Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data

Antonio G. Chessa* and Robert Griffioen**

Abstract – Statistical institutes are considering web scraping of online prices of consumer goods as a feasible alternative to scanner data. The lack of transaction data generates the question whether web scraped data are suited for price index calculation. This article investigates this question by comparing price indices based on web scraped and scanner data for clothing and footwear in the same webshop. Scanner data and web scraped prices are often equal, with the latter being slightly higher on average. Numbers of web scraped product prices and products sold show remarkably high correlations. Given the high churn rates of clothing products, a multilateral method (Geary-Khamis) was used to calculate price indices. For 16 product categories, the indices show small overall differences between the two data sources, with year on year indices differing only by 0.3 percentage point at COICOP level (men's and women's clothing). It remains to be investigated whether such promising results for web scraped data will also be found for other retailers.

JEL Classification: C43, E31
Keywords: CPI, scanner data, web scraping, multilateral methods, Geary-Khamis method

** Statistics Netherlands, Team CPI (ag.chessa@cbs.nl)*
*** Statistics Netherlands, Team CPI at the time this research was carried out*

This research was funded by a Eurostat grant. The authors want to express their gratitude to Eurostat for the possibility of carrying out the research under the grant assigned.

Received on 31 July 2017, accepted after revisions on 1st April 2019

To cite this article: Chessa, A. G. & Griffioen, R. (2019). Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 49–68. https://doi.org/10.24187/ecostat.2019.509.1984

The use of scanner data to measure the Consumer Price Index (CPI) is gradually expanding. Indeed, scanner data offer an almost ideal alternative to traditional survey data because these data sets contain transaction data. Both prices and expenditures are available for every sold item, with each item identified by its barcode (officially known as GTIN, the Global Trade Item Number, which is issued and administered by the international company GS1). The expenditures by item available in scanner data can then be used to construct weighted price indices, which gives scanner data a big advantage over survey data.

In Europe until 2014, four National Statistical Institutes (NSIs) used scanner data in their CPI and this number has increased to ten by January 2018 (see also Leclair *et al.*, this issue). NSIs are allowed to develop their own method for processing scanner data and calculating price indices for elementary aggregates. Comparability of methods across countries is nevertheless desirable, also for elementary aggregates, so in an attempt to guide NSIs to start with the processing of scanner data, Eurostat has set up guidelines and a description of current practices (Eurostat, 2017).

Obtaining scanner data may be a lengthy process. Different factors are involved, such as finding out which persons to get in touch with in a retailer's organisation, the willingness of a retailer to cooperate, and the available time in order to prepare a data set according to a format usable by a statistical institute. Several countries, such as the Netherlands, benefit from a statistical law when requesting scanner data, but countries without such a law may face difficulties in the acquisition of scanner data, and some NSIs are focusing on collecting online data (e.g., see Breton *et al*., 2016). The use of web scraping for collecting online prices and information about item characteristics has greatly gained in popularity in recent years (Breton *et al*., 2016; Cavallo, 2016; Griffioen & ten Bosch, 2016). Web scraping of online prices opens new possibilities for official statistics. Like scanner data, sample sizes can be drastically increased and data collection and processing can be automated to a large extent. Automated online data collection also allows to decrease the administrative burden of price collection, not only for NSIs themselves but also for retailers. Statistical institutes therefore consider the replacement of sample surveys by automated collection of online price data as a big opportunity and challenge.

Given the increasing popularity of web scraping, it is important to explore the possibilities and limits of using online prices for price index calculation. Web scrapers only collect online prices; expenditures for items offered on a website can obviously not be collected online. Of course, this also holds for traditional price collection. However, now that scanner data have become available, it is possible to quantify the consequences on a price index of having or missing certain information. For example, using expenditure-based weights or equal weights for products in an index number formula may result in quite different price indices (Chessa *et al*., 2017).[1]

Finding such differences leads to the following important question: do the numbers of web scraped product prices correlate well with the numbers of sales contained in scanner data? In case of an affirmative answer, price indices that are exclusively based on web scraped prices and quantities are expected to give good approximations to price indices based on scanner data. The outcome obviously depends on different factors, such as the policy of online shops and the design of their websites (e.g., which products are promoted and found more often on a website) and the scraping strategy (is a whole site scraped, how often and at which times). Of course, a sensible comparison between price indices based on scanner data and web scraped data can only be made if the same metadata about items can be used in price index calculations.

Statistics Netherlands (CBS) receives scanner data from a large Dutch online department store since several years. In October 2012, CBS started to collect online prices and metadata from the same store with a web scraper. The scanner data and web scraped data therefore offer an excellent opportunity for comparing product prices, quantities and price indices between the two data sources. Price indices calculated using scanner data can be used as benchmarks in order to assess the accuracy of price indices calculated with web scraped data. The objective of this paper is to compare price indices based on the two data sources.

The paper is organised as follows. The next section briefly describes the information contained in the scanner data and web scraped

1. We use the term "product" as a more generic concept alongside "item", which refers to GTIN. A product is equivalent to an item when GTINs have low rates of churn, that is, when assortments are stable over time. When assortments are not stable, for instance, when GTINs have rather short lifetimes because relaunches occur, then GTINs should be linked and combined into groups. The GTINs in each group have the same set of item characteristics. We call such groups "products". How characteristics are selected, and whether GTINs are suitable as products, is a complex issue that would deserve a separate study.

data of the Dutch online store. Then in the third section we describe the method applied to the scanner data and web scraped data of the online shop, which we call the "QU-method" (Quality adjusted Unit value method). The price indices calculated for the two data sources are then compared at category and COICOP level[2] in the fourth section. The paper concludes with the main findings of this study and some suggestions for further research.

Scanner Data and Web Scraped Data of a Dutch Web Store

In the first years of its web scraping development programme, which was initiated more than five years ago, CBS focused on clothing and footwear, as part of its policy to reduce the use of traditional surveys for these product categories in the CPI. Consequently, the comparisons between prices, quantities and price indices for web scraped and scanner data will focus on clothing and footwear items. Results of a data analysis are also presented, in which product prices and quantities are compared for the two data sources.

Scanner Data

CBS receives scanner data from the Dutch online department store since January 2011. The retailer specifies and sends the data on a weekly basis, an agreement that is also made with other retailers. The scanner data cover the transactions of the entire assortment of the department store. The assortment is very broad; besides clothing and footwear, the department store sells electronics, products for house and garden, products for recreational activities, etc.

For every item (GTIN), the scanner data sets contain the following information, delivered as separate fields:

- Year and week of sales (combined in one field);
- GTIN;
- Item number, a retailer specific 6-digit code of an item;
- A text string with a (short) description of the item;
- Group according to which an item is classified by the retailer;
- Group number;
- Number of items sold;
- Turnover (expenditure);
- Number of items sent back;
- Turnover for returned items;
- VAT.

Since the end of 2013, numbers of returned items and the corresponding turnover are also included by the retailer in the data, and are available every week in the data since March 2014. Returned values are subtracted from the fields "Number of items sold" and "Turnover", so that these values are net values. "Number of items sold" and "Turnover" can therefore take negative values, when "Number of returned items" and their corresponding turnover are larger than the numbers of items sold originally and the associated turnover.

Web Scraped Data

Product types like clothing may exhibit high rates of churn. New items have to be linked to exiting items of the same or comparable quality in order to capture "hidden" price changes when calculating price indices. Such replacements of items are also known as "relaunches". Items can be linked according to a set of common characteristics. It is therefore important that scanner datasets contain such information about items.

However, statistical institutes depend on what retailers are able to deliver, so that scanner data may not always contain sufficient metadata for linking items. Unfortunately, this is the case for the scanner data of the online store treated in this paper (see later in this section). Statistical institutes may contact retailers and request more information. Web scraping offers an interesting alternative for supplementing item information in scanner data.

The web scraper built for the Dutch online store collects data every day since the first day it was run (6 October 2012). The following information is collected for each item:

- One field with year, month and day to which the scraped data applies;
- The retailer specific item number;
- An item description;

2. *By this we mean the COICOPs men's clothing and women's clothing.*

- Brand name;

- Three levels of item classification;

- Item price;

- The item's regular price.

The item descriptions collected by the web scraper contain more information than that from the scanner data. A typical item description in the scanner data is, for instance, 'Men's trousers'. The web scraped text strings also contain the item's brand name, package content (e.g. number of single items in a multi-pack item), and the size, fabric and type of fit are specified for some clothing items. The brand name is also available as a separate field.

The item level on the website can be reached by navigating from the main menu through two submenus, so that items are classified according to three group levels. As was mentioned at the beginning of this section, the assortment of the online store is quite broad. The main focus of the web scraper is to collect information about clothing and footwear items. The three levels of item classification that apply to clothing and footwear can be summarised as follows:

- The upper level (main menu) subdivides clothing and footwear items into five groups: 'Men's clothing', 'Women's clothing', 'Children's clothing', 'Premium selection' and 'Sale'. We will refer to the upper level as "main group" in this paper;

- The intermediate level is called "category". The scraper has collected information from 145 categories during the period investigated in this study (March 2014 – December 2016);

- The most detailed level is called "type", which contains 1,131 groups.

