Discussion Paper

Elementary price indices for internet data

2017 | 08

Leon Willenborg

Content

Summary

The paper assumes data of various web shops to be available. These data are used to calculate price index information. The assumption is that at the lowest level no turnover information is available. At a higher level of aggregation proxy turnover information may be available, so that price developments for various articles can be weighted according to economic importance. The paper concentrates on methods to calculate (unweighted) elementary price indices, using stratification of the articles. The method is applied at the subgroup level and yields price indices at the group level. These can in turn be aggregated to the group level, using weights. This step is not considered here, however. The method starts with price comparisons at the subgroup level and aggregates them to the group level. The price indices calculated in this fashion need not be transitive. They can be transitivized, however, as is shown in a companion paper (Willenborg, 2017). The methods described in the present paper are a first step in the process to produce transitive price indices for web shop data.

Keywords

Elementary price indices, web data, internet shops, supply prices

1. Introduction

COICOPs[1] for clothing are fairly crudely defined: women's clothes, men's clothes, children's clothes and baby's clothes. This subdivision of clothing is too crude for the calculation of index numbers. In fact they are more suitable for publication of index numbers, and as building blocks for the CPI. [2] To calculate index numbers there should be a more refined subdivision of the four COICOPS especially if one wants to apply stratification to calculate price indices.

In a project at Statistics Netherlands to improve the data collection for the CPI[3] a stratification method was proposed, that we shall refer to as the group method, to calculate price indices for clothing, using data collected from a web shop A. This method is based on an internal classification CBS uses for clothing. This was especially designed for the group method. The group method was chosen as it is able to handle the dynamics of the clothing population, that is, new and disappearing articles. The method is fairly easy to use in practice, provided one is able to automatically classify the clothing items. Due to the large volume of data, automatic coding is indispensable for processing the data. In principle the group method is generally applicable, provided suitable metadata, that is, descriptions of the clothing items, is available. But the group method as such, irrespective of the groups being used, is general and fairly simple to apply. Also the internal classification of clothing is independent of a particular (web) shop. Because of this, the classes of the internal classification of clothing have been defined rather broadly, but still a lot more detailed than the COICOPS for clothing.

When the group method was proposed it was intended as a preliminary method that allowed to get the processing of the internet data started. In order to avoid complications in the processing the majority of the groups were purposely made big enough to be nonempty every month with high probability. This means that for each month of the year (with great confidence) prices will be available for most of these groups. The web scraper collected information from the web shop that was 'targeted' on a daily basis. The method chosen was to be simple to allow price indices to be calculated easily. This was the group method. Later we would look for refinements. In fact the current report is an attempt at such a refinement, staying in the same spirit of the original method: instead of working with the classes of the internal classification of clothing the idea is to use subgroups thereof, in order to employ more

[1] COICOP = Classification of individual consumption according to purpose is a reference classification published by the UN Statistics Division.
[2] CPI = Consumer Price Index.
[3] The aim of the project was to reduce the traditional data collection using price takers visiting brick-and-mortar shops to collect price quotes for a selection of goods ('a basket'). Instead scanner data provided by shops and prices scraped from web shops were to replace these data.

homogeneous strata. But the method is still a stratification method, requiring no parametric model.

Such a refinement of strata requires the use of more metadata associated with the clothing products. Only part of this information that was actually available was used in the original method, in order to keep it simple. All this meta-information is alpha-numeric. It should be noted that a consequence of using the subgroup method is that there are quite a lot of items that do not exist the whole year round. They may include seasonal items, such as summer jackets.

The subgroup method can be applied for various clothing web shops. This may result in different refinements of the internal classification of clothing across the web shops, as they may in fact use a (partly) different selection of garments and different ways to describe them. So this in fact would lead to classifications that in their greatest detail are dependent on the various web shops. But from a certain intermediate level upwards they are all the same and are equal to the internal classification of clothing. So the internal classification of clothing acts as the common 'backbone'.

It should be noted that the traditional method to link items (on the basis of a primary key, such as the EAN)[4] has been sidestepped with the application of the group method. This was a deliberate action. The reason was that no information was available to link similar items on the basis of this information (exact matching). There is no information available to link a relaunch to one (or more) previous items. Due to the large amount of data available it is impossible to do this linking by hand, or even semi-automatically; this would be far too laborious.[5]

The group method is also markedly different from the method traditionally applied in the CPI in particular for clothing, which is based on collection of price quotes of articles in (physical) shops by price takers. The idea behind the traditional method is to observe the prices of a selective number of articles in consecutive months. This allows to compute the elementary price indices at the article level, from which they can be aggregated upward. If an article has disappeared (is not for sale anymore, or is not observed in the store when the price taker visits) a replacement has to be found, that will be observed from then on as long as possible, until another replacement is needed. In this case the selection of the articles and possible replacements of articles that have disappeared temporarily of for ever, is done by the price takers. This intelligence can be applied because a relative small number of articles is involved and observations are made only once a month. Compared to the amount of web data, this is only a tiny amount of information that can be handled and processed quite differently, in a labour intensive, non-transparent, irreproducible process.

The present report is ultimately is about the subgroup method. But it presents build-up towards that method by showing several other methods to calculate monthly

[4] EAN = European Article Numbering, now referred to as International Article Numbering (but retaining the abbreviation EAN), is a 13 digit barcoding standard (12 digits for data and 1 for checking) .

[5] This is not to imply that the use of EANs to link items is superfluous. For products that exist a quite long time there is every reason to use it.

prices as well, including discussions of their advantages and disadvantages for dealing with web data.

The paper is organised as follows. In Section 2 the role of elementary price indices for internet data are discussed. Several methods are presented that can be used to calculate monthly price averages from these internet data. Section 3 is about several methods to calculate price averages of items in a web shop observed at different months. Section 4 deals with some methods that use exact matching, using keys like EANs or web-ids,[6] to link items observed in different months, so that their prices can be compared. Section 5 discusses the group method, which in fact is a method based on a (rough) classification to compare average prices for different months. This method can be seen as derived from the ones in the previous section by relaxing the condition that exactly the same items are compared with respect to their prices. Only the fact that the items are in the same class in the internal classification of clothing is used, which means that the items are comparable. This property is a favourable one of the group method. In this respect it is superior to the matched items methods of Section 4, for clothing data which is about a highly dynamic population. The group method also is fairly easy to apply. Besides the method yields transitive price indices. Automatically classifying the clothing items (with respect to the internal classification of clothing) is the most challenging part computationally. A possible drawback is that certain classes in the internal classification of clothing may be too heterogeneous. Therefore, in Section 5, more homogeneous subgroups are considered as a basis for index calculations. Several methods are discussed that can be used to compare average prices between different months. The subgroups are formed by using characteristics of the items, used as the dimensions that underlie the internal classification of clothing, and that can be found on the website. Properties like type of clothing, fabric, colour, brand, etc. The subgroups are refinements of the classes of the internal classification of clothing, and they are web shop dependent. The methods in this section may yield elementary indices that are not quite consistent, in the sense that they may not be transitive. Such inconsistencies are no problem, however, as they can be mended. See Willenborg (2017) for a method to produce transitive price indices from nontransitive ones. Section 6 concludes the main text of the present paper with a discussion of the main results and a with a view on possible future research. A list of references completes the present document.[7]

2. Initial thoughts

We discuss several methods to calculate elementary price indices. The application we have in mind (but it is not the only one to which the methods proposed may be

[6] Web-ids are shop dependent. They are local keys, that is, unique within a limited time window. They may be reused when a product has been withdrawn from the market. A web-id may be re-used for a completely different product.

