A Comparison of Price Index Methods for Scanner Data

Antonio G. Chessa[1], Johan Verburg[2] and Leon Willenborg[3]

1 May 2017

Abstract

Scanner data offer new opportunities and challenges for price index calculation compared to traditional price collection, for instance in the use of transaction data for constructing product weights within elementary aggregates. This raises the question how Jevons indices, which are still used for supermarket scanner data, compare with more sophisticated methods. This paper compares weighted and unweighted bilateral and multilateral methods on department store scanner data of different product groups. The results show that: (1) applying equal weights to products within elementary aggregates may lead to considerable differences compared to weighted methods, (2) bilateral methods either do not capture the full population dynamics expressed by scanner data or may suffer from chain drift, and (3) differences among multilateral methods are smaller but cannot be ignored. Differences between hedonic and other multilateral methods can be ascribed to the lack of interaction terms between item attributes in traditional hedonic models. Inclusion of pairwise interactions improves model fits and gives results that are practically the same as for the Geary-Khamis (GK) and time product dummy (TPD) method. We show that the GEKS has a number of issues. One of these is a downward bias in cases with dump prices for disappearing products; GK and TPD indices are insensitive to such prices. Differences between window updating methods for including information from a new month are rather small in this study, but the use of rolling windows combined with splice methods has shown signs of drift. Methods that calculate direct indices with the most recently updated set of parameter values are free of chain drift. An extensive follow-up study is being prepared for further analyses and to verify the findings with scanner data of different retailers. More attention will then be given to the problem of product definition and the impact of different degrees of product differentiation on price indices.

Keywords: Scanner data, dynamic universe, bilateral methods, Geary-Khamis method, Time Product Dummy method, GEKS, hedonic models, transitivity, updating problem.

JEL Classification: C43, E31.

[1] Division of Economic and Business Statistics and National Accounts, CPI department, Statistics Netherlands, email: ag.chessa@cbs.nl.

[2] Division of Economic and Business Statistics and National Accounts, CPI department, Statistics Netherlands, email: j.verburg@cbs.nl.

[3] Division of Corporate Services, IT and Methodology, Statistics Netherlands, email: lrjc.willenborg@cbs.nl.

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

1. Introduction

The popularity and availability of electronic sales data for the compilation of the Consumer Price Index (CPI) has increased over the past 10-15 years. Switching from traditional surveys to electronic data delivery reduces administrative burden for both statistical agencies and retailers, and the availability of expenditure data in scanner data sets can be exploited to increase the accuracy of CPI figures. A transition from traditional survey data to scanner data is a lengthy process, which covers different stages from establishing the first contact with a retail chain until testing a method for calculating price indices in a CPI production environment. The awareness among statistical agencies about the issues and potential problems that may be encountered during this process is growing, which is aided by the recent issuing of Eurostat guidelines for scanner data acquisition and processing and by yearly scanner data workshops. The number of countries that make use of scanner data in their CPI is gradually growing, which in Europe has increased to seven countries in 2016 (Belgium, Denmark, Iceland, the Netherlands, Norway, Sweden and Switzerland).

Scanner data have clear advantages over survey data. In traditional surveys, prices are collected for relatively small samples of goods, while scanner data contain both prices and quantities sold for every item that has been purchased by consumers. The differences in types and amount of information between survey and scanner data pose new challenges to statistical agencies. One of these challenges is the choice of index method, which may be reconsidered when switching to scanner data because of different reasons.

Scanner data contain expenditure information at the item level (i.e., at the barcode or GTIN level), which makes it possible to use expenditure shares of items as weights for calculating price indices at the lowest (also called "elementary") aggregate level. In traditional surveys, prices are collected each month for the same 'basket of goods'. Scanner data contain information of all items sold, so that these data sets reveal the actual dynamics of the entire population of sold items. Items that are sold throughout a year may only be a subset of the entire population. Items may also disappear, forever or temporarily (e.g., for seasonal goods). Disappearing items may be replaced by new ones, while totally new items may also enter an assortment in the course of a year.

The additional sources of information and the subsequent challenges offered by scanner data legitimate the question to what extent price indices will be affected when switching from survey to scanner data, or, put more directly, from unweighted fixed-basket methods with a relatively small number of items, and infrequently observed, to weighted 'dynamic basket' index methods, which are based on frequently observed large populations. Most statistical agencies that are using scanner data in their CPI still make use of a monthly chained Jevons index, which makes the above question even more relevant.

A previous paper provides an extensive overview of methods for calculating price indices based on scanner data (de Haan et al., 2016). In the present study, the majority of these methods is applied and compared using scanner data from a Dutch department store. Price indices for bilateral and multilateral methods were already compared in other studies (Chessa, 2016a, 2017a). These studies compare a relatively small number of methods on a large number of scanner data sets. In the present study, we extend the number of methods, which will be compared on a small number of scanner data sets.

This paper is organised as follows. The scanner data sets that are used for comparing price indices for different methods are described in Section 2. A characterisation of the data will be given in terms of the dynamics of the assortments, such as the numbers of new and disappearing items in time and the number of items that are sold each month.

Sections 3.1 and 3.2 give a summary of bilateral and multilateral index methods, which are applied to the scanner data sets in Section 4. Unweighted and weighted bilateral methods have been applied; in this paper we focus on the Jevons and Törnqvist index. The multilateral methods encompass the hedonic method, and methods that were originally developed for international price comparisons: the Time Product Dummy (TPD) method, the Geary-Khamis (GK) method and the GEKS method.

Multilateral methods yield simultaneous estimates of price indices for different time periods. The computational procedure is repeated for every new publication period in order to update the sequence of price indices. Prices and quantities sold of the new period are added, which may alter parameter values and price indices of past periods. However, published index numbers cannot be revised in the CPI, apart from exceptional situations. Different approaches have been proposed for calculating updated price indices, which are described in Section 3.3.

Results are presented in Section 4. Price indices of the different methods are compared, and the impact of weighted versus unweighted indices is shown. The treatment of revisions is an important practical problem, so that special emphasis will be given to comparing the different updating methods. The differences among index methods and choice aspects are discussed and analysed in Section 5. Section 6 lists the main conclusions and topics for further research.

2. Scanner data sets and population dynamics

Besides the Jevons for supermarket scanner data, Statistics Netherlands uses a Laspeyres type method for scanner data of other retailers in its CPI. The Laspeyres methods make use of samples of items from the scanner data. The objective is to replace these methods with another method, which is able to process all data. That is, the future method should not only be able to process existing and disappearing items, but also new ones, preferably as soon as these are sold. The Geary-Khamis method is a candidate method, as this method has already been introduced into the Dutch CPI (Chessa, 2016a).[4]

Statistics Netherlands receives scanner data from a Dutch chain of department stores every week since 2009. The present study uses a part of these data to compare a broad range of different index methods. Scanner data of four product groups are selected: bed clothing, pastries, office supplies and men's T-shirts. The data cover the period February 2009-March 2013.

For each item sold, the data sets contain an EAN (the European version of the GTIN), year and week of sales, quantities sold in each week and the corresponding turnover (expenditure). The data also contain a description of each item, from which characteristics were derived by applying a semi-automatic text mining method.[5] Throughout this paper, we use the term "attribute" to denote a set of one or more characteristics of the same type. In other words, by "characteristic" we mean a specific value of an attribute. For example, 'black' and 'white' are values (characteristics) of the attribute 'colour'. The attributes that were derived from the item descriptions of the four product groups are shown in Table 1.

[4] We used the term "Quality adjusted Unit value method" ("QU method") in previous papers instead of "Geary-Khamis method" (GK method). The QU method is a family of unit value based index methods, with the GK method as special case. As we consider the GK method as the only member of the 'QU family' in this paper, we will use the term GK here.

[5] By "semi-automatic" we mean a method that starts with a visual inspection of the text strings and a first list of key words for item characteristics. The list is subsequently used to link the GTINs to the characteristics, which is the automatic part. Next, the GTINs that are still not linked are inspected, which may lead to additional characteristics and attributes, so that the list of search terms is expanded. This process was repeated until a satisfactory turnover coverage of linked GTINs within a product group was reached.

Table 1. Item attributes for each of the four product groups.

Bed clothing	Pastries	Office supplies	Men's T-shirts
Item type	Item type	Item type	Neck shape
Size	Size	Size	Sleeve length
Fabric	Taste	Fabric	Fabric
Colour		Colour	Colour
Package quantity			Package quantity
Stretch/not stretch			Stretch/not stretch

The number of item attributes extracted with text mining varies between three and six. This is due to the varying length of the item descriptions across the product groups. The data set for pastries has the shortest item descriptions, from which three attributes could be retrieved. Item type refers to product category; for instance, sheets and pillows are two different item types that are contained in the same group of bed clothing items. There is no subdivision into different item types for men's T-shirts, as T-shirt is the only item type. Package quantity denotes the content of a package, which is measured in the case of the four product groups as the number of single items per package (e.g., two-pack T-shirts and single T-shirts). Apart from package quantity and size, the other attributes are categorical variables, which mostly consist of a limited number of categories (characteristics). For instance, sleeve length and neck shape consist of two categories (long and short, and V- and O-shape, respectively).

Scanner data contain information of all items sold. The numbers of items are much larger than the numbers of products in a traditional survey. The numbers of items sold for the four product groups in the four-year period range from 311 items for bed clothing to 1,953 items for men's T-shirts. With such numbers, it is reasonable to expect that not all items will be available each month. Some items may be available over a longer period, while other items will leave an assortment, either temporarily or for good, which may be replaced by new items.

In order to have a first idea of the dynamics of the assortments in the four product groups, we calculated three statistics (cf. Willenborg, 2017b): 1) the number of items that are sold in two successive months ("flow"), 2) the number of items that are sold in one month, but not in the next month ("outflow"), and 3) the number of items that are sold in one month, but not in the previous month ("inflow"). The three statistics are thus based on sales in two successive months.[6] This means that "inflow" does not only contain new items, but also items that are temporarily unavailable. Similarly, "outflow" does not only capture items that disappear forever and the subset of "flow items" does not merely contain items that are sold in each month of the four-year period. The three statistics are shown in Figure 1.

The plots show that the percentage of items that are available in two successive months roughly lies between 70 and 90 for bed clothing, pastries and office supplies, but is much smaller for T-shirts. The inflow percentages for T-shirts range between 10 and 30 for almost all months, which makes this product group especially interesting when comparing bilateral and multilateral index methods.

The flow statistics in Figure 1 only take into account whether an item is sold in one or two months. There may be a large number of new and disappearing items, but if the corresponding expenditure shares are small, their impact on a price index is expected to be limited. Therefore,

[6] The flow concepts can be extended to cases with more than two months. As such extensions give rise to more 'states', we decided to include only two months in the flow dynamics for reasons of simplicity.

we computed the three flow statistics also by including the sold quantities of each item in two successive months.

Figure 1. Flow, inflow and outflow for the four product groups over the four-year period, expressed as percentages of the total number of items in two successive months. The first month on the horizontal axis denotes February 2009.

