs are combined. The following possibilities can be thought of for this purpose:
    a. old and new GTINs could be matched through the retailer's internal product codes or SKUs (Stock Keeping Units). Retailers usually assign the same SKU to follow-up items as for items that leave the assortment;[4]
    b. if SKUs are not available, or cannot be used for some reason, then different GTINs could be matched through a set of product characteristics.

An obvious question is how to verify whether relaunches occur or not, or to an extent that price indices are hardly affected by relaunches. It is important to calculate basic statistics such as lifetimes of GTINs, the expenditure shares for GTINs that form the stable part of an assortment and also for outgoing GTINs. Such statistics can be used to make first assessments about the possible impact of hidden price increases on a price index due to relaunches.

The availability of SKUs resolves important problems: price increases at relaunches will be captured, while techniques for identifying product characteristics from product information become superfluous. The latter is not an issue when characteristics are contained in separate fields in a scanner data set. However, if product characteristics are contained in text strings of item descriptions, then text mining techniques could be used in order to identify characteristics. Search terms need to be stored in lists, which have to be monitored each month.

If SKUs are not available, then GTINs can be matched by product characteristics. GTINs are matched and combined into the same product when they share the same characteristics. The question is which characteristics should be selected. Chessa (2016) suggests a sensitivity analysis, which quantifies the impact of adding characteristics on a price

[4] Statistics Netherlands expects to receive an extended scanner data set from a do-it-yourself store chain, to which SKUs will be added. Moreover, the data will also include the exact dates when the old item will be replaced by the follow-up item.

index. Such an analysis could be used to set up minimum sets of characteristics, which could also be helpful to restrict the time spent during monthly maintenance of these lists.

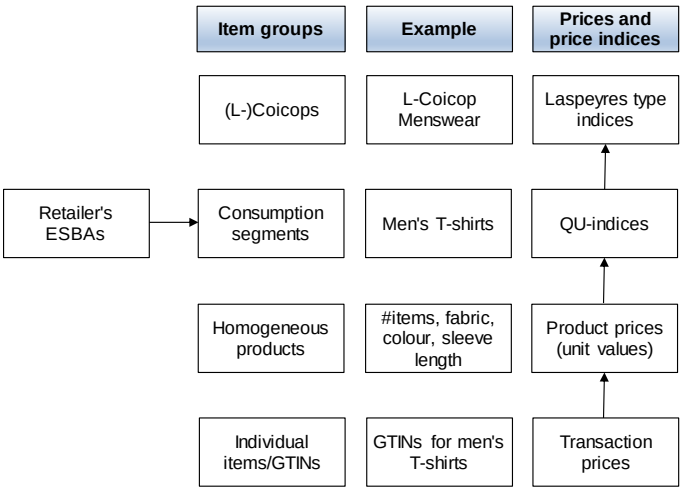
3. The index method

The index method described in this section is intended to be applied to consumption segments. Price indices calculated for consumption segments will be aggregated to Coicop levels according to conventional Laspeyres type methods.[5] Figure 3-1 synthesises the price index methods and price notions used at different item group levels within the CPI.

The choice of index method at consumption segment level departs from the simplified case where products within a consumption segment are equivalent. In this case, we can calculate unit values for consumption segments, that is, total turnover divided by the sum of the quantities sold over all products within segments (see CPI-manual (2004, p. xxii)). The price index then equals a unit value index, which can be written as a ratio of a turnover index divided by an unweighted quantity index.

When items in a consumption segment are not homogeneous, the segment must be differentiated into homogeneous products. The unit value index cannot be applied in that case and must be refined as well. The refinement merely applies to the quantity index (of course, turnover remains unaffected). We preserve the additive form of the quantity measures, but the quantities sold $q_{i,t}$ of product *i* in every month *t* are now weighted by a factor v_i.

Figure 3-1
The new methodology ("QU") within the CPI.