The main groups Premium selection and Sale may contain items on discount. An item may therefore be reached from the main group 'Sale' or from one of the three main groups 'Men's clothing', 'Women's clothing' or 'Children's clothing'. The web scraper "navigates" through each of the five main groups, which means that the same item can be scraped more than once on one day. Multiply scraped items are recorded as separate counts.

Obviously, 'Sale' does not only contain clothing and footwear items, but also other items on discount. The web scraper therefore also collects the above-listed information for electronics, house and garden, beauty and care products, etc. The web scraped data contain two prices for items on discount: the item's actual (i.e., discount) price and the item's regular price. The regular prices of items on discount are collected together with the discounted prices; regular price in fact refers to the price just before the discount period. In our index calculations, we use of course the discounted prices for items on discount – not the regular prices.

Data Analysis

In this subsection, we investigate several aspects of the scanner data and web scraped data that are of direct interest to price index calculation. Our primary focus is obviously on comparing prices calculated from the two data sources. Quantities sold are used to calculate unit product values and, together with prices, they constitute a source for deriving product weights. A second interesting question therefore is how the quantities sold compare with the numbers of web scraped product prices.

Properties of the Two Data Sets

A first key step before using large electronic data sets in the CPI or for research purposes is to subject these to a number of checks. The articles on data quality by Daas & van Nederpelt (2010) and Daas & Ossen (2010) propose a number of "quality dimensions" on which data can be checked. Below, we summarise our findings on some of the dimensions that we investigated for scanner data and web scraped data.

- Completeness: The variables (i.e. the columns or fields) in both data sets show a high degree of completeness. All records of the scanner data are filled, except for the GTIN code, which has a high percentage of missing values (46.4%). The reason for this large number of missing values is unknown. This could be due to the fact that the retailer has its own product codes, which are available for each record. Item descriptions are available for every record as well. The web scraped data also have a high degree of completeness. Prices and item descriptions are missing in 21 records, which is negligible on several millions of records.

- Stability: Stability is another essential factor that needs to be checked before using a data set for regular statistical production. CPI production

will be hampered when, in one month, the total number of records appears to be much lower than usual. Both scanner data and web scraped data do not reflect rapid increases or decreases in the total number of records per month. The number of records increases over time, which can be ascribed to the extended assortment.

- Degree of detail: The amount of metadata in the scanner data of the webshop is limited. The following figures serve as an indication: 25% of the item descriptions contain one word and 62% consist of two words at most.

The web scraper collected information for 385,833 items during the period March 2014-December 2016. This number is quite close to the number of 407,253 sold items in the scanner data, although the scanner data cover the whole assortment (in contrast to the web scraped data). The large number of web scraped items is partly due to the fact that the web scraper also collects information about items other than clothing in 'Premium selection' and 'Sale'. Another reason for the large number is that the website may also contain items that were not sold.

If we combine brand name with the three levels of item classification in order to group or link items, then the 385,833 web scraped items are subdivided into 59,588 of these item groups. The ratio of the number of items to item groups is thus fairly small. It is much smaller than for the scanner data (1,635 groups for 407,253 items), which highlights the greater level of detail in the metadata collected by the web scraper. This benefits the homogeneity of products when item characteristics are used to define products.

- Timeliness: CBS receives scanner data on a weekly basis for all retailers, and the data are usually received on time. The web scraper collects data on a daily basis, during the night so as not to interfere with busy shopping hours. The data are available as soon as the data have been collected on the website. However, situations may arise that affect timeliness. One of these occurs when a website is unreachable or when it has changed. Based on our experience, the first case has rarely taken place. The second situation is more frequent and for this reason we set up a "DevOps team" (Development and Operations team) to adapt and maintain the web scrapers (for more information on how CBS implemented this, see Griffioen *et al.*, 2016).

Price Comparisons

It is important to note that scanner data enable us to compute transaction prices, that is, the prices actually paid by consumers, and may include different components like for instance, special discounts, for card holders or customers with coupons. This is not so with web scraped prices, which are not transaction prices, but the prices offered by a retailer on a website.

The price for a set of different transactions of the same item, or items of the same quality, can be calculated as a unit value: the ratio of total expenditure divided by the sum of the quantities sold (ILO *et al*., 2004, p. xxii). Usually, this boils down to a straightforward exercise. However, complications may arise when consumers return items frequently. The online store has a customer-friendly return policy, which allows consumers to return items within 14 days after delivery and free of charge within this period.

Returned quantities and corresponding expenditures are subtracted from the quantities sold and the expenditures in the week in which items are returned and processed by a retailer. Quantities sold and expenditures therefore represent net values in the scanner data. The processing week may differ from the week of purchase. This has two important implications: net quantities and expenditures may be negative; unit values derived from the two net values will differ from the original price paid when the price at which items were bought differs from the price in the week in which items were returned. In addition, consumers tend to buy more of an item when it is on discount. This means that the first weeks after a discount deserve special attention when comparing prices based on scanner data and web scraped data.

CBS asks for separate information about quantities returned and the corresponding expenditures when requesting scanner data. The scanner data of the Dutch online department store contain this information since week 12 of 2014. We are thus able to quantify the impact of item returns on net expenditures, quantities sold and unit values.

Figure I shows prices based on scanner data and web scraped data for a single item during one year. The prices derived from scanner data (Figure I-A) include item returns; that is, quantities and expenditures of returned items are subtracted from the sales values in the weeks in

which the items were returned in order to yield net values (dotted line). Prices were calculated only when both net expenditures and quantities are greater than zero. Three very high peaks appear. Each of these peaks follows a week with lower prices. Unit values that are calculated from net expenditures and quantities result in prices that are higher than the prices in the week in which items are returned. High price peaks occur when the quantities of returned items are close to the quantities sold in the week in which items are returned.

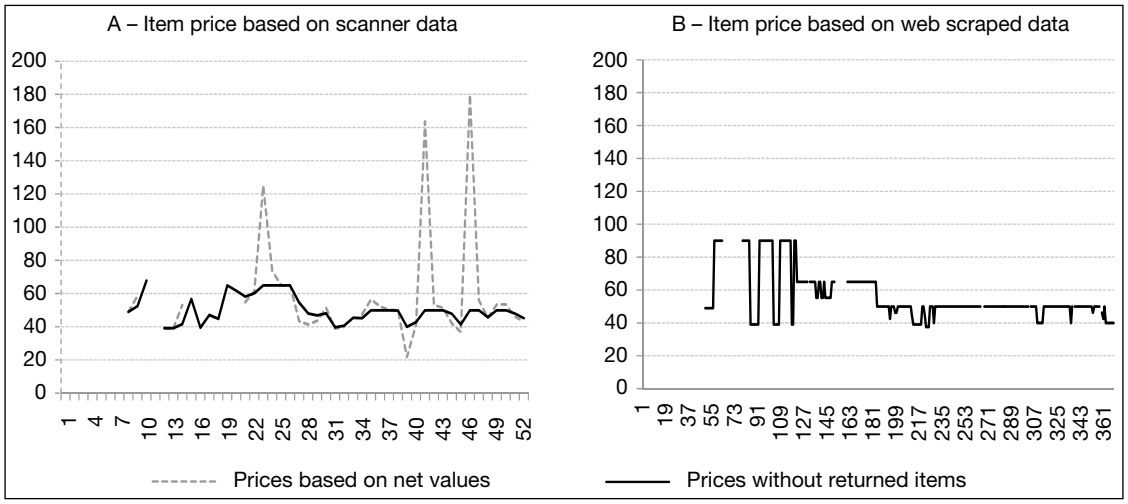
The subtraction of these values allows calculating the "true", original transaction prices (black line in Figure I-A). This highlights the importance of requesting separate information about expenditures and quantities of returned items. The corrected prices compare much better with the web scraped prices shown in Figure I-B. Web scraped prices are higher, on average, in the first weeks (i.e., until week 19, or day 109 on the right). The item was sold for the first time in week 8 of 2015. Apparently, the item entered the assortment at high prices, but the black line in Figure I-A suggests that the consumers mostly bought the item when it was on discount. After the initial period, the differences between the prices for the two data sets become smaller.

Given the impact that returned items may have on net expenditures and quantities, we decided to exclude returned items from expenditures and quantities in order to compare prices and quantities sold with web scraped prices and quantities. We computed two basic statistics for prices and quantities: ratios of web scraped prices to prices based on scanner data, and correlations between numbers of sold products and the numbers of web scraped product prices over time. We computed correlations in the second case, because a one to one comparison between numbers of sold products and web scraped numbers is difficult to make.

Histograms for ratios are shown in Figure II for the combined categories "Trousers and jeans" for women and women's shoes. We combined in the same group items with the same brand name and the most detailed level of item classification (Type). We also made this choice for price index calculation (see below). Items from the main groups 'Premium selection' and 'Sale' were included as well in order to take into account discount prices. An example of a [Brand×Type] group is "Jeans bermuda" of, say, brand X. A combination of [Brand×Type] will be referred to as "product" in this paper.

