An Overview of Price Index Methods for Scanner Data

Jan de Haan[a], Leon Willenborg[b] and Antonio G. Chessa[c]

27 April 2016

(Preliminary draft; please do not quote)

Abstract: A number of approaches to constructing price indexes using scanner data are currently available. This paper provides an overview of the various approaches, shows how they relate to each other and summarizes empirical applications by both statistical agencies and academic researchers. We focus on "big data" index number methods that use all the price and quantity information in scanner data sets, in particular multilateral methods, rather than traditional small-sample methods. The paper also identifies topics for future research.

Keywords: big data, dynamic universe, hedonic regressions, multilateral index number methods, quality adjustment, transitivity.

JEL Classification: C43, E31.

---

[a] Division of Corporate Services, IT and Methodology, Statistics Netherlands, and OTB, Faculty of Architecture and the Built Environment, Delft University of Technology; email: j.dehaan@cbs.nl.

[b] Division of Corporate Services, IT and Methodology, Statistics Netherlands, email: l.r.j.c.willenborg@cbs.nl.

[c] Division of Economic abd Business Statistics and National Accounts, Statistics Netherlands, email: a.g.chessa@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

Since its very beginning, it has been customary in the Consumer Price Index (CPI) to base price index calculations for consumer goods on a fixed basket of goods. Prices of each good are traditionally collected through surveys, for instance, by recording shelf prices in shops or by telephone surveys. Statistical agencies plan such surveys in one, two or three weeks of each month before publication of CPI figures. Baskets of goods, and the corresponding expenditure weights, are often updated by the end of each year in order to capture shifts in consumption patterns over time.

The 70s of the last century marked an important change in processing payments of goods in shops, with the introduction of the barcode scanner at counters (the first one was introduced in the US in 1974). Transaction data by barcode (“scanner data”) were recorded electronically, which could be used by retailers for their own administration. This new source of data also opened up new possibilities for academic research and for statistical agencies. Scanner data contains the transactions of all goods that have been sold, the prices actually paid by consumers, and the quantities sold for each item. In contrast, prices collected in traditional surveys are prices offered, quantities sold are not available, and the traditional basket is a relatively small sample of the complete universe of goods. Electronic data also reduces administrative burden for the outlets where prices are collected and may help increase the efficiency of CPI production.

In spite of their potential, scanner data is still used by only a small number of statistical agencies in compiling their CPI. The number of European countries that are using scanner data has increased to six in 2016 (Belgium, Denmark, the Netherlands, Norway, Sweden and Switzerland). Such a small number may be explained by different factors. Some countries have a statistical law, which facilitates statistical agencies in obtaining data from retailers. Another factor is the rather long process from establishing first contacts with a retailer until data delivery, data analysis and its eventual processing. Recent developments suggest, however, that the number of countries that will be using scanner data is likely to increase during the coming years.[1]

A shift from survey data to scanner data brings about important changes with regard to data processing in the CPI. Scanner data sets contain large numbers of items, which may even amount to more than 100,000 items per retailer. This means that items and associated item information need to be assigned in an efficient way to the relevant product categories. Switching from survey data to scanner data also implies a transition

[1] 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.

from a fixed basket of goods to a dynamic universe that contains all goods that have been sold. While classical bilateral index formulas may be adequate for a fixed basket, a transition to a dynamic universe raises the question as to whether traditional methods are still valid. Not only offers scanner data weighting information (in terms of turnover) at the individual item level, the universe of goods includes all goods that disappear and goods that are introduced into an assortment in the course of a year.

GTINs provide the tightest definition of a product, i.e. the most homogeneous one, which is the level of the individual item. The availability of barcodes in scanner data sets might therefore suggest that the problem of setting up product descriptions as in traditional surveys does no longer exist. However, barcodes of existing items may change, even when unimportant exterior aspects of the packaging of items are modified. Old and new GTINs have to be linked in order to capture possible price changes after a change in barcode. It is therefore important that scanner data sets contain information about item characteristics and/or other information for linking GTINs. More details are given in Section 2.

The remainder of this paper gives an overview of index methods, discusses some issues and gives a summary of methods used in different countries based on scanner data. Section 3 starts with a review of classical bilateral indexes. These methods may become restrictive as the dynamics of product assortments tend to grow. Multilateral index number methods are potentially better suited to deal with such situations. Hedonic methods will be described but also other types of multilateral methods, including the time product dummy method, the GEKS-method and the Geary-Khamis method, which were originally developed for international price comparisons.

Multilateral methods yield simultaneous estimates of price indexes for different time periods. The computational procedure is repeated for every new publication period. Prices and quantities sold of the new period are added, which may alter product weights and price indexes of past periods. However, once published, the CPI cannot be revised, apart from exceptional situations. Different approaches have been proposed to deal with this “no-revisions constraint” for calculating price indexes in a new period, which are discussed in Section 4.

A summary of price index methods currently used by statistical agencies and related applications by academic researchers is presented in Section 5.

Section 6 concludes and suggests topics for further research.

2. Properties of the population, data, and methods

2.1 Properties of the population

This paper focuses on index methods that can be applied to scanner data. By scanner data we mean data on turnover or expenditures and quantities sold of products specified by Global Trade Item Number (GTIN), displayed as a barcode. A GTIN is an identifier for trade items. The system comprises, among others, the Universal Product Code and the European Article Number (EAN) and is developed and managed by GS1, a private company. Since a GTIN refers to trade items, it may apply to both physical goods and services upon which there is a need to retrieve pre-defined information and that may be priced, ordered or invoiced at any point in a supply chain.[2]

Scanner data and barcodes are almost naturally linked to consumer goods. This paper focuses on consumer goods and the Consumer Price Index as the primary field of application. However, the index number methods described in this paper can be applied in a broader context. Examples are producer prices for goods and services, such as health care and education, provided of course that detailed information is available on prices and quantities sold.

The main classification system of goods and services used for CPI purposes is COICOP (Classification Of Individual Consumption by Purpose). COICOP is published by the UN Statistics Division and distinguishes between expenditures of households, non-profit institutions serving households and general government. The CPI typically relates to expenditures of households (see ILO et al., 2004). The COICOP classification system consists of 14 so-called Divisions, which are subdivided into Groups and then Classes. For the European Harmonized Index of Consumer Prices (HICP), these Classes are further subdivided into "Subclasses", which are the lowest publication levels of the HICP.

While COICOP, or COICOP/HICP, provides a useful reference classification, it needs to be detailed further for CPI construction as the Classes or Subclasses are still too heterogeneous in the sense that they will typically contain many different products and product varieties in terms of brand, package content, etc., which have to be treated as separate items for making price comparisons. A next step could be to subdivide the (sub)classes even further into more homogeneous strata, depending on how the scanner data is obtained. Often the data provider – either a retail chain or a market research company – uses its own detailed product classification system, which might be helpful

[2] More details can be found at: http://www.gs1.org/gtin.

in this respect, as long as this system can be linked to COICOP. But even these strata will generally contain different product varieties.

An essential question that needs to be answered when compiling price indexes is how to define items that will be “homogeneous enough” for making price comparisons. At a particular point in time and at a particular outlet, GTINs represent homogeneous items. However, this does not necessarily mean that GTIN constitutes the suitable level of product differentiation; this level can sometimes be too detailed for CPI purposes. GTINs are assigned according to allocation rules which may hamper comparisons over time. Existing items may receive a new GTIN, for instance, in order to fit them into a new product line. Changes to packaging and product formulation may give rise to a new GTIN, even if the item in question remains the same from the consumer’s perspective in terms of content and other relevant attributes. When fully comparable items (but with different GTINs) return to the store shelves at different prices, their price changes will obviously not be captured if GTIN would be used as item identifier.

The extent of the problem of such “disguised” price changes for “relaunches” of fully comparable items may differ across different types of consumer goods. Disguised price increases of relaunches occur frequently in the Dutch drugstore market (Chessa, 2013). Yearly assortment renewals are not uncommon in the clothing market, especially in trendy segments, and price indexes constructed at the GTIN level may drop to almost zero within a few years (Chessa, 2015). These examples illustrate that the dynamics or item churn of product categories is a key property when describing the population of consumer goods, and should be a point of attention when analyzing and using scanner data for price measurement.

The above examples also illustrate the need for having additional information, in particular on the price-determining characteristics, in order to be able to compare “like with like”, which is fundamental to the construction of (quality-adjusted) price indexes. Without characteristics information it will be difficult if not impossible to determine whether or not two items with different GTINs are fully comparable, i.e. to notice and properly deal with relaunches, or adjust for quality change if the items have different characteristics.

Product definition is an essential part of price index construction, but we do not go into this topic in detail as this paper focuses on methods for the treatment of scanner data in the CPI rather than conceptual issues. Yet, the considerations mentioned above are relevant for defining requirements for scanner data sets, which will be discussed in the next section.

2.2 Properties of the data

Scanner data is a form of transactions data. The core data is values (in terms of turnover or expenditures) and quantities sold, usually at the GTIN level. The appropriate concept of price for an item sold in some time period is the unit value, which is defined as the ratio of total value and total quantities sold in that period, in other words, the average price actually paid by consumers.[3] Scanner data allows statistical agencies to calculate unit values and unit value indexes at the item level and to aggregate these up using any index number formula.

Most scanner data sets also contain short item descriptions. As we argued in section 2.1, additional information – also referred to as meta-data – may be required for the construction of quality-adjusted price indexes. Ideally, the statistical agency would have the following information at its disposal for each GTIN.

- The time period to which the transactions data pertain.
  Values and quantities sold are usually aggregated across some time period by the data supplier. The aggregated data should be available at least by week so that the statistical agency can do a further aggregation over time according to its own rules.
- An outlet indicator.
  Transaction data is preferably specified at the individual store level, allowing the agency to choose the appropriate aggregation across stores. If the data cannot be supplied for individual stores, a distinction by retail chain (type of outlet) should at least be available, including a distinction between purchases in physical stores and online purchases.
- Descriptive information.[4]
  As mentioned earlier, short textual descriptions (at the GTIN level) are usually part of the scanner data sets. Although these descriptions contain information on the characteristics of the items, it may prove difficult to extract this information in an automated fashion and, more importantly perhaps, the information may not cover all details needed, including brand, content, unit of measurement, color, etc. Thus, information on the most important characteristics, preferably given in a format that enables automated processing, would be worthwhile.

---

[3] See Balk (1998) and *Consumer Price Index Manual: Theory and Practice* (ILO et al., 2004). Bradley (2005) and Triplett (2003) discussed a number of potential problems when dealing with unit values from scanner data.

[4] Most countries in Europe also compile an HICP excluding changes in tax rates (VAT and excise duties). It would therefore be helpful if the required information would be available as well.

- Other relational information.
  It was also mentioned earlier that having a classification of items into groups as used by the retailer (or market research company) for its own purposes might be useful to increase homogeneity of product categories. Internal product codes, or "Stock Keeping Units" (SKUs), can also be beneficial.