The price index P_t in month *t* with respect to some base or reference month 0 takes the following form:

$$P_t = \frac{\sum_{i \in G_t} p_{i,t}\, q_{i,t} / \sum_{i \in G_0} p_{i,0}\, q_{i,0}}{\sum_{i \in G_t} v_i\, q_{i,t} / \sum_{i \in G_0} v_i\, q_{i,0}}, \qquad (1)$$

where $p_{i,t}$ denotes the price (unit value) of product *i* in month *t* and G_t denotes the set of products sold in month *t* for a consumption segment G. The sets G_t and G_0 in month *t* and the reference month 0 may be different.

Formula (1) can be rewritten by dividing the numerators and denominators of the turnover index and the weighted quantity index by the sums of the quantities sold over all products in months *t* and 0 respectively, so that:

[5] It is also possible to extend the aggregation method used within consumption segments to higher levels. This would give rise to a more consistent method. However, we leave this alternative aggregation method as a subject of further research.

$$P_t = \frac{\bar{p}_t/\bar{p}_0}{\bar{v}_t/\bar{v}_0}. \quad (2)$$

The numerator of (2) is equal to the unit value index. This is the price index under homogeneity of the consumption segment G. Otherwise the unit value index has to be adjusted, and the adjustment term is given by the denominator of (2). The latter is a ratio of average values of the v_i in months *t* and 0, which are weighted by the shares of the quantities sold of each product. The ratio $\bar{v}_t/\bar{v}_0$ can be considered as a quality index. Its values are determined by two factors: the relative values of the v_i of the products and the shares of the quantities sold of each product in two months. For instance, a shift towards products with higher v_i (higher quality) in month *t* results in a quality index greater than 1, which leads to a downward adjustment of the unit value index. Because of this property of index formula (2), we refer to it as a "quality-adjusted unit value index", which is shortened to "QU-index".

Expression (1) can be considered as a family of index methods. Different choices for the v_i lead to different index formulas. For example, if we set the v_i equal to the product prices in the reference month, then (1) becomes a Paasche index; if the v_i are set equal to the prices in month *t*, then (1) turns into a Laspeyres index; if we take the average of the prices from months 0 and *t*, then the Fisher index results as a special case (see also Chessa (2016, p. 18)).

The three examples are all bilateral indices. Monthly chained indices are not capable of dealing with price changes for products that are not available temporarily, while bilateral direct indices cannot process new products timely, unless some form of price imputation is carried out. However, price imputations are not needed when the product prices from each month in some time interval *T* are used to construct the v_i. This allows us to set up an index method according to which all transactions can be processed. The dynamics of an assortment, where items may disappear, new ones enter, while other items persist, can then be fully reflected in the price and quantity index behaviour over time.

Since the v_i are part of the quantity index, price changes need to be removed from the v_i. The product prices in the v_i are therefore deflated with the price index itself. Obviously, numerous ways can be thought of to combine the deflated prices from multiple months into the product weights v_i. In this paper, we consider weighted arithmetic means:

$$v_i = \sum_{z \in T} \varphi_{i,z} \frac{p_{i,z}}{P_z}, \quad (3)$$

where the weights $\varphi_{i,z}$ are non-negative and sum to 1 over all $z \in T$.

Different questions need to be answered before price indices can be computed:

1. How should the weights $\varphi_{i,z}$ be chosen?
2. How should the interval *T* be chosen?
3. How should the v_i be updated each month?

Our base choice for the weights $\varphi_{i,z}$ is the share of the quantities sold in month *z* in the sum of the quantities over an entire period *T*, that is:

$$\varphi_{i,z} = \frac{q_{i,z}}{\sum_{s \in T} q_{i,s}}. \quad (4)$$

If we replace time by country, then expressions (1)-(4) are equivalent to the Geary-Khamis method known in international price comparisons (Khamis, 1972). This method has been the subject of some debate in international price comparisons, concerning a phenomenon known as "substitution bias" (Balk, 2001). However, we expect that this phenomenon will be hardly an issue in intertemporal comparisons. Consumers tend to buy larger quantities at lower prices, for instance, at discounts. Expression (4) assigns larger weights to the corresponding prices in such cases. In Section 4, alternative choices for the weights $\varphi_{i,z}$ will be considered in order to quantify the impact of different choices on the results.