The graphs in Figure II show the combined price ratios of all products in each month. The graphs show high peaks around 1 (equal prices), and both are skewed towards ratios larger than 1. Web scraped prices tend to be higher, on average, than transaction prices. The same was already noted for the prices of the single item (cf. Figure I). Lower scanner data prices may be

Figure I
Weekly prices based on scanner data and daily prices based on web scraped data for a single item (men's jeans) in 2015



Notes: Two price calculations are shown for scanner data, with returned items (i.e. based on net values) and without. Prices are in euro. The horizontal axes denote week number (scanner data) and day number (web scraped data).
Sources: Scanner data for prices on clothing (left) and web scraped prices (right).

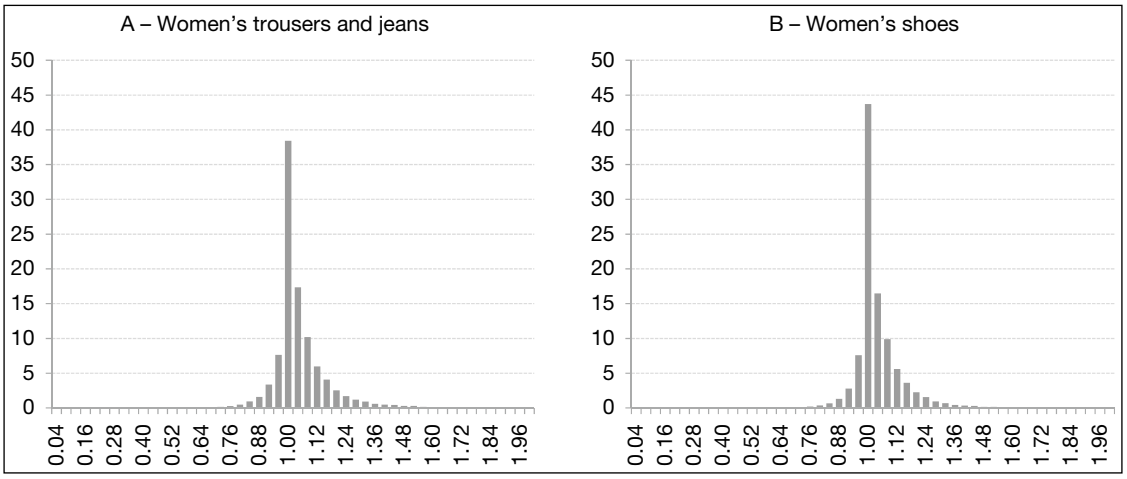
caused by shifts in sales towards cheaper items, for instance, when such items are on discount ("quantity effect").

Quantity Comparisons

We calculated correlations between the numbers of sold products and the numbers of web scraped product prices. For each product, a correlation was calculated from the pairs of sold numbers and numbers of scraped prices of all months in the time series. Both graphs show remarkably high correlations, with the highest frequencies occurring for the largest correlation classes (Figure III). Such patterns would not be obtained if the web scraped numbers were independent of the numbers of sold products. This would lead to distributions centred on zero correlation. The small bumps for the smallest correlation class can be attributed to a large extent to products for which prices are observed in only two months. Removing these products from the calculations eliminates the bumps.

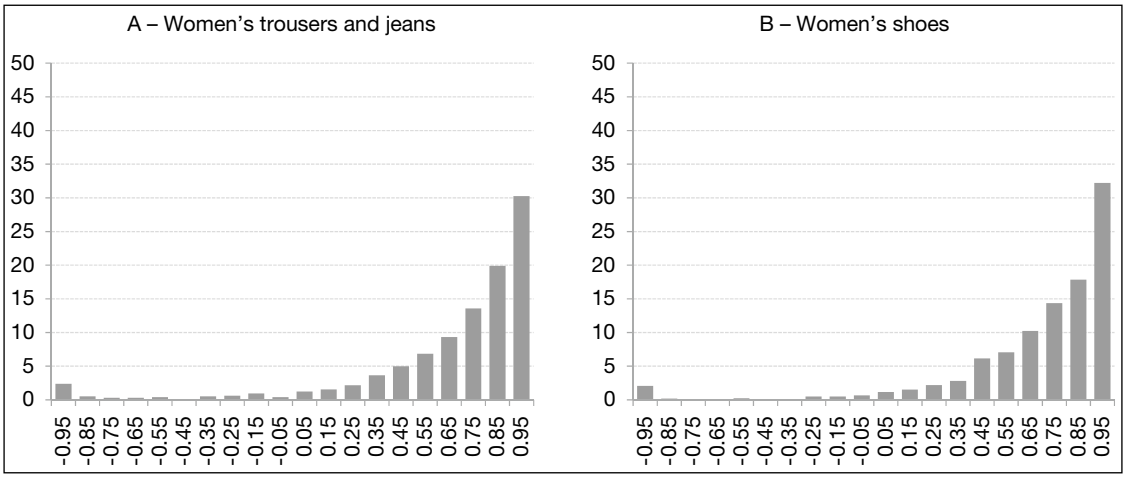
The frequencies by which items can be found across different menus of a website over time

Figure II
Frequency distributions of ratios of web scraped product prices to unit values for scanner data, for two product categories



Notes: The frequencies in a graph sum to 100 per cent. The price ratios on the horizontal axes are centred class values, using a class width of 0.04.
Sources: Scanner data and web scraped data on clothing and shoes.

Figure III
Frequency distributions of correlations between the numbers of web scraped product prices and the numbers of products sold



Notes: Frequencies sum to 100 per cent. The correlations on the horizontal axes are centred class values, using a class width of 0.1.
Sources: Scanner data and web scraped data on clothing and shoes.

seem to correspond quite well with the quantities sold. This may be traced back to the retailer's policy to promote items that are sold more often on the website. Other product categories yield similar results, both for prices and quantities, which constitute favourable conditions for the price index comparisons between the two data sets. It is therefore important to be in contact with the retailer in order to find out more about its strategy behind organising the website.

Assortment Dynamics

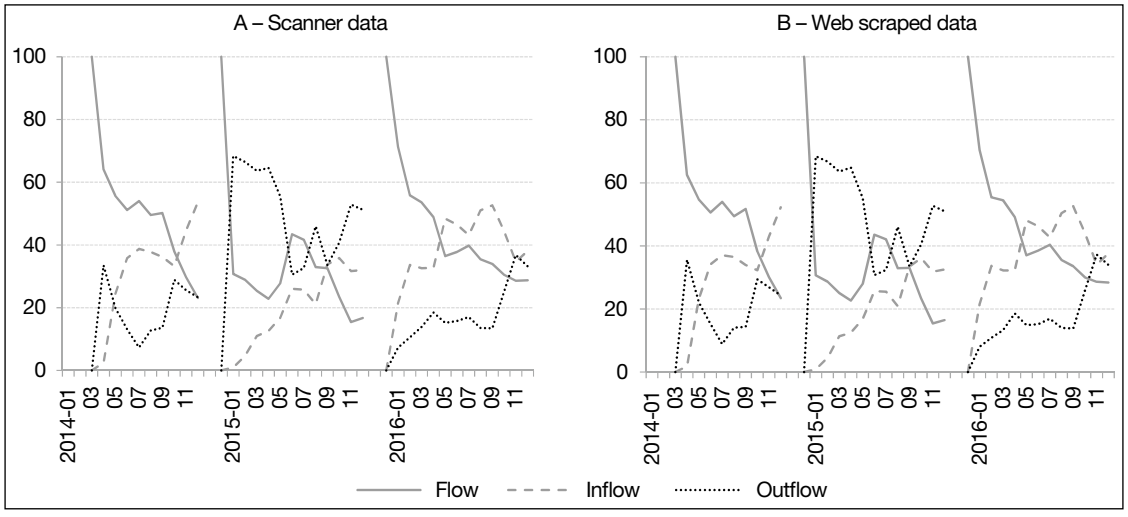
Clothing and footwear are usually characterised by high churn rates. We investigated the dynamics of the assortments of different product categories for scanner data and web scraped data. We quantified the dynamics by introducing three measures: (i) the share of products that are sold or are available over longer periods, referred to as "flow", (ii) the share of products that enter an assortment during a year, or "inflow", and (iii) the share of products that leave an assortment, or "outflow". We calculated the three flow measures as bilateral statistics, that is, for pairs of months. The first month was kept fixed (chosen as the base month). Products that are sold or are available both in the base month and in the second, or current, month are counted as flow, products that are not sold/available in the base month but only in the current month are counted as inflow, while products that are available in the base month but not in the current month are counted as outflow.[3]

The three flow statistics are calculated for every month in the period March 2014-December 2016. This is done for each year separately, using March 2014, December 2014 and December 2015 as base months for the three years. The statistics are calculated by performing counts at product level, that is, for [Brand×Type] groups. Figure IV shows the three flow statistics for men's trousers and jeans.