Some of the above points need further elucidation. Consumers may return items to a store. Turnover and quantities sold should therefore include returns. An important point of attention is that retailers may process item returns in the week in which they are returned, which may differ from the week in which items were purchased. Prices in the two weeks may be different. In order to compute correct transaction prices, turnover and quantities sold that apply to returned items should be included separately in the data.

In most countries, the CPI is published as a monthly statistic.[5] Strict publication deadlines apply, which have consequences for the timeliness of data deliveries. Weekly data deliveries offer clear advantages in this respect over data that is aggregated by the supplier across several weeks and supplied by the end of a month. Statistical agencies themselves can convert weekly prices into monthly prices by time aggregation. Due to timeliness constraints, less than four weeks of data are generally used for this purpose.[6] The number of weeks differs across countries, e.g., midweek in Norway, the first 14 days in Switzerland and the first three full weeks in The Netherlands.

A distinction of transaction data by store, or type of store, is relevant because the service offered may differ across stores in terms of opening hours, terms and conditions for returns and reimbursements, etc. Service therefore has to be taken into account when defining homogeneous products.

As was suggested in section 2.1, one way to deal with the "relaunch problem" would be to define items by the most important price-determining characteristics rather than by GTIN. Put differently, different GTINs that share the same set of characteristics are deemed to represent the same (fully comparable) item, and turnover and quantities sold of the various GTINs can be simply aggregated. Statistical agencies may want to ask the data suppliers to provide the information about item characteristics in separate fields; if the information is provided as text strings, some form of "text mining" must be applied, which can be time consuming. The question arises, of course, as to whether the characteristics information in a scanner data set is sufficient to guarantee homogeneity

[5] Australia and New Zealand are the only two OECD countries where the CPI is published at a quarterly frequency.

[6] However, this can lead to bias in the index; see Fox, Diewert and De Haan (2016).

within groups of GTINs. If the data set contains only a limited amount of information on characteristics, statistical agencies should consider complementing the information one way or another.

An alternative and perhaps more efficient approach to reducing the “relaunch problem” and defining homogeneous items might be the use of the retailer’s SKUs, its internal product codes. It is important to carefully study, for each scanner data set, the relationship between SKUs and GTINs. Is it a one-to-one relationship, is there at least one GTIN for each SKU, or do we encounter different SKUs for a single GTIN?

2.3 Properties of methods

It is customary to evaluate the accuracy of a sample statistic by the mean squared error, the sum of its variance and squared bias. This may not be very helpful for assessing the accuracy of a time series of price (and quantity) indexes, in particular when the indexes are based on scanner data. A scanner data set for a product category typically covers the whole population or universe of items so that there is no sampling of items needed. It should be mentioned though that some statistical agencies have implemented scanner data based on small samples.

This is not to say that price indexes from scanner data are without error. A time series of indexes can be deemed biased if its trend differs systematically from the “true” underlying trend. An obvious difficulty is that the “true” trend is unknown. Fortunately, index number theory provides some guidance as to the appropriate measurement target. Nevertheless, as will be explained in section 4, many choices still have to be made, and it is therefore not always immediately clear what the appropriate target or benchmark price index is.

Volatility can also be seen as an aspect of quality of a time series. Price indexes from scanner-data can be quite volatile because of temporary price reductions in case of promotional sales. But the volatility due to sales prices reflects a genuine phenomenon and should not be smoothed out. More generally, volatility of a price index series at the product level is much less important than bias since we expect volatility to be reduced considerably when aggregating across products. So it is different types of bias we will mainly focus on to characterize the various methods.

Three potential sources of index bias are: a lack of weighting, high-frequency chaining, and poor adjustment for quality changes. Index number theory suggests that items should be *weighted according to economic importance*. As weighting information (quantities or expenditures) is always available in scanner data, it ought to be used. The two best-known weighted index number formulas are those of Laspeyres and Paasche. Their geometric average defines the Fisher index, which is superlative (Diewert, 1976). Superlative indexes have satisfactory properties (ILO et al., 2004). Another example of a superlative index is the Törnqvist index. In spite of the available detailed weighting information, several statistical agencies have actually implemented scanner data without weighting at the item level.

Standard index number theory is concerned with bilateral comparisons in a static universe of items, where the prices of a fixed (static) set of items in one or more periods are directly compared with the prices in some earlier (fixed) base period. In the static-universe case, the use of a bilateral superlative price index formula is probably the best way to measure aggregate price change. However, scanner data, or other highly detailed transactions data in general, usually shows a big churn in terms of new and disappearing items. Put differently, we are typically faced with a dynamic universe, and this requires special attention.

There are two issues involved. First, it is important to maximize the number of matches in the data set: the matched-model principle ensures that we compare like with like and is the basic idea behind price measurement. So in a dynamic-universe context, some form a chaining is called for to construct the time series. Given the big churn rate often observed in scanner data, high-frequency chaining, in particular period-on-period chaining, seems reasonable. However, period-on-period chaining of weighted indexes, including superlative indexes, can lead to significant drift as a result of non-symmetric effects on quantities sold and expenditure shares before and after storable goods are on sale; see e.g. Ivancic (2007), Ivancic, Diewert and Fox (2011), and De Haan and Van der Grient (2011). There are methods available that are transitive, hence *free from chain drift*, including weighted multilateral indexes. In section 3.2. a number of these methods will be discussed.

The second issue is that adhering to a matched-model approach may not suffice, even in a multilateral framework. This approach assumes that the unmatched new and disappearing items do not contribute to aggregate price change or, in other words, that the implicit price change of new and disappearing items does not differ from that of the matched items. For products that exhibit significant quality changes, for example high-technology goods such as consumer electronics, this is a very restrictive assumption. A proper price index is explicitly *adjusted for quality change*, which requires information on the price-determining characteristics of the items. Typically, hedonic regressions are used, which make these methods model-based and prone to *model error*. This is also true for non-hedonic multilateral regression methods. When only a subset of all relevant characteristics is available, hedonic indexes can still be estimated, but omitted variables bias will occur. It remains to be seen whether the results are acceptable.

It was mentioned earlier that the issue of what constitutes an item is an important one. Readily available item identifiers (or keys) such as GTIN may be too detailed for CPI purposes since different GTINs can represent items which are perfect substitutes from a consumer's perspective (Reinsdorf 1999; De Haan, 2002). Item churn will then be overestimated and matched-model price indexes will be based on fewer matches than desirable. Most importantly, price changes of fully comparable items, i.e. items whose GTIN has changed but which have stayed the same in terms of their characteristics, will not be captured by a method based on exact matching of GTINs (but they will be picked up by hedonic regression methods). There is some evidence that such "disguised" price changes have become more common in recent years.

Another issue concerns the sample period, particularly for multilateral methods where the sample or estimation period is (initially) fixed. Of course in reality the sample period is extended every month or quarter. When estimating multilateral indexes on the extended data set, the results for the earlier periods will differ from those previously estimated (albeit often slightly). Statistical agencies do not accept *revisions* of their CPI. A rolling window approach has been proposed to deal with the problem, but there may be other ways to circumvent revisions. Details will be given in section 4.

3. An overview of methods

3.1 Matched-model indexes and period-on-period chaining

3.1.1 Index number formulas for a static universe

We assume that the transactions at the item level have been aggregated across time (and individual consumers) into monthly or quarterly expenditures, quantities purchased and unit values. The unit values can relate to a single retail store, a chain of stores, such as a supermarket chain, or even more than one chain, although the latter should probably be advised against; see Ivancic and Fox (2013). Because scanner data is often characterized by a big churn in terms of new and disappearing items, at least when items are identified at the most detailed level, i.e. by GTIN, we focus on methods for a dynamic universe. Yet, to set the stage, we briefly describe two standard index number formulas and a less conventional one for the static-universe case.

We denote the fixed set of items belonging to some product category by S and its size (the number of items) by N. The aim is to construct price indexes which compare the base period 0 – the starting period of our index series – to, say, T periods t. The prices (unit values) of each item $i \in S$ in periods 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 "direct" or bilateral ratio of unweighted geometric means or *Jevons price index*

$$P_J^{0,t} = \prod_{i \in S} \left(\frac{p_i^t}{p_i^0} \right)^{1/N} = \frac{\prod_{i \in S} (p_i^t)^{1/N}}{\prod_{i \in S} (p_i^0)^{1/N}} \tag{1}$$

as it satisfies more tests than other unweighted indexes.

Since weighting information is always available in scanner data, the construction of superlative indexes is possible. Our focus is on the Törnqvist rather than the Fisher index because its geometric form simplifies decomposition analyses. In most cases, the two indexes will lead to very similar results. The bilateral *Törnqvist price index* is given by

$$P_T^{0,t} = \prod_{i \in S} \left(\frac{p_i^t}{p_i^0} \right)^{(s_i^0 + s_i^t)/2}, \tag{2}$$

where $s_i^0 = p_i^0 q_i^0 / \sum_{i \in S} p_i^0 q_i^0$ and $s_i^t = p_i^t q_i^t / \sum_{i \in S} p_i^t q_i^t$ denote the expenditure shares in period 0 and period t; q_i^0 and q_i^t are the quantities purchased.

Von Auer (2014) showed that many conventional price index formulas can be interpreted as generalized unit value indexes. A generalized unit value index between periods 0 and t is equal to the value index $V^{0,t} = \sum_{i \in S} p_i^t q_i^t / p_i^0 q_i^0$ divided by a quantity index given by the ratio of standardized quantities. This method requires standardization factors $\lambda_{i/b}$ to express the quantity purchased of each item i, q_i^t $(t = 0,...,T)$, in terms of a quantity of an arbitrary base item b. The standardized quantities may be added up, and $Q^{0,t} = \sum_{i \in S} \lambda_{i/b} q_i^t / \sum_{i \in S} \lambda_{i/b} q_i^0$ is a useful measure of quantity change. The *generalized unit value index* is defined as

$$P_{GUV}^{0,t} = \frac{V^{0,t}}{Q^{0,t}} = \frac{\sum_{i \in S} p_i^t q_i^t \Big/ \sum_{i \in S} p_i^0 q_i^0}{\sum_{i \in S} \lambda_{i/b} q_i^t \Big/ \sum_{i \in S} \lambda_{i/b} q_i^0} = \frac{\left[\sum_{i \in S} s_i^t (\tilde{p}_i^t)^{-1} \right]^{-1}}{\left[\sum_{i \in S} s_i^0 (\tilde{p}_i^0)^{-1} \right]^{-1}}. \tag{3}$$

Note that if the product category is perfectly homogeneous, then $\lambda_{i/b} = 1$ for all i and (3) simplifies to the ordinary unit value index. For later use, the last expression of (3) writes the generalized unit value index as the ratio of harmonic averages of 'quality-adjusted prices' defined as $\tilde{p}_i^0 = p_i^0 / \lambda_{i/b}$ and $\tilde{p}_i^t = p_i^t / \lambda_{i/b}$.