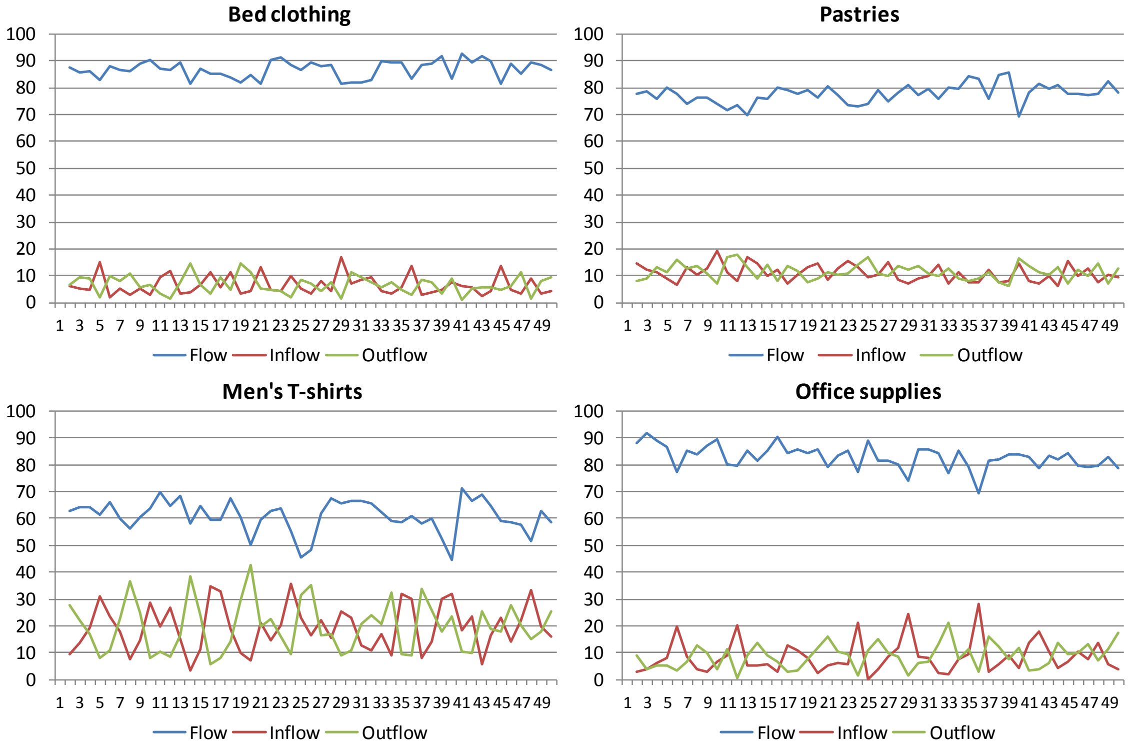
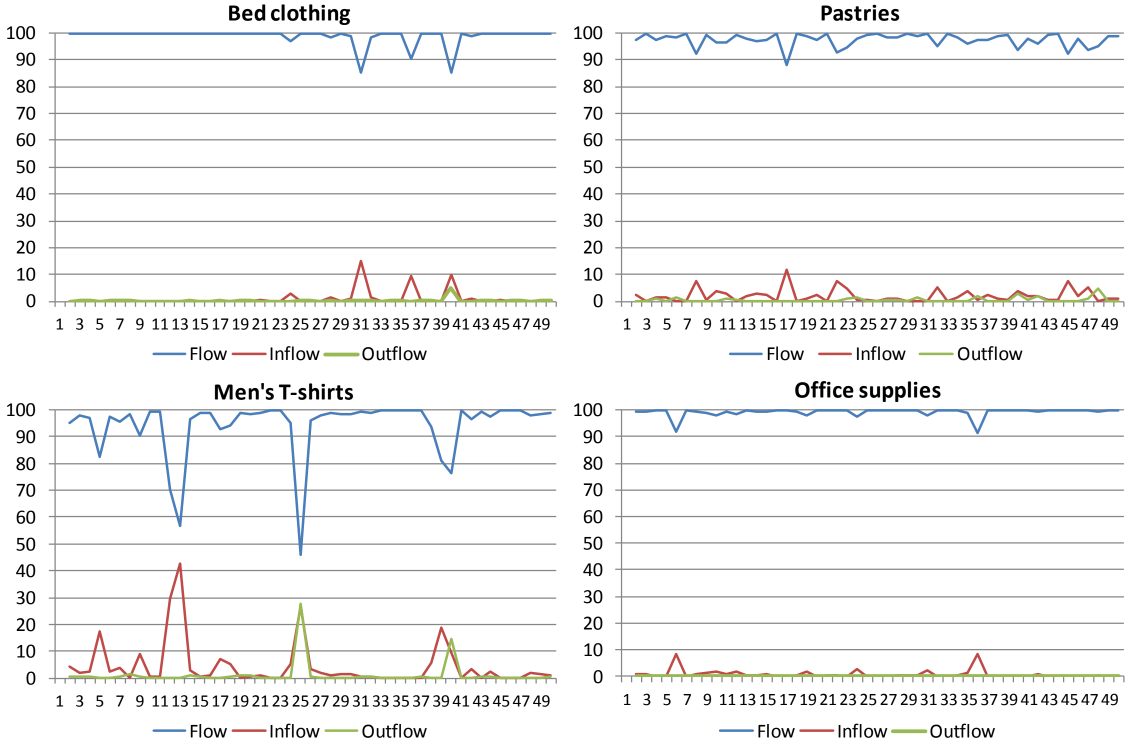


Figure 2. Percentages of flow, inflow and outflow for the four product groups, adjusted for the quantities sold in two successive months.



For inflow and outflow we simply took the quantities in the month in which the items are sold. How to include quantities sold in two successive months for flow is less trivial. In that case, we took the harmonic mean of the quantities sold (based on an analogy with the conductance of electrical resistances connected in series). Other types of mean did not affect the results significantly. The quantity adjusted flow statistics are shown in Figure 2.

If we compare the three statistics with Figure 1, it is immediately clear that the percentages for items that are sold in two successive months are considerably higher than in the unadjusted version. Apparently, the shares of quantities of items sold are larger for items that are available in both months than for items that (re-)enter or disappear from an assortment. The percentages of "flow items" are close to 100 in the majority of the months for bed clothing and office supplies and are somewhat less extreme for pastries. The percentages of quantity adjusted flow for T-shirts are also higher than for the unadjusted version, but Figure 2 still shows months with high percentages of inflow and outflow.

The above flow statistics can be helpful to formulate first expectations about the differences among the index methods that are applied to the four data sets (Section 4).

3. Index methods and choice aspects

3.1 Bilateral index methods

We start from a situation where the transactions at the item level have been aggregated across time and individual consumers into monthly expenditures, quantities purchased and unit values. Scanner data are often characterised by a big churn in terms of new and disappearing items, at least when items are identified at the most detailed level, that is, by GTIN. The plots of the flow dynamics in Section 2 show that the degree of churn may differ considerably across product groups. For the least dynamic product groups bilateral index methods may give results that are comparable with multilateral methods. In order to verify this, we decided to include bilateral methods as well in the comparison of methods in Section 4. In this section we give a concise overview of the methods that have been compared, starting with the bilateral methods. For more details about the methods, see de Haan et al. (2016).

We consider both weighted and unweighted (i.e., equally weighted) and direct and monthly chained bilateral index methods. We start with direct methods, which generate price indices where prices in a base month 0 are compared to, say, prices in T months. We denote the sets of *homogeneous products* belonging to some product group in months 0 and t by G_0 and G_t. A *product* may refer to a single item (GTIN) or to a group of GTINs that share the same set of characteristics that are found to be relevant in order to achieve homogeneity. We further denote the set of matched products in the two months by $G_{0,t}$ and the number of matched products by $N_{0,t}$. The prices (unit values) of each product i in months 0 and t are denoted by $p_{i,0}$ and $p_{i,t}$.

If quantity or expenditure information is not available, the international CPI Manual (ILO et al., 2004) recommends the Jevons price index. The direct Jevons index can be written as follows:

$$P_{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_{i,t}}{p_{i,0}} \right)^{1/N_{0,t}}. \qquad (1)$$

As scanner data contain expenditure data, the construction of weighted bilateral indices is possible. We applied Fisher and Törnqvist indices to the four data sets. As these index formulas have given comparable results, we treat only one of the two in this paper (Törnqvist). The direct Törnqvist price index is given by

$$P_{0,t} = \prod_{i \in G_{0,t}} \left(\frac{p_{i,t}}{p_{i,0}} \right)^{(s_{i,0}+s_{i,t})/2}, \qquad (2)$$

where $s_{i,0}$ and $s_{i,t}$ denote the expenditure shares of the matched products in months 0 and t.

Given the high rate of item churn often encountered in scanner data, an alternative to direct indices could be the use of chained indices. However, high-frequency chaining of weighted price indices could lead to a drifting time series. The extent of chain drift is therefore one of the aspects that will be investigated in the second part of this paper.

A simple way to avoid chain drift could be, for instance, to use equal weights and construct a time series by chaining period-on-period matched-model Jevons indices. But the lack of weighting may be unsatisfactory as scanner data contain expenditure data at the most detailed product level. Different index methods can be thought of that use product weights, based on expenditures, that are free of chain drift. One class of such methods are weighted multilateral index methods.

3.2 Multilateral index methods

Multilateral price index methods are typically applied to compare price levels across countries or regions. These methods yield transitive price comparisons. Transitivity is a desirable property for spatial comparisons because the results will be independent of the choice of base country. Well-known methods are the GEKS method (Gini, 1931; Eltetö and Köves, 1964; Szulc, 1964), the Geary-Khamis (GK) method (Geary, 1958; Khamis, 1972), and the Country-Product Dummy (CPD) method proposed by Summers (1973). For details on the various methods, see Balk (1996, 2001), chapter 7 in Balk (2008), Diewert (1999) and Deaton and Heston (2010).

Multilateral spatial price comparisons can be easily adapted to comparisons over time. The GEKS, GK and CPD are applied to the four data sets in Section 4. Also the hedonic method is applied, so that four methods are compared. The methods are briefly described below. Other methods contained in the overview paper of de Haan et al. (2016) are not considered here. Among these methods is the so-called "Cycle Method", a method developed by Willenborg (2010). The application of the method is still under investigation. Although some experimental results are available (Willenborg and van der Loo, 2016; Willenborg, 2017a), results will be reported when the method has been studied more thoroughly.

3.2.1 The GEKS method

The GEKS method starts from a set of matched-model bilateral indices and then "transitivises" the bilateral price comparisons. Suppose we want to calculate GEKS indices on a time interval $[0, T]$. The GEKS price index between months 0 and t can be calculated as an unweighted geometric average of $T + 1$ ratios of matched-model bilateral price indices $P_{0,z}$ and $P_{t,z}$, which are both constructed with the same index number formula, with month z running through $[0, T]$. GEKS indices thus result from equally weighted $T + 1$ time paths. The index can be written as follows (see Ivancic, Diewert and Fox (2011) or de Haan and van der Grient (2011)):[7]

[7] This expression is used in Willenborg (2017c) to find a way to calculate the GEKS index with the help of Excel. This paper also contains various generalisations of the GEKS index. Another paper shows how the GEKS method can be generalised to the Cycle Method (Willenborg, 2017d).

$$P_{0,t} = \prod_{z=0}^{T} \left(\frac{P_{0,z}}{P_{t,z}} \right)^{\frac{1}{T+1}}. \tag{3}$$

The bilateral indices should satisfy the time reversal test. In its standard form, the GEKS method uses bilateral Fisher indices as inputs. Other choices are possible as well, such as bilateral Törnqvist indices, which is the choice made in this paper. The window length $T+1$ is a point of concern. According to Ivancic, Diewert and Fox (2011), a 13-month window is probably optimal because it is the shortest window that can deal with seasonal goods. A longer window would lead to a loss of "characteristicity". The choice of window length is a problem that occurs with other multilateral methods as well. We will come back to this in Section 3.3.

3.2.2 The Geary-Khamis method

The Geary-Khamis (GK) method is entirely constructed upon the unit value concept. Prices of homogeneous products are equal to the ratio of expenditure and quantity sold. Quantities of different products cannot be added together as in the case of homogeneous products. The prices $p_{i,t}$ of different products $i \in G_t$ in month t are transformed into "quality adjusted prices" $p_{i,t}/v_i$. The adjustment factors v_i of the products convert quantities sold $q_{i,t}$ into "common units" $v_i q_{i,t}$. The price and quantity transformations allow us to define and calculate a quality adjusted unit value $\tilde{p}_t$ for a set of products G_t in month t:

$$\tilde{p}_t = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t}}{\sum_{i \in G_t} v_i q_{i,t}}. \tag{4}$$

Note that the total expenditure in the numerator of (4) is not affected by the transformations.

Expression (4) can be used to define a price index by dividing the quality adjusted unit values in two months:

$$P_{0,t} = \frac{\tilde{p}_t}{\tilde{p}_0} = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t} / \sum_{i \in G_0} p_{i,0} q_{i,0}}{\sum_{i \in G_t} v_i q_{i,t} / \sum_{i \in G_0} v_i q_{i,0}}. \tag{5}$$

Notice that the numerator of the expression on the right-hand side of (5) is an index that measures change in turnover or expenditure between two months. The denominator is a weighted quantity index. Expression (5) makes clear why the GK index is transitive: both the turnover index and the weighted quantity index are transitive.

The weights v_i are defined as follows in the GK method:

$$v_i = \frac{\sum_{z=0}^{T} q_{i,z} p_{i,z} / P_{0,z}}{\sum_{z=0}^{T} q_{i,z}}. \tag{6}$$

Expression (6) shows that the v_i are unit values as well. For each product, the expenditures are summed over the entire interval $[0, T]$, which are divided by the quantities sold of a product over the same interval. In order to exclude price changes from the v_i and the weighted quantity index, the product prices are deflated by the price index. The v_i are also known as "reference prices" (usually referred to as international prices in the spatial context).

Since the GK index itself acts as the deflator in (6), equations (5) and (6) must be solved simultaneously. This can be done iteratively (Maddison and Rao, 1996; Chessa, 2016a) or as the solution to an eigenvalue problem (Diewert, 1999). For more details about the GK method, see Geary (1958), Khamis (1972), Auer (2014), Chessa (2016a) and the aforementioned references.

3.2.3 Time Product Dummy method

The adapted version of the CPD method for spatial price comparisons to the time domain was called the Time Product Dummy (TPD) method by de Haan and Krsinich (2014). Suppose there are N different items observed during a time interval $[0, T]$. The form of the TPD index can be derived by departing from a stochastic model for the pooled price data of all months $0, \dots, T$:

$$\ln p_{i,t} = \alpha + \delta_t + \gamma_i + \varepsilon_{i,t}. \tag{7}$$

The parameters γ_i are known as "item fixed effects" and the δ_t as "time dummy parameters". The $\varepsilon_{i,t}$ are residuals and α denotes the intercept.

We notice that i in (7) represents an item (GTIN). If products are defined as combinations of item characteristics, then the parameters must be estimated under the restriction that the γ_i of different items, but with the same characteristics, have the same value. We estimated the parameters by Weighted Least Squares (WLS) regression with expenditure shares as weights.