The flow rate in the base months is, by definition, equal to 100%. The rapid decline of the flow rates and the increase and high values of inflow indicate a highly dynamic assortment. The two graphs clearly show that there is hardly any difference between the flow statistics for scanner data and web scraped data. This means that items that are not sold anymore are quickly removed from the website. It is also worth noting that the high degree of dynamics is evidenced at product level, that is, at a less detailed level than the item/GTIN level. The high dynamics at product level play an important role in the choice of index method.

3. The choice for bilateral measures was made in order to keep calculations tractable. Extensions to additional months are obviously possible, but the characterisation of the dynamics becomes more complex. See Willenborg (2017) for more details.

Figure IV
Flow dynamics for men's trousers and jeans, per year, for scanner data and web scraped data



Notes: The three flow measures are expressed as percentages, which sum to 100% in each month.
Sources: Scanner data and web scraped data on clothing.

The QU Method

Clothing is a notoriously complex field in price index calculation, because product categories may be characterised by high churn rates. Bilateral index methods may be problematic: direct bilateral methods do not include new products in the index calculations in the course of a year, but only at the next base month, while monthly-chained index methods may suffer from chain drift. The comparative study in Chessa *et al*. (2017) shows that weighted bilateral indices may significantly differ from transitive indices, contrary to the condition that price index methods should satisfy in order to exclude chain drift.

In contrast with bilateral methods, which use information from two periods in index calculations, multilateral methods use information from multiple periods. A big advantage of multilateral methods over bilateral methods is that transitive, drift free indices can be calculated using different weights across products, which are even allowed to vary from month to month. However, certain methods, among which the GEKS method (GEKS for Gini-Eltetö-Köves-Szulc), are sensitive to downward biases when applied to dynamic assortments where products leave an assortment under clearance prices (Chessa *et al*., 2017). Such situations are not uncommon for clothing (Chessa, 2016a). We therefore selected a method that does not have the afore-mentioned problems, which we call the "QU method" (Quality adjusted Unit value method), for the scanner data and web scraped data of the online shop. This method was introduced into the Dutch CPI in January 2016 (Chessa, 2016a). When applied to price comparisons over countries, it is also known as the Geary-Khamis (GK) method, which is in fact a special case of the broader class of QU methods. For this reason, we prefer to use the latter term or, more specifically, also "QU-GK".

Index Formula

Chessa *et al*. (2017) compare weighted and unweighted bilateral and multilateral index methods on scanner data sets of four product categories of a different Dutch department store than the one considered in the present paper. The use of weights in index formulas may lead to substantially different results compared to equal weights methods. But the use of weights in bilateral methods may be problematic, in particular when used to calculate monthly chained indices. Such indices may lead to severe drift, which directly results from the intransitivity of monthly-chained bilateral indices.

Direct bilateral indices do not timely capture new products, which are included only at the next base month, unless prices are imputed in months before the month of introduction to an assortment. The comparison for clothing shows that the contribution of new products to an index may be considerable (Chessa *et al*., 2017). Multilateral methods are free of chain drift, allow a timely inclusion of new products and price imputations are not needed.

The assortment dynamics justify the choice for a multilateral method also for the scanner data and web scraped data of the Dutch online shop. The differences among price indices for different multilateral methods are not very large in Chessa *et al*. (2017), but may be significant. The GEKS method, and also the CCDI method recently proposed by Diewert & Fox (2017), are sensitive to clearance prices of outgoing items, which lead to downward biases (Chessa *et al*., 2017). Other methods, like the QU method and the Time Product Dummy method, do not have this drawback.

The QU method was introduced into the Dutch CPI in January 2016; its first application in the CPI was on mobile phones. Since July 2017, it is also applied to scanner data of the Dutch department store referred to above. The QU method can be considered as a family of methods, which also covers some well-known bilateral methods, such as the Laspeyres, Paasche and Fisher indices (see also Auer, 2014). But its primary aim is to construct multilateral, transitive indices. In fact, the method extends the concept of unit value to sets of heterogeneous goods. In order to accomplish this, we have to account for quality differences between products. For this reason, we refer to the method as "Quality adjusted Unit value method", which we abbreviate to "QU method". Other authors, like Auer (2014), speak of Generalised Unit Value.

In order to explain the idea behind the QU method, we first introduce some notation. Let G_0 and G_t denote sets of products that belong to some product category G, for a base month 0 and, say, current month t. The sets of products in 0 and t may be different. Let $p_{i,t}$ and $q_{i,t}$ denote the prices and quantities sold for product $i \in G_t$, respectively, in month t. We want to find scaling factors, say v_i, that transform the prices of different products in month t into "quality

adjusted prices" $p_{i,t} / v_i$. This transformation implies that quantities sold $q_{i,t}$ of each product are converted into quantities $v_i q_{i,t}$. In expression (3) below, the v_i of the products are defined as average deflated prices over a time interval. The v_i could be interpreted as "reference prices" and $v_i q_{i,t}$ as quantities valued at the reference prices of the products.

The price and quantity transformations allow us to define and calculate a "quality adjusted unit value" $\tilde{p}_t$ for a set of products G_t in month t:

$$\tilde{p}_t = \frac{\sum_{i \in G_t} \left(p_{i,t} / v_i \right) \left(v_i q_{i,t} \right)}{\sum_{i \in G_t} v_i q_{i,t}} = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t}}{\sum_{i \in G_t} v_i q_{i,t}} \qquad (1)$$

Note that $\sum_{i \in G_t} p_{i,t} q_{i,t}$, the total expenditure, is not affected by the transformations.

Expression (1) can be used to define a price index by dividing the quality adjusted unit values in two months:

$$P_t = \frac{\tilde{p}_t}{\tilde{p}_0} = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t} / \sum_{i \in G_0} p_{i,0} q_{i,0}}{\sum_{i \in G_t} v_i q_{i,t} / \sum_{i \in G_0} v_i q_{i,0}} \qquad (2)$$

The numerator on the right-hand side of (2) is an index that measures change in turnover or expenditure between two months. The denominator is a weighted quantity index. Expression (2) makes clear why the price index is transitive: both the turnover index and the weighted quantity index are transitive.

The weights v_i are defined as follows over some time interval $[0,T]$:

$$v_i = \sum_{z=0}^{T} \frac{q_{i,z}}{\sum_{s=0}^{T} q_{i,s}} \frac{p_{i,z}}{P_z} \qquad (3)$$

Expression (3) in fact says that the v_i are unit values as well. For each product, the expenditures are summed over the interval $[0,T]$ and divided by the quantities sold of a product over the same time interval. In order to exclude price changes from the v_i and the weighted quantity index, the product prices of different months are deflated by the price index of the product category. The v_i are also known as "reference prices" (usually referred to as international prices in the spatial context). Expression (3) is the choice made for these prices in the Geary-Khamis (GK) method.

Average deflated prices over a period are thus used to obtain the transformed quantities $v_i q_{i,t}$. The product prices of all months in a time interval $[0,T]$ are used, as is usually done in practice, also with other multilateral methods. Nevertheless, it may be worth to consider refinements of (3); for example, discount prices might be excluded from the v_i in order to obtain values that represent quality more closely. This could be investigated in future research.

The choice of prices for defining the v_i is quite common in index theory. The QU method can be regarded as a family of index methods, in the sense that different choices for the v_i lead to different index formulas. In order to illustrate this with several examples, we simply consider the set of products that are sold in both months, that is $G_0 \cap G_t$. If we set $v_i = p_{i,0}$ for each product $i \in G_0 \cap G_t$, then expression (2) turns into a Paasche price index. If we set $v_i = p_{i,t}$ for each product i, then formula (2) becomes a Laspeyres price index. If the v_i are equal for all products, then (2) simplifies to a unit value index. This is precisely what we would expect for products of the same quality, since their quantities sold can be summed without transforming these.

Since the price index acts as a deflator in (3), equations (2) and (3) must be solved simultaneously. Chessa (2016a) describes an iterative algorithm, which starts with arbitrary initial values for the price indices $P_1,\ldots,P_T$, with $P_0 = 1$ (see also Maddison & Rao, 1996). These price indices are substituted in expression (3), so that initial values can be calculated for each v_i. These values are entered in expression (2) to yield updates of the initial price indices. These two steps are repeated until the differences between the price indices in the last two iteration steps satisfy a stop criterion set by the user. More details about the QU or GK method can be found in Geary (1958), Khamis (1972), Auer (2014) and Chessa (2016a).

Before applying the method, a number of questions need to be dealt with, firstly the length of the time interval $[0,T]$, and the way to include additional data, since new data becomes available each month. We address later the issue of the definition of the products included in the sets of goods G_t.

Length of the Time Window

For the choice of the time interval or window we use a fixed base month (December of the

previous year), which is in line with the HICP regulations. The Dutch CPI uses a window length of 13 months and we do the same here.