While the standardization factors are preferably estimated using information on both prices and item characteristics, they can be estimated using prices only. We could assume that relative prices for broadly comparable items reflect quality differences. For example, setting $\lambda_{i/b} = p_i^0 / p_b^0$ turns the generalized unit value index into the Paasche price index; the corresponding quantity measure then is the Laspeyres quantity index. A disadvantage of the generalized unit value approach is that, except for $\lambda_{i/b} = p_i^0 / p_b^0$, the index violates the identity test (but the quantity index does satisfy this test). This test, or axiom, requires a price index to equal 1 if the prices of all items in period t are equal to those in period 0. It can be argued, however, that this axiom is not relevant here because standardization makes the 'composite item' fully homogeneous.

The generalized unit value index and the Jevons index are *transitive*: the results of price comparisons between any time periods are independent of the choice of base period, which in our case is the starting period 0. Transitivity implies that an index can be expressed as a period-on-period chained index. Like other superlative indexes, the Törnqvist index is not transitive. In the static-universe case, a violation of transitivity is not overly important since bilateral indexes can be used and chaining is not required. Transitivity is important for (weighted) indexes in a dynamic-universe context, to which we now turn.

3.1.2 Period-on-period chaining

We denote the sets of items in period t $(t = 0,...,T)$ by S^t, with size N^t. Index number methods for a dynamic universe can be classified as to whether or not they explicitly take into account new and disappearing items using information on item characteristics. Methods that do are referred to as *quality-adjusted methods* and will be discussed in section 3.3. Methods that do not explicitly account for unmatched new and disappearing items are referred to as *matched-model methods*.

Given the high rate of item churn often encountered in scanner data, at least at the GTIN level, maximizing the number of matches is useful. The CPI Manual (ILO et al., 2004) recommends chaining matched-model superlative price indexes, e.g. chaining period-on-period Törnqvist price indexes

$$P_T^{t-1,t} = \prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}} \right)^{\frac{s_{iM}^{t-1} + s_{iM}^t}{2}}, \tag{4}$$

where s_{iM}^{t-1} and s_{iM}^t are the expenditure shares in the two periods with respect to the set $S_M^{t-1,t} = S^{t-1} \cap S^t$ of matched set of items available in period t-1 and period t. However, as we know now, high-frequency chaining of weighted price indexes should be avoided as that often leads to a drifting time series.

Because it is systematic patterns in the weights that cause the problem, a simple way to avoid *chain drift* would be not to weight and construct a time series by chaining period-on-period matched-model Jevons indexes

$$P_J^{t-1,t} = \prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}} \right)^{\frac{1}{N_M^{t-1,t}}}, \tag{5}$$

where $N_M^{t-1,t}$ is the corresponding number of matched items. From a theoretical point of view, the lack of weighting is problematic. A solution to the chain-drift issue might be the construction of weighted transitive multilateral price indexes, which are free from chain drift by definition.

3.2 Multilateral indexes: transitivizing bilateral indexes

Multilateral price index methods are typically applied to compare price levels across countries or regions to make the comparisons transitive. Transitivity is a very desirable property for spatial comparisons because the results will be independent of the choice of base country. Well-known methods are the GEKS (Gini, 1931; Eltetö and Köves; 1964; Szulc, 1964) method, the Geary-Khamis 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. We distinguish between two types of multilateral methods. The first type, which is discussed in the present section, starts from a set of matched-model bilateral indexes, i.e. price comparisons between pairs of time periods, and then "transitivize" the bilateral price comparisons. The second type of multilateral methods, described in section 3.3, attain transitivity in some other way.

3.2.1 *GEKS and weighted GEKS*

The intertemporal GEKS method relies on all the matches in the data across the entire sample period $0,...,T$ (instead of only adjacent-period matches). The GEKS price index between period 0 and period *t* is calculated as the geometric average of the ratios of the matched-model bilateral price indexes $P^{j,l}$ and $P^{k,l}$, constructed with the same index number formula, where each period *l* is taken as the base. In case the bilateral indexes satisfy the time reversal test, the GEKS index can be written as (see Ivancic, Diewert and Fox, 2011 or De Haan and Van der Grient, 2011)[7]

$$P_{GEKS}^{0,t} = \prod_{l=0}^{T} \left[\frac{P^{0,l}}{P^{t,l}} \right]^{\frac{1}{T+1}} = \prod_{l=0}^{T} \left[P^{0,l} \times P^{l,t} \right]^{\frac{1}{T+1}} . \qquad (6)$$

The time reversal test requires that when the base period and the comparison period are reversed, the index should be the reciprocal of the original index. In its standard form, the GEKS method uses bilateral Fisher indexes as inputs, but other choices are possible as well, including bilateral Törnqvist indexes. The window length is a point of concern as different choices will give different answers. Ivancic, Diewert and Fox (2011) argue that a 13-month window may be optimal because it is the shortest window that can deal with strongly seasonal goods. Enlarging the window would increasingly lead to a loss of characteristicity, i.e. short-term price movements will be increasingly affected by price changes in the distant past.

It is possible to formulate weighted GEKS indexes, which may take into account the reliability of the bilateral price indexes; see e.g. Rao (1999) (2001). Melser (2015), proposed a weighted GEKS approach where the weights are dependent on the degree of matching of the items, for example in terms of their expenditure shares. In this case, the choice of window length may not be a big issue after all because less reliable bilateral indexes – those with a lower degree of matching – will be suitably down-weighted. This potentially enables the use of a longer window.

3.2.2 *Methods based on spanning trees*

Hill (1999a,b) developed methods for spatial price comparisons using spanning trees.[8] A spanning tree is a concept from graph theory. For background material on some basic

[7] According to Lamboray and Krsinich (2015), the GEKS method suffers from a form of asymmetry and they proposed a modified version, referred to as the intersection GEKS index. A potential problem with their approach is that the intersection GEKS index is not transitive while transitivity is the main reason for using a multilateral index number method for a dynamic universe.

[8] Balk (2008, section 7) provides a useful outline of Hill's spanning tree approaches.

concepts and notions from graph theory, see for example Balakrishnan (1997) or Van Steen (2010). The key-idea of Hill's approach is that a spanning tree is a supplier of paths which is used to calculate price indexes between two states (two countries, or in our context two time periods). Which spanning tree actually to use is, to a large degree, a matter of taste. Hill proposes two methods to find optimal spanning trees, but there are other ways to do so.

The essence of the methods is that they use chaining of certain bilateral indexes, which are associated with the spanning tree chosen,[9] to calculate a price index between any pair of states. If there are n different states, $n-1$ bilateral indexes are required. If two states, say v and w, are compared, the spanning tree provides the (unique) path connecting v and w, say $v = v_0, v_1, \dots, v_k = w$. Hence, a chain index $\prod_{a=0}^{k-1} P^{v_a,v_{a+1}}$ can be calculated, where $P^{v_a,v_{a+1}}$ is the bilateral index between the ordered pair of states (v_a, v_{a+1}). This bilateral index is an arc on the (oriented) spanning tree.

The results depend on the choice of spanning tree; two different spanning trees generally produce different results. In section 3.2.3 below, we discuss a method which avoids this kind of dependency.

3.2.3 Cycle method

Willenborg (2010) developed the "cycle method" in reaction to the methods discussed in the previous section.[10] The cycle method provides an adjustment method which does not depend on the choice of spanning tree. That is, while the method uses a spanning tree to calculate the adjustments, the actual choice of spanning tree does not matter; any spanning tree would yield the same result.[11]

Assume we have n states and assume they are assembled in a set V. Suppose that for certain ordered pairs of states (v, w) price indexes have been calculated; v denotes the reference state and w the reporting state. We refer to such an ordered pair of states as an arc. Assume there are m such arcs. The set of arcs A together with the set V define a directed graph, or digraph, $G = (V, A)$. In this context we call it a price index digraph

[9] A spanning tree for a graph $G = (V, E)$ where V is the set of points and E the set of edges, which are sets consisting of two points from V, is a subgraph $T = (V', E')$ such that $V' = V$ and $E' \subseteq E$ and it is without any cycles. A cycle in a graph is a closed path in this graph.

[10] The method is based on Willenborg (1993), an unpublished paper which was inspired by problems in a totally different area, namely land surveying.

[11] For an application to price indexes based on data from web shops, see Willenborg (2015a,b). Willenborg and Van der Loo (2016) is another application of the cycle method to clothing data collected from a web shop.

(PIDG).[12] For arc $j = (v, w)$, let $P^j = P^{v,w}$ denote the corresponding price index. It is understood that if $P^j = P^{v,w}$ is the price index associated with arc j, then $\left(P^j\right)^{-1} = (P^{v,w})^{-1} = P^{w,v}$ is associated with the reversed arc. Let $y_j = \ln\left(P^j\right) = \ln(P^{v,w})$ be the log price index of $P^j = P^{v,w}$, where $j = (v, w)$ is an ordered pair of states. Let y denote the vector with the log price indexes as components. We assume that the initial PIDG G is not transitive, and that the cycle method is used to produce a transitive adjustment. We also assume that an $m \times m$ weight matrix $\boldsymbol{W}$, to be specified by the practitioner, is available which is nonsingular, diagonal, and nonnegative. The j-th diagonal element of $\boldsymbol{W}$, i.e. W_{jj}, denotes the weight associated with arc j of G. $\boldsymbol{W}$ regulates how much the price indexes associated with the respective arcs in G can be perturbed.

The PIDG G is supposed to have at least one cycle (otherwise there would be no problem with transitivity). To calculate a cycle matrix for G, some spanning tree for G should be chosen. Willenborg (2010) explains how this can be done efficiently. Let $\boldsymbol{C}$ denote the resulting cycle matrix. It is a matrix of order $(m - n + 1) \times m$, where m is the number of edges in G and n the number of its points. $\boldsymbol{C}$ is actually a (-1, 0, 1) matrix. If $\boldsymbol{C} = \left(c_{ij}\right)$, i indexes the elementary cycles and j the arcs. Suppose that $j = (v, w)$, then $c_{ij} = 0$ if arc j is not an arc on cycle i, $c_{ij} = 1$if j is part of cycle i, and $c_{ij} = -1$ if the reverse of arc j, i.e. (w, v), is part of cycle i.