[7] The present document was reviewed by Sander Scholtus.

applied) is internet data on clothing. We assume that for each item there is a key that identifies an article uniquely. This is used by the shops for inventory and ordering purposes. We furthermore assume that for each article there are characteristics available to describe the items. These are used by the customers to get informed about some characteristics of the items so they can select and make purchase decisions.

In standard price index theory the population of items (or articles) is assumed to be static. That is no existing items ever disappear from the market, nor are new items introduced into the market. This implies that existing items are immutable in their qualities.

The static item universe is an idealization that never really existed, not even in the past, although things changed less rapidly than now. It is a good starting point for the theory, however. However, a more realistic assumption is of a dynamic population of items. In this, items are constantly removed from the market, and new items are introduced, sometimes only as guises of previous items ("relaunches"). But sometimes truly innovative items arrive, that have no precursors.[8]

Starting point is a group of articles that are observed at different times ($t = 1,2$), both the prices of the articles that are present (including their keys). The articles are supposed to belong to the same class in the internal classification of clothing. This implies that these articles have a certain "kinship", although the classes of the internal classification of clothing can be rather heterogeneous. For the internal classification of clothing proxy-weights for groups at the most detailed level in the internal classification of clothing are available. These weights are proportional to sales of the items in a certain period. There are (theoretically) at least three sources for these weights in case of clothing: the web shop itself (web shop A in our case), GfK[9] and the Budget Survey. Within a class of the internal classification of clothing all articles have the same weight.

In the sections that follow we deal with a series of methods to calculate price indices for clothing, and more in particular for web shop A internet data. We start with a matching model approach and introduce a number of variations and alternatives. For each method advantages and disadvantages are discussed. In particular we deal with the group method that was proposed for web shop A. Although this method yields seasonal patterns, and qualitatively at least, gives reasonable results, the classes in the internal classification of clothing on which it is based, are sometimes quite broad. This was done with a purpose: in this way nearly every month should yield average prices for each class in the internal classification of clothing, and the group method is then easily applied to calculate price indices. It would be attractive to look for smaller groups which increase homogeneity, and for an alternative for the group method in case some months have no average prices. The information to produce these refined

[8] Smart phones as a fairly recent example of a class of such items. The clothing area seems to be less innovative than electronics, which is probably partly due to a certain conservatism of the customers.

[9] GfK SE, Gesellschaft für Konsumforschung (Society for Consumer Research) is Germany's largest market research institute and the fourth largest market organization in the world.

stratifications should be available from the website, for each clothing article (its meta-information).

Before we turn to methods to calculate elementary price indices, we first consider ways to calculate average prices for groups of articles observed during an entire month.

3. Calculation of monthly prices

This section is about the phase preceding the making of elementary price indices. These are the building blocks for the aggregated price indices.

3.1 Various methods for price calculations

Consider a group of articles, whose price are observed daily, during an entire month. We define three indicator functions δ, η, ζ as follows:

$$\delta_{ij} = \begin{cases} 1 & \text{If article i is observed on day j,} \\ 0 & \text{If article i is not observed on day j,} \end{cases}$$

$$\eta_{ij} = \begin{cases} 1 & \text{If article i is observed on day j,} \\ -\infty & \text{If article i is not observed on day j,} \end{cases}$$

$$\zeta_{ij} = \begin{cases} 1 & \text{If article i is observed on day j,} \\ \infty & \text{If article i is not observed on day j,} \end{cases}$$

Each of these functions is handy in a particular situation with missing values. These missing values may occur because an article's price is not available in the data, either because the article was not present on the website on a particular day, for whatever reason (out-of-stock / temporarily not available, permanently not available), or because the web scraper did not pick up the information because the webpage was changed, etc. The values $-\infty$ and ∞ are to be interpreted as 'missing'. They do not refer to different missing values, however. Only, one is convenient to use in one situation, and the other in another. Both values should be added to $\mathbb{R}$ so we get $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, with the understanding that $-\infty \leq x \leq \infty$ for all $x \epsilon \overline{\mathbb{R}}$, and $x \cdot (-\infty) = -\infty$ for all $x \geq 0$, $x \cdot \infty = \infty$ for all $x \geq 0$, and $x + (-\infty) = -\infty$ for all $x \epsilon \mathbb{R} \cup \{-\infty\}$, $x + \infty = \infty$ for all $x \epsilon \mathbb{R} \cup \{\infty\}$.

In Table 3.1.1 we have indicated the presence and missingness of certain prices and their values. This uses the δ-indicator function, which is appropriate in some cases. For other cases (see below) one should use the η or ζ indicator functions instead. But

we do not show similar tables with these indicator functions. For instance, if $p_{11}\delta_{11} = 0$ because article 1 is missing in month 1 (δ_{11}=0) , this is not the same as $p_{11}\delta_{11} = 0$ because $p_{11} = 0$ and $\delta_{11} = 1$.

3.1.1 Matrix with monthly prices of articles in one group and for one month.

		Days				
		Day 1	**…**	**Day j**	**…**	**Day l**
Articles	A_1	$p_{11}\delta_{11}$	…	…	…	$p_{1l}\delta_{1l}$
	…	…				…
	A_i	…		$p_{ij}\delta_{ij}$		…
	…	…				…
	A_k	$p_{k1}\delta_{k1}$	…	…	…	$p_{kl}\delta_{kl}$

Note that the δ's in Table 3.1.1 act as quantities.[10]

The aim now is to calculate from these numbers a monthly average. This can be done in a number of ways, some of which we discuss here.

The methods we consider are the following:

1. First calculate the (time-)average price per article, and then average these article-averages over the articles.
2. Calculate the lowest and the highest price per article in a month. Then average each of these over the articles. This may be used to calculate intervals for the average prices.
3. First calculate the average price for each day, yielding a day-average. Then calculate the averages of the day-averages.
4. Calculate an average over all the observed prices (various averages are possible, as well as robust versions such as the median and midrange, based on point 2. above).

The averaging can be done in many different ways. See for instance the interesting article in Wikipedia on this topic (en.wikipedia.org/wiki/Average). Popular in price index theory and applications are the arithmetical and geometric mean. Also in use is the harmonic mean. In addition, robust averages should be mentioned, like the median, the winsorized / trimmed mean, the midrange, although they are not used in traditional price index applications.[11] In big data applications they might, however, prove to be useful. Other averages than the ones treated below are possible, for instance based on geometric means. However, as we do not wish to dwell on this

[10] Supply quantities, not sales quantities. For internet data they cannot be retrieved from the webpages of the web shops. It is possible that some information of turnover is available in the data: when diversification of clothing item is linked to their turnover (high turnover ↔ highly diversified supply of group of clothing items).