The impact of changing the window length on price indices has been a subject of investigation in Chessa *et al*. (2017) and more extensively in Chessa (2017a). The first study compared windows of 13 months and the entire period of 50 months for four product categories. Substantial differences were found in one of the categories. In Chessa (2017a), the differences were also quantified at COICOP level. The differences between windows of 13 months and 4 years are in the order of tenths of percentage point in the year on year indices or even negligible for quite a number of COICOPs. There was no difference between the two window lengths at retailer level for a large Dutch supermarket chain.

Weight Updating and Index Calculation

With new data becoming available each month, the inclusion of additional data may lead to different values of the v_i, and the price indices that were calculated until the previous month may change. However, price indices cannot be revised in the CPI, apart from exceptional situations. How can we calculate a price index for a next month, given this "revision problem"?

In theory, the solution of equations (2) and (3) provides us with a set of 13 transitive index numbers for any year $[0,T]$, where the base month 0 denotes December of the previous year and $T = 12$ represents December of the current year. Price indices and product weights or reference prices v_i are calculated for all 13 months of the year simultaneously, so that the v_i have the same value in every month. We could publish the resulting indices if we had the possibility to revise price indices of previous months each time new data of a next month are included in the index calculations. The v_i calculated for December of the current year eventually gives the desired set of values for the product weights, which could be used in each month to obtain transitive indices.

In practice, we cannot forecast the prices in future months, so that the aim of constructing transitive indices will remain, at most, an ideal theoretical benchmark. The inclusion of data of a next month changes the values of the v_i and in turn also the price indices of previous months. Price indices of previous months can usually not be revised in the CPI, which raises the question of how a price index for a next month could be computed.

Different methods have been proposed for updating the v_i and for calculating price indices of a next month. Updating methods are constructed upon choices about three aspects[4]:

- The use of a fixed base month or a moving reference month;

- The adoption of a rolling window against a monthly expanding window. The latter can only be used in combination with a fixed base month;

- The use of a direct index method, a monthly-chained method or a splicing method.

Chessa (2016a) proposed a fixed base month method, a monthly expanding window and a direct method for calculating a price index for a next month. The method uses data from different numbers of months throughout a year (two months in January, three in February, until reaching the maximum number of 13 months in December), and does not require historical data. The direct index method calculates price indices for the current month with respect to the base month by making use of the most recent set of values for the v_i.

The method ensures that the price indices of December are equal to the transitive price indices that would be obtained by making use of the full data of 13 months in every month of the year. This means that the "fixed base monthly expanding window" (FBEW) method is free of chain drift. The use of a direct index method allows us to bypass chain drift. Index series longer than one year are constructed by chaining the series of the current year to the index of December of the previous year, so that some form of chaining is eventually used. But it is a less frequent form of chaining and, moreover, the use of 13-month windows means that the theoretical values of the v_i are allowed to differ from year to year for each product. This is an explicit choice, which could be made to reflect gradual quality changes over time.

The monthly expanding window could also be replaced by a 13-month rolling window, while still calculating price indices with a direct method with respect to a fixed base month. This alternative method is compared with the FBEW

4. Notice that these choices, and therefore the type of updating method, can be applied in combination with any multilateral method. An illustration of this can be found in Chessa et al. *(2017).*

method in Chessa (2017a) and in Lamboray (2017). Differences between the two methods turned out to be very small or negligible. The indices calculated with the updating methods and the transitive "benchmark" indices turned out to be almost the same or even equal in each of the cases studied (Chessa, 2016a; 2017a; 2017b). Large differences occurred occasionally and mostly in short time periods.

A different class of methods uses a moving reference month instead of a fixed base month. A natural choice is to combine a moving reference month with a rolling window of fixed length, as this allows the inclusion of data from a next month in an elegant way. Different methods can be thought of in order to calculate a price index for the current month, which are known as "splicing methods"; see de Haan *et al*. (2016) for an overview and Chessa *et al*. (2017) and Krsinich (2014) for applications.

The "movement splice" (MS) method chains the month on month index of the most recent rolling window to the index of the previous month, while Krsinich's (2014) "window splice" (WS) method chains the year on year index of the most recent, full window to the index of 12 months ago. The MS method is a monthly-chained method, which, as such, is sensitive to chain drift. Although the WS method uses a kind of direct method, it is also a high-frequency chaining method. Empirical results indicate potential drift, which may be substantial (Chessa, 2016b).

Price Indices for Web Scraped and Scanner Data

Preparation of the Data and Methodological Choices

We calculated price indices with the QU method for men's and women's clothing of the Dutch online shop, based on scanner data and exclusively with web scraped data. In order to make meaningful comparisons, we supplemented the scanner data with the metadata from the web scraped data. This was done by linking the two data tables with the retailer specific item codes as linking key. We calculated price indices for eight product categories in both men's clothing (trousers and jeans, coats and jackets, underwear and pyjamas, shirts, shoes, sportswear, sweaters and cardigans, T-shirts and polo shirts) and women's clothing (trousers and jeans, coats and blazers, dresses and skirts, lingerie, shoes, sportswear, sweaters and cardigans, T-shirts and tops).

The eight categories cover about 85 per cent of the total expenditure for men's clothing over the period March 2014 – December 2016, and about 80 per cent for women's clothing. Sale items and 'Premium selection' items were also included.[5]

Product definition is the first important step that has to be made before price index calculation. While this is not the primary focus of this study – aimed at the comparison of scanner data and web scraped data – it is clear that this should be carefully dealt with, as price indices may be very sensitive to variations in the degree of product differentiation (Chessa, 2016a; 2017b).

Clothing items usually show a high degree of churn, which was also evidenced at a less detailed level than the item or GTIN level (cf. Figure IV). Exiting items and new items of the same or similar quality have to be linked in order to prevent indices from a downward bias, the extent of which may be severe when items leave an assortment under clearance prices (Chessa, 2016a). Exiting and new items can be linked by common characteristics, here brand name and "Type", i.e. the most detailed level of item classification.

Items are thus combined into the same group when they are of the same [Brand×Type] groups, which we call "products". Products should be homogeneous, that is, the items in a group should be of the same or comparable quality. This issue should be further explored in a future study, in particular when considering online store data to become part of the CPI. The average size of the products ranges between 7 and 16 items. Considering the fact that item codes and GTINs are usually different for clothing items of different sizes, which can be said to be of the same quality, the above-mentioned range suggests that the product definitions are not broad.

The following choices were made in order to apply the QU method to the scanner data and the web scraped data:

- For scanner data, unit values were calculated for every product in each month in which it was sold. Expenditures and quantities of the items

5. Non-clothing items contained in these two groups were excluded during the extraction of the data for each of the above categories.

sold in a product were summed, and sales values for returned items were excluded;

- For web scraped data, average monthly prices were calculated for each product. The quantities sold were replaced by the total number of web scraped prices for a product in a month, summed over all items. Items may be scraped more than once: multiple numbers are retained in average prices and quantities;

- The QU method was applied with a fixed base month. This is December of each year, as is done in the Dutch CPI. The base month in 2014 is March of the same year, as it is the first month of the period chosen in this study. Window lengths of 13 months were used (of course except for 2014). We did not apply updating methods, but we calculated the weights v_i and the price indices using the complete data of all the months in a year.

We first provide in Table 1 an example of how product prices and quantities are calculated for scanner data and web scraped data.

The price of the product in Table 1, A is calculated from scanner data as a unit value, that is, as the ratio of the summed expenditures over the six items and the summed quantities. Expenditures and quantities for returned items are excluded, which means that these values are summed with the net expenditures and quantities. A product price is calculated from the web scraped data as the ratio of the sum of the scraped item prices over the days in the month and the total number of scraped item prices, summed over the six items (last column of Table 1, B).

Price Indices

Figures V and VI below show price indices computed with each data source for two categories of men's and women's clothing. The price indices from web scraped data follow those computed from scanner data quite well, even the peaks and dips of the scanner data indices. The high correlations between scanner data and web scraped prices and numbers are reflected in the comparison of the price indices. The close match between the price indices for the two data sets is evidenced in the entire set of 16 product categories (see in Appendix the price indices for all the product categories).