Let $\hat{x}$ denote an adjustment of y that satisfies the cycle condition $\boldsymbol{C}\hat{x} = 0$. We assume that $\hat{x}$ is obtained by minimising the expression $(x - y)'\boldsymbol{W}^{-1}(x - y)$ under the condition $\boldsymbol{C}x = 0$. The result can be calculated using the Lagrangian multiplier method.[13] In our case we have:

$$\hat{x} = y - \boldsymbol{W}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}')^{-1}\boldsymbol{C}y = (\boldsymbol{I}_m - \boldsymbol{W}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}')^{-1}\boldsymbol{C})y = \boldsymbol{P}y, \tag{7}$$

where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix and with $\boldsymbol{P} = \boldsymbol{I}_m - \boldsymbol{W}\boldsymbol{C}'(\boldsymbol{C}\boldsymbol{W}\boldsymbol{C}')^{-1}\boldsymbol{C}$, which is an $m \times m$ matrix. Note that $\boldsymbol{C}\boldsymbol{P} = 0$, which implies that $\boldsymbol{C}\hat{x} = 0$, as required. It can be proved that the adjustment $\hat{x}$ in (7) does not depend on the choice of spanning tree that is used to calculate the cycle matrix.

The above analysis is in terms of log indexes. By exponentiating the components of $\hat{x}$ one finds the adjusted price indexes. They have the property that the product of all the adjusted indexes belonging to an arc on any cycle in the PIDG equals 1. From this it

---

[12] In fact it is more adequate to call the triple (V, A, J), where J is a vector of price indexes, or of their logarithms, a PIDG (see below).

[13] In a statistical context we would be dealing with a Restricted Generalized Least Squares (RGLS) estimator. Our setting is deterministic rather than stochastic because there are no observations with errors, yielding random variables, but the results are the same as in a stochastic setting.

follows that if one wants to calculate the price index between a base point in the PIDG and a reference point, one can take any path in the PIDG and multiply the price indexes corresponding to the arcs on the path. The result is independent of the path chosen, and this is exactly the purpose of the cycle method.

Just like with weighted GEKS, the cycle method can give different weights, via the weight matrix $\boldsymbol{W}$, to the indexes for the different pairs of states. The cycle method is more general than (weighted) GEKS, however. In particular, it does not require indexes to be available for all ordered pairs of states, i.e. the PIDG does not necessarily have to be complete.[14] Moreover, the initial price indexes need not satisfy the time reversal test. The choice of window length is not a problem either when applying the cycle method; it is up to the statistician to decide what arcs to take into account. If she wants to consider only pairs of months of not more than, say, 12 months apart, this would be fine. The PIDG will then simply have less arcs than the maximum possible number of months in the time window.

3.3 Other multilateral approaches

3.3.1 Geary-Khamis method

The Geary-Khamis (GK) method, when applied to comparisons over time, gives rise to the following price index:

$$P_{GK}^{0,t} = \frac{\sum_{i \in S^t} p_i^t q_i^t \Big/ \sum_{i \in S^t} \widehat{p}_i q_i^t}{\sum_{i \in S^0} p_i^0 q_i^0 \Big/ \sum_{i \in S^0} \widehat{p}_i q_i^0} = \frac{\left[\sum_{i \in S^t} s_i^t \left(\frac{p_i^t}{\widehat{p}_i} \right)^{-1} \right]^{-1}}{\left[\sum_{i \in S^0} s_i^0 \left(\frac{p_i^0}{\widehat{p}_i} \right)^{-1} \right]^{-1}}. \tag{8}$$

The numerator of (8) is a price index with “reference prices” $\widehat{p}_i$ (usually referred to as international prices in the spatial context). We require the value to be 1 in the starting period 0, so it will be necessary to normalize the index by dividing by its period 0 value, the numerator of (8). The reference prices in (8) are calculated as follows:

$$\widehat{p}_i = \frac{\sum_{\tau \in S_i} q_i^\tau \left(\frac{p_i^\tau}{P_{GK}^{0,\tau}} \right)}{\sum_{\tau \in S_i} q_i^\tau}, \tag{9}$$

[14] If there are n states, the PIDG has a maximum of n^2 ordered pairs, or arcs in the corresponding maximum PIDG.

where S_i is the set of time periods in which item i is actually sold and for which prices are available. Equation (9) shows that $\widehat{p}_i$ is equal to a weighted arithmetic average of the deflated observed prices, with each period's share in the total number of sales of the item (across the entire sample period) serving as weights. Since the GK index itself acts as the deflator, equations (8) and (9) must be solved simultaneously. This can be done iteratively or as the solution to an eigenvalue problem (Diewert, 1999).

The GK index can alternatively be written as

$$P_{GK}^{0,t} = \frac{\sum_{i \in S^t} p_i^t q_i^t \Big/ \sum_{i \in S^0} p_i^0 q_i^0}{\sum_{i \in S^t} \widehat{p}_i q_i^t \Big/ \sum_{i \in S^0} \widehat{p}_i q_i^0}. \qquad (10)$$

i.e., as the value index $\sum_{i \in S^t} p_i^t q_i^t / \sum_{i \in S^0} p_i^0 q_i^0$ divided by $\sum_{i \in S^t} \widehat{p}_i^0 q_i^t / \sum_{i \in S^0} \widehat{p}_i^0 q_i^0$. The latter should be viewed as a quantity index if we require the product of the (GK) price index and the quantity index to be equal to the value index. Expression (10) makes clear why the GK price index is transitive: both the value index and the quantity index are transitive. Transitivity of the quantity index holds because the index can be expressed as a period-on-period chained Lowe-type index

$$\frac{\sum_{i \in S^t} \widehat{p}_i q_i^t}{\sum_{i \in S^0} \widehat{p}_i q_i^0} = \frac{\sum_{i \in S^1} \widehat{p}_i q_i^1}{\sum_{i \in S^0} \widehat{p}_i q_i^0} \times \frac{\sum_{i \in S^2} \widehat{p}_i q_i^2}{\sum_{i \in S^1} \widehat{p}_i q_i^1} \times .... \times \frac{\sum_{i \in S^t} \widehat{p}_i q_i^t}{\sum_{i \in S^{t-1}} \widehat{p}_i q_i^{t-1}}, \qquad (11)$$

with reference prices $\widehat{p}_i^0$ that are fixed across the sample period $t = 0,...,T$.

The GK method has been criticized in spatial comparisons for giving too much weight to very large countries (in terms of quantities sold). This so-called Gerschenkron effect resembles substitution bias in the intertemporal context. Alternative weighting schemes have therefore been proposed, using the items' turnover (expenditure) shares in the periods they are sold, which likely mitigates the impact of large countries, or in our case time periods, with many sales. Also, instead of an arithmetic average, a harmonic average of the deflated prices could be chosen. Iklé (1972) proposed the expenditure-share weighted harmonic average as an alternative to (9):

$$\widehat{p}_i = \left[\sum_{\tau \in S_i} s_i^\tau \left(\frac{p_i^\tau}{P_{IKL}^{0,\tau}} \right)^{-1} \right]^{-1}, \qquad (12)$$

where $P_{IKL}^{0,\tau}$ is defined as the second expression on the right-hand side of (8) and thus has the same form as the GK index.

3.3.2 Time-product dummy method

The Country Product Dummy (CPD) method known from spatial price comparisons is a regression-based approach. The version for temporal comparisons was referred to by De Haan and Krsinich (2014) as the Time Product Dummy (TPD) method. Assuming there are N different items observed during the sample period, most of which will typically not be purchased in every time period, the regression model for the pooled data of all periods $0,...,T$ can be written as

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t , \tag{13}$$

where D_i is a dummy variable that has the value of 1 if the observation relates to item i and 0 otherwise (a dummy for item N is excluded to identify the model), and D_i^t is a time dummy that has the value 1 if the observation pertains to period t and 0 otherwise. The parameters γ_i are item fixed effects. Exponentiating the estimated time dummy parameter $\hat{\delta}^t$ yields the TPD index going from period 0 to period t, i.e. $P_{TPD}^{0,t} = \exp(\hat{\delta}^t)$. The TPD index is transitive because the regression results are independent of the choice of base period (which in this case is the starting period 0).

Following Diewert's (2004) proposal, we assume that model (13) is estimated by Weighted Least Squares (WLS) regression with expenditure shares as weights. It turns out that the TPD method can then be written as a system of equations which is similar to the Iklé system, the main difference being that geometric averages rather than harmonic averages are taken; see also Rao (1990) (2005). A useful way of writing the TPD index is (De Haan and Hendriks, 2013)

$$P_{TPD}^{0,t} = \frac{\prod_{i \in S^t} \left(\dfrac{p_i^t}{\exp(\hat{\gamma}_i)} \right)^{s_i^t}}{\prod_{i \in S^0} \left(\dfrac{p_i^0}{\exp(\hat{\gamma}_i)} \right)^{s_i^0}} , \tag{14}$$

and it can be shown that the exponentiated fixed effects estimates are equal to

$$\exp(\hat{\gamma}_i) = \prod_{\tau \in S_i} \left(\frac{p_i^\tau}{P_{TPD}^{0,\tau}} \right)^{\frac{s_i^\tau}{\sum_{\tau \in S_i} s_i^\tau}} \tag{15}$$

That is, we can interpret the $\exp(\hat{\gamma}_i)$ as reference prices, which in this case are equal to expenditure-share weighted geometric averages of the deflated prices, where the time-product dummy index now serves as deflator.[15] Given the similarity of the TPD system and the Iklé system, we would expect them to yield quite similar results.

GK, Iklé and TPD are essentially matched-model methods; to affect the price index, an item should be observed at least twice during the sample period. This is easy to understand: an item can only contribute to aggregate price change if a price relative is available for that item. Put otherwise, items with a single observation during the sample period are zeroed out. One implication is that items which are new in the most recent period T are ignored.

3.3.3 Discussion

It should be noted that the usefulness of matched-model multilateral methods, including GEKS, may depend on how items are identified, as pointed out in section 2.2. Using GTINs to identify items, "disguised" price changes of homogenous products with changing GTINs – or "relaunches" as we called them – will not be noticed. This was the main reason why Greenlees and McClelland (2010) and De Haan and Hendriks (2013) found severe downward biases in, respectively, GEKS and (unweighted) TPD indexes for apparel. Downward bias would probably also result for strongly seasonal products if "the same" items had different GTINs in different years. After adjusting for relaunches, multilateral index number methods are perfectly suitable for treating strongly seasonal products; for early applications, see Balk (1980) (1981).

One property of all weighted multilateral methods should be mentioned here: they violate the identity test. That is, in the static-universe case, where there are no new or disappearing items across the sample period, the price index going from period 0 to period t will in general not be equal to 1 when the prices of all items stay the same. We do not worry too much about this failure. Transitivity, or more generally, a properly measured trend, is of greater importance in a time series context.

[15] We can interpret the reference prices in both the TPD, Iklé and GK methods as estimates of the period 0 prices because the period t prices are deflated and expressed "in period 0 values". Hence, the numerator of equation (8) for the GK and Iklé indexes can be viewed as a Paasche-type index where the base period prices for all items sold in period t are imputed according to (9) or (12). Similarly, the TPD index can be viewed as a geometric Paasche-type index with imputed period 0 prices based on (15); the exponentiated fixed effects are equal to the predicted period 0 prices, and the denominator of (14) is equal to 1 since the regression residuals sum to zero in each period. Note that when all items match across the sample period, these multilateral indexes will differ from the true (non-transitive) Paasche or geometric Paasche indexes.