[11] But here filters are used to screen the data and eliminate outliers. This is also a form of using robust estimation, when used in combination with traditional estimators.

topic too long, we have chosen to concentrate on arithmetical means, as the group method uses such a mean.

In order to make the above verbal descriptions more concrete we cast them in formulas. We assume that we work with a single group of articles, in a particular month of some year. We introduce some notation and along the way we introduce some interesting estimators, not all of which will actually be used in the sequel of the present report. They are introduced simply because they should be tested in applications or in experiments.

We start with some counters:

- $d_i = \sum_{j=1}^{l} \delta_{ij}$, the number of days in the reference month that article i is present;
- $a_j = \sum_{i=1}^{k} \delta_{ij}$, the number of articles present (or observed by the web scraper) on day j of the reference month;
- $t = \sum_{i=1}^{k} \sum_{j=1}^{l} \delta_{ij}$, the total number of prices observed of articles in the reference group in the reference month;

We then proceed with some estimates that can be used to represent "the price" of the articles in the reference (sub)group in the reference month. We use $a \wedge b$ to denote the minimum of the numbers a and b, and $a \vee b$ to denote the maximum of the numbers a and b. If a or b are missing then this missing is represented as a $-\infty$ (in case $\vee$ is used) or ∞ (in case $\wedge$ is used). In practice there is only one symbol to indicate missingness,[12] such as 'NA' in the R language. Depending on the operator being used below, we assume that the missingness indicator is interpreted either as $-\infty$ or as ∞. In case δ is used the value of the missing price is not important, as long as it is agreed that $\odot \cdot 0 = 0$, where $\odot$ is a missing value, which is nothing but a real value which happens to be unknown.

- $\bar{p}_{i.} = \frac{1}{d_i} \sum_{j=1}^{l} p_{ij} \delta_{ij}$, the average price of article i in the reference month;
- $p_i^{\vee} = p_{i1}\eta_{i1} \vee \ldots \vee p_{il}\eta_{il}$, the highest price for article i in the reference month;
- $p_i^{\wedge} = p_{i1}\zeta_{i1} \wedge \ldots \wedge p_{il}\zeta_{il}$, the lowest price for article i in the reference month;

It should be noted that in the expressions above, replacing the employed indicator function by any of the two other indicator functions yields results that are incorrect.

The last two estimators yield upper and lower bounds of the price of article I in the reference month, and we have: $p_i^{\wedge} \leq \bar{p}_{i.} \leq p_i^{\vee}$. Instead of $\bar{p}_{i.}$ we can also use the mid-price, defined as follows:

- $p_i^{mid} = \frac{1}{2}(p_i^{\wedge} + p_i^{\vee})$, the midrange price for article i in the reference month.

It can be used as an average price for article i in the reference month.

[12] There may be different reasons for value being missing: refused to answer, original answer was wrong and replaced by a missing value, etc. But that is not the point here.

We can also use $p_i^{\wedge}$ and $p_i^{\vee}$ to define lower and upper bounds for the average price of the reference group in the reference month. First we introduce some extra indicators.

- $\delta_i = \delta_{i1} \vee \ldots \vee \delta_{il}$, the indicator for article *i* in the reference month, indicating whether *i* is present or not in this month;
- $\delta = \sum_{i=1}^{k} \delta_i$, the number of articles observed in the reference month, i.e. for which at least one price was observed in this month;

With these indicators we define the equivalents of $p_i^{\wedge}$ and $p_i^{\vee}$ for the entire reference group:

- $p^{\vee} = \frac{1}{\delta}\sum_{i=1}^{k} p_i^{\vee}$ if $\delta > 0$, the average highest price of articles in the reference group in the reference month;
- $p^{\wedge} = \frac{1}{\delta}\sum_{i=1}^{k} p_i^{\wedge}$ if $\delta > 0$, the average lowest price of articles in the reference group in the reference month;

$p^{\vee}$ and $p^{\wedge}$ are only well defined in a month for a particular group of articles if there is at least one article observed in that month for that group of articles.

Using $p^{\vee}$ and $p^{\wedge}$ we can define the equivalent of the mid-range price p_i^{mid} for the entire group:

- $p^{mid} = \frac{1}{2}(p^{\wedge} + p^{\vee})$, the mid-range price for the reference group in the reference month.

When considering all available prices in the reference month, one can define the average of these prices as follows:

- $\bar{p} = \frac{1}{t}\sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij}\delta_{ij}$, the average price of all articles in the reference group in the reference month. t is a counter defined above.

There is another type of estimator that may be of interest, at the article level. It can be most easily described in words. For article *i*, it is the first price observed for this article in the reference month. Let M^i be the set of days in the reference month for which article *i* is observed. Let Δ_i denote the first day in M^i. In case the days are indicated by numbers, it is the smallest number in M^i. So $p_{i\Delta_i}$ is an estimator for the price of article *i* in the reference month.

This estimator is interesting if a different strategy is used for collecting price information on the internet. Instead of collecting as many prices as possible during a month (as in the approach used for web shop A), the goal could be to collect only one price for each article in the group, and as soon as possible. Once a price has been obtained for an article no further prices will be obtained anymore for that article in that month. The first price observed for this article in the reference month will be its estimate for that month. This would require an observational approach that is different from the one that is currently used. The current approach can be likened to trawling the sea floor by a fishing boat with drag net. Instead the alternative method

is a targeted search for prices. It would require a web scraper that operates quite differently from the currently used ones, based on trawling.

These 'search estimators' can in turn be used to calculate an average monthly price for the reference group, assuming $\delta > 0$:

$$p^1 = \frac{1}{\delta}\sum_{i=1}^{k} p_{i\Delta_i}\,\delta_i.$$

So far our attention was focused on the articles, with the exception of the estimator $\bar{p}$. Instead we could focus on the days in the month and define the average daily price of the articles in the reference group, assuming $\delta > 0$:

$$\bar{p}_{.j} = \frac{\sum_{i=1}^{k} p_{ij}\delta_{ij}}{\sum_{i=1}^{k} \delta_{ij}}.$$

3.2 Monthly prices and elementary price indices

The monthly prices are ingredients for the elementary price indices. They are the building blocks for the aggregated price indices (at COICOP level) that we are actually interested in. There are various possibilities to use average monthly prices to calculate elementary price indices. The methods may either use a parametric model are not. If such a model is not used, one can use approaches based on exact matching of articles or on statistical matching. Several such methods are discussed later. Each method has its own advantages and disadvantages. None seems to be clearly superior.

4. Methods using exact matching

4.1 Exactly matched items: static population

Characteristic for this method is that the prices of exactly the same articles at two different points in time are compared, by calculating the price ratios. In Table 4.1.1 an example is shown featuring 4 articles. The method aims at comparing prices of the same items at different points in time. The 'sameness' of the articles is guaranteed by using a primary key (like EAN or web-id). So this method finds exactly the same elements, by what is generally known as exact matching in the literature on data linkage.[13]

[13] Or object identity matching as it is called in the Memobust handbook, describing methods used in modern business statistics. See the contribution: Microfusion - Object identifier matching (https://ec.europa.eu/eurostat/cros/content/object-identifier-matching-method_en).