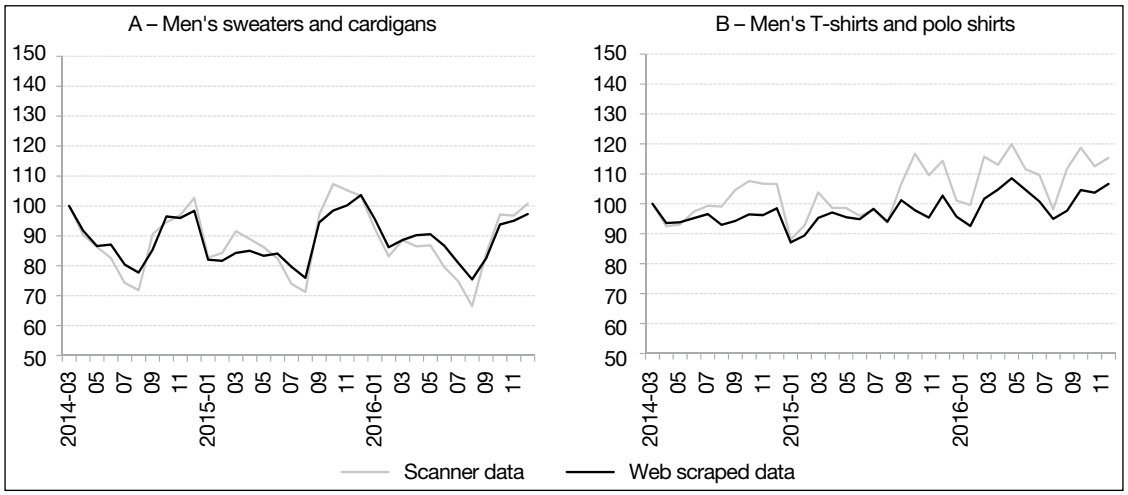
The price indices of the product categories were combined by applying the usual Laspeyres type method. The resulting price indices for the COICOPs men's clothing and women's clothing are shown in Figure VII. We used annually fixed weights for the product categories in the case of scanner data. The category weights were set

Table 1
Computation of product prices and quantities

Item	Nr 1	Nr 2	Nr 3	Nr 4	Nr 5	Nr 6	Product
A – Scanner data							
Net expenditure	0	118	13,201	2,711	25,108	13,009	-
Expenditure returns	75	3,377	7,174	2,257	7,481	15,004	-
Net quantity	0	0	899	186	1,643	986	-
Quantity returns	5	198	372	124	434	812	-
Expenditure	75	3,495	20,375	4,968	32,589	28,013	89,515
Quantity	5	198	1,271	310	2,077	1,798	5,659
Price	14.95	17.65	16.03	16.03	15.69	15.58	15.82
B – Web scraped data							
Number of scraped prices	5	22	31	31	31	29	149
Sum of scraped prices	74.75	392.21	523.22	626.02	523.22	557.57	2,696.99
Price	14.95	17.83	16.88	20.19	16.88	19.23	18.10

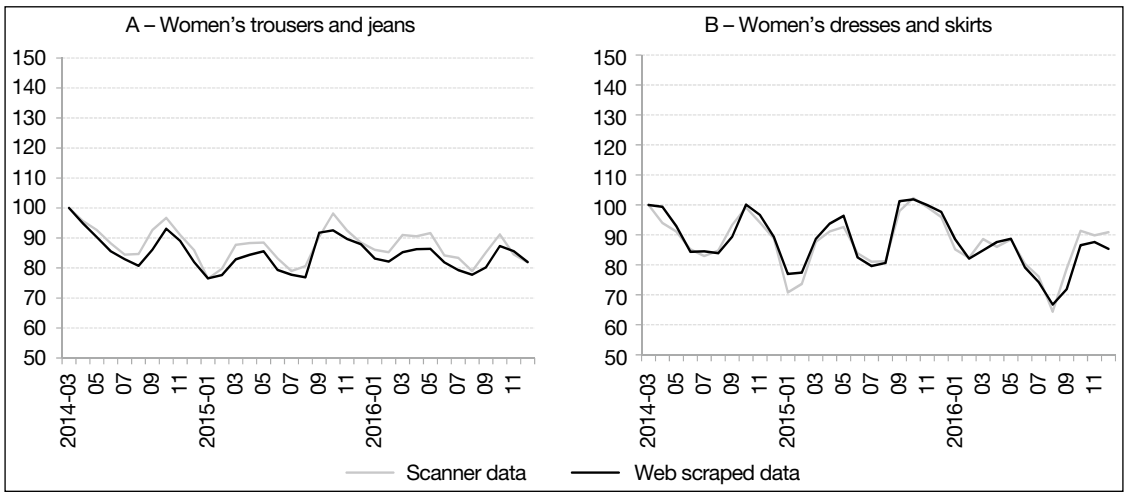
Notes: For scanner data: Expenditures and quantities, both for net values and for returned items of short-sleeve T-shirts of the same brand. The six items have different item codes (indicated as Nrs 1-6), which are combined into the same product based on common characteristics. Total expenditure, total quantity and price (unit value, in euro) of the product are also shown. The values are taken from the scanner data of the online store and apply to one month. For web scraped data: Numbers of scraped prices and the sum of these prices for the same items and month as for scanner data. These values are also shown for the product, which are obtained as sums over the six items.
Sources: Scanner data and web scraped data of clothing products.

Figure V
QU-Indices for two categories of men's clothing, for scanner data and web scraped data (March 2014 = 100)



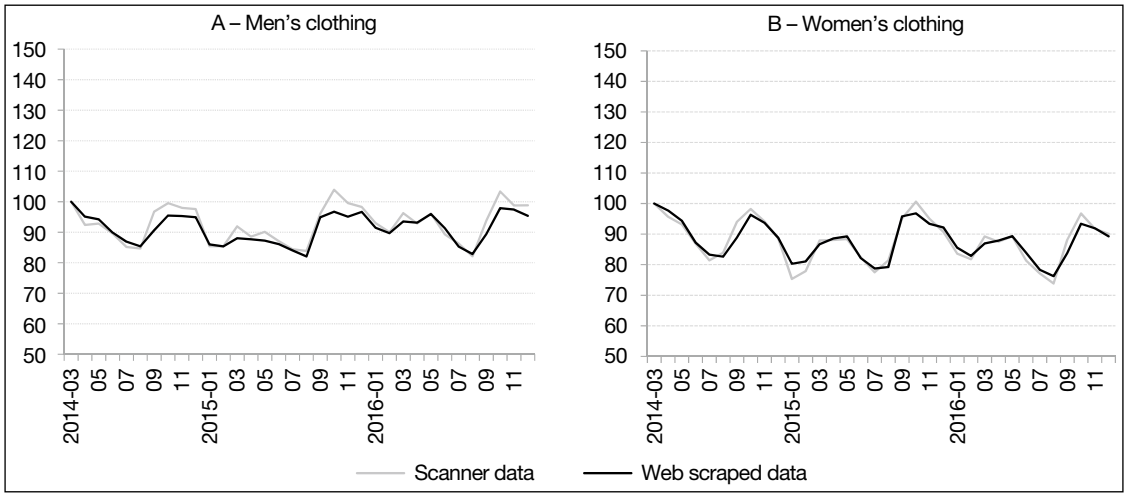
Sources: Scanner data and web scraped data of clothing products.

Figure VI
QU-Indices for two categories of women's clothing, for scanner data and web scraped data (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

Figure VII
Aggregate price indices for men's and women's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

equal to the annual expenditure shares of the categories of the preceding year, except for 2014, as this is the first year in the series. In the latter case, we took the annual expenditure shares of 2014.

For the web scraped data, we replaced expenditure by average price times the number of web scraped product prices, summed over all products in a category over a year. The differences between the scanner data and web scraped indices are very small for the two COICOPs. The differences between year on year indices are only 0.3 percentage point, on average, for both COICOPs.

Sensitivity Analysis

The above results show that using the numbers of web scraped product prices instead of numbers of products sold yields reliable price indices. This finding is consistent with the results of the data analysis presented in the first part of this paper. In order to go further, we investigated whether replacing the numbers of web scraped product prices by numbers that ignore the correlations with the numbers of products sold, would affect the price indices. We replaced the numbers of web scraped prices by 0 or 1, with 0 meaning that no prices were found by the web scraper for a product in a month, while 1 denotes that prices were found, but the exact numbers are ignored. The impact of this change on the price indices is shown below (Figure VIII). The results are only shown at COICOP level.

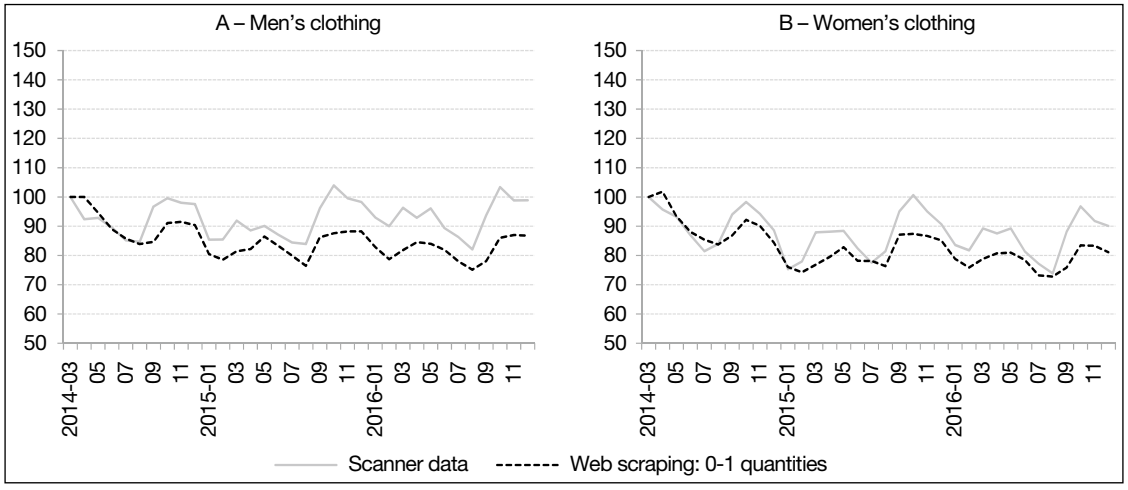
Replacing the numbers of web scraped product prices by 0 or 1 has a big impact on the web scraped indices, which is clearly visible at COICOP level. The results for the 16 product categories are not shown, but we merely mention that similar differences were found in 13 of the 16 categories. Each of these cases shows a downward behaviour of the index (as in Figure VIII).