3.4 Quality-adjusted methods

Hedonic regression has become the default approach to adjusting for quality change. It is probably also the only explicit quality-adjustment approach that can be automated to a considerable degree, which is a prerequisite for feasible treatment of big data. So if a method explicitly controls for quality change, we assume that hedonic adjustments have been performed. In this section, we will discuss weighted methods only.

3.4.1 Hedonic imputation indexes

The set of items sold in the base period, S^0, can be split into a set of matched items which are still sold in period *t*, S_M^{0t}, and a set of unmatched disappearing items S_D^{0t} which are no longer available. Likewise, the period set of items, S^t, can be split into S_M^{0t} and a set of unmatched new items, S_N^{0t}, which were not sold in the base period. From an index number perspective, the period *t* prices for all $i \in S_D^{0t}$ and the period 0 prices for all $i \in S_N^{0t}$ are 'missing' and have to be imputed by $\hat{p}_i^t$ and $\hat{p}_i^0$. As mentioned above, we assume that the imputations are based on hedonic regressions. The (single) *hedonic imputation Törnqvist price index* reads[16]

$$P_{IT}^{0,t} = \prod_{i \in S_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{(s_i^0 + s_i^t)/2} \prod_{i \in S_D^{0t}} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{s_i^0/2} \prod_{i \in S_N^{0t}} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{s_i^t/2} . \tag{16}$$

The natural approach would be to run hedonic regressions for each time period separately to estimate the imputed values, for instance using a log-linear model:

$$\ln p_i^t = \alpha^t + \sum_{k=1}^{K} \beta_k^t z_{ik} + \varepsilon_i^t , \tag{17}$$

where z_{ik} denotes (the quantity of) characteristic *k* for item *i* and β_k^t the corresponding parameter; the errors ε_i^t are independently distributed with zero mean. To save degrees of freedom, a pooled regression might be considered. In the present bilateral case, the following model could be estimated on the pooled data of the item sets S^0 and S^t in the two periods 0 and *t*:

$$\ln p_i^t = \alpha + \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t , \tag{18}$$

[16] A double imputation approach, where observed prices for unmatched items are replaced by estimated prices as well, might reduce bias due to omitted hedonic variables (Syed, 2010). Feenstra (1995) proposed "exact" hedonic superlative indexes where all prices, including those for the matched items, are estimated using hedonic regression; for an empirical application, see Ioannides and Silver (2003).

where the time dummy variable D_i^t has the value 1 if the observation pertains to period t and 0 otherwise; the parameters for the characteristics are now constrained to be fixed across time. The quality-adjusted bilateral time dummy hedonic index between period 0 and period t is found by exponentiating the estimated time dummy parameter, i.e. by $\exp(\hat{\delta}^t)$. De Haan (2004) proposed the use of weights $(s_i^0 + s_i^t)/2$ for $i \in S_M^{0t}$, $s_i^0/2$ for $i \in S_D^{0t}$ and $s_i^t/2$ for $i \in S_N^{0t}$ in a WLS regression to estimate model (18). This yields an imputation Törnqvist price index (16) where the imputed prices are based on the pooled regression. Notice that the regression weights are equal to the item weights in (16).

A disadvantage of a bilateral hedonic imputation price index such as (16) is that *i*) unlike multilateral indexes, it does not make use of all the matches in the data across the sample period, and *ii*) due to item churn, the index will increasingly rely on model-based imputations as the percentage of matched items decreases over time. Thus, some form of low-frequency chaining, e.g. annual linking in December of each year, cannot be avoided. Again, period-on-period chaining should be avoided due to potential chain drift in weighted indexes.

3.4.2 Multilateral hedonic indexes

One approach to incorporating new and disappearing items in a multilateral context is the use of bilateral hedonic imputation rather than matched-model indexes as inputs in the GEKS index. De Haan's (2004b) proposal for estimating weighted bilateral time dummy hedonic indexes could be used. The resulting Imputation Törnqvist GEKS price index is quality-adjusted, transitive and preserves all the matches in the data across the sample period $0,...,T$.

An alternative, and easier to implement, method for estimating quality-adjusted multilateral price indexes is the multi-period weighted time dummy hedonic approach. Instead of using data for just two periods, the log-linear hedonic model

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t, \qquad (19)$$

is estimated on the pooled data of all periods; the time dummy D_i^t has the value 1 if the observation pertains to period t $(t = 0,..,T)$ and 0 otherwise. The resulting time dummy hedonic (TDH) index, $P_{TDH}^{0,t} = \exp(\hat{\delta}^t)$, is transitive because the regression results do not depend on the choice of base period, similar to the pooled time-product dummy model given by (13).[17]

[17] Due to the non-linear transformation, the time dummy indexes are not unbiased. Kennedy (1981) and Van Garderen and Shah (2002) discussed bias-correction terms. Usually, the bias will be small and can be

If a WLS regression is run with expenditure shares as weights, the TDH index can be written as (De Haan and Krsinich, 2014a)

$$P_{TDH}^{0,t} = \frac{\prod_{i \in S^t} (p_i^t)^{s_i^t}}{\prod_{i \in S^0} (p_i^0)^{s_i^0}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k \left(\sum_{i \in S^0} s_i^0 z_{ik} - \sum_{i \in S^t} s_i^t z_{ik} \right)\right] = \frac{\prod_{i \in S^t} (\hat{\tilde{p}}_i^t)^{s_i^t}}{\prod_{i \in S^0} (\hat{\tilde{p}}_i^0)^{s_i^0}} . \tag{20}$$

The exponentiated factor in (20) adjusts the ratio of weighted geometric average prices for changes in the weighted average characteristics. The second expression on the right-hand side of (20) writes the index as the ratio of expenditure-share weighted geometric averages of estimated quality-adjusted prices, defined as $\hat{\tilde{p}}_i^0 = p_i^0 / \exp[\sum_{k=1}^{K} \hat{\beta}_k z_{ik}]$ and $\hat{\tilde{p}}_i^t = p_i^t / \exp[\sum_{k=1}^{K} \hat{\beta}_k z_{ik}]$. By comparing the items i with an arbitrary base item b, the quality-adjusted prices are written as $\hat{\tilde{p}}_i^0 = p_i^0 / (\hat{p}_i^0 / \hat{p}_b^0) = p_i^0 / \exp[\sum_{k=1}^{K} \hat{\beta}_k (z_{ik} - z_{bk})]$ and $\hat{\tilde{p}}_i^t = p_i^t / (\hat{p}_i^t / \hat{p}_b^t) = p_i^t / \exp[\sum_{k=1}^{K} \hat{\beta}_k (z_{ik} - z_{bk})]$.

3.4.3 Quality-adjusted unit value indexes

It is straightforward to adapt the generalized unit value index (3) to a dynamic universe. Assuming we may add up standardized quantities, the quantity index defined on S^0 and S^t becomes $Q^{0t} = \sum_{i \in S^t} \lambda_{i/b} q_i^t / \sum_{i \in S^0} \lambda_{i/b} q_i^0$, and the dynamic counterpart of (3) is

$$P_{QAUV}^{0,t} = \frac{\sum_{i \in S^t} p_i^t q_i^t \Big/ \sum_{i \in S^0} p_i^0 q_i^0}{\sum_{i \in S^t} \lambda_{i/b} q_i^t \Big/ \sum_{i \in S^0} \lambda_{i/b} q_i^0} = \frac{\left[\sum_{i \in S^t} s_i^t (\tilde{p}_i^t)^{-1} \right]^{-1}}{\left[\sum_{i \in S^0} s_i^0 (\tilde{p}_i^0)^{-1} \right]^{-1}} , \tag{21}$$

where $\tilde{p}_i^0 = p_i^0 / \lambda_{i/b}$ and $\tilde{p}_i^t = p_i^t / \lambda_{i/b}$ are quality-adjusted prices, as before, and where b is again an arbitrary base item. De Haan (2004b) referred to (21) as a quality-adjusted (rather than generalized) unit value index because the effects of new and disappearing items are now explicitly taken into account.

Provided that information on item characteristics is available, the standardization or quality-adjustment factors $\lambda_{i/b}$ in (21) can be estimated using hedonic regression. De Haan (2015a) proposed the use of the weighted pooled time dummy hedonic regression, which is equivalent to using the quality-adjusted prices $\hat{\tilde{p}}_i^0$ and $\hat{\tilde{p}}_i^t$ defined below (20) as estimates of $\tilde{p}_i^0$ and $\tilde{p}_i^t$ in (21). This may seem a bit strange as we would now have two different measures of quality-adjusted price change: the time dummy hedonic index

ignored. For a comparison of time dummy hedonic and hedonic imputation indexes, see Diewert, Heravi and Silver (2009) and De Haan (2010).

(20) and an quality-adjusted unit value index, based on the same quality-adjusted prices. How should we choose between the two measures?

The quality-adjusted unit value index simplifies to the ordinary unit value index if the product category is perfectly homogeneous, i.e. if all the items have the same set of characteristics, or $\lambda_{i/b} = 1$ for all *I*. This property of the quality-adjusted unit value is certainly useful. The practical relevance may be limited though, for two reasons. First, product categories (in scanner data) are seldom perfectly homogeneous. Second, the two methods are surprisingly similar: the time dummy hedonic index is equal to the ratio of weighted geometric means of quality-adjusted prices and the quality-adjusted unit value index is equal to the ratio of weighted harmonic means of (the same) quality-adjusted prices. So we would expect them to produce similar results.

The last point can be illustrated by the following relation:

$$P_{QAUV}^{0,t} = P_{TDH}^{0,t} \left[\frac{\sum_{i \in S^0} s_i^0 \exp(u_i^0)}{\sum_{i \in S^t} s_i^t \exp(u_i^t)} \right] \cong P_{TDH}^{0,t} \left[\frac{1 + \frac{1}{2}(\sigma^0)^2}{1 + \frac{1}{2}(\sigma^t)^2} \right]. \tag{22}$$

Equation (22) says that the quality-adjusted unit value index (based on quality-adjusted prices derived from the pooled hedonic regression) is equal to the time dummy hedonic index multiplied by a factor (in square brackets), which is the ratio of expenditure-share weighted means of the exponentiated residuals $u_i^0 = \ln(\hat{p}_i^0 / p_i^0)$ and $u_i^t = \ln(\hat{p}_i^t / p_i^t)$ in periods 0 and *t*. In other words, the bracketed factor changes the time dummy hedonic index into a quality-adjusted unit value index. The second expression of equation (22) is an approximation that depends upon the expenditure-share weighted variances, $(\sigma^0)^2$ and $(\sigma^t)^2$, of the regression residuals in periods 0 and *t*. De Haan and Krsinich (2014a) argued (and showed empirically) that this weighted residual variance will be stable over time so that the two methods produce similar results.