Using the arithmetic mean yields a Carli price index, which in itself is not a very attractive price index, as it does not satisfy, the time reversibility test and the transitivity test. But it is still mentioned because it can be used to produce transitive price indices (which are also time reversible). So the transitivization method to be applied can mend some defects of a price index. But it seems preferable to use geometric means in this situation (yielding a Jevons index), which has some nice properties, despite the fact that it uses weights derived from turnover.

Note that the geometric mean in Table 4.1.1 equals the ratio of the geometric means of the prices measured at different points in time (t=1, 2). So when this formula is used, it is not required to consider prices of the same article at different points in time. In case of the arithmetic mean, however, in the first formula, such a pairing is necessary. The resulting method is, however, not recommended. This relaxation for the geometric mean can be used to generalize the method illustrated in Table 4.1.1 to the one illustrated in Table 4.1.2.

In Table 4.1.2 the arithmetic mean leads to a Dutot index, and the geometric mean to a Jevons index. Both are acceptable elementary price indices, being time reversal and transitive.

Remark In web shops web-ids may be available as identifiers for articles. When the web scraper collects information from a website on a daily[14] basis, and such web-ids are available, it seems appropriate and convenient to use them. However, one should also keep an eye on the items that do not appear in two consecutive months, i.e. the new ('inflowing') or disappearing ('outflowing') articles. They should not be ignored. When a group or subgroup method is used this is automatically taken into account. ■

Note that the method in Table 4.1.2 presupposes that the articles can be distinguished on the basis of their key values, but that this property is not used in fact, when group averages are calculated. Price changes are 'pure', that is, are not due to different compositions of the groups at the two moments in time. Furthermore, it should be noted that the geometric average is the same as in Table 4.1.1, but that this usually is not the case for arithmetic averages (i.e. $F_1 \neq F_3$ in general).

From a mathematical (more in particular, algebraic) point of view, the geometric average is more attractive than an average that uses arithmetic means. The mathematical operations of multiplication and its generalization (raising to a power) are better suited for the definition of price indices as a ratio of prices (in this case without weights). A ratio involves a division, which is directly related to the operation of multiplication.[15]

[14] Or even less frequently, such as weekly or perhaps even monthly.

[15] However, a statistical view may throw a somewhat different light on this remark. In particular when prices are small, or even zero, problems may arise without special precautions. But such values are to be considered outliers for geometric means, as big values are for arithmetic averages. This is apparent if the (natural) logarithm of the prices is considered.

4.1.1 Pairs of articles.

Key	Price at t=1	Price at t=2	Price ratios
K1	p_{11}	p_{12}	$f_1 = \frac{p_{12}}{p_{11}}$
K2	p_{21}	p_{22}	$f_2 = \frac{p_{22}}{p_{21}}$
K3	p_{31}	p_{32}	$f_3 = \frac{p_{32}}{p_{31}}$
K4	p_{41}	p_{42}	$f_4 = \frac{p_{42}}{p_{41}}$
Arithmetic mean (Carli index)			$F_1 = \frac{1}{4}(f_1 + f_2 + f_3 + f_4) = \frac{1}{4}\left(\frac{p_{12}}{p_{11}} + \frac{p_{22}}{p_{21}} + \frac{p_{32}}{p_{31}} + \frac{p_{42}}{p_{41}}\right)$
Geometric mean (Jevons index)			$F_2 = \sqrt[4]{f_1 f_2 f_3 f_4} = \sqrt[4]{\frac{p_{12}p_{22}p_{32}p_{42}}{p_{11}p_{21}p_{31}p_{41}}}$

4.1.2 Group averages, with matched pairs of articles.

Key	Price at t=1	Price at t=2	Price ratios
K1	p_{11}	p_{12}	
K2	p_{21}	p_{22}	
K3	p_{31}	p_{32}	
K4	p_{41}	p_{42}	
Arithmetic mean (Dutot index)			$F_3 = \frac{\frac{1}{4}(p_{12} + p_{22} + p_{32} + p_{42})}{\frac{1}{4}(p_{11} + p_{21} + p_{31} + p_{41})} = \frac{p_{12} + p_{22} + p_{32} + p_{42}}{p_{11} + p_{21} + p_{31} + p_{41}}$
Geometric mean (Jevons index)			$F_4 = \frac{\sqrt[4]{p_{12}p_{22}p_{32}p_{42}}}{\sqrt[4]{p_{11}p_{21}p_{31}p_{41}}} = F_2$

4.2 Exactly matched items: dynamic population

This method is the same as the exact matching method, but this time the population of articles is dynamic rather than static. It is now suited to cope with relaunches. A relaunch is the introduction of the variant of an article that was on sale earlier, possibly with some modifications, which may be very superficial, such as a different packaging. But also the contents of the package may have changed (think of a bigger bottle of shampoo, or a bigger tube of toothpaste), which may require actual adjustments of the prices. The problem with relaunches is that in our applications it is often not known which articles they replace. For the traditional method where small samples of articles are followed through time, a few relaunches can easily be handled. But when dealing with 'big data' such as internet data or scanner data this is catastrophic for the application of the exact matching method, as it cannot be done by hand or even semi-automatically, due to the sheer volume of data.

We now consider several examples.

Example with price comparisons considering only 'flow'.

In this example we have disappearing and new articles, for only the articles present in both periods are used for price comparisons. Table 4.2.1 describes a simple situation at two time periods.

4.2.1 Matched model with 'inflow' and 'outflow'

Key	Price at t=1	Price at t=2	Price ratios
K1	p_{11}	-	-
K2	p_{21}	p_{22}	$f_2 = \frac{p_{22}}{p_{21}}$
K3	p_{31}	p_{32}	$f_3 = \frac{p_{32}}{p_{31}}$
K4	-	p_{42}	-
Arithmetic mean (Carli index for matched pairs only)			$F_5 = \frac{1}{2}(f_2 + f_3) = \frac{1}{2}\left(\frac{p_{22}}{p_{21}} + \frac{p_{32}}{p_{31}}\right)$
Geometric mean (Jevons index for matched pairs only)			$F_6 = \sqrt[2]{f_2 f_3} = \sqrt[2]{\frac{p_{22}p_{32}}{p_{21}p_{31}}}$

In Table 4.2.1 we have three cases: 'outflow'/disappearance of an 'old' article (article K_1), 'flow' / continuing articles (articles K_2, K_3), 'inflow' / appearance of a new article (article K_4). If matched articles are used, only continuing articles are used to calculate an index for t=1→t=2. The selection of continuing articles is only possible because the keys are used for both periods. The situation in Table 4.2.1 generalizes that of Table 4.1.1. The inflowing and outflowing articles are excluded from the index calculations. This is the case for the transition $t = 1 \rightarrow t = 2$. Inflow, outflow and continuity of articles are local properties, depending on two consecutive months, $t = i$ and $t = i + 1$.Note that if article K_4 is actually a relaunch of article K_1 then it is absent in the article matching method. In the example below, this situation is treated.