The time interval *T* represents the current year of publication, which covers a period of 13 months that also includes December of the previous year as base month. This choice implies that the v_i are constant during a year, at least in theory, and are allowed to vary from year to year. The choice of the current year has to do with the aim of including

new products as well into the index calculations. A length of one year and December as base month are in agreement with CPI practice.

The use of product prices and quantities from the current year introduces an additional problem, as the prices and quantities of 13 months are only known at the end of a year. The v_i must therefore be updated with prices and quantities of the next month. The use of a rolling 13-month window has given volatile and apparently biased results (Chessa, 2015). Instead, we use a time window that is enlarged each month with respect to an annually fixed base month and update the v_i according to expressions (3) and (4), with T replaced by the interval $[0, t]$, where month 0 denotes the base month and month t the current month. The monthly updated v_i are used to compute a price index for month t with respect to the base month. The price indices in the final month of a year are based on v_i in which the product prices and quantities of all 13 months are used. If these v_i were used in every month of a year, which is impossible in practice, then the resulting indices would be transitive. As a consequence, the "real time indices" with monthly updated v_i are free of chain drift by construction. An interesting question is how the real time and transitive "benchmark indices" compare throughout a year. This will be considered in Section 4 as well.

As the price index appears in expression (3) for the v_i, the question is how price indices can be computed. A method that is simple to implement makes use of an iterative scheme, which starts with arbitrary initial values for the price indices in each month between the base month and the current month. The initial values are then used to calculate values for the v_i according to expression (3), which are subsequently used to recalculate the price indices according to formula (1). This procedure converges to a unique solution, when it exists,[6] and is repeated until the differences between the prices index values of the last two iteration steps are "sufficiently small" (a stop criterion has to be set by the user). It is important to note that price indices of previous months are recalculated, but these are not used to revise price indices that are already published.

Instead of using arbitrary initial values for the price indices, initial values can be constructed by using a two-stage method that leaves out prices and quantities of the current month at the first stage. The resulting 'first-stage' price index is then used to deflate the current month prices and quantities, which are used to update the v_i and the price index. This method has given very accurate results (Chessa, 2015), which allows to reduce computation times. In contrast with the complete iterative method, price indices of previous months are not recalculated with the initial, two-stage method.

4. Impact of different choices on price indices

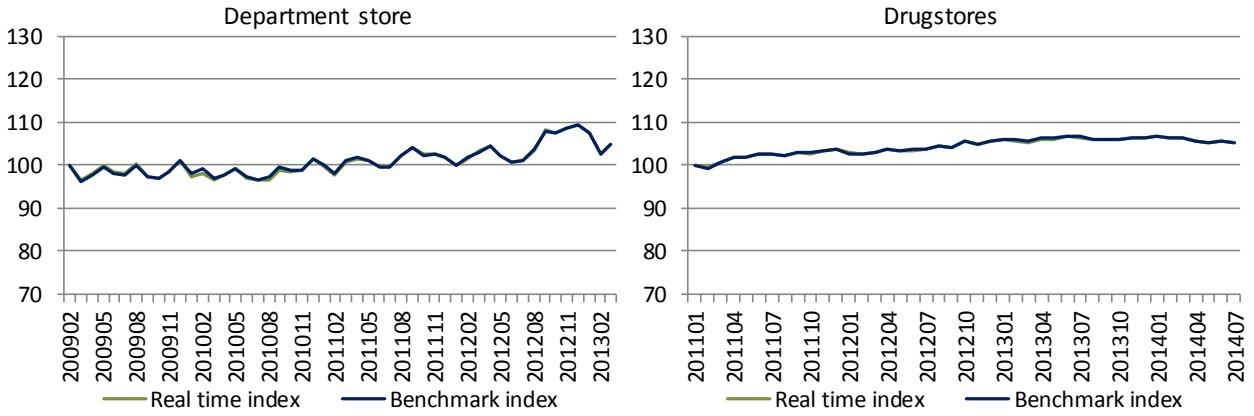
This section compares different methodological choices, as already referred to in the previous section, and quantifies their impact on the price indices. For this purpose, we have used scanner data sets of a Dutch department store and a drugstore chain. Historical data have been used for the index calculations, which cover periods of 4 years (department store) and 3 years and 7 months (drugstores).