The differences in the year on year indices are much larger than with the original numbers of web scraped prices. For men's clothing, the average difference with the scanner data indices increases to almost 5 percentage points and to almost 4 percentage points for women's clothing. These results suggest that the original numbers of web scraped prices should be used when calculating price indices from web scraped data. Manipulation of these numbers, like removing double prices, should be discouraged.

* *
*

To our knowledge, the study presented here is the first to compare price indices calculated from scanner data and web scraped data. The comparison was possible because both data sources are available from the same retailer. These first results look very promising, given the remarkable accuracy of the web scraped indices, especially at COICOP level. This is

Figure VIII
Price indices for men's and women's clothing, with the numbers of web scraped product prices replaced by binary values (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

especially valuable, as web scraping is rapidly gaining popularity for official statistics. Scanner data remain the preferred option since it contains transaction data, but not all NSIs have easy access to scanner data.

The positive and valuable outcomes put even greater emphasis on the question why the price indices calculated with only web scraped data are so close to the indices computed with scanner data. At this stage, we can only hint at possible reasons, among which one that comes to mind is related to the fact that the retailer is an online store and does not have physical outlets. Because of this, the retailer may be more inclined towards promoting parts of the store's assortment with high sales. Such items could be made easier to find by the consumer, by placing these under different main groups or categories of the website. For example, the same items could appear both in "Sales" and one of the conventional main groups. This could contribute to explain the high correlations between the numbers of products sold and web scraped product prices. Contacts with the retailers about their strategy to organise their website would help verifying whether they are more likely to promote items on high sales.

More generally, a number of lessons can be derived from this study:

- The method of sampling prices from a website clearly matters. This study shows, at least for the retailer considered here, that scraping an entire website benefits the accuracy of price indices calculated from web scraped data. Sampling entire websites may be time consuming, but statistical institutes could consider sampling on specific days instead of every day.

- The website in this study was scraped by site navigation, the first generation of scrapers built at CBS. This is also a rather time consuming technique, which was an important reason behind our decision to scrape during the night. Online stores make use of dynamic pricing. Prices during shopping hours could be decreased, so missing these prices could explain a part of the differences between web scraped prices and scanner data prices. Meanwhile, we have developed a second generation of scrapers, which extract prices and metadata from the code behind the product overview pages. This is a much faster scraping technique, which makes it possible to scrape even very large websites at various times during a day. In the future, this will allow us to study the impact of dynamic pricing on price indices and to focus on new applications, such as constructing real time indices. The impact of dynamic pricing on price indices is, of course, impossible to quantify here. However, the small differences between the price indices for scanner data and web scraped data suggest that the impact of dynamic pricing would be small in this case.

- This study also suggests to use the original numbers of web scraped prices in price index calculations with web scraped data. Deduplication of prices should be discouraged. The results show that the web scraped indices lose their accuracy when removing multiple prices (cf. Figure VIII), as the difference with the year on year indices based on scanner data increases up to five percentage points per year. In addition, all deviating indices show a downward drift. At the same time, we admit that the removal of multiple prices was done in a rather extreme way, leaving only one observation per product in a month. Nevertheless, the results show that the numbers of originally scraped prices should be treated carefully.

- In spite of the positive findings obtained from this study, it is always worth trying to request expenditure data from retailers, also when retailers cannot, or are not willing to, deliver complete scanner data sets.

At the same time, we should be cautious with our conclusions. Web scraped data are not transaction data and the results of the present study apply to a single retailer. We therefore suggest a number of directions for future research.

This study could be repeated with other online stores whose websites have a similar structure as the one investigated here, that is, where items on discount are promoted more often than other items and where popular items are easier to find. Statistics Netherlands' CPI unit is currently developing web scrapers for retailers of consumer electronics for which scanner data are available. This would provide us with an interesting test case, even more since these retailers have physical outlets. Do they promote items on high sales more often than less popular items on their website? Or do they follow a different strategy, such as publicising new items?

Web scraping is a valuable means for supplementing information about items in scanner data, which may be limited. Combining the two data sources provides the opportunity of using the best from both worlds: transaction data from scanner data and additional information

about item characteristics from web scraped data. In principle, this provides an ideal setting for applying and testing methods for selecting item characteristics and defining homogeneous products and, consequently, for handling relaunches. However, when using web scraped metadata for supplementing the metadata in scanner data sets of physical stores, it should be noted that it may not be possible to supplement all GTINs in scanner data with web scraped data. The assortments of physical and online stores may be different if, for instance, retailers want to include only a part of the items offered in physical outlets on their website.

Finally, we are well aware that comparative studies like the one presented in this paper may be difficult to repeat, as the availability of both scanner data and web scraped data from the same retailer is rare. This is even more difficult for NSIs that encounter problems with the acquisition of scanner data. We therefore encourage NSIs that are in the more fortunate position of possessing scanner data to invest in statistical research on scanner data. Is it possible, through statistical analyses and tests, to obtain a characterisation of scanner data? Is it possible to derive specific patterns, for instance how prices and quantities correlate over time? Applying the same analyses to web scraped data could give indications on the extent of similarity with scanner data and a better idea of the suitability of web scraped data for price index calculation. We therefore suggest that more attention is given to time series analyses and other statistical analyses of scanner data. □

BIBLIOGRAPHY

Auer, L. von (2014). The Generalized Unit Value Index Family. *Review of Income and Wealth*, 60, 843–861. https://doi.org/10.1111/roiw.12042

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, M., Payne, C., Smith, T., Winton, J. & Woods, A. (2016). Research indices using web scraped data: May 2016 update. Office for National Statistics, internal report, 23 May 2016. https://www.ons.gov.uk/releases/researchindicesusingwebscrapedpricedatamay2016update

Cavallo, A. F. (2016). Are online and offline prices similar? Evidence from large multi-channel retailers. NBER, *Working Paper* N° 22142. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_MIT_are_online_and_offline_prices_similar.pdf

Chessa, A. G. (2016a). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2016, 49–69. https://ec.europa.eu/eurostat/cros/content/new-methodology-processing-scanner-data-dutch-cpi-antonio-g-chessa_en

Chessa, A. G. (2016b). Comparisons of the QU-method with other index methods for scanner data. Paper prepared for the first meeting on multilateral methods organised by Eurostat, Luxembourg, 7-8 December 2016. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017a). Comparisons of QU-GK indices for different lengths of the time window and updating methods. Paper prepared for the second meeting on multilateral methods organised by Eurostat, Luxembourg, 14-15 March 2017. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017b). The QU-method: A new methodology for processing scanner data. *Statistics Canada International Symposium Series : Proceedings*. https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X201700014752

Chessa, A. G., Verburg, J. & Willenborg, L. (2017). A comparison of price index methods for scanner data. Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/A%20comparison%20of%20price%20index%20methods%20for%20scanner%20data%20-Antonio%20Chessa,%20Johan%20Verburg,%20Leon%20Willenborg%20-Paper.pdf

Daas, P. J. H. & van Nederpelt, P. W. M. (2010). Application of the object oriented quality management model to secondary data sources. Statistics Netherlands, *Discussion paper* N° 10012.

Daas, P. J. H. & Ossen, S. J. L. (2010). In search of the composition of data quality in statistics and other research areas. Statistics Netherlands, *Discussion paper*.