It should also be noted that the arithmetic mean used in this situation leads to Carli indices, which is not very desirable, as they do not behave well under time reversal and exhibit other weaknesses. The geometric mean does not yield such an index; it is a Jevons index that does not have this drawback, and is preferable. This was already remarked in case of Table 4.1.2.

Remark In case an item is missing or introduced at the second period in a traditional setting with a small sample of articles in real shops visited by price takers, an option is to impute missing prices. This procedure is understandable as the items concerned may have simply been missed by a price taker at the second, or first period. However, in a big data setting, with (nearly) complete observation of the population, such as the one we are considering, it is more likely that a missing article is actually not available, temporarily or permanently. Imputing a value ignores this circumstance. So it is not so obvious to impute missing prices in this case. ■

Example with relaunches

In Table 4.2.2 we have a situation with three articles on t=1 and three articles on t=2, but only two of these exist in both periods. The new article at t=2 is a relaunch of the article that disappeared after t=1. So the new and old article are comparable but they are not the same, as they have a different key (EAN or web-id).[16]

4.2.2 Matched model with relaunches.

Key	Price at t=1	Price at t=2	Price ratios
K1	p_{11}	-	-
K2	p_{21}	p_{22}	$f_2 = \frac{p_{22}}{p_{21}}$
K3	p_{31}	p_{32}	$f_3 = \frac{p_{32}}{p_{31}}$
K4	-	p_{42}	-
Assumption			K_4 is known to be a relaunch of K_1
Arithmetic mean (Carli price index also for relaunch)			$F_7 = \frac{1}{3}\left(f_2 + f_3 + \frac{p_{42}}{p_{11}}\right) = \frac{1}{3}\left(\frac{p_{22}}{p_{21}} + \frac{p_{32}}{p_{31}} + \frac{p_{42}}{p_{11}}\right)$
Geometric mean (Jevons index also for relaunch)			$F_8 = \sqrt[3]{f_2 f_3 \frac{p_{42}}{p_{11}}} = \sqrt[3]{\frac{p_{22}p_{32}p_{42}}{p_{21}p_{31}p_{11}}}$

In order to apply the matched model with relaunches of Table 4.2.1 the link between articles K_4 and K_1 has to be established. We assume here that it was established in one way or another (noting that it may be a source of error). This is the tricky bit.

[16] A web-id for web shop A is strictly speaking not a key, as it is not unique over time. A key value is being reused if an article has faded out. It is temporarily a key, in a time period of a few consecutive months. When compiling the CPI, and only using data of the current month and comparing it with the previous month using a chain index, there should be no problem with a web-id when articles are matched on the basis of key equality.

Once the link is established we are in the same position as before (exact matching in a static environment) and the price index can be calculated accordingly.[17]

If we consider the formulas for geometric averages (for example F_2) then we see again that the 1:1 pairing of articles at t=1 and t=2 is not needed at all. Again it is about the comparison of price averages. In other words, we can apply this method also for a group method. Table 5.1.1 illustrates the situation. The prices used are those that are present in the group at that time (t=1 or t=2). No link between these articles at different periods in time is established. So any degree of overlap is possible between those two sets of articles.

The same remarks concerning the use of arithmetic and geometric averages, as made above, hold. The ones based on the geometrical mean are to be favoured. ■

5. Methods using classifications

5.1 Group method: using a 'broad' classification

The groups used in this method are defined by the internal classification of clothing. The groups correspond to the most detailed level of this classification, that was designed as an internally used refinement of the COICOPs for clothing. The core of this method is to divide the items into groups that are big enough, in terms of expected 'filling' every month; they are expected to be nonempty quite often. The division is done on the basis of characteristics associated with the items. These can be found in the descriptions of the items on the website, sometimes directly, sometimes after applying text mining. So the groups are standard. For a second (web) shop with the same kind of metadata, the same classification would be used and the same groups would be obtained.

In case of the group method the calculation of the price indices at the group level is easy (in case the prices for both months being compared exist). We then have that the price index with i as the basis month and j as the reporting month equals $I_{ij} = \frac{\bar{p}_j}{\bar{p}_i}$, with $\bar{p}_i, \bar{p}_j$ the average price for a given class in the internal classification of clothing, for months i and j.[18] The groups may be a bit "coarse" or heterogeneous in some cases. But some experiments have shown that this method yields very reasonable results and that it is easy to implement (which was the main reason to opt for this

[17] Here we assume we do not have to correct for 'quality differences' that might exist between a product and its relaunch. But in some cases this may be necessary.

[18] This is the 'happy flow'. For the 'exceptional flow', that is, in case $\bar{p}_i$ or $\bar{p}_j$ is missing, one should find an alternative solution. This usually involves broadening the group by looking at higher level classes of the internal classification of clothing.

method).[19] The idea is to replace the group method by one of the other methods described in this paper (hopefully). In particular the aim is to use subgroups that are more homogeneous.

Example illustrating the group method

We now consider an example that concerns sets of (similar) articles on the web site of a web shop at different months. It is assumed that keys (EANs) are either not present or are not used. As a result we do not know which articles are actually the same and which are different, at different points in time. At one point in time the different prices are assumed to belong to different items.

In case of Table 5.1.1 one is dealing with two groups of articles with a different composition, that one abstracts from. This approach may result in price change due to a change of composition of the group and not due to a price change of the same articles. To avoid this effect one should standardize the composition of a class in the internal classification of clothing, perhaps not in terms of the clothing items themselves, but in terms of their properties, and consider the price development of such a standardized group.

The notation in Table 5.1.1 has been adapted accordingly. There are no keys (EANs) anymore to identify articles. The group sizes may also be different, as in fact is supposed in this table. The sizes of the groups are 4 at t=1 and 3 at t=2 in this particular case, but in general any other combination of sizes is also acceptable (neglecting the problem of precision; a group may be required to have a minimum number of members). Because one does not know which articles one is dealing with, the method may not be 'pure', in the sense that changes in average price can be attributed to price changes only. It may be possible that changes are due to changes in group composition, so changes in the (supply) quantities.

5.1.1 Group method, general situation: different (number) of articles at different periods are possible.

Price at t=1	Price at t=2	Price ratios
p_1^1 p_2^1 p_3^1 p_4^1	p_1^2 p_2^2 p_3^2	
Arithmetic mean (variant of Dutot price index)		$F_7 = \frac{\frac{1}{3}(p_1^2 + p_2^2 + p_3^2)}{\frac{1}{4}(p_1^1 + p_2^1 + p_3^1 + p_4^1)}$
Geometric mean		$F_8 = \frac{\sqrt[3]{p_1^2 p_2^2 p_3^2}}{\sqrt[4]{p_1^1 p_2^1 p_3^1 p_4^1}}$

[19] That is, in case of the 'happy flow' when the average prices for a group in the two months being compared are both present. In case of the 'exceptional flow', when at least one of these prices is missing, it requires a bit more work.

This method (based on arithmetic means) is applied for web shop A internet data and seems to yield quite acceptable results, which is reassuring. In order to apply it one only has to link the articles available on the website to this class in the internal classification of clothing. This is done by automatic coding, but this time on secondary keys, that is, using secondary key matching. A computer program was especially written for this task. It is able to do its job quite well. Of course, in order to apply this method, enough descriptive information of the articles should be available on the website. It should be noted that this method is transitive and hence the resulting (chain) price index is drift-free.