Model (7) and the aforementioned assumptions yield the classical TPD index, with quality-adjusted price of a set of products G_t in month t that can be written as

$$\tilde{p}_t = \prod_{i \in G_t} \left(\frac{p_{i,t}}{v_i} \right)^{s_{i,t}}, \tag{8}$$

where $v_i = \exp(\gamma_i)$. The TPD index can be expressed as follows:

$$P_{0,t} = \frac{\tilde{p}_t}{\tilde{p}_0} = \frac{\prod_{i \in G_t} \left(\frac{p_{i,t}}{v_i} \right)^{s_{i,t}}}{\prod_{i \in G_0} \left(\frac{p_{i,0}}{v_i} \right)^{s_{i,0}}}. \tag{9}$$

The quality adjustment factors satisfy the following set of equations:

$$v_i = \prod_{z=0}^{T} \left(\frac{p_{i,z}}{P_{0,z}} \right)^{w_{i,z}} \tag{10}$$

for all $i \in G_t$, where

$$w_{i,z} = \frac{s_{i,z}}{\sum_{\tau=0}^{T} s_{i,\tau}} \tag{11}$$

denotes the share of the expenditure share in month z in the sum of the expenditure shares of product i over the interval $[0, T]$. From the above expressions it follows immediately that the TPD index is transitive on $[0, T]$.

Note the similarities between expressions (8)-(10) for the TPD method and expressions (4)-(6) for the GK method. The quality adjusted unit value (4) for the GK method can be written as a weighted harmonic mean of the quality adjusted prices of each product, where the weights are equal to the expenditure shares of the products as well. The deflated prices for the TPD method have different weights in (10) compared to the GK method, which uses the share of the quantities sold in a month in the sum of the quantities sold over the entire interval $[0, T]$.

3.2.4 Hedonic method

The multilateral hedonic method is closely related to the TPD method. It can be considered as a special case of the latter: instead of estimating parameters γ_i for items i, parameters are defined and estimated for the characteristics of the items.

Let A denote a set of attributes and let B_a denote a set of characteristics for attribute $a \in A$. For categorical attributes a we introduce parameters β_b for every $b \in B_a$. In a traditional hedonic model, the parameter γ_i in expression (7) is replaced by $\sum_{a \in A} \sum_{b \in B_a} \beta_b \, \delta_{i,b}$, where $\delta_{i,b} = 1$ if b is a characteristic of item i and 0 otherwise. Numerical attributes can be treated as categorical variables as well. For instance, one may want to treat different package contents as separate values.

But one could also treat such attributes as numerical variables and estimate a single parameter β_a per unit of content. In that case, the parameter γ_i in expression (7) is replaced by $\sum_{a \in A} \beta_a \, \delta_{i,a}$, where $\delta_{i,a}$ denotes the value of attribute a for item i. This reduces the number of free parameters to be estimated. On the other hand the question is whether this model choice gives a better fit to the data compared to the previously described model form, where the numerical values of an attribute are treated as separate categories.

The above specifications in terms of parameters for characteristics or attributes represent traditional specifications of hedonic models. Attributes are treated separately; that is, it is unusual to include interactions between different attributes in a hedonic model.

Apart from the difference in the specification of model terms for item attributes, the other terms in expression (7) for the TPD model also apply to hedonic models. The parameters of a hedonic model are also estimated by minimising WLS, with expenditure shares as weights. The resulting index formula therefore has the same form as TPD index (9).

3.3 Issues and choices in price index calculation

The methods summarised in sections 3.1 and 3.2 are applied to the four scanner data sets described in Section 2. The results will give some insight into the impact of the choice of index formula on a price index. However, the choice of formula is not the only issue statistical agencies are faced with in the quest of a method when introducing new data sources into the CPI.

The choice of index formula raises a range of other questions, some of which have already been mentioned in the past two sections. We list the issues below, followed by the choices made that will be compared in Section 4. The next section will therefore not only provide insight into the choice of the index formula on a price index, but also into the impact of other choice aspects on the results. In Section 4, we will focus on the following aspects when comparing price indices:

1. *Product weighting*. Comparisons will be made between weighted and unweighted methods.
2. *Index formula and transitivity*. How do the index methods compare? In particular, how do methods that are not transitive compare to methods that produce transitive price indices?

For example, does a monthly chained bilateral method drift in situations with regular inflow and outflow of items?

3. *Updating problem*. For bilateral comparisons over time it is obvious how to proceed with index calculations from month to month. Direct methods use a fixed base month and the current period is simply shifted each month. In monthly chained index methods, the base and current month are pairs of adjacent periods that are both moved one month. For multilateral methods it is less obvious how to proceed with the next month. Adding new information may change the values of quality adjustment parameters and, consequently, of previously calculated price indices. How could a price index for the next month be calculated without leading to drift, as price indices of previous months will change? Three different methods are summarised below, which are compared in Section 4.
4. *Window length* in multilateral methods. Should the adjustment factors v_i in the GK and TPD method, the parameters for the item characteristics in the hedonic method and the bilateral indices used as inputs for the GEKS be calculated from windows of 13 months or from longer windows? Does window length have an impact on price indices?
5. *Level of product differentiation*. Hedonic methods represent items in terms of a number of characteristics. The other three multilateral methods, and also bilateral methods, may either use single items as the most detailed level of product differentiation or combine items into groups at a less detailed level, based on common characteristics. The latter may offer a solution to capturing hidden price changes when GTINs of items change ("relaunches"). Do the price indices at GTIN and group level differ? And are the price indices for the GK, TPD and the GEKS at group level comparable to the hedonic indices?

The results in Section 4 are presented in the same order as listed above, with the fifth point treated within each of the first four aspects.

As referred to under point 3, we considered three updating methods, which are used for calculating a price index for the next month when using multilateral methods. The three methods, which are described in more detail in de Haan et al. (2016), differ in the adaptation of the time window and in the calculation of a price index when price and quantity information of a new month are included. The three methods make the following choices on the time window and price index calculation:

- The entire time window is shifted one month ("rolling window"). Price indices are calculated for each month within the shifted window. The *window splice method* proposed by Krsinich (2014) calculates a price index for the new month by chaining the indices of the shifted window to the index of 12 months ago (for windows of 13 months);
- The second method also uses a rolling window and calculates price indices for the shifted window in the same way as in the window splice method. But a price index for the new month is calculated by chaining the month-on-month index for the last month of the shifted window to the index of the previous month (i.e., the last month of the previous window). This method is referred to as the *movement splice method*;
- The third method, which is proposed by Chessa (2016a), has elements in common with current CPI practice. Instead of using a monthly rolling window, it uses a time window with a fixed base month, which is shifted each year to the next base month. The time window is enlarged each month in order to include information from a new month. The window thus contains two months in January, three in February and will eventually reach the full length of 13 months in December of each year. Price indices are calculated with

respect to the base month with the most recent set of parameter values; this ensures that the indices are free of chain drift. We will refer to this updating method as a *fixed base monthly expanding window method*.

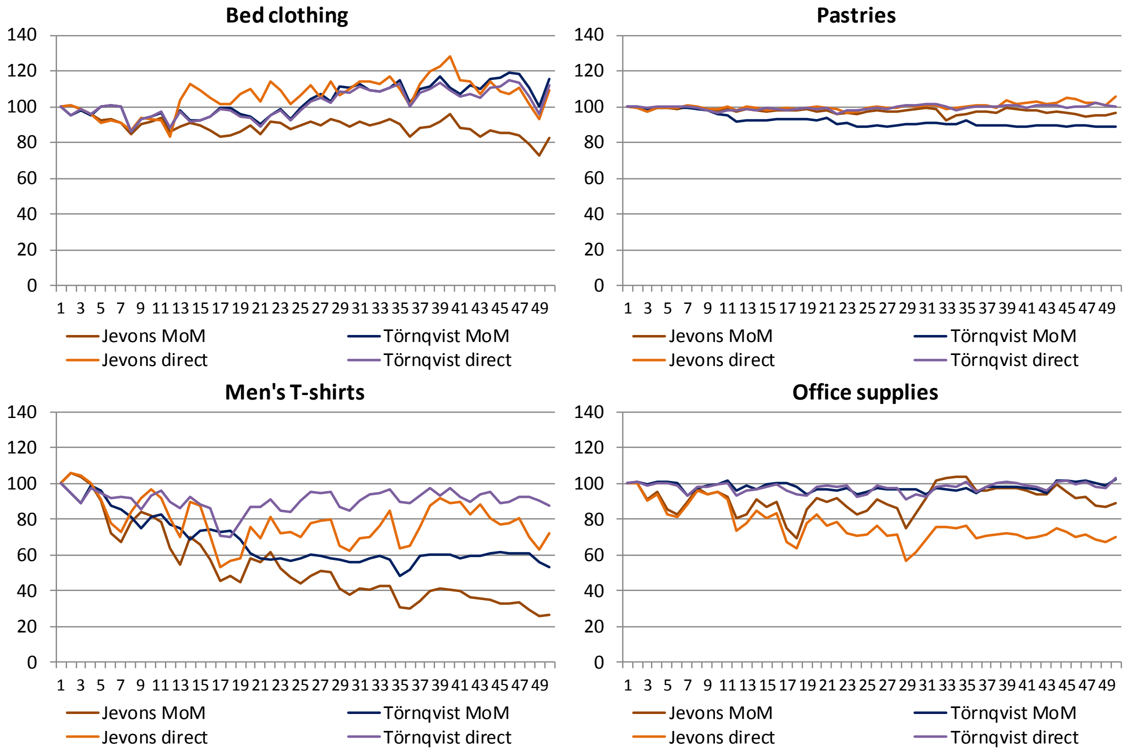
4. Results

4.1 Weighted and unweighted indices

We will first show results for weighted and unweighted indices for each of the four product groups. We will merely do this for bilateral indices, the Jevons and the Törnqvist index, which are both calculated as monthly chained indices and as direct indices, with February of each year as base month in the latter case (February 2009 is the first month of the time series). Direct indices were thus calculated for periods of 13 months. The yearly indices are chained in February of each year.

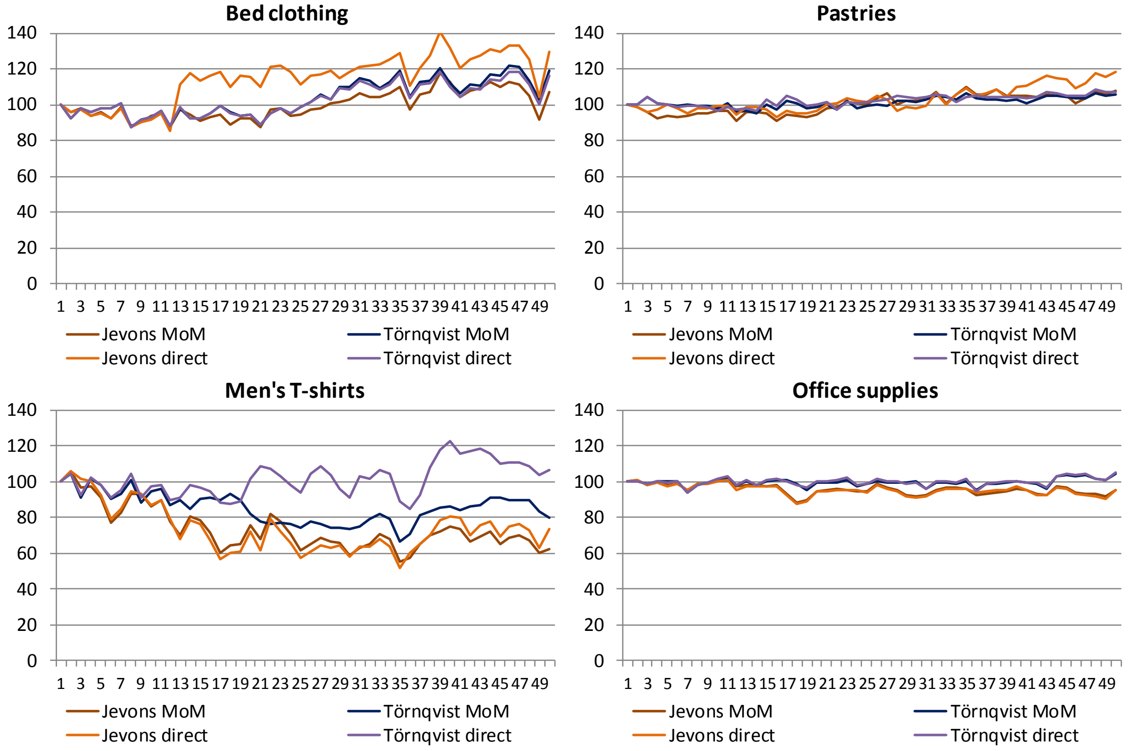
Results are presented for two cases of product differentiation: in one case, each GTIN is treated as a separate homogeneous product, while in the second case GTINs that share the same characteristics are combined into the same group, which is assumed to be a homogeneous product. All characteristics that were extracted from the item descriptions were used for this purpose (see Table 1). The results for these two levels of product differentiation are shown in Figure 3 and Figure 4.