Both scanner data sets contain weekly prices and quantities sold at GTIN level for different product groups. The Dutch CPI uses the first three full weeks of a month to compute monthly average product prices (turnover divided by quantities sold over the three weeks). The data sets also contain characteristics of each item. SKU's are not available in the two data sets. For the department store, products are defined according to a set of characteristics for clothing (type of item, fabric, colour, number of items in a package), while GTINs were taken as products for the rest of the assortment. GTINs were not used to characterise drugstore products because of the extent of relaunches in the assortment. Characteristics were therefore used to differentiate products for the entire assortment.

Below, we compare price indices on different methodological aspects. Given the short length of this paper, it was decided to show price indices at aggregate level for both the department store and the drugstore chain. Figure 4-1 compares the real time indices with the transitive benchmark indices for the department store and the drugstore chain. The differences are negligible, and are small or negligible also at lower aggregate levels (Chessa, 2016). The results show that the real time indices practically behave as if these were transitive indices.

[6] Translated into CPI practice, this boils down to checking each publication month whether a product exists that has been sold both in the current month and in one of the previous months. If this is not the case, then the price index of the consumption segment will be imputed in the publication month (e.g., from the corresponding L-Coicop).

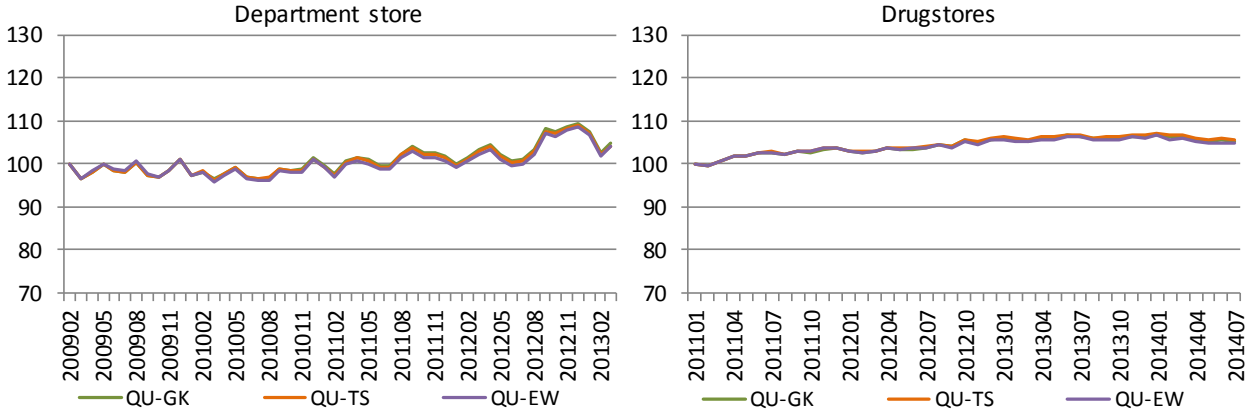
Figure 4-1
Real time and benchmark indices for a Dutch department store (Feb. 2009 = 100) and a drugstore chain (Jan. 2011 = 100) at overall level.



In Figure 4-2, the real time indices are compared with QU-indices that are calculated with two alternative weighting schemes for the deflated prices in the v_i. One scheme uses equal weights (QU-EW) for the $\varphi_{i,z}$ in expression (3). Prices are left out when quantities decrease by more than 90%. The second scheme makes use of turnover shares (QU-TS) instead of quantities sold as in (4). These two variants result in price indices that hardly differ from the indices in Figure 4-1 (which we refer to as QU-GK). The results thus appear to be robust under different weighting schemes in the v_i. This observation also holds at lower aggregate levels (Chessa, 2016).