If the month-to-month composition of the groups does not change too much, this is an attractive method to calculate price indices. The method is simple to use, it deals with the inflow and outflow of articles automatically, and yields plausible results. On a month-to-month basis the population of clothing articles available in a web shop typically does not change dramatically, except perhaps around sales periods.

We already remarked before that the classes in the internal classification of clothing can be fairly broad. It would be attractive to work with smaller subgroups to calculate the elementary price indices. These subgroups are not part of the internal classification of clothing, but should be defined for each outlet or web shop separately, provided the information (meta-information) about the articles allows such a subdivision. The subdivision depends on the descriptions available, which may differ from shop to shop. We have considered two web shops for the research in the present paper, denoted by web shop A and web shop B. For both web shops such a subdivision is possible. Here the meta-information for each article is available in the form of

- short descriptions (web shop A) ,
- long descriptions (web shop A and web shop B), or,
- characteristics (web shop B).

It should be noted that the details of the descriptions as well as their layout are different for both web shops.

That the meta-information about the articles is present is one thing, and an important one. But using it is another matter. For the short descriptions (for web shop A) this is not a problem. The long descriptions are not structured, so extracting information from this source is not trivial. It requires some effort to be able to mine these texts and extract the useful information from them. Apart from this “unstructuredness”, another problem with this information is that it is not systematically provided. In some cases the colour of an item is mentioned, whereas in another case it is not (for no apparent reason). And similarly for other characteristics. This implies that one should be able to handle missings in this sort of information. Or one should use information that is (almost) always available, like brand in the data of web shop B.[20]

[20] Web shop A is a shop as well as a brand, so there is no problem of stratifying the items into brands. Web shop A only sells its own brand.

It should be remarked that the effect of the substratification has yet to be investigated. It is not clear that it will lead to serious improvements, or even improvements at all. Once this has been investigated, one is in a position to judge whether the extra effort needed to achieve (some of the) subdivisions was well spent. For if a simpler method yields satisfactory results, it is to be preferred to a more complicated one, with, perhaps, only a marginal advantage.

Table 5.1.2 generalizes Table 5.1.1. The keys of the articles (EANs or web-id's) are not known, or even if they are known, they are not used. Like in Table 5.1.1 the price indices can be determined by the average prices at both periods. The only difference is that one does not know what the identity is of the articles. In extreme cases there is no overlap, that is, there are no articles present in both months.

5.1.2 Method based on group averages.

Price at t=1	Price at t=2	Price ratios
p_{11} p_{21} p_{31} p_{41}	p_{12} p_{22} p_{32} p_{42}	(Notation from Table 4.1.2 is used)
Arithmetic mean		$F_3 = \frac{\frac{1}{4}(p_{12}+p_{22}+p_{32}+p_{42})}{\frac{1}{4}(p_{11}+p_{21}+p_{31}+p_{41})} = \frac{p_{12}+p_{22}+p_{32}+p_{42}}{p_{11}+p_{21}+p_{31}+p_{41}}$
Geometric mean		$F_4 = \frac{\sqrt[4]{p_{12}p_{22}p_{32}p_{42}}}{\sqrt[4]{p_{11}p_{21}p_{31}p_{41}}}$

Price indices are the same as those in Table 4.1.1, keeping in mind that one takes the averages of the articles that are present (have been observed by the web scraper) at that time, i.e. t =1, 2. As was remarked before the articles in different periods can be totally different. If the periods are close, this is unlikely to happen, although the groups may be slightly different. If they are more distant this could be the case, however. That does not render the method useless, as we are dealing with articles that belong to the same class in the internal classification of clothing so they have a certain "kinship". As was discussed before, improvement might be possible (although this has yet to be shown) by using more homogeneous subgroups.

Comparison of Table 5.1.1 with Tables 4.2.1 and 4.2.2 suggests that the same articles belong to the groups at different periods in time. That is not necessarily the case, as was indicated above. But if the composition of the groups can be different, so also the number of articles in different months. The situation depicted in this table is not the one that is typical when the group method is applied. Typically, the composition and the group sizes are different. See Table 4.2.1.

5.2 Subgroup methods based on a refined classification

In this section we look at elementary indices at a subgroup level, as a refinement of the group level. At the group level, every category (group) is assumed to be suffi-

ciently filled with articles. At the subgroup level this may not be the case, at least for some months of the year. Which months this concerns depends on the groups that are refined by subgroups.

As an example of a classification to yield groups, we consider the internal classification of clothing. So a group is then a class in the internal classification of clothing. We consider periods of a year, and sometimes of two consecutive years (in case of seasonal articles). The building blocks for these periods are typically months. These choices are just convenient for the intended immediate application, but are otherwise somewhat arbitrary. Other choices are possible as well. The methods that we discuss do not really depend on a particular choice.

The idea is to use the finer level of detail to calculate elementary price indices, and use these to calculate aggregate indices at the group level. Aggregation over the subgroups is possible, but also over the months in the period considered.

Aggregating over the subgroups in a group

We assume that the basic periods considered are months. Within a class of the internal classification of clothing a month would be a state. Likewise for subgroups. It is possible that average prices for each group (or subgroup) are calculated, from which by comparison a price index can be calculated. It is also possible that price indices at the subgroup level are determined from which the aggregated price indices at the group level are calculated. We assume that these aggregated price indices calculated at the group level need not be transitive. But in Willenborg (2017) a general method is presented to calculate transitive indices from these 'rough' ones.

Comparing subgroup prices across time

This section considers the problem of how to compare prices for a subgroup of articles at different months. There are several options. One option is like the subgroup method applied to web shop A: compare the (average) prices of the articles in the subgroup for the two reference periods. This does not consider the contents of the subgroups at all. There may even be no overlap at all. Another variant is obtained by allowing comparison only if the subgroups have a minimum overlap in articles. This property is not transitive, in the sense that if SG_1, SG_2, SG_3 are three sets of articles corresponding to a specific subgroup at months 1, 2 and 3, respectively and $|SG_1 \cap SG_2| \geq \vartheta$ and $|SG_2 \cap SG_3| \geq \vartheta$ this does not imply that $|SG_1 \cap SG_3| \geq \vartheta$, where $\vartheta > 0$ is a threshold value. In fact it is possible that $SG_1 \cap SG_2 \cap SG_3 = \emptyset$, implying that there is no item present in all three sets.

In practice this implies that they are not too far apart in time. Basically this means that the same method is applied as in the first option, provided the overlap is large enough, and that no price ratio is calculated if the overlap is not large enough. A third option for a price ratio would be to consider the overlapping articles not only as a criterion to calculate a price ratio or not, but to actually base the price ratio on the overlap. If so, an additional requirement could be that the overlap is large enough. That in turn would imply that the months that can be directly compared should not be too distant.

The choices for calculating the elementary price indices at the subgroup level are presented in Table 5.2.1. Here we have a subgroup SG considered at two months s and t, i.e. $SG_t$and SG_s.