3.4.4 Implicit quality adjustment

Aizcorbe, Corrado and Doms (2003) as well as Krsinich (2014) pointed to the fact that the TPD model (13) arises from the time dummy hedonic model (19) by replacing the hedonic effects $\sum_{k=1}^{K} \beta_k z_{ik}$ by item-specific fixed effects γ_i. They argue that the TPD method should therefore control for quality changes. A similar argument could be made for the quality-adjusted unit value index based on the "quality-adjusted" prices from the TPD model and, although not regression-based, also for the Iklé and GK indexes.

Since no information on characteristics of items is used, the quality adjustment is implicit. To obtain truly quality-adjusted price indexes, the relative "reference prices" (the exponentiated fixed effects in case of the TPD index) should act as proper quality-adjustment factors. Further research is needed to explore how the implicit adjustment works and to describe the necessary conditions under which it will produce satisfactory results. The quality adjustment is also partial because items sold in one period only are ignored, including newly introduced items in the last period *T*. Depending on how the time series is updated, this may not be such an important issue (see section 4 below).

4. Dealing with revisions in multilateral indexes

A problem with the multilateral price indexes discussed in section 3 is that they suffer from revisions. When the sample period is extended, data for period $T+1$ is added and price indexes for the new sample period $0,...,T+1$ are estimated, previously estimated indexes will change. Statistical agencies do typically not accept such revisions. In this section, we will outline four methods for dealing with the revisions problem. Two of them are based on a rolling window approach, and we discuss these first.

4.1 Rolling window approaches

With a *rolling window approach*, the estimation window is shifted forward one period (keeping the length fixed at $T+1$ periods), and the indexes are re-estimated on the data of periods $1,...,T+1$. We will discuss two ways of extending the time series for periods $1,...,T$ to period $T+1$: standard splicing method, which we also refer to as movement splicing, and window splicing proposed by Krsinich (2014). We borrow heavily from De Haan (2015b) to illustrate the two methods.

Suppose the length of the estimation window is 13 months. The *movement splice* works as follows: after moving forward the window one month and re-estimating the indexes, the most recent estimated month-on-month movement of the index is spliced on to the existing time series. The *window splice* splices the entire newly estimated 13-month series on to the index level of 12 months ago. To explain this, we introduce some additional notation. In particular, we will use (x) for results coming from the estimation window starting in period x. For example, $P^{0,t}(0)$ denotes the multilateral price index (constructed according to one of the methods discussed earlier) going from period 0 to period *t*, estimated on the data of the sample period $0,...,T$. After moving forward the estimation window by one period, the price index between periods 1 and t is denoted by $P^{1,t}(1)$.

The standard *movement splice* extends the existing time series $P^{0,1}(0)....P^{0,T}(0)$ by multiplying $P^{0,T}(0)$ by the movement $P^{1,T+1}(1)/P^{1,T}(1)$. Thus, the price index with a movement splice (MS) pertaining to the 'new' period $T+1$ and index reference period 0 is calculated as

$$P_{MS}^{0,T+1} = P^{0,T}(0) \times \frac{P^{1,T+1}(1)}{P^{1,T}(1)} = P^{0,T}(0) \times P^{T,T+1}(1) = P^{0,1}(0) \times P^{1,T}(0) \times P^{T,T+1}(1), \qquad (23)$$

using the transitivity property of multilateral indexes. Indexes with a movement splice are also known as rolling year indexes. This name is a bit ambiguous because window splicing is based on a rolling window approach as well.

The *window splice* method extends the time series by multiplying the level of the index for period 1, $P^{0,1}(0)$, by the index going from period 1 to period $T+1$, $P^{1,T+1}(1)$, based on the new estimation window. Hence, the index with a window splice (WS) for period $T+1$ with index reference period 0 is calculated as

$$P_{WS}^{0,T+1} = P^{0,1}(0) \times P^{1,T+1}(1) = P^{0,1}(0) \times P^{1,T}(1) \times P^{T,T+1}(1). \qquad (24)$$

The difference between $P_{MS}^{0,T+1}$, given by (21), and $P_{WS}^{0,T+1}$, given by (24), is the use of $P^{1,T}(0)$ rather than $P^{1,T}(1)$ in the decomposition. That is, the standard splice measures the price change across the overlapping period $t = 1,...,T$ of the two estimation windows based on the initially estimated model instead of the re-estimated model.

The two ways of splicing can be applied to any multilateral index. It should be mentioned, however, that any splicing method impairs transitivity; chain drift in spliced multilateral indexes cannot be completely ruled out. As long as the estimation window is not too short, say at least 13 months, this is unlikely to be a big problem.

4.2 Three other approaches

Revisions of published index numbers could of course be circumvented by continuously extending the estimation window without publishing the revised figures. The problem is the loss of characteristicity mentioned earlier. The rolling window approaches discussed above deal with this issue, but European statistical agencies in particular are reluctant to use these approaches; they prefer annual chain linking in a fixed calendar month rather than monthly chain linking with a movement or window splice. Annual chain linking (in combination with resampling) in December is also recommended by Eurostat for the HICP.

To comply with HICP recommendations, and following traditional choices made in the Dutch CPI, Chessa (2015) proposed to construct short-term index series, starting in December and ending in December of the next year, i.e. with a length of 13 months, and chain link them in December of each year. The short-term indexes are based on the multilateral GK method where the length of the estimation window is extended each month and the revised index numbers are not published. So the index for January in the short-term series is estimated on two months of data (a bilateral rather than multilateral, comparison), the index for February is estimated on three months of data (a trilateral comparison), and so forth, until in December thirteen months of data is used. Similar to rolling year approaches, the sample is updated every month with new items affecting the index from the second month after their introduction.

Chessa's (2015) proposal has two potential drawbacks. The indexes for the first months of each year are estimated on sparse data, and so we expect them to be relatively volatile. The second potential problem is that December, being the link month, is given special importance. This seems to be at odds with the idea behind multilateral methods of making the results independent of the choice of base or link period. It would thus be useful to perform empirical research into the effect of the choice of link month within Chessa's framework for dealing with revisions.

Melser (2015) recently advocated the use of weighted GEKS and a constrained least squares criterion for updating the existing time series, the constraint being that all previously estimated multilateral indexes are left unchanged. The weighted GEKS index for the new period $T+1$ turns out to be a weighted geometric average of the inflated bilateral indexes $P^{t,T+1}$ going from period *t* to period $T+1$, i.e.

$$P_{GEKS}^{0,T+1} = \prod_{t=0}^{T} \left[P_{GEKS}^{0,t} \times P^{t,T+1} \right]^{w^{t,T+1} / \sum_{t=0}^{T} w^{t,T+1}} . \tag{25}$$

The weights $w^{t,T+1}$ in (25) are those used to calculate the weighted GEKS index, which become smaller as the estimation window grows, because they depend on the degree of matching. The choice of window length may thus be less unimportant here. Melser's (2015) approach to updating weighted GEKS indexes can also be applied to the more usual unweighted GEKS indexes by simply setting the weights in (25) equal to 1.

The cycle method described in section 3.2.3 is suited to deal with the constraint that previously published price indexes cannot be revised. When data for period $T+1$ is added, a set of new indexes emerges – the set corresponding to arcs ending in $T+1$, i.e. indexes with respect to the pairs of time periods $(1, T+1), (2, T+1), \dots, (t, T+1)$ – that can be adjusted. The original cycle method can be modified by adding constraints to fix previously adjusted price indexes. The modified method then produces adjusted values of the newly added price indices that extend the PIDG for T periods to one for $T+1$ periods. Details about updating the cycle index and a few numerical examples can be found in Willenborg (2015b).

5. Applications

5.1 Small-sample methods

Before describing scanner-data applications of the methods outlined in sections 3 and 4 we want to pay some attention to sample-based methods a number of statistical agencies have implemented. While these are not "big-data" methods and thus not the main topic of our paper, it will be useful to discuss them briefly and mention the reasons why the agencies are using them.

In Australia, Belgium, Denmark and Switzerland, and perhaps in other countries as well, scanner data for supermarkets was implemented in the CPI using sample-based methods. Typically, prices observed by price collectors visiting the stores were replaced by prices (unit values) from scanner data without affecting the sampling design and the index number formula used, at least during the initial stages. The statistical agencies of those countries, like most countries with traditional price observation, are still using the Jevons index number formula. Their methods differ in several other respects though, for example in the identification of items, which is either by GTIN or SKU, the number of weeks of data used to calculate the unit values (see also section 2.2), and whether or not manual replacements and quality adjustments are carried out. Why have these agencies chosen to stick to an unweighted (Jevons) index number formula, notwithstanding the availability of weighting information at the individual item level and use only a small proportion of the available prices information? To shed some light on this issue, we will take the case of the Australian Bureau of Statistics (ABS) as an example.

The ABS adopted a phased approach, "given the lack of international consensus on the best approach to fully integrate transactions data" (Howard et al., 2015).[18] Phase 1, referred to as price replacement, is similar to what was mentioned earlier: in essence, the method of price collection was changed without changing any of the compilation

[18] It is true that international consensus has not emerged, in spite of more than twenty years of empirical research on the use of scanner data for price measurement, which (as far as we know) originated in 1994 with the work of Saglio (1994). Over the last five to ten years substantial progress has been made though, in particular regarding multilateral methods.

practices. In phase 2, the number of price observations was increased and expenditure information from scanner data was used for weighting purposes (above the elementary aggregation level). In phase 3, new methods will be utilized. "Contemporary research suggests that the application of multilateral methods represent the most promising way in which this can be achieved."

The third phase is interesting because the ABS agrees that multilateral methods are most promising, but the second phase is interesting as well. "This phase needs to be carefully considered and managed due to resource implications. For example, mapping products from transactions datasets to CPI classifications can be very resource intensive and may result in not all products being able to be mapped to the Elementary Aggregate level of the index". "Another barrier to implementing the larger samples for phase 2 is the workload implications, with current manual processes taking some time for analysts to complete each quarter. To address this, processes for selecting new items [....] have been automated."

The above suggests three reasons why statistical agencies may not want to use a "big data solution" when they receive scanner data but stick to a small-sample approach, at least for some time: a lack of international consensus on the best approach, a potential increase in workload, and resource implications. There may actually be a fourth reason: a lack of data on characteristics of all the items in scanner data sets required for making explicit quality adjustments or for defining items by their characteristics (rather than by readily available identifiers such as GTIN or SKU).

Our paper focuses on methodology and can only touch upon these issues. They do remind us, however, of the fact that big data methods and processes must be efficient and largely automated to be acceptable for implementation. Nevertheless, we do believe that many of the issues can be resolved, certainly when metadata, including information on item characteristics, becomes increasingly available in the future.