The QU-GK method has also been compared to the so-called "time product dummy method" (de Haan and Krsinich, 2014). Like the QU-GK method, the time product dummy method (TPD-method for short) is an adaptation of a method for international price comparisons ("country product dummy method") to intertemporal comparisons. Price indices according to the TPD-method can be expressed as a ratio of weighted geometric means of "quality-adjusted prices" $p_{i,t}/v_i$ and $p_{i,0}/v_i$ for two months t and 0, which are weighted by the turnover shares of each product in the two months. The QU-GK method can be written in a similar fashion, which results in a ratio of weighted harmonic means of quality-adjusted prices in two periods, which are weighted by the turnover shares of each product as well.

Figure 4-2
Real time indices of Figure 4-1 (QU-GK) compared to the QU-TS and QU-EW variants of the QU-method.



The QU-GK and TPD-indices are shown in Figure 4-3, as overall price indices for the Dutch department store and drugstores. The price indices for the two methods hardly differ, which also applies for most Coicops and underlying consumption segments. Figure 4-4 compares the price indices according to our choices made concerning product differentiation for the department store and the drugstores with two extremes of product differentiation. On the one

hand, consumption segments are taken as homogeneous products, for which unit values are calculated. The other extreme takes every GTIN as a distinct product.

Figure 4-3
The QU-GK method compared with the TPD-method.

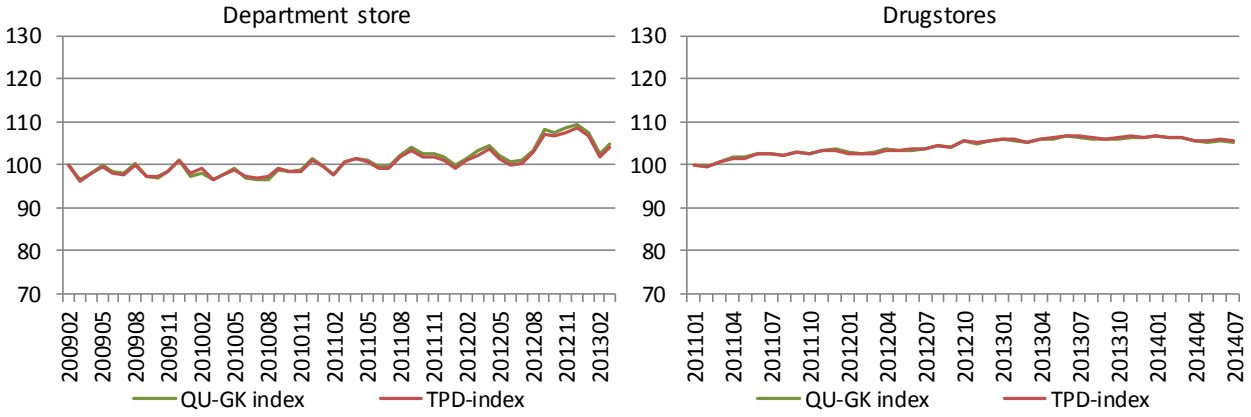


Figure 4-4
Price indices compared with two extreme levels of product differentiation.