Furthermore, $\Omega_{st} = SG_s \cap SG_t$, which is the overlap between $SG_t$and SG_s, i.e. the articles these sets have in common. Also, $\Delta_{st} = (SG_s \backslash SG_t) \cup (SG_t \backslash SG_s)$ which is the symmetric difference of $SG_t$and SG_s , i.e. the articles not in the overlap Ω_{st}. In Table 5.2.1 the overlap is considered large enough, for a parameter $0 < \alpha < 1$, if $|\Omega_{st}| \geq \alpha|\Delta_{st}|$. Finally there are various prices that occur in Table 5.2.1. They have the following meaning. p^{SG_t} is an average price for the articles in SG_t; likewise p^{SG_s}; $p^{\Omega_{st,t}}$ is the average price of the articles in Ω_{st}in month t; likewise $p^{\Omega_{st,s}}$ is defined. The tacit assumption is that $p^{\Omega_{st,s}} > 0$ so that the reciprocal value also exists.

5.2.1 Various ways of comparing prices across time.

Method	Conditions of use	Comment
Subgroup price ratios, without restrictions	$\frac{p^{SG_t}}{p^{SG_s}}$ if $p^{SG_s} > 0$ irrespective of the size of the overlap	No restrictions on the overlap between the subgroups at different periods in time.
Subgroup price ratios, with restrictions	$\frac{p^{SG_t}}{p^{SG_s}}$ if $p^{SG_s} > 0$ and the overlap is large enough	As the previous method except that the ratio is only defined if the subgroups have enough overlap.
Price ratios based on overlap in subgroups	$\frac{p^{\Omega_{st,t}}}{p^{\Omega_{st,s}}}$ if $p^{\Omega_{st,s}} > 0$ and the overlap is large enough	Contrary to the previous method only using the prices of the articles in the overlap. This overlap should also be sufficiently large.

Note that the price ratios in Table 5.2.1 satisfy the time reversibility test. Also note that the methods in Table 5.2.1 differ in computational complexity. The first one does not require any knowledge of the composition of the subgroups at the different times compared. For the second one, one needs to know he relative size of the overlap between the subgroups for different months. For the third one, one needs to know the prices of the articles in the overlap, for both months.

Comparing prices across subgroups

In order to apply this method we assume that the subgroups within a group are ordered in some way or another, and that this ordering is kept fixed. Contrary to time, the subgroups do not have a natural ordering. But one can easily make one (any one will do). It is important to keep it fixed over the entire period of, say, a year.

The overlap between any two subgroups in a single month is always empty as the subgroups are disjoint in that case. So Methods 2 and 3 of the previous section cannot be applied. But a variant of the first method can. It simply compares average prices of two subgroups a and b. Let $p^{SG_t^a}$ denote an average price of the articles in subgroup a at month t. Likewise $p^{SG_t^b}$.

5.2.2 Comparing prices over the subgroups.

Method	Explanation and requirements	Comment
Subgroup price ratios, without restrictions	$\frac{p^{SG_t^a}}{p^{SG_t^b}}$ if $p^{SG_t^a} > 0,\ p^{SG_t^b} > 0$	Subgroups a and b, in month t should both be nonempty. There are no additional conditions concerning the sizes of these subgroups
Subgroup price ratios, without restrictions	$\frac{p^{SG_t^a}}{p^{SG_t^b}}$ if $p^{SG_t^a} > 0,\ p^{SG_t^b} > 0$ and the size of each of the subgroups should be large enough	Subgroups a and b, in month t should both be larger in size than some threshold value

Form of the elementary price indices

Looking at the form of the indices that the subgroup methods uses, we can observe that they are of a special form, viz

$$\frac{\Pi_i p_i^{\alpha_i}}{\Pi_j p_j^{\beta_j}}, \tag{5.1}$$

where $\alpha_i, \beta_j \geq 0$ en $\sum_i \alpha_i = 1$ en $\sum_j \beta_j = 1$.[21] Price indices of this type are generalisations of Cobb-Douglas indices (see e.g. Balk, 2008, p.97), in the sense that the number of factors in denominator and numerator may be different, whereas in Cobb-Douglas indices they are the same.

The multiplicative nature of these price indices makes them very attractive as price indices, from an algebraic point of view at least.

The effect of the subgroup pairing method, compared to the group method, is that prices within a group may get different weights. The exact form of the weights depends on the method used. Contrary to higher level weights, these elementary weights are not directly related to turnover.

5.3 Examples of subgroup methods

In this section we consider an important application for price indices: their development over time. We illustrate the main points by considering a set of examples.

Example with 6 subgroups and a time window of 4 months

We consider a situation of a group consisting of 6 subgroups, observed in a period of four consecutive months. In Table 5.3.1 six subgroups of a class of the internal classification of clothing are presented, denoted as SGi for $i = 1, \ldots, 6$. These subgroups are dependent on the web shop, whereas the class of the internal classification of

[21] We discard the fact that in practice the α's and β's are rational as well.

clothing to which they belong is shop independent. Indicated (by an "x") are the months at which average prices are known (because prices of products in these subgroups were available, i.e. collected by the web scraper). An '-' indicates that they are not present.

5.3.1 Subgroups of a class in the internal classification of clothing and their presence in various months.

Subgroup	t=1	t=2	t=3	t=4
SG1	×	×	×	-
SG2	×	×	-	-
SG3	×	-	-	x
SG4	-	×	×	×
SG5	-	×	×	-
SG6	-	-	×	×

On the basis of the information in Table 5.3.1, Table 5.3.2 is calculated. It contains the chain indices that can be calculated from Table 5.3.1 by using corresponding pairs of subgroups. The indices I_{ij} are at the group level and pertain to base month i compared with reporting month j. The indices I_{ij}^{k} are for subgroup k and pertain to base month i compared to reference month j. We have assumed that $I_{ii} = 1$ for $i = 1, \ldots, 4$, and that $I_{ji} = 1/I_{ij}$ for $j > i$. At the moment we are not interested in how the indices are calculated from the prices observed for certain subgroups at certain months.

5.3.2 Chain indices calculated for the situation of Table 5.1.2.

		Reporting months			
		1	**2**	**3**	**4**
Base months	**1**	$I_{11} = 1$	$I_{12} = \sqrt{I_{12}^{1} I_{12}^{2}}$	$I_{13} = I_{13}^{1}$	$I_{14} = I_{14}^{3}$
	2	$I_{21} = 1/I_{12}$	$I_{22} = 1$	$I_{23} = \sqrt[3]{I_{23}^{1} I_{23}^{4} I_{23}^{5}}$	$I_{24} = I_{24}^{4}$
	3	$I_{31} = 1/I_{13}$	$I_{32} = 1/I_{23}$	$I_{33} = 1$	$I_{34} = \sqrt{I_{34}^{4} I_{34}^{6}}$
	4	$I_{41} = 1/I_{14}$	$I_{42} = 1/I_{24}$	$I_{43} = 1/I_{34}$	$I_{44} = 1$

Transitivity does not hold. For example $I_{13} = I_{13}^{1}$ and $I_{12}I_{23}$=$\sqrt{I_{12}^{1} I_{12}^{2}}\sqrt[3]{I_{23}^{1} I_{23}^{4} I_{23}^{5}}$, so in general $I_{13} \neq I_{12}I_{23}$. This means that the resulting index when comparing two states (months) would depend on the path connecting the states, which is undesirable. So we cannot use these indices directly. However, we can apply the method in Willenborg (2010) to adjust these values, so that transitivity holds for the adjusted values. Willenborg (2017) applies this method to the situation discussed here.