Diewert, W. E. & Fox, K. J. (2017). Substitution bias in multilateral methods for CPI construction using scanner data. Vancouver School of Economics, The University of British Columbia, *Discussion paper* N° 17-02.
https://irs.princeton.edu/sites/irs/files/Diewert%20and%20Fox%20Substitution%20Bias%20and%20MultilateralMethodsForCPI_DP17-02_March23.pdf

Eurostat (2017). *Practical Guide for Processing Supermarket Scanner Data*. September 2017.
https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

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.
https://doi.org/10.2307/2342991

Griffioen, A. R. & ten Bosch, O. (2016). On the use of internet data for the Dutch CPI. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_Netherlands_on_the_use_of_internet_data_for_the_Dutch_CPI.pdf

Griffioen, A. R., ten Bosch, O. & Hoogteijling, E. H. J. (2016). Challenges and solutions to the use of internet data in the Dutch CPI. Paper presented at the *UNECE Workshop on Statistical Data Collection,* The Hague, The Netherlands, 3-5 October 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2016/mtg1/WP2-3_Netherlands_-_Griffioen_ap.pdf

de Haan, J., Willenborg, L. & Chessa, A. G. (2016). An overview of price index methods for scanner data. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_1_room_doc_Netherlands_an_overview_of_price_index_methods.pdf

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications.
https://doi.org/10.5089/9787509510148.069

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.
https://doi.org/10.2307/2345041

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 *UNECE-ILO Meeting of the group of experts on consumer price indices*, Geneva, Switzerland, 26-28 May 2014.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf

Lamboray, C. (2017). The Geary Khamis index and the Lehr index: how much do they differ? Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/The%20Geary%20Khamis%20index%20and%20the%20Lehr%20index%20how%20much%20do%20they%20differ%20-%20Claude%20Lamboray%20-Paper.pdf

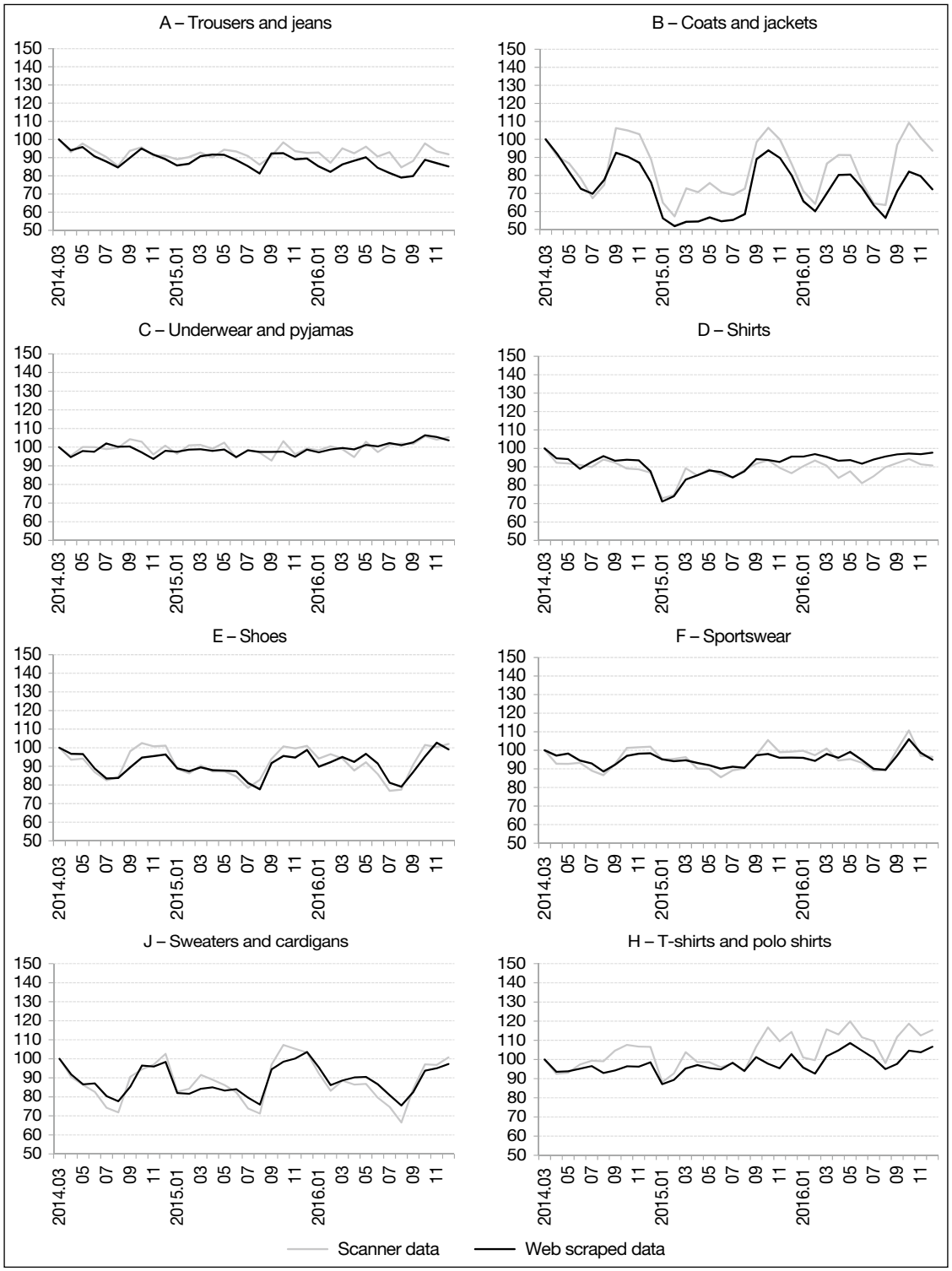
Maddison, A. & Rao, D. S. P. (1996). A generalized approach to international comparison of agricultural output and productivity. Groningen Growth and Development Centre, Research memorandum GD-27.
https://www.rug.nl/research/portal/files/3258249/GD-27.pdf

Willenborg, L. (2017). Quantifying the dynamics of populations of articles. Statistics Netherlands, *Discussion Paper* N° 2017/10.
https://www.cbs.nl/en-gb/background/2017/25/quantifying-the-dynamics-of-populations-of-articles

APPENDIX

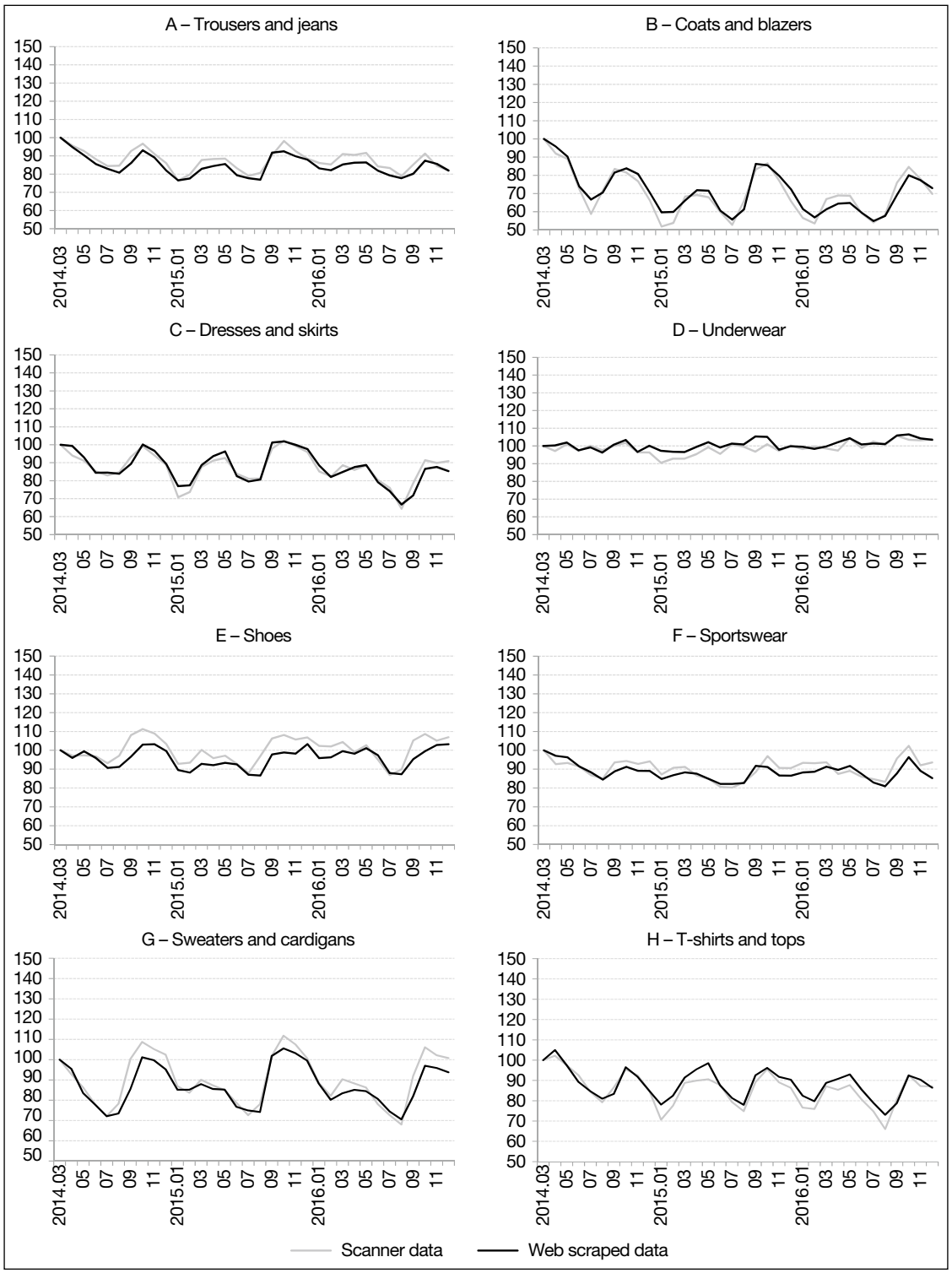
SCANNER DATA AND WEB SCRAPED DATA PRICE INDICES FOR 16 PRODUCT CATEGORIES OF MEN'S CLOTHING AND WOMEN'S CLOTHING

Figure A-I
Men's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

Figure A-II
Women's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.