Figure 3. Monthly chained and direct Jevons and Törnqvist indices for the four product groups at GTIN level. February is chosen as the base month for the direct indices each year.



Both figures show big differences between the Jevons and Törnqvist indices, which shows that weighting has a big impact on a price index. Substantial differences are found both for the monthly chained and direct indices, at both levels of product differentiation.

Figure 4. Monthly chained and direct Jevons and Törnqvist indices for the four product groups at GTIN group level. February is chosen as the base month for the direct indices each year.



There are also notable differences between the monthly chained and direct Jevons indices, which are larger than the differences between the chained and direct Törnqvist indices. The item flow dynamics in Figure 1 show quite high inflow and outflow rates, so that the chained and direct Jevons indices can differ. In Figure 1, the rates are not corrected for quantities sold, a situation that applies to unweighted indices like the Jevons index. Correcting for sales yields much smaller inflow and outflow rates (Figure 2), which offers an explanation for the smaller differences between the monthly chained and direct Törnqvist indices. Differences between the two indices may be attributed to possible chain drift in the monthly chained index and to the high inflow and outflow rates in certain months, especially for T-shirts (see Figure 2). Chain drift can be verified when the monthly chained indices are compared with a multilateral index. This will be done in the next section.

4.2 Index formula and transitivity

In this section, we include the multilateral methods in the comparison. We will set up the comparison in two parts: first, we make a comparison between bilateral and multilateral indices, and next we make a comparison among multilateral methods only. With regard to bilateral methods we focus on the monthly chained and direct Törnqvist indices, so we leave out the Jevons indices. The results for the multilateral methods are shown for the fixed base monthly expanding (FBME) window method.

The monthly chained and direct Törnqvist indices are compared with the TPD method in Figure 5 and Figure 6. The Törnqvist indices deviate considerably from the TPD indices, which is more pronounced for the monthly chained indices. The TPD indices with FBME updating are free of chain drift. The monthly chained Törnqvist indices show signs of chain drift for T-shirts and pastries at GTIN level. Drift is reduced at GTIN group level, possibly because the asymmetry in the weights is reduced in high frequency chaining in comparison with GTIN level.

Figure 5. Monthly chained and direct Törnqvist indices, and time product dummy (TPD) indices for the four product groups at GTIN level. The latter two indices use February of each year as base month. FBME = fixed base monthly expanding window.

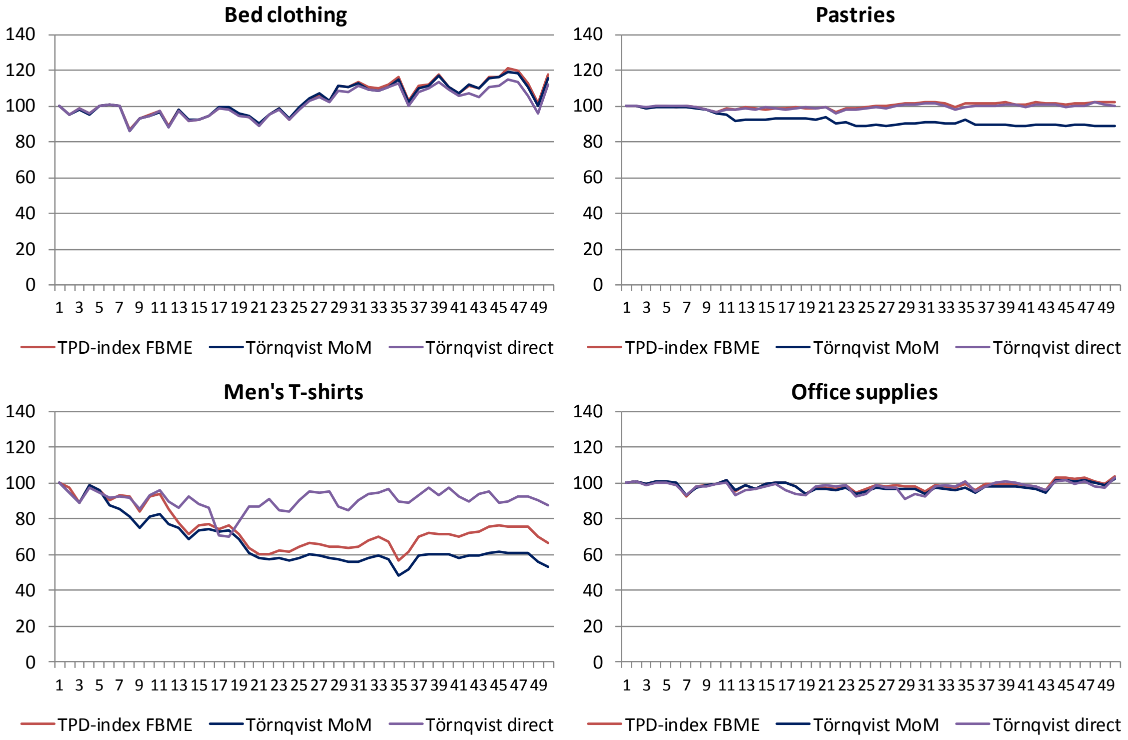
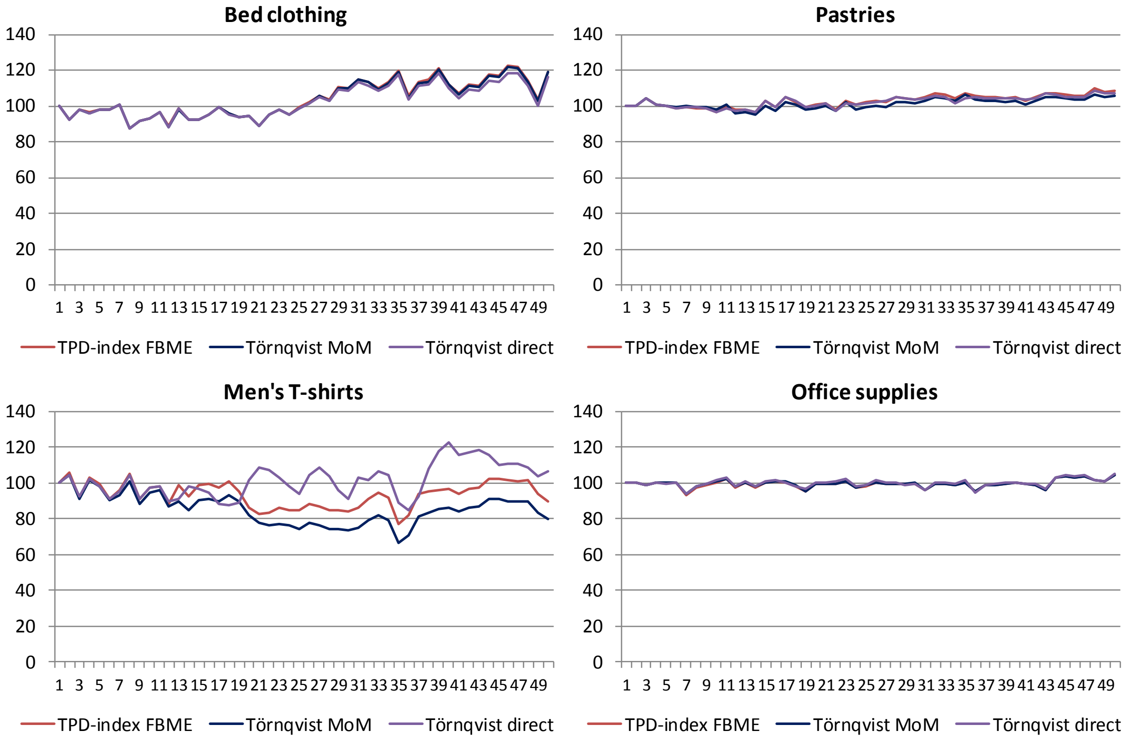


Figure 6. Monthly chained and direct Törnqvist indices, and the TPD indices for the four product groups at GTIN group level. The latter two indices use February of each year as base month. FBME = fixed base monthly expanding window.



Direct indices do not apply high frequency chaining, but this does not exclude the possibility of drifting in such indices. The direct Törnqvist indices show smaller differences with the TPD indices than the monthly chained indices, and even give very good matches for bed clothing, pastries and office supplies, at both levels of product differentiation. The direct indices fail, however, for T-shirts. This is the most dynamic data set of the four, that is, it is the one with the highest inflow and outflow rates.

The indices of the four multilateral methods are shown in Figure 7 and Figure 8. There is hardly any difference between the TPD and GK indices, while the GEKS indices deviate somewhat from these two methods. The hedonic indices show bigger differences with respect to the other three methods.

An important difference in the application of the four methods at GTIN level lies in the fact that the hedonic method makes use of the items' characteristics at this level, while the other three methods are applied at GTIN level without taking the characteristics into account. That is, the GEKS, the GK and TPD indices are calculated by assuming each GTIN to represent a distinct product. The differences between the hedonic and the other three methods are largest for T-shirts, while smaller differences emerge for pastries.

There is a clear relation between these findings and the flow dynamics shown in Figure 2. The inflow rates are highest for T-shirts. Pastries show considerably lower inflow rates, but inflow of items is evidenced throughout almost the entire period of 50 months. The inflow rates for bed clothing and office supplies are virtually non-existent, apart from a number of spikes in a few months.

When inflow applies to new items, differences between the hedonic indices and the other three methods at GTIN level may be caused by relaunches of existing items, which are assigned a new GTIN. If such relaunches are accompanied by price increases, then these price changes will be missed at GTIN level. A better comparison between the hedonic indices and the indices for the other three methods is therefore offered by Figure 8, where the GEKS, the TPD and the GK method are applied at GTIN group level.

Figure 7. Time dummy hedonic (TDH), time product dummy (TPD), Geary-Khamis (GK) and GEKS indices for the four product groups at GTIN level. FBME = fixed base monthly expanding window.

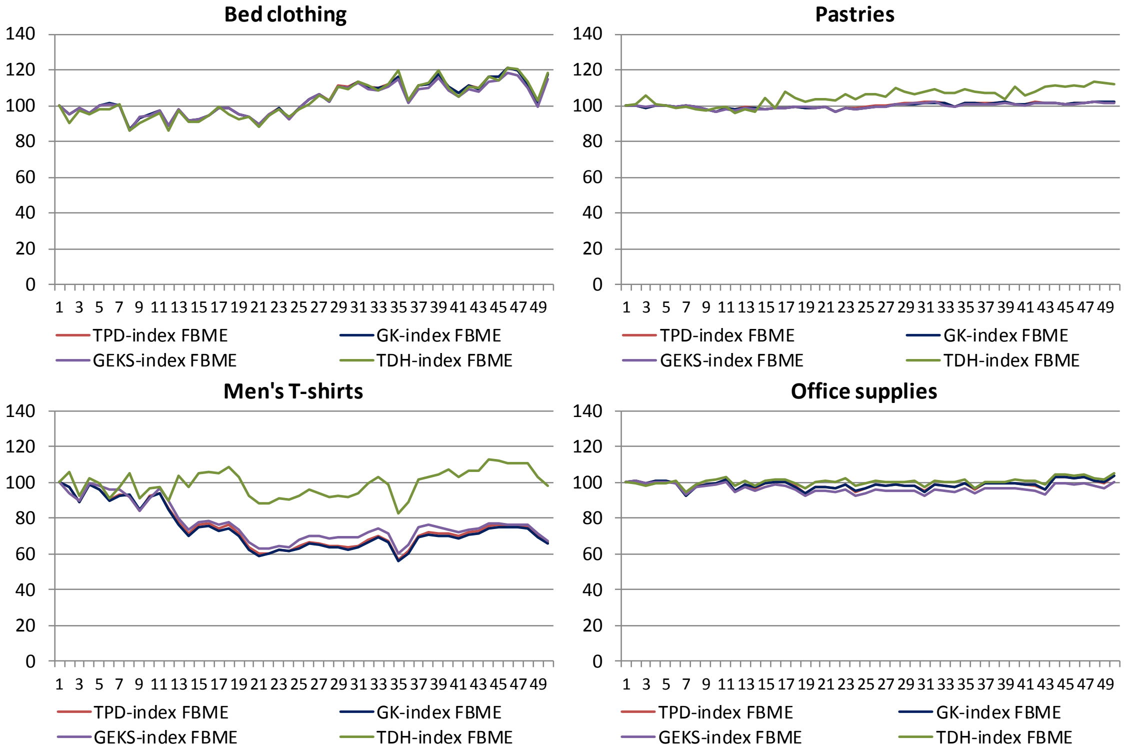
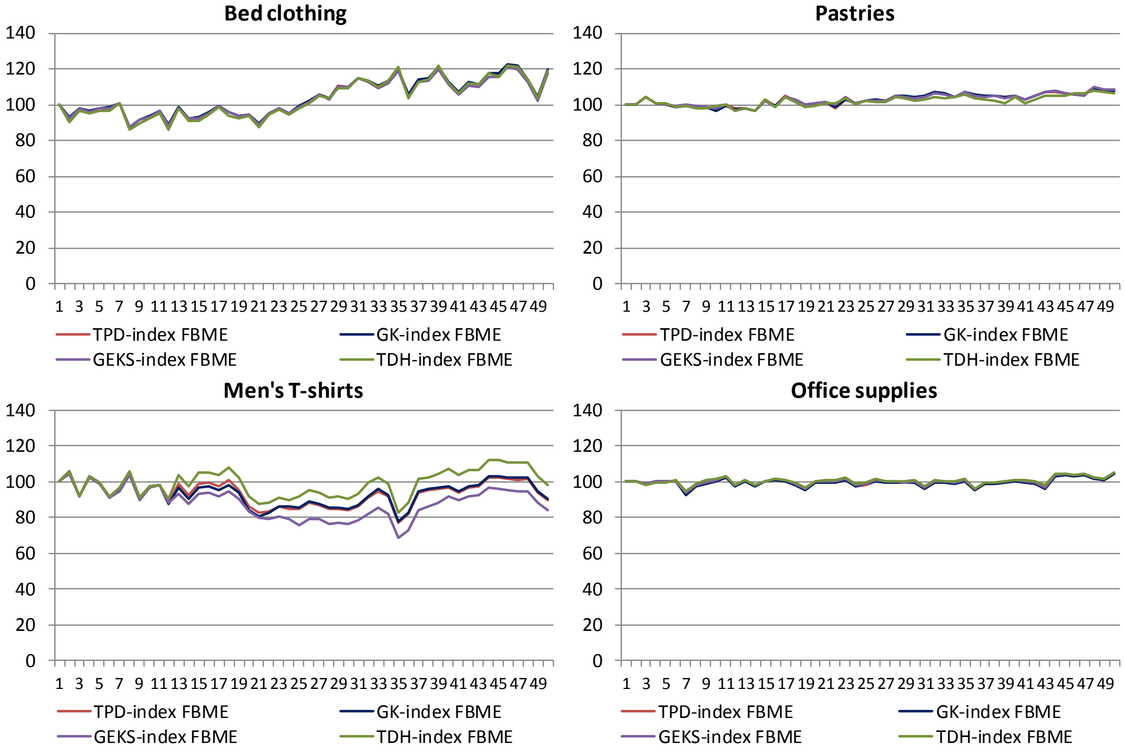