5.2 Matched-model methods

5.2.1 Period-on-period chained indexes

Matched-model price index methods can be easily automated when GTINs or SKUs are used for exact matching of items. In January 2010, Statistics Netherlands implemented monthly chained matched-model Jevons indexes at the (chain-specific) product category level, given by expression (5), for scanner data from supermarkets. All major chains are presently included, and price collection by visiting the stores is no longer taking place.[19] An unweighted index number formula was chosen because evidence became available that high-frequency chaining of superlative indexes could lead to significant chain drift. To maximize the number of matches, month-on-month chaining was preferred. The use of a multilateral method (Rolling Year GEKS; see section 5.2.2) was considered but not implemented given the lack of consensus.

Expenditures within a product category are usually highly skewed. Computing a Jevons index for all the items would mean that the many low-expenditure items receive the same weight as high-expenditure items. It was therefore decided to introduce a cut-off rate, or threshold, based on average expenditure shares across two adjacent months. Items below the threshold are excluded from the computation, and so a crude form of weighting is applied. Nevertheless, the absence of explicit weighting is obviously a weakness of the current Dutch method.

A potential weakness is that items which are "dumped" at unusually low prices, e.g. due to clearance sales, may lead to downward bias in the index. To counteract this bias, thresholds on monthly price and expenditure decreases were put in place. Details about the Dutch method, including the exact threshold definitions and settings, can be found in Van der Grient and De Haan (2010) (2011).

There are many studies on scanner data confirming that high-frequency chaining of matched-model superlative indexes can lead to significant drift and we will mention only a few. Feenstra and Shapiro (2003) used ACNielsen scanner data for the US on canned tuna and found upward bias in weekly chained Törnqvist price indexes. They attributed the bias to consumers' inventory behavior, although advertising and special displays also seemed to have an impact. The upward direction of the bias, or chain drift, is surprising as most other studies find downward bias when storable goods are on sale. Ivancic (2007), using ACNielsen scanner data on different products sold in Australian supermarkets, found downward drift in chained Fisher price indexes; see also Ivancic, Diewert and Fox (2009) (2011).

Drift in chained matched-model superlative price indexes has been documented for durable goods as well. Here, the drift is likely due to seasonal fluctuations in prices and quantities, such as Christmas sales. De Haan and Krsinich (2014b) found downward

[19] Scanner for some supermarket chains were already introduced in 2001 using a fixed-basket approach, i.e. using a Lowe index; see Schut et al. (2002). The idea was to mimic traditional processes on a sample of about 10,000 items (GTINs) for each chain. Not surprisingly, this method turned out to be extremely demanding, in particular due to the manual selection of items as replacements of disappearing items and quality adjustments made when deemed necessary.

drift in chained Törnqvist indexes for several consumer electronics goods New Zealand using scanner data from GfK. Silver and Heravi (2005), who used GfK scanner data on televisions for the UK, presented evidence of downward bias in chained Fisher indexes, though this could also be due to quality-change bias.

5.2.2 Multilateral indexes

The issue of how to resolve the chain drift problem while at the same time maximizing the number of matches in the data was addressed by Ivancic, Diewert and Fox (2011). They proposed the Rolling Year GEKS (RYGEKS) method, in which the GEKS index described in section 3.2.1 is updated using a standard movement splice (see section 4.1) to ensure that previously published index numbers will not be revised. A limitation of their data set was that it covered only 15 months, which is a bit short to assess rolling year indexes. However, the GEKS procedure as such worked as expected.

De Haan and Van der Grient (2011) estimated RYGEKS indexes for a number of goods sold by a Dutch supermarket chain. Instead of bilateral Fisher indexes, they used bilateral Törnqvist indexes as inputs in the (RY)GEKS index. Their scanner data set covered 44 months; in general, the difference between RYGEKS indexes and GEKS indexes estimated on the 44 months of data was small. Except for detergents, significant downward drift in chained matched-model Törnqvist indexes was found, up to 30% for toilet paper. They also estimated Jevons-RYGEKS indexes, which differed substantially from the Törnqvist-RYGEKS indexes. This result once more underlines the importance of weighting at the individual item level.

Melser (2015) used U.S supermarket scanner data to compare different ways to circumvent revisions in GEKS indexes, including his constrained least squares criterion for updating the existing time series (see section 4.2). His method seemed to perform quite well. He also found that window splicing produced slightly more volatile indexes than standard movement splicing, albeit without affecting the trend. A similar result was found by Chessa (2015) who used scanner data for a Dutch department store. De Haan (2015b), on the other hand, using the same New Zealand scanner data set for consumer electronics as in De Haan and Krsinich (2014b), did not find any difference in volatility between TPD indexes with a window splice or a movement splice. Our conclusion is that that the issue of how best to update a time series of multilateral indexes when using a rolling window approach is not yet settled.

Lamboray and Krsinich (2015) compared GEKS indexes with the "Intersection GEKS" (IntGEKS) indexes they proposed; their data set on consumer electronics goods was the same as the one used by De Haan and Krsinich (2014b). The empirical results showed that the IntGEKS was less volatile than the GEKS. Also, the IntGEKS tended to more closely match the multilateral time dummy hedonic index (see section 5.3.2). The reasons for the last result are somewhat unclear. There is likely to be significant quality improvement in electronic products, and so a priori we would expect to find differences in trends between the explicitly quality-adjusted hedonic index and the matched-model IntGEKS index.

A potential problem with all matched-model approaches, including multilateral methods, is that if exact matching at the GTIN level is used, "disguised" price changes during relaunches will not be captured. The use of SKU rather than GTIN for matching items, which is done by e.g. the ABS, could be a solution, but unfortunately SKUs are not always available in scanner data sets. Another option is to define items by a limited set of characteristics extracted from product descriptions. This approach has been taken by Statistics Netherlands for the introduction of transactions data on mobile phones in January 2016 and which will also be followed for the introduction of scanner data of a Dutch department store by mid-2016.

Not only is there a difference with the current GTIN-based treatment of scanner data for supermarkets in the identification of items, another method was also developed. This "QU method" is an instance of the multilateral Geary-Khamis method (see section 3.3.2) where the time series is updated according to Chessa's (2015) approach described in section 4.2. Using scanner data from the department store mentioned above, Chessa, Boumans and Walschots (2015) checked for differences between the proposed monthly GK indexes and monthly GK benchmark indexes estimated on a full 13 month window. Of course the indexes for the 13-th month (December) according to the two methods coincide by construction. The proposed index numbers appeared to agree well with the benchmark index numbers, which is reassuring.

5.3 Quality-adjusted methods

De Haan and Krsinich (2014b) compared a number of rolling year multilateral methods, including the Rolling Year Time Dummy Hedonic (RYTDH) method and the (hedonic) Imputation Törnqvist RYGEKS (ITRYGEKS) method, both with a standard movement splice (see section 3.4.2). Since the two multilateral methods are weighted and explicitly quality adjusted, we would expect them to produce similar results. This was confirmed for desktop and laptop computers, televisions and camcorders, but for other products the evidence was a bit mixed. Across all the products though, the RYTDH index was closer to the ITRYGEKS than the matched-model RYGEKS. The ITRYGEKS method (with a movement splice) was introduced for consumer electronics products in the New Zealand CPI in the September 2014 quarter; see Statistics New Zealand (2014).

De Haan and Krsinich (2014a) used quality-adjustment factors estimated from multilateral time dummy hedonic regressions in order to calculate quality-adjusted unit value indexes (see section 3.4.3), again on the same consumer electronics data set as in De Haan and Krsinich (2014b). The resulting quality-adjusted unit values were similar to the original time dummy hedonic indexes, as expected. In fact, the differences were hardly noticeable.

6. Conclusions and future work

The overview of index methods in this paper serves as a starting point for an empirical study at Statistics Netherlands where the focus is on a comparison of different methods. Price indexes will be calculated on a number of scanner data sets, which will allow us to gain more insight into the impact of different choices on aggregate price change, e.g. weighted versus unweighted methods, monthly chained versus direct price indexes, and bilateral versus multilateral methods.

The present paper also motivates further fundamental research. Below, we list topics that could be addressed in subsequent papers.

- It would be interesting to explore how the GEKS method and other multilateral methods relate to each other.
- In particular, since the cycle method generalizes the GEKS method, how do the weighted GEKS and cycle methods compare?
- The comparisons between hedonic and TPD (fixed effect) methods described in this paper call for further research into implicit quality adjustment in multilateral indexes.
- Price indexes according to the QU-method are calculated with December of each calendar year as link (or base) month. This choice is in agreement with current practice in the CPI/HICP. Chessa (2015) reported results of a preliminary study about the impact of the choice of link month. This study could be extended in future research.
- The monthly updating of the weights in the quantity index of the QU-method leads to estimates in the first months of a year that are based on relatively little information. Techniques could be developed to improve these weight estimates and the corresponding price indexes.

- Chessa (2015) investigated alternatives to the functional form of (8) and found hardly any differences between the corresponding price indexes. A follow-up study (Chessa, 2016) that considered a number of alternative weighting schemes suggested that substitution bias is not a big issue in intertemporal comparisons. More fundamental research on this phenomenon could be done in future studies.
- Scanner data sets may suffer from missing information on item characteristics. This could lead to problems, for instance, when applying hedonic methods. The problem of missing values for auxiliary variables applies to regression models in general (Haitovsky, 1968; Little, 1992). It may be worthwhile to find ways for dealing with missing item information using some scanner data sets.

There are other topics besides theoretical issues that deserve attention. As was mentioned in the first part of the present paper, the availability of information on item characteristics and product identifiers is of crucial importance. Items need to be defined before price indexes can be computed in a meaningful way. We identify the following additional topics for further research.

- In cases where relaunches occur, the question arises what information could be used to define items such that "disguised" price increases will be captured. One interesting option would be the use of the retailer's own product codes (SKUs). The availability and applicability of SKUs is worth investigating.
- If SKUs are not available, or cannot be applied for some reason, information on item characteristics becomes even more important. There may be different ways to define (homogeneous) products when relaunches occur, based on aggregating across GTINs with the same characteristics. For example, this approach could be applied to an entire product category, but it could also be applied solely to items within the product category which undergo relaunches (while sticking to GTIN as item identifier for other items). The question is whether the two approaches yield similar results.
- If relevant item characteristics are found missing in a scanner data set, and if a retailer does not have, or is not willing to supply, additional information, then it is important to develop techniques for complementing missing information, for instance, by using web scrapers. Techniques from text mining, machine learning and computational linguistics could be considered for this purpose. Photo feature extraction could be an interesting alternative, possibly also in combination with the aforementioned techniques.

References

Aizcorbe, A., C. Corrado and M. Doms (2003), "When Do Matched-Model and Hedonic Techniques Yield Similar Price Measures?", Working Paper no. 2003-14, Federal Reserve Bank of San Francisco.