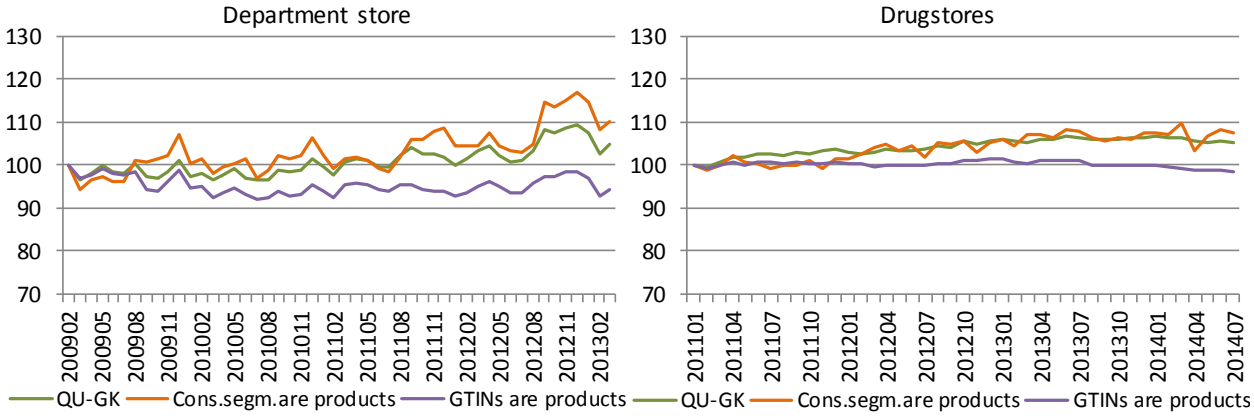


Figure 4-4 shows considerable differences between the three indices. The price indices with GTINs as products deviate most from the QU-GK indices. Taking GTINs as products yields indices that lie below the QU-GK indices for both retailers, as expected, since price increases after relaunches are not taken into account. The unit value based index lies above the QU-GK index for the department store. This implies that the denominator of expression (2) has increased, which means that consumption patterns have shifted towards higher quality products over time.

5. First CPI experiences and final remarks

A new methodology for processing electronic transaction data sets and calculating price indices has been proposed in this paper. The approach has several advantages. It enables processing of all transactions, with a timely inclusion of new products in the course of a year. As a consequence, the index method does not require price imputations. Artificial filters for dump prices and small turnover shares are not needed. The index method makes use of prices and quantities sold to compute product weights v_i and old and new GTINs are linked, either by item characteristics or by SKUs.

The results in Section 4 have shown that the price indices are remarkably robust under different choices for the weights in the v_i and the form of the price index formula. The real time index, with monthly updated v_i, hardly differs from a transitive benchmark index. The proposed method is free of chain drift and the use of price and quantity information from a small number of months in the beginning of each year is not an issue.

The comparative study in Section 4 points out that the problem of defining homogeneous products is by far the most sensitive aspect and therefore deserves particular attention. The results show how important it is that scanner data sets contain information about relevant product characteristics (preferably in separate columns in order to limit time consuming text mining techniques). Statistical agencies should ask retailers whether they can supply their own internal product codes. These codes may provide an efficient way of linking outgoing GTINs to follow-up items, thus capturing price increases after relaunches without using product characteristics (Section 2).

The new methodology is now part of the Dutch CPI. In January 2016, the new approach was applied for the first time to mobile phones. Monthly production is proceeding efficiently: it takes 30 to 45 minutes to prepare the input data for the index calculations (mainly to collect characteristics of devices that are missing in the data). The new approach has resulted in a huge time saving, compared to the 2 or 3 days required by previous methods.

Scanner data of the department store will be processed with the QU-method in the CPI within a few months. The drugstore data and data of do-it-yourself stores will follow at a later stage. In the course of 2016, a research project for supermarket scanner data will be started that is aimed at comparing the QU-method with the method currently used for supermarkets in the Dutch CPI.

References

Balk, B.M. (2001), "Aggregation Methods in International Comparisons: What Have We Learned?", paper originally prepared for the Joint World Bank - OECD Seminar on Purchasing Power Parities, 30 January - 2 February 2001, Washington DC.

Chessa, A.G. (2015), "Towards a Generic Price Index Method for Scanner Data in the Dutch CPI", Ottawa Group Meeting, 20-22 May 2015, Urayasu City, Japan.

Chessa, A.G. (2016), "Processing Scanner Data in the Dutch CPI: A New Methodology and First Experiences", paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland. (To be published in *EURONA*, June 2016.)

CPI-manual (2004), *Consumer price index manual: Theory and practice*, Geneva, Switzerland: ILO/IMF/OECD/UNECE/Eurostat/The World Bank.

de Haan, J., and H.A. van der Grient (2011), "Eliminating Chain Drift in Price Indexes Based on Scanner Data", *Journal of Econometrics*, 161, pp. 36-46.

de Haan, J., and F. Krsinich (2014), "Time Dummy Hedonic and Quality-Adjusted Unit Value Indexes: Do They Really Differ?", paper presented at the Society for Economic Measurement Conference, 18-20 August 2014, Chicago, U.S.

Khamis, S. H. (1972), "A New System of Index Numbers for National and International Purposes", *Journal of the Royal Statistical Society A*, 135, pp. 96-121.