Figure 8. TDH, TPD, GK and GEKS indices for the four product groups at GTIN group level, calculated with the FBME updating method.



Each of the four methods operates on the basis of item characteristics, although in a different way for the hedonic method than for the other three methods. Figure 8 only shows differences between the hedonic index and the other three indices for T-shirts. The GEKS index also differs from the other three indices for T-shirts at GTIN group level. The differences for the GEKS at this level of product differentiation are somewhat surprising, since no systematic difference was found at GTIN level with the TPD and GK indices (Figure 7). We will continue the discussion of the results and differences between the indices in Section 5.

4.3 Impact of updating method

In this section we compare indices for the three window updating methods, which were described at the end of Section 3.3. We have chosen the Geary-Khamis method in order to show the differences among the updating methods, but we could have chosen any of the other three multilateral methods. The multilateral methods have led to similar findings, so we decided to leave out the TDH, TPD and GEKS indices.

The results are shown in Figure 9 and Figure 10 at GTIN and GTIN group level. The three updating methods hardly show any difference for bed clothing, pastries and office supplies. The differences are larger for T-shirts. The movement splice (MS) method shows a structural difference when compared to the fixed base monthly expanding (FBME) window method at both levels of product differentiation. The FBME method is, by construction, free of chain drift. The MS method is a high frequency chaining method, which uses re-calculated parameters each month. The MS method is therefore susceptible to chain drift.

The window splice (WS) method behaves erratically for T-shirts, but follows the trend of the FBME method. However, also the WS method is a high frequency chaining method, although it operates in a different way compared to the MS method. We continue the discussion on the comparison of the updating methods in Section 5.

Figure 9. Geary-Khamis (GK) indices for the window splice (WS), the movement splice (MS) and the fixed base monthly expanding (FBME) window method at GTIN level.

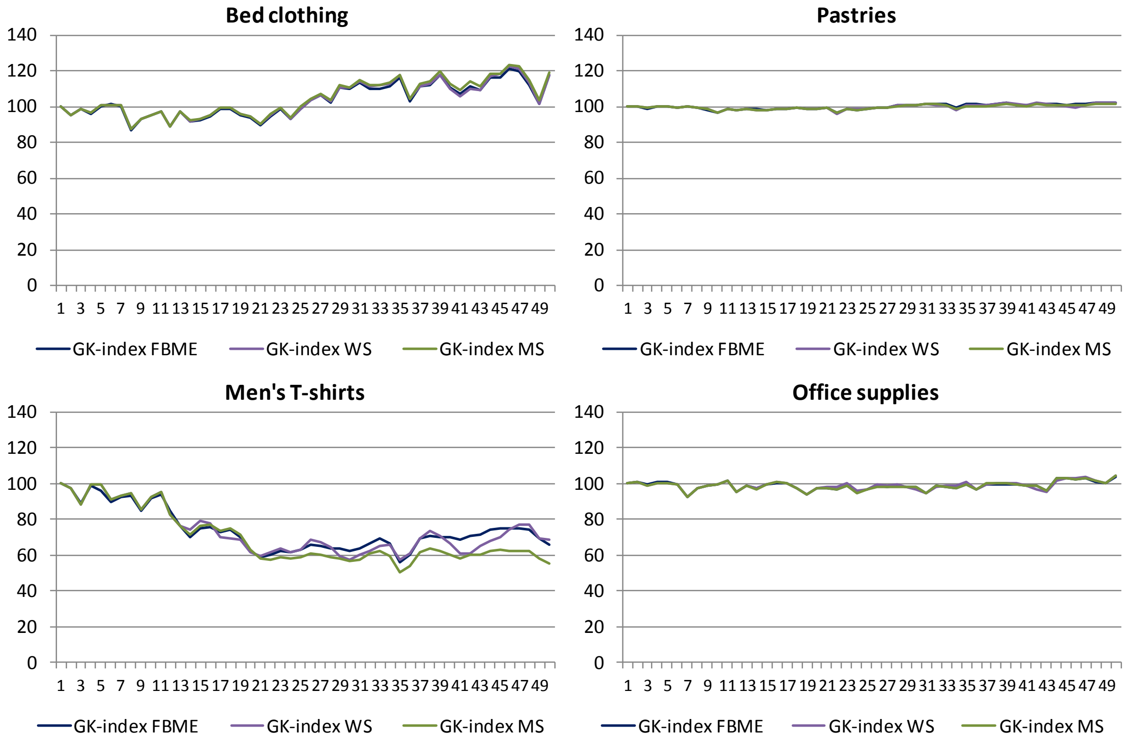
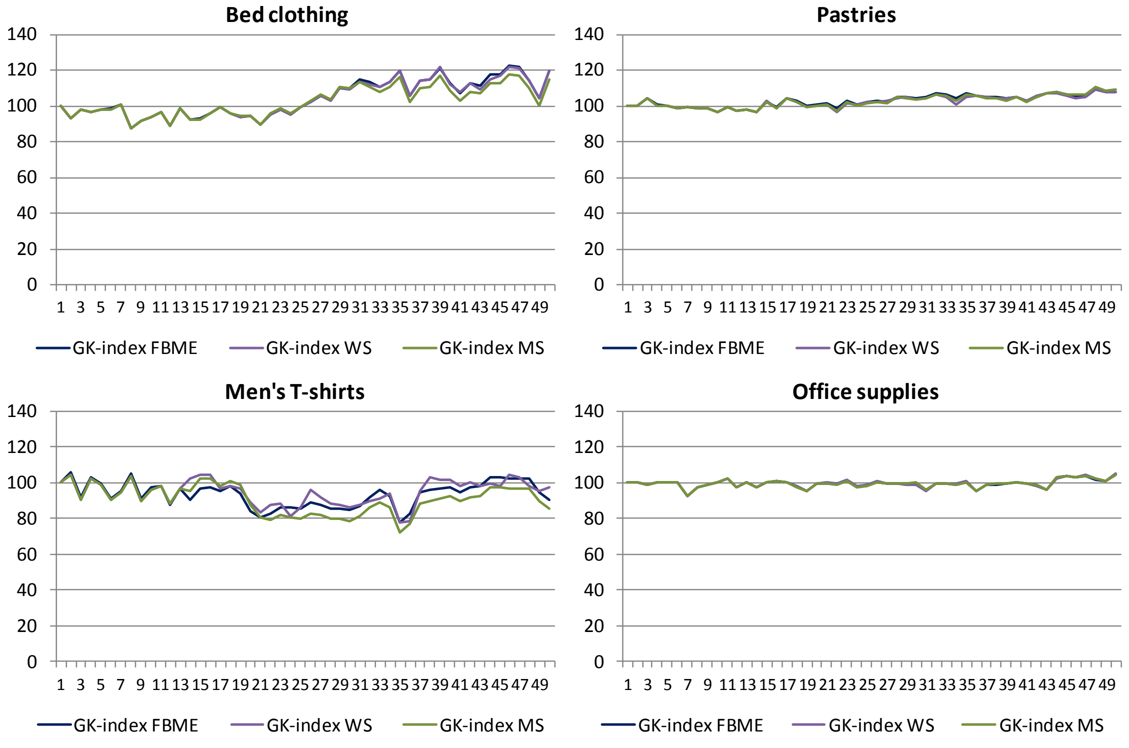


Figure 10. Geary-Khamis (GK) indices for the window splice (WS), the movement splice (MS) and the fixed base monthly expanding (FBME) window method at GTIN group level.



4.4 Length of the time window

The window updating methods compared in the previous section make use of time windows of 13 months. The FBME method employs a monthly increasing window size, but eventually reaches the full length of 13 months. Previously in this paper we have mentioned that longer windows could lead to a loss of "characteristicity". Nevertheless, to our knowledge the question how price indices compare for different window lengths has not been addressed in previous studies.

In this study, price indices for the four multilateral methods were also calculated on the full window of 50 months. This means that the parameters and indices of the TDH, TPD and GK method were estimated from the combined price and quantity data of the full window, while for the GEKS input bilateral Törnqvist indices were calculated from the combined data of 50 months.

Price indices for the full window are shown for the GEKS method in Figure 11 and Figure 12. These indices are compared with the GEKS indices that were calculated with the FBME window method, which is applied to windows of 13 months.

Figure 11. GEKS indices at GTIN level for the full window of 50 months compared with the GEKS indices calculated each year with the FBME window updating method.