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

Balakrishnan, V.K. (1997), *Graph Theory*, Schaum's Outline Series, McGraw-Hill.

Balk, B.M. (1998), "On the Use of Unit Value Indices as Consumer Price Sub-indices", Paper presented at the fourth meeting of the Ottawa Group, 22-24 April 1998, Washington D.C., USA.

Balk, B.M. (1980), "A Method for Constructing Price Indices for Seasonal Commodities", *Journal of the Royal Statistical Society A* 142, 68-75.

Balk, B.M. (1981), "A Simple Method for Constructing Price Indices for Seasonal Commodities", *Statistische Hefte* 22, 1-8.

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.

Bradley, R. (2005), "Pitfalls of Using Unit Values as a Price Measure or Price Index", *Journal of Economic and Social Measurement* 30, 39-61.

Chessa, A.G. (2013), "Comparing Scanner Data and Survey Data for Measuring Price Change of Drugstore Articles", Paper presented at the Workshop on Scanner Data for HICP, 26-27 September 2013, Lisbon.

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

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

Chessa, A.G., S. Boumans and J. Walschots (2015), "Towards a New Methodology for Processing Scanner Data in the Dutch CPI", Paper presented at the Workshop on Scanner Data, 1-2 October 2015, Rome, Italy.

Dalén, J. (2014), "The Use of Unit Values in Scanner Data", Paper presented at the Scanner Data Workshop, 2-3 October 2014, Vienna, Austria.

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

Diewert, W.E. (1976), “Exact and Superlative Index Numbers”, *Journal of Econometrics* 4, 115-145.

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

Diewert, W.E. (2004), “On the Stochastic Approach to Linking the Regions in the ICP”, Discussion Paper no. 04-16, Department of Economics, The University of British Columbia, Vancouver, Canada.

Diewert, W.E., S. Heravi and M. Silver (2009), “Hedonic Imputation versus Time Dummy Hedonic Indexes”, pp. 87-116 in W.E. Diewert, J. Greenlees and C. Hulten (eds.), *Price Index Concepts and Measurement*, Studies in Income and Wealth, Vol. 70. Chicago: University of Chicago Press.

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

Feenstra, R.C. (1995), “Exact Hedonic Price Indices”, *Review of Economics and Statistics*, vol. LXXVII, 934-654.

Feenstra, R.C. and M.D. Shapiro (2003), “High-Frequency Substitution and the Measurement of Price Indexes”, in R.C. Feenstra and M.D. Shapiro (eds.), *Scanner Data and Price Indexes*, Studies in Income and Wealth, Volume 61. University of Chicago Press, Chicago.

Fox, K.J., W.E. Diewert and J. de Haan (2016), “A Newly Identified Source of Potential CPI Bias: Weekly versus Monthly Unit Value Price Indexes, *Economics Letters* (forthcoming).

van Garderen, K.J. and C. Shah (2002), “Exact Interpretation of Dummy Variables in Semilogarithmic Equations”, *Econometrics Journal* 5, 149-159.

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.

Greenlees, J. and R. McClelland (2010), “Superlative and Regression-Based Consumer Price Indexes for Apparel Using U.S. Scanner Data”, Paper presented at the Conference of the International Association for Research in Income and Wealth, 27 August 2010, St. Gallen, Switzerland.

van der Grient, H.A. and J. de Haan (2010), “The Use of Supermarket Scanner Data in the Dutch CPI”, Paper presented at the Joint ECE/ILO Workshop on Scanner Data, 10 May 2010, Geneva, Switzerland.

van der Grient, H. and J. de Haan (2011), “Scanner Data Price Indexes: The “Dutch” Method versus Rolling Year GEKS”, Paper presented at the twelfth meeting of the Ottawa Group, 4-6 May 2011, Wellington, New Zealand.

de Haan, J. (2004a), “The Time Dummy Index as a Special Case of the Imputation Törnqvist Index”, Paper presented at the eighth meeting of the Ottawa Group, 23-25 August 2004, Helsinki, Finland.

de Haan, J. (2004b), “Direct and Indirect Time Dummy Approaches to Hedonic Price Measurement”, *Journal of Economic and Social Measurement* 29, 427-443.

de Haan, J. (2004c), “Estimating Quality-Adjusted Unit Value Indexes: Evidence from Scanner Data”, Paper presented at the SSHRC International Conference on Index Number Theory and the Measurement of Prices and Productivity, 30 June – 3 July 2004, Vancouver, Canada.

de Haan, J. (2010), “Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and ‘Re-pricing’ Methods”, *Jahrbücher fur Nationalökonomie und Statistik* 230, 772-791.

de Haan, J. (2015a), “A Framework for Large Scale Use of Scanner Data in the Dutch CPI”, Paper presented at the fourteenth Ottawa Group meeting, 20-22 May 2015, Tokyo, Japan.

de Haan, J. (2015b), “Rolling Year Time Dummy Indexes and the Choice of Splicing Method”, Unpublished paper, Statistics Netherlands, The Hague, The Netherlands.

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

de Haan, J. and R. Hendriks (2013), “Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes”, Paper presented at the Economic Measurement Group Workshop, 28-29 November 2013, Sydney, Australia.

de Haan, J. and F. Krsinich (2014a), “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.

de Haan, J. and F. Krsinich (2014b), “Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes”, *Journal of Business & Economic Statistics* 32, 341-358.

Haitovsky, Y. (1968), “Missing Data in Regression Analysis”, *Journal of the Royal Statistical Society B* 30, 67-82.

Hill, R.J. (1999a), “Comparing Price Levels across Countries using Minimum-Spanning Trees”, *The Review of Economics and Statistics*, Volume 81, 135-142.

Hill, R.J. (1999b), “Comparing Price Levels across Countries using Spanning Trees”, in: *International and Inter-Area Comparisons of Income, Output and Prices*, A. Heston and R.E. Lipsey (eds.), Studies in Income and Wealth, Volume 61, University of Chicago Press.

Howard, A., K. Dunford, J. Jones, M. van Kints and K. Naylor (2015), “Using Transactions Data to Enhance the Australian CPI”, Paper presented at the fourteenth meeting of the Ottawa Group, 20-22 May 2015, Tokyo, Japan.

Iklé, D.M. (1972), "A New Approach to the Index Number Problem", *Quarterly Journal of Economics* 86, 188-211.

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

Ioannides, C. and M. Silver (1997), "Chained, Exact and Superlative Hedonic Price Changes: Estimates from Micro Data", *Applied Economics* 35, 1005-1014.

Ivancic, L. (2007), *Scanner Data and the Construction of Price Indices*, PhD thesis, University of New South Wales, Sydney, Australia.

Ivancic, L., W.E. Diewert and K.J. Fox (2009), "Scanner Data, Time Aggregation and the Construction of Price Indexes", Discussion Paper no. 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.

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

Ivancic, L. and K.J. Fox (2013), "Understanding Price Variation Across Stores and Supermarket Chains: Some Implications for CPI Aggregation Methods", *Review of Income and Wealth* 59, 629-647.

Kennedy, P.E. (1981), "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations", *American Economic Review* 71, 801.

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 Indexes with No Characteristic Information", Paper presented at the meeting of the group of experts on consumer price indices, 26-28 May 2014, Geneva, Switzerland.

Lamboray, C. and F. Krsinich (2015), "A Modification of the GEKS Index When Product Turnover is High", Paper presented at the fourteenth Ottawa Group meeting, 20-22 May 2015, Tokyo, Japan.

Little, R.J.A. (1992), "Regression with Missing X's: A Review", *JASA* 87, 1227- 1238.

Melser, D. (2015), "Scanner Data Price Indexes: Addressing Some Unresolved Issues", Unpublished paper, School of Economics, Finance and Marketing, RMIT University, Melbourne, Australia.

Rao, D.S.P. (1990), "A System of Log-Change Index Numbers for Multilateral Comparisons", pp. 127-139 in: Salazar-Carillo and Rao (eds.), *Comparisons of Prices and Real Products in Latin America*, Contributions to Economic Analysis Series. North-Holland, Amsterdam.

Rao, D.S.P. (1999), "Weighted EKS Method for Aggregation Below Basic Heading Level Accounting for Representative Commodities", Note prepared for the ICP Unit, OECD.

Rao, D.S.P. (2001), "Weighted EKS and Generalized CPD methods for Aggregation at the Basic Heading Level and Above Basic Heading Level", Joint World Bank –

OECD Seminar on purchasing Power Parities, February 2001, *Recent Advances in Methods and Applications*, Washington DC, U.S.

Rao, D.S.P. (2005), "On the Equivalence of Weighted Country-Product Dummy (CPD) Method and the Rao-System for Multilateral Price Comparisons", *Review of Income and Wealth* 51, 571-580.

Reinsdorf, M.B. (1999), "Using Scanner Data to Construct CPI Basic Component Indexes", *Journal of Business & Economic Statistics* 17, 152-160.

Saglio, A. (1994), "Comparative Changes in Average Price and Price Index: Two Case Studies", Paper presented at the first meeting of the Ottawa Group, 31 October -4 November 1994, Ottawa, Canada.

Schut, C. et al. (2002), "The Use of Scanner Data from Supermarkets in the Consumer Price Index" [in Dutch], Statistics Netherlands, Voorburg, The Netherlands.

Silver, M. and S. Heravi (2005), "A Failure in the Measurement of Inflation: Results from a Hedonic and Matched Experiment Using Scanner Data", *Journal of Business & Economic Statistics* 23, 269-281.

Statistics New Zealand (2014), "Measuring Price Change for Consumer Electronics Using Scanner Data". Available from www.stats.govt.nz.

van Steen, M. (2010), *Graph Theory and Complex Networks*.

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

Syed, I. (2010), "Consistency of Hedonic Price Indexes with Unobserved Characteristics", Discussion paper, School of Economics, University of New South Wales, Sydney, Australia.

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

Triplett, J.E. (2003), "Using Scanner Data in Consumer Price Indexes: Some Neglected Conceptual Considerations", in R.C. Feenstra and M.D. Shapiro (eds.), *Scanner Data and Price Indexes*, Studies in Income and Wealth, Volume 61. University of Chicago Press, Chicago.

Willenborg, L. (1993), "An Adjustment Method Based on Graph Homology", Report, Statistics Netherlands, Voorburg, The Netherlands.

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

Willenborg, L. (2015a), "Elementary Price Indexes for Internet Data". Report, Statistics Netherlands, The Hague, The Netherlands.

Willenborg, L. (2015b), "Transitivizing Elementary Price Indexes for Internet Data", Report, Statistics Netherlands, The Hague, The Netherlands.

Willenborg, L. and M. van der Loo (2016), "Transitivizing Price index numbers using the cycle method: some empirical results", Report, Statistics Netherlands, The Hague, The Netherlands.