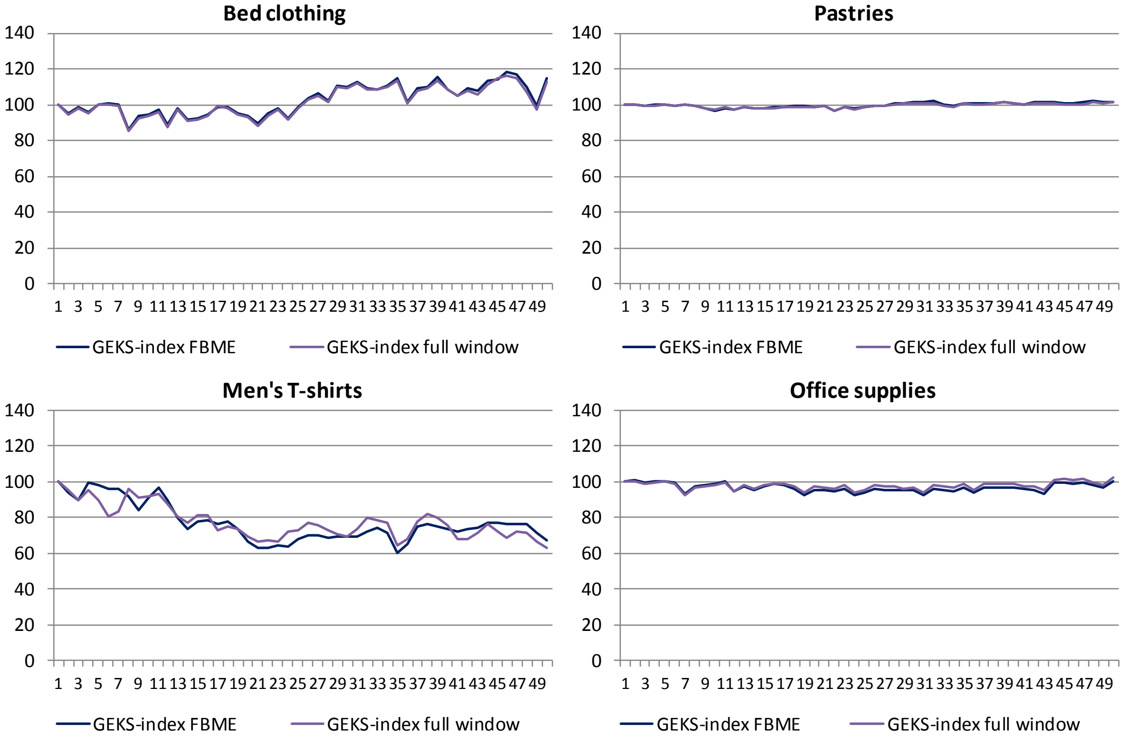
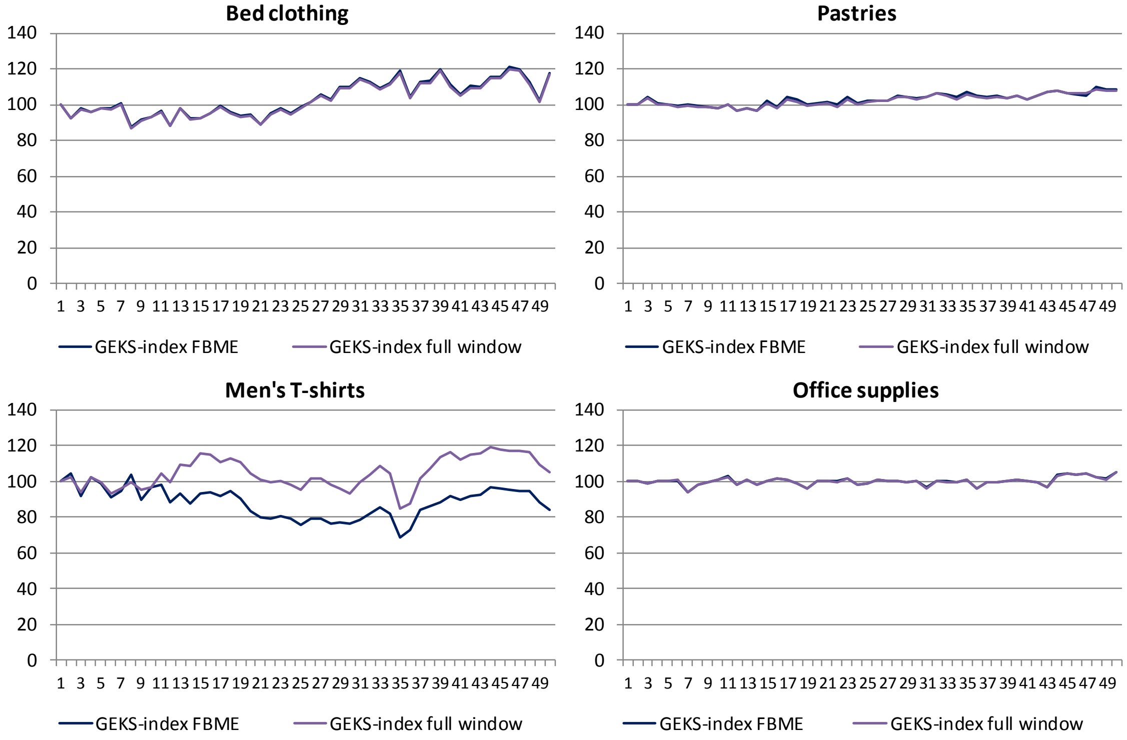


Figure 11 shows hardly any difference between the indices for the two window sizes. The indices in Figure 12 at GTIN group level are practically the same for bed clothing, pastries and office supplies. However, the indices for T-shirts at GTIN group level differ considerably.

The TDH, TPD and GK methods lead to similar findings; differences between the two window lengths are only found for T-shirts at GTIN group level. The differences between the two window lengths for T-shirts are smaller than for the GEKS. Given the size of the difference shown in Figure 12, there is a need of further analysing the results, which will be done in Section 5.

Figure 12. GEKS indices at GTIN group level for the full window of 50 months compared with the GEKS indices calculated each year with the FBME window updating method.



5. Discussion

In this section, we discuss the results of Section 4 in more detail for each of the choice aspects treated. We will verify whether it is possible to rank the choice aspects from most to least influential in terms of their impact on price indices. Such a ranking could be used to inform decisions about the choice of an index method, for instance when statistical agencies intend to replace survey data by scanner data or when they are already using scanner data but intend to review their current practices.

5.1 Weighted versus unweighted indices

The differences between the Jevons and Törnqvist indices in Figure 3 and Figure 4 show that the choice between weighted and unweighted (i.e., equally weighted) indices may have a big impact on a price index. Different products may have different price changes over time, so that different weighting schemes will give different indices. The differences can become quite large, both at GTIN level and at GTIN group level. The monthly chained indices may suffer from chain drift, so that the impact of weighting alone is more difficult to assess for these indices.

5.2 Index formula and transitivity

5.2.1 Bilateral versus multilateral methods

In Figure 5 and Figure 6, direct and monthly chained Törnqvist indices are compared with Time Product Dummy indices. The latter are calculated by applying a monthly expanding window method with a fixed base month, which is free of chain drift by construction. Multilateral methods are able to incorporate existing, new and disappearing items directly into index calculations.[8] This means that we can use the 'TPD-FBME' indices as benchmarks for assessing a number of effects when applying bilateral indices: (1) the extent to which monthly chained bilateral indices may suffer from chain drift, and (2) the extent to which the exclusion of new items until the next base month in direct bilateral methods affects a price index.

The results in Figure 5 and Figure 6 show that the direct Törnqvist and TPD indices give comparable results for bed clothing, pastries and office supplies. The results for T-shirts differ significantly. The weighted flow statistics for the first three product groups in Figure 2 show zero inflow and outflow rates in most months, with only a few spikes in several months. The assortment of T-shirts is more dynamic. There are several large spikes in the inflow rates. One of these spikes occurs at month 13, which marks the entrance of a new collection of T-shirts, with a new type of fabric (organic cotton).

This newly added characteristic of fabric gave rise to a range of new products of T-shirts at the beginning of the second year of the time series, which immediately generated a significant expenditure share (as can also be seen from the huge inflow spike in Figure 2). The direct Törnqvist method includes the new products only in the next base month, while the TPD and other multilateral methods directly include new products. The existing part of the assortment has an increasing price trend, while the organic cotton T-shirts entered at high prices, which started to decrease in the course of the year of introduction. This explains why the direct Törnqvist index rises above the TPD index in the course of the second year. The difference with the TPD index becomes quite large.

The monthly chained Törnqvist indices differ from the TPD indices for T-shirts as well and also for pastries at GTIN level (Figure 5). For both product groups, the monthly chained indices lie below the TPD indices. This is due to the high-frequency chaining of the monthly chained bilateral index, which clearly exhibits chain drift. This is an undesirable property. A necessary condition for index methods is that they should be transitive/free of chain drift. Methods that do not satisfy this property should not be allowed.

5.2.2 Comparisons among multilateral methods

The TPD, GK and GEKS indices give comparable results at GTIN level, but the hedonic indices differ from these three methods for pastries and T-shirts (Figure 7). At group level (Figure 8), the four methods give practically the same results for bed clothing, pastries and office supplies, but the GEKS and the hedonic method differ from the TPD and GK indices for T-shirts. We analyse the differences below.

GEKS method

The GEKS index for T-shirts at GTIN group level lies below the other three indices (Figure 8). In this subsection we try to find an explanation for this behaviour. Below, we make a formal comparison between the GEKS and TPD index formulas. For ease of presentation, we simplify the problem by considering cases where prices of each product are available in every month $0, 1, \dots, T$. We denote the set of products by G.

In Section 4, we used Törnqvist indices in order to compute GEKS indices. Expression (3) for the GEKS index is thus equal to

[8] New products may have an impact on a hedonic index from the first month in which they are sold, while for fixed effects multilateral methods (GK, TPD) price indices will show an effect of new products from the second month.

$$P_{0,t} = \prod_{z=0}^{T} \left(\frac{P_{0,z}}{P_{t,z}}\right)^{\frac{1}{T+1}} = \frac{\prod_{z=0}^{T} \left(\prod_{i \in G} \left(\frac{p_{i,t}}{p_{i,z}}\right)^{\frac{s_{i,t}+s_{i,z}}{2}}\right)^{\frac{1}{T+1}}}{\prod_{z=0}^{T} \left(\prod_{i \in G} \left(\frac{p_{i,0}}{p_{i,z}}\right)^{\frac{s_{i,0}+s_{i,z}}{2}}\right)^{\frac{1}{T+1}}}, \tag{12}$$

which we shorten to $\tilde{p}_t/\tilde{p}_0$. The numerator of (12) can be rewritten in the following form:

$$\tilde{p}_t = \left\{\prod_{i \in G} \left(\frac{p_{i,t}}{v_i'}\right)^{s_{i,t}}\right\}^{\frac{1}{2}} \left\{\prod_{i \in G} \left(\frac{p_{i,t}}{v_i''}\right)^{\frac{1}{T+1}\sum_{z=0}^{T} s_{i,z}}\right\}^{\frac{1}{2}}, \tag{13}$$

where

$$v_i' = \prod_{z=0}^{T} p_{i,z}{}^{\frac{1}{T+1}}, \tag{14}$$

$$v_i'' = \prod_{z=0}^{T} p_{i,z}{}^{s_{i,z}/\sum_{\tau=0}^{T} s_{i,\tau}}. \tag{15}$$

The same form can be obtained for the denominator. Expression (13) shows that the GEKS index can be rewritten as the geometric mean of two TPD like indices: one in which the quality adjusted prices $p_{i,t}/v_i'$ are weighted by the expenditure shares of the products in the month t of observation, while the quality adjusted prices $p_{i,t}/v_i''$ in the rightmost term of (13) are weighted by the average expenditure shares of each product i over the entire period $[0, T]$.[9]

Expression (15) is equal to the adjustment factors in the TPD index, apart from the deflator that is used in the latter (see expressions (10) and (11)). In contrast to the weighting scheme in (15), expression (14) attaches equal weight to the monthly prices. Note how the two types of weighting in the GEKS translate into the weighting of product prices in the two adjustment factors: the equal weighting of the different time paths in (14), and the expenditure based weighting of the products in the bilateral Törnqvist indices in (15).

Expression (14) may cause problems to the index. A situation in which problems arise is where products are taken out of an assortment and a small number of remaining products are sold at dump prices. Such prices receive a relatively large weight in the adjustment factor (14) compared to the weighted version (15). It is easily verified that the first term within brackets in (13) pulls the index down, as the index can be written as follows:

$$P_{0,t} = \frac{\tilde{p}_t}{\tilde{p}_0} = \left\{\prod_{i \in G} \frac{p_{i,t}{}^{s_{i,t}}}{p_{i,0}{}^{s_{i,0}}} (v_i')^{s_{i,0}-s_{i,t}}\right\}^{\frac{1}{2}} \left\{\prod_{i \in G} \left(\frac{p_{i,t}}{p_{i,0}}\right)^{\frac{1}{T+1}\sum_{z=0}^{T} s_{i,z}}\right\}^{\frac{1}{2}}. \tag{16}$$

Notice that the adjustment factors v_i'' drop out in the second term within brackets in (16). The index will have a downward bias since the values of the v_i' for disappearing products decrease

[9] The general case, in which product prices may not be available in one or more months, can be written in a similar form. The weighting is more complex, because product matches for different pairs of months may lead to different monthly expenditure shares.

when dump prices are assigned higher weights and because $s_{i,0} - s_{i,t} > 0$ for exiting products. The v_i' of the regular part of an assortment are expected to be affected to a much smaller extent by the definition and weighting in the adjustment factors.

Disappearing products that are sold at much lower prices than the regular prices occur in the data for men's T-shirts. Consequently, it does not come as a surprise that the GEKS index lies below the other three indices. On the other hand, the question is why this does not occur at GTIN level (Figure 7). Possibly there is an interaction with the level of product differentiation.

Products that leave an assortment at dump prices are not uncommon. Consumer goods in which this occurs are clothing and drugstore items. In the present study we applied different methods to scanner data of only four product groups. It would therefore be interesting to extend the number of data sets, and also to cover different market segments and types of consumer goods, in order to form a better idea of how widespread the downward bias of the GEKS is. We come back to this point in the final section. But apart from this, it should be clear that the GEKS has shortcomings that cannot be accepted for price index calculation.

In addition to the above observations, there is more to say about the GEKS. The fact that GEKS-Törnqvist indices can be written in terms of TPD like indices (see expression (13)) raises the question whether the GEKS is really needed. The TPD method suppresses dump prices in a better way than the GEKS, since the weighting of product prices in different months is done in better agreement with the frequency of occurrence of the different transaction prices throughout the period $[0, T]$. Because the GEKS-Törnqvist and the TPD index have the same functional form, and because the weighting of the product prices in the adjustment factors v_i of the TPD index is better to justify, the GEKS can be considered to be superfluous.

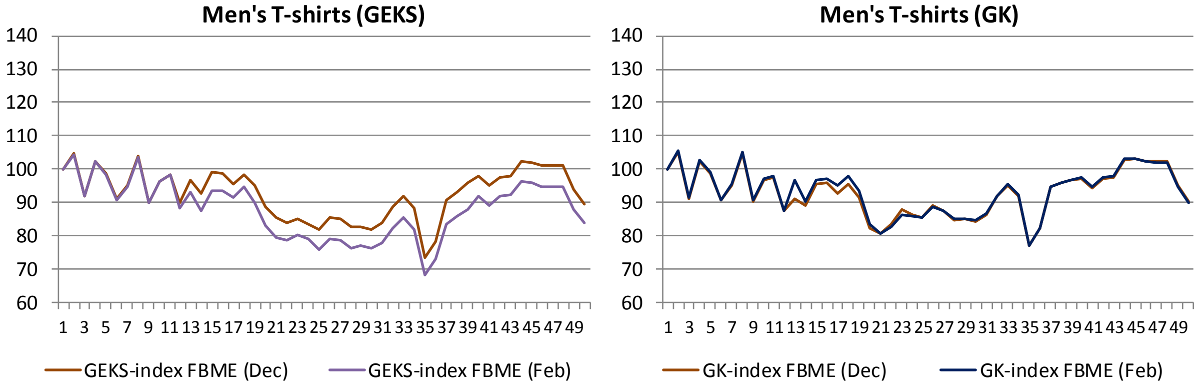
Another questionable property of the GEKS-Törnqvist index is the summation of expenditure shares of a product over different time periods. This occurs in the second term within brackets in (13). Product expenditures are normalised with respect to the total expenditure within a specific time period in order to calculate expenditure shares. One cannot even ensure that shares of different periods for the same product can be ranked on an ordinal scale, leave alone that shares can be summed. Such operations are therefore not meaningful from measurement scale theoretic considerations. Note that the same holds for the weighting in the v_i of the TPD index (see expression (11)). The weighting should therefore be changed in applications of the TPD method.

Let us return to the relatively large weight that the GEKS assigns to dump prices. In our analyses of the results, this property of the GEKS turned out to have a big impact on the results when changing the base month. Price indices with the FBME window updating method were calculated with February as base month, which is the first month in the time series. The CPI uses December as base month. We also calculated price indices with December as base month. Figure 13 shows the GEKS and GK indices for men's T-shirts at GTIN group level for the two base months.

The results for the GEKS method show that changing the base month from February to December raises the index by about 5-6 percentage points, while the GK index is hardly affected by the change of base month. The same holds for the TPD index, which is not shown here. Interestingly, the choice of December as base month gives similar indices for the GEKS and GK method.

The data show a couple of products that started to be sold at dump prices after January of the third year, while sales dropped rapidly after that month. This is a period in which the difference between the GEKS indices for the two base months starts to increase. The choice of February as base month implies that the dump prices already affect the GEKS index in the second year. The choice of December as base month leaves the GEKS index unaffected in the second year.

Figure 13. GEKS and GK indices at GTIN group level for men's T-shirts, calculated with the FBME window updating method for December and February as base months.



Note that there is hardly any impact of the dump prices on the GEKS index when December is chosen as base month, as the corresponding products are hardly sold from the beginning of the third year and consequently hardly contribute to the index. Because of this, the GEKS index with December as base month appears to be more plausible. Note the similarity of the GEKS index for December as base month and the GK indices. As dump prices are suppressed in the adjustment factors of the GK index, and also in the TPD index, their impact on the indices is negligible, whatever the choice of base month.

Hedonic method

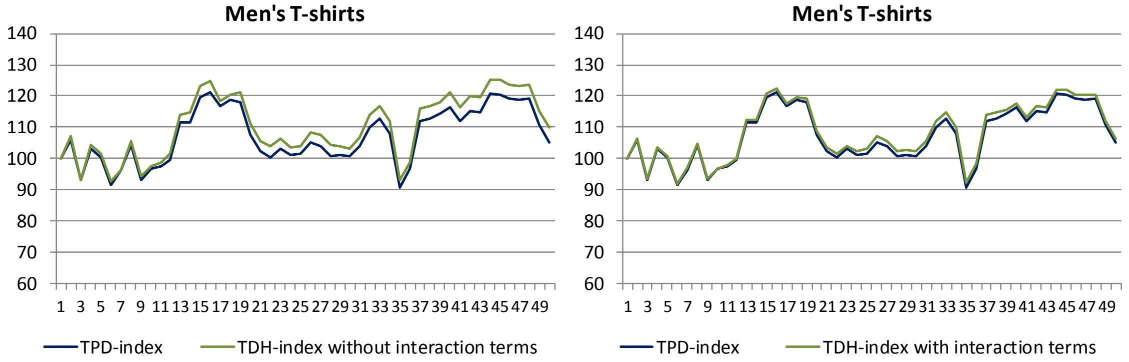
The differences with the hedonic method for pastries and T-shirts have different causes. We first discuss the differences for pastries. Information about item characteristics in the scanner data of the department store is limited for Coicop 01 items. The information is contained in text strings, so that an extraction procedure was developed for the item characteristics. Besides the limited number of item attributes in the data, it turned out to be difficult to create an exhaustive list of search terms for the characteristics, which made the results sensitive to errors. This explains the differences between the hedonic method and the other three methods, since the TPD, GK and GEKS method are applied at GTIN level in Figure 7. If we use the characteristics to combine the GTINs into groups, then the four methods give practically the same results (Figure 8).

The differences between the hedonic index and the other three methods for T-shirts is caused by differences in index formulas. In order to understand this, we turn our attention to Figure 8, where the GK, the TPD and the GEKS are applied by making use of item characteristics as well. In particular, we focus the attention on the comparison between the hedonic method and the TPD method, since the price indices for both methods can be expressed in the same form, that is, as ratios of geometric averages of quality adjusted prices. What makes the difference between the two methods at GTIN group level is how the quality adjustment parameters are related to the item characteristics. Conventional hedonic methods estimate a parameter for each characteristic separately, while in the GK, the GEKS and the TPD method this is done by combining all item characteristics selected.

Although it is not common use to specify adjustment factors for combinations of attributes in hedonic methods, it is instructive to consider refinements of these methods in this direction in order to understand the differences with the other three methods. For this purpose, we merely specified pairwise interaction terms for the attributes. We selected only 4 from the 6 available attributes in order to limit the amount of work in specifying interaction terms. The selected attributes are the four most influential ones with regard to their impact on the price index for T-shirts (fabric, number of items in a package, sleeve length and colour).

The TPD and TDH indices with and without pairwise interaction terms are shown in Figure 14. The differences between the two methods practically disappear after including pairwise interaction terms in the hedonic method. The marginal differences that remain are probably due to the two attributes that are not paired with the other four.

Figure 14. TPD and TDH indices for men's T-shirts, with the latter applied with and without pairwise interaction terms for four item attributes. The parameters and price indices are calculated on the full length of the time period (50 months).



The specification of pairwise interaction terms implies that additional parameters are included in a hedonic model. One cannot limitlessly add parameters to a model, since this may overfit the data. Different methods exist in the literature for comparing models with different numbers of parameters. Examples of well-known measures are the Akaike and Bayesian Information Criterion (e.g., see Claeskens and Hjort (2008)). These information criteria are adjusted maximum likelihood functions, in which the number of free parameters acts as a penalty term. More complex models (i.e., with more free parameters to estimate) may give better fits in terms of maximum likelihood, but are penalised more than models with less parameters.

The BIC has a more severe penalty term than the AIC (for sample sizes larger than 7). It is therefore especially interesting to compare the BIC values for the hedonic models with and without interaction terms. The BIC turned out to be better for the model with interaction terms. The same analysis was performed in a similar study for men's socks as well, where the difference between the hedonic index and the TPD index increased to about 20 percentage points in less than four years (Chessa, 2016b). Also in this case the hedonic model fits improved after adding pairwise interactions and the differences between the two indices disappeared. The differences between the hedonic indices and the TPD and GK indices therefore appear to be very small after refinements of the hedonic models. Further details about this analysis can be found in Chessa (2016b).

5.3 Updating methods

The differences between the three window updating methods turn out to be small (Figure 9 and Figure 10). The movement splice method deviates from the other two methods for T-shirts and to a smaller extent for bed clothing (Figure 10). The two splice methods that are applied with a rolling window are both high-frequency chaining methods, so that chain drift cannot be excluded. The window splice method gives results that are comparable with those for the fixed base monthly expanding window. The indices calculated with the window splice method show more variability for T-shirts.

Also the window splice method is a high-frequency chaining method, so that it was decided to extend its application to other data sets. The additional comparisons in Chessa (2016b) show large differences with the FBME method, on which direct indices are calculated with respect to the base month. The latter are free of chain drift and were calculated for different choices of the base month in the cited study. The results for the window splice method can therefore be said to be sensitive to chain drift as well.

5.4 Window length

Multilateral price indices were calculated for windows of 13 months and on the total length of the time period (50 months). The results for the two window lengths were shown in Figure 11 and Figure 12. The impact of window length on a price index turns out to be negligible, except for T-shirts at GTIN group level, which shows large differences between the two window lengths (Figure 12). The question is whether the results for T-shirts constitute an exceptional case.

This was examined by extending the index calculations to more data sets. GK and TPD indices were calculated in recent studies for department store, drugstore and supermarket scanner data, which was done for window lengths of 1, 2 and 4 years (Chessa, 2016b, 2017b). The differences turned out to be very small and even negligible for supermarkets (Chessa, 2017b).

5.5 Product differentiation

There is an additional choice element, which, because of the scope of this study, has remained undiscussed so far. As we calculated price indices at GTIN level and GTIN group level, it is interesting to make a direct comparison between the indices at both levels of product differentiation. Price indices for the four product groups at the two levels are shown in Figure 15 for the GK method.

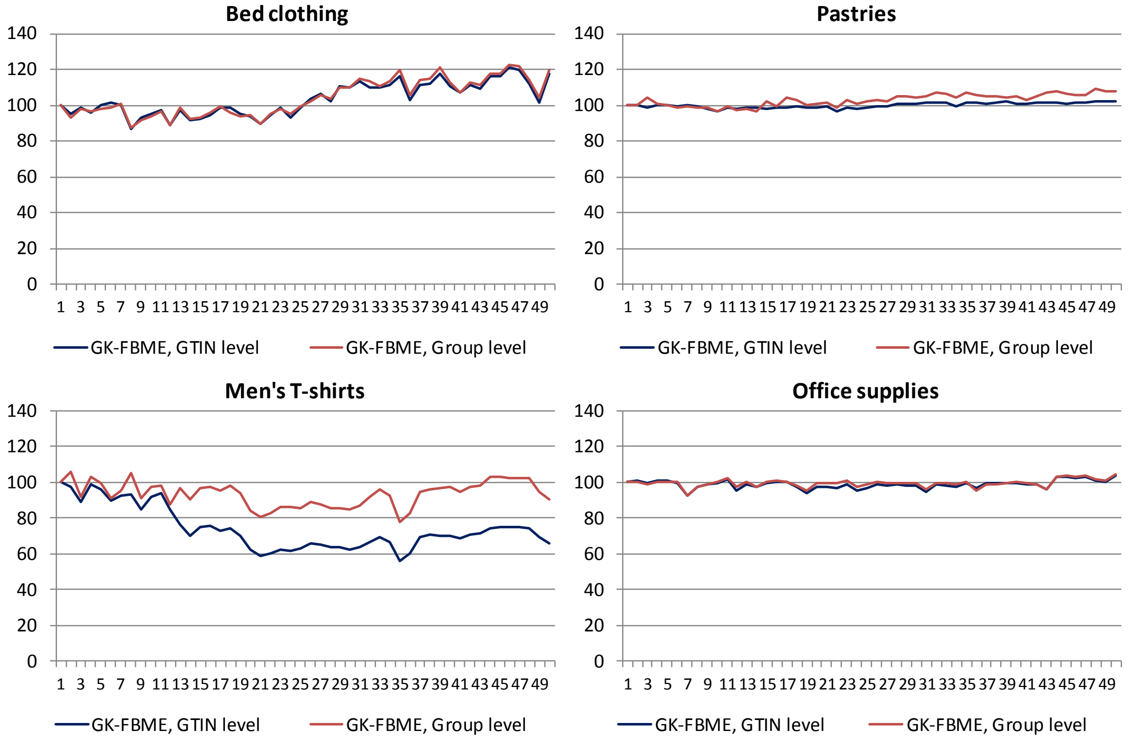
The differences between the indices at the two levels are negligible for bed clothing and office supplies. The differences are larger for pastries, but, as was already pointed out previously in this section, the indices at GTIN group level are influenced by errors in the extraction of item characteristics from the short item descriptions and possibly also by the limited number of attributes contained in the text strings (one or two in the majority of the strings).

The department store scanner data contain information about more attributes for clothing items, so that more refined indices can be calculated at GTIN group level for clothing. The differences between the price indices at GTIN and GTIN group level are very large for men's T-shirts. The differences shown in Figure 15 are not uncommon for clothing. New collections are regularly introduced, even each year depending on product group and age group (e.g., for teenagers). A pattern that is often observed when new clothing collections are introduced is that the prices and numbers sold of existing items rapidly decrease, while the new items enter at higher, regular prices. Outgoing and new items that share the same characteristics have to be linked in order to capture "hidden" price increases associated with such relaunches. An example of a rapidly decreasing price index when outgoing and new items are not linked is shown in Chessa (2016a).

Differences between price indices at GTIN and GTIN group level may also be substantial at retailer level, as is shown in Chessa (2017a) for department stores, which can be ascribed to clothing, and also for drugstores. The differences between the year on year indices lie between

1.5 and 3.5 percentage points at overall retailer level. For both retail chains, the indices at GTIN level lie below the indices at GTIN group level.

Figure 15. GK indices at GTIN and GTIN group level for the four product groups, calculated with the fixed base monthly expanding window method.



These findings show the importance of developing sound methods for defining products. It is also important to identify types of consumer goods where relaunches are more likely (fashionable goods, such as clothing and beauty products). The availability of information for linking outgoing to new items is indispensable. These links can be established by making use of item characteristics or, ideally, of links between retailers' own product codes, which retailers use for their own purposes (stock keeping).

It is beyond the scope of the present paper to treat the problem of product definition. Some first ideas for a method that could be used for selecting item attributes and to link outgoing to new items are summarised in Chessa (2016a). Different methods could be used for selecting attributes, from sophisticated statistical methods that are rooted in model selection (Claeskens and Hjort, 2008) and regression to approaches that are easier to understand and visualise, such as sensitivity analysis. This topic is still under study at Statistics Netherlands.

5.6 Summary of findings

The above analyses allow us to distinguish between choice aspects that are influential and hardly influential in terms of their impact on price indices. Weighting versus equal weighting of products turned out to have a big impact in this study, followed by the choice between weighted bilateral and multilateral methods. The choice of the level of product differentiation has proven to have a big impact on the results, as was shown in cited studies.

The choice of updating method turned out to have a small impact on the indices, with the movement splice method showing signs of drift. However, results of an extended study clearly suggest that the window splice method is sensitive to chain drift as well (Chessa, 2016b). Both methods are high-frequency chaining methods. The aforementioned cited recent studies have shown that window length has a small or negligible impact on the results.

The comparisons of the four multilateral methods show a varied picture. The TPD and GK methods give similar results. But the analyses in Section 5.2.2 have evidenced issues with the hedonic method and the GEKS, which may lead to big differences with the TPD and GK method. Possible causes for the differences have been identified: the lack of interactions among item attributes in traditional hedonic models and the erroneous weighting of monthly prices in the adjustment terms or "reference prices" in the rewritten, TPD like form of the GEKS, which may lead to downward biases of the index.

6. Conclusions and future work

Statistical agencies have to make different choices when selecting a method for price index calculation. This makes the compilation of inflation and growth figures a complex subject. This study has shed some light on the impact of different choice aspects on price indices and on the suitability of different index methods. The findings and analyses presented in Section 5 allow us to draw a number of conclusions:

- Weighted and equally weighted indices show big differences. This is an essential finding since statistical agencies still use a Jevons index for supermarket scanner data in their CPI. Jevons indices are calculated for elementary aggregates. Although different types of thresholds are used, for instance, in order to eliminate GTINs with expenditure shares below a threshold, differences between 'filtered' Jevons indices and weighted methods can be quite large, also at overall retailer level (Chessa, 2016b). Much can therefore already be gained by using expenditure based weights within elementary aggregates. In addition, thresholds for expenditure shares and dump prices can be removed when switching to weighted methods. In general, unweighted methods should not be used when scanner data are available.
- Also the comparison between bilateral and multilateral indices has evidenced substantial differences. A major problem with the bilateral methods studied in this paper is that these are not transitive. High-frequency chaining methods should not be applied as these lead to drift. Direct methods are not suited for dynamic populations.
- Multilateral methods are able to process all data and can timely include new products into index calculations, without waiting until the next base month. Such methods are therefore the preferred choice. However, the analyses in Section 5.2.2 have clearly shown that considerable care is still needed when choosing among multilateral methods. Hedonic methods turn out to be fine when accounting for interactions among item attributes. When this aspect is taken care of, the results are practically the same as for the GK and TPD method.
- The suitability of the GEKS is a different story. The weighting is problematic for different reasons mentioned in Section 5.2.2. The equal weighting in one of the two "reference prices"—v_i' in expression (13)—generates a downward bias in cases with dump prices for outgoing items, and possibly also with seasonal goods. As expression (13) for the GEKS-Törnqvist index has the same form as the TPD index, the GEKS is in fact superfluous, also in view of the fact that the TPD index uses a more appropriate weighting.

- A recent paper by Diewert and Fox (2017) proposes the use of the CCDI method. This method is very closely related to the GEKS-Törnqvist. In fact, it is the same as expression (13), with the difference that the unweighted reference price v_i' is used in both terms on the right-hand side of (13). Given the statements in our previous point, we therefore expect that the CCDI will perform even worse than the GEKS in the cases mentioned. In addition to this, the CCDI leaves a lot of essential questions unanswered when applying it to scanner data: How should the method be applied when product prices are missing? And what about the consistency of the method across different time units for prices? If we would switch, say, from monthly to daily prices, does the method yield consistent results for the two choices?
- Drift cannot be excluded for window splice methods, which makes these methods undesirable. Therefore, we will use updating methods that work with a fixed base month. Rolling windows combined with a fixed base month are an alternative to the (asymmetrical) FBME updating method. Both methods give practically the same results (Chessa, 2017b).

The analyses in Section 5.2.2 allow us to reduce the range of multilateral methods that we consider to be applicable to scanner data. The methods applied and discussed in this paper can be reduced to GK, TPD and hedonic, with the latter as a special case of the TPD method if one considers interactions among item attributes. The impact of window length on the results has turned out to be small or even negligible. The aforementioned methods can therefore be applied with a 13-month window and a fixed base month, so that indices are free of chain drift. This choice is in agreement with current CPI practice.

Whatever method is chosen, statistical agencies should always keep in mind that the reliability of the results depends on the availability, quality and relevance of information about item characteristics. Such information is needed for defining suitable levels of product differentiation, in the sense that old and new GTINs of relaunched items have to be linked in order to capture 'hidden' price changes. The results in Chessa (2017a) show that different levels of product differentiation may lead to substantially different price indices, even at aggregate retailer level.

Finally, we list some topics for future research:

- As was stated in this paper, we will extend this study to data sets of different retailers and types of consumer goods. Beside the department store scanner data, we will also include scanner data from drugstore and supermarket chains. Additional research has already been carried out and reported (Chessa, 2016b, 2017a, 2017b). But we aim at applying all multilateral methods studied in the present paper to the additional data sets in order to further verify the findings reported here and to extend the analyses of Section 5.
- We have just started with a similar comparative study for consumer electronics, which is restricted to multilateral methods. First results for mobile phones and televisions are summarised in Chessa (2016b). More extensive studies with additional product categories will be carried out in future research. This will be done when the limited amount of metadata about product characteristics from scanner data can be supplemented with additional metadata collected by web scrapers.
- The problems with the GEKS encourage further application and software development of the Cycle Method (cf. Willenborg, 2017d). In particular the choice of suitable weights to obtain useful results is a topic that needs attention. Also the behaviour of the CM for different dynamics of item populations will be investigated.

- An important question within the context of the CPI is at which product group level multilateral methods should be applied. European regulations prescribe the use of Laspeyres type methods for combining the price indices of elementary aggregates to indices for higher aggregates. As the latter methods make use of weights based on a previous year, it is essential not to set the elementary aggregate level "too low". By this we mean that when entire product groups leave an assortment at dump prices, the rapid price decreases combined with the larger weights of a previous year will result in downward biases of the indices for higher aggregates. This has to be avoided. In an attempt to reduce the risk of such biases even further, it would also be highly recommended to consider alternatives to the Laspeyres type methods for higher aggregates.
- Price collection from web sites, for instance, by using web scrapers, is rapidly gaining popularity at statistical agencies (Breton et al., 2016; Griffioen and ten Bosch, 2016). The use of internet prices for price index calculation poses different challenges compared to scanner data, such as finding suitable weights for products and product groups. A preliminary study that compares price indices based on scanner data and web scraped data from the same retailer can be found in Chessa and Griffioen (2016). This research will be continued in the near future.
- In addition to the previous point, index methods that use scanner data may highly benefit from additional metadata collected by web scrapers. The number of item attributes and characteristics included in scanner data sets can be small, as retailers may not always be able or willing to provide the information requested by statistical agencies. Possibilities of combining scanner data with web scraped data are also studied at Statistics Netherlands, currently for consumer electronics.

References

Auer, L. von (2014). The Generalized Unit Value Index Family. *Review of Income and Wealth*, 60, 843-861.

Balk, B.M. (1996). A Comparison of Ten Methods for Multilateral International Price and Volume Comparison. *Journal of Official Statistics*, 12, 199-222.

Balk, B.M. (2001). Aggregation Methods in International Comparisons: What Have We Learned? ERIM Report, Erasmus Research Institute of Management, Erasmus University Rotterdam.

Balk, B.M. (2008). *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. New York: Cambridge University Press.

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, M., Payne, C., Smith, T., Winton, J., and Woods, A. (2016). Research Indices Using Web Scraped Data: May 2016 Update. Office for National Statistics, UK, internal report, 23 May 2016.

Chessa, A.G. (2016a). A New Methodology for Processing Scanner Data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, issue 1/2016, 49-69.

Chessa, A.G. (2016b). Comparisons of the QU-method with other Index Methods for Scanner Data. Paper prepared for the first meeting on multilateral methods organised by Eurostat, Luxembourg, 7-8 December 2016. Statistics Netherlands.

Chessa, A.G. (2017a). The QU-method: A New Methodology for Processing Scanner Data. *Proceedings of Statistics Canada 2016 International Methodology Symposium*. Available at: http://www.statcan.gc.ca/eng/conferences/symposium2016/program

Chessa, A.G. (2017b). Comparisons of QU-GK Indices for Different Lengths of the Time Window and Updating Methods. Paper prepared for the second meeting on multilateral methods organised by Eurostat, Luxembourg, 14-15 March 2017. Statistics Netherlands.

Chessa, A.G., and Griffioen, R. (2016). Comparing Scanner Data and Web Scraped Data for Consumer Price Indices. Report, Statistics Netherlands.

Claeskens, G., and Hjort, N.L. (2008). *Model Selection and Model Averaging*. UK: Cambridge University Press.

Deaton, A., and Heston, A. (2010). Understanding PPPs and PPP-Based National Accounts. *American Economics Journal; Macroeconomics*, 2, 1-35.

Diewert, W.E. (1999). Axiomatic and Economic Approaches to International Comparisons. In A. Heston and R.E. Lipsey (eds.), *International and Interarea Comparisons of Income, Output and Prices*, Studies in Income and Wealth, Vol. 61, pp.13-87. Chicago: University of Chicago Press.

Diewert, W.E., and Fox, K.J. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. Discussion paper 17-02, Vancouver School of Economics, The University of British Columbia, Vancouver, Canada.

Eltetö, Ö, and Köves, P. (1964). On an Index Computation Problem in International Comparisons. *Statiztikai Szemle*, 42, 507-518. [in Hungarian]

Geary, R.C. (1958). A Note on the Comparison of Exchange Rates and Purchasing Power between Countries. *Journal of the Royal Statistical Society A*, 121, 97-99.

Gini, C. (1931). On the Circular Test of Index Numbers. *International Review of Statistics*, 9, 3-25.

Griffioen, A.R., and ten Bosch, O. (2016). On the Use of Internet Data for the Dutch CPI. Paper presented at the UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland.

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

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

de Haan, J., Willenborg, L., and Chessa, A.G. (2016). An Overview of Price Index Methods for Scanner Data. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 2-4 May 2016, Geneva, Switzerland.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva.

Ivancic, L., Diewert, W.E., and Fox, K.J. (2011). Scanner Data, Time Aggregation and the Construction of Price Indices. *Journal of Econometrics*, 161, Issue 1, 24-35.

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

Krsinich, F. (2014). The FEWS Index: Fixed Effects with a Window Splice – Non-Revisable Quality-Adjusted Price Indices with No Characteristic Information. Paper presented at the meeting of the group of experts on consumer price indices, 26-28 May 2014, Geneva, Switzerland.

Maddison, A., and Rao, D.S.P. (1996). A Generalized Approach to International Comparison of Agricultural Output and Productivity. Research memorandum GD-27, Groningen Growth and Development Centre, Groningen, The Netherlands.

Summers, R. (1973). International Price Comparisons Based Upon Incomplete Data. *Review of Income and Wealth*, 19, 1-16.

Szulc, B. (1964). Index Numbers of Multilateral Regional Comparisons. *Przeglad Statystyczny*, 3, 239-254. [in Polish]

Willenborg, L. (2010). Chain Indexes and Path Independence. Report, Statistics Netherlands.

Willenborg, L. (2017a). Transitivizing Elementary Price Indexes for Internet Data using the Cycle Method. Discussion Paper, Statistics Netherlands.

Willenborg, L. (2017b). Quantifying the Dynamics of Populations of Articles. Discussion Paper, Statistics Netherlands.

Willenborg, L. (2017c). From GEKS to Cycle Method. Discussion Paper, Statistics Netherlands.

Willenborg, L. (2017d). Price Indexes and Transitivity. Discussion Paper, Statistics Netherlands.

Willenborg, L., and van der Loo, M. (2016). Transitivizing Price Index Numbers Using the Cycle Method: Some Empirical Results. Report, Statistics Netherlands.