Remark An alternative way to calculate chain indices for the entire group would be by first averaging the available prices for each month, and then taking the ratios of these averaged prices. However, this approach does not really use the existence of different subgroups within a group. It is in fact the group method, explained in Section 4. The resulting index, however, is transitive. ■

Remark Yet another way to calculate a consistent set of chain indices is to choose a unique path connecting any pair of states (defined by subgroup and month combinations). In that way the choice of a path to calculate the elementary price indices for a base state and another state is not an issue. For instance we could take as the path the subsequent months in the reference period. So we would have as a so-called spanning tree[22] the set V={1,2,3,4} of vertices and E= {(1,2), (2,3), (3,4)} the set of arcs, with corresponding elementary price indices $I_{12} = \sqrt{I_{12}^1 I_{12}^2}$, $I_{23} = \sqrt[3]{I_{23}^1 I_{23}^4 I_{23}^5}$ and $I_{34} = \sqrt{I_{34}^4 I_{34}^6}$. We also assume to hold: $I_{21} = 1/I_{12}$, $I_{32} = 1/I_{23}$, $I_{43} = 1/I_{34}$, $I_{11} = 1$, $I_{22} = 1$, $I_{33} = 1$ and $I_{44} = 1$. From this we can calculate by using transitive closure[23] $I_{13} = I_{12}I_{23} = \sqrt{I_{12}^1 I_{12}^2}\sqrt[3]{I_{23}^1 I_{23}^4 I_{23}^5}$, $I_{14} = I_{13}I_{34} = \sqrt{I_{12}^1 I_{12}^2}\sqrt[3]{I_{23}^1 I_{23}^4 I_{23}^5}\sqrt{I_{34}^4 I_{34}^6}$, $I_{24} = I_{23}I_{34} = \sqrt[3]{I_{23}^1 I_{23}^4 I_{23}^5}\sqrt{I_{34}^4 I_{34}^6}$. Also: $I_{31} = 1/I_{13}$, $I_{41} = 1/I_{14}$, $I_{42} = 1/I_{24}$.

It should be borne in mind that the resulting elementary indices are in fact path dependent. We could, for instance, also have chosen the spanning tree with arc set: {(1,2), (1,3), (1,4)}, with the corresponding assumptions about the indices associated with these arcs, etc. In this case we would have obtained, e.g., $I_{23} = I_{21}I_{13} = I_{13}/I_{12} = I_{13}^1/\sqrt{I_{12}^1 I_{12}^2}$, which is (typically) different from the value calculated above. This illustrates the path dependence of the values obtained. But each choice yields a consistent set of indices, respectively index values.

The essence of the method is that a spanning tree is defined on the set of states (months). Hill suggested a similar method for higher aggregates using special weights (the so-called Paasche-Laspeyres spread) to calculate an optimal spanning tree (cf. Balk, 2008, p. 256). ■

Remark In this report we also use spanning trees and optimal spanning trees, but in a different way than Hill suggested. Hill's method amounts to selecting an 'optimal' spanning tree from an inconsistent PIDG, and using this as a basis to deduce the values of other arcs by applying transitive closure. The method in Willenborg (2017) produces a consistent (transitive) PIDG from an inconsistent one. For the consistent PIDG one can then use any spanning tree to generate the other values, associated with the other arcs. But we will also have our optimal spanning tree, namely the linear digraph on a set of consecutive months. This is optimal in terms of overlap of articles of subgroups. ■

We now go one level deeper in detail and look at the prices observed per subgroup for each month. Table 5.3.3 gives an example of the observed prices that Table 5.3.1 hides.

[22] This is a tree with the same set of vertices as the original price index digraph (PIDG) , and with the edges taken from the original PIDG, in such a way that we obtain a (single) tree. This tree is connected and has (by definition) no loops, which the original PIDG may have. See Willenborg (2017, Appendix D) for more information on PIDGs and spanning trees, etc.

[23] This concept is also explained in Willenborg (2017, Appendix D).

5.3.3 Observed prices per subgroup.

Subgroup	t=1	t=2	t=3	t=4
SG1	$p_{11}^1, p_{12}^1, p_{13}^1, p_{14}^1$	p_{21}^1	p_{31}^1, p_{32}^1	-
SG2	p_{11}^2, p_{12}^2	$p_{21}^2, p_{22}^2, p_{23}^2$	-	-
SG3	p_{11}^3, p_{12}^3	-	-	p_{41}^3
SG4	-	p_{21}^4, p_{22}^4	$p_{31}^4, p_{32}^4, p_{33}^4$	p_{41}^4, p_{42}^4
SG5	-	$p_{21}^5, p_{22}^5, p_{23}^5$	p_{31}^5	-
SG6	-	-	p_{31}^6	$p_{41}^6, p_{42}^6, p_{43}^6$

To illustrate the calculation of the indices I_{ij}^k we consider two examples in Table 5.3.4, viz I_{12}^1 and I_{12}^2, which are the elementary indices for the months t=1 and t=2. Also the group chain index calculated from these indices is given.

By comparison, in Table 5.3.5 the index J_{12} for the group method has been calculated using geometric averages instead of arithmetic.[24] The subgroup structure is not used at all. Note also that more prices are used than for the subgroup pairing method used in Table 5.3.2. However the indices belong to the same class of indices (generalized Cobb-Douglas indices). Of course, this special form of the indices is due to the geometric averaging that is used.

[24] For application to web shop A arithmetic means have actually been used instead of geometric. They are a bit easier to apply and under normal circumstances gives comparable results.

5.3.4 Observed prices for the subgroups on t=1 and t=2.

Subgroup	t=1	t=2	Chain indices
SG1	$p_{11}^1, p_{12}^1, p_{13}^1, p_{14}^1$	p_{21}^1	$I_{12}^1 = \frac{(p_{21}^1)^1}{(p_{11}^1)^{1/4}(p_{12}^1)^{1/4}(p_{13}^1)^{1/4}(p_{14}^1)^{1/4}}$
SG2	p_{11}^2, p_{12}^2	$p_{21}^2, p_{22}^2, p_{23}^2$	$I_{12}^2 = \frac{(p_{21}^2)^{1/3}(p_{22}^2)^{1/3}(p_{23}^2)^{1/3}}{(p_{11}^2)^{1/2}(p_{12}^2)^{1/2}}$
Group chain index calculated from the subgroup chain indices			$\sqrt{I_{12}^1 I_{12}^2} = \frac{(p_{21}^1)^{1/2}(p_{21}^2)^{1/6}(p_{22}^2)^{1/6}(p_{23}^2)^{1/6}}{(p_{11}^1)^{1/8}(p_{12}^1)^{1/8}(p_{13}^1)^{1/8}(p_{14}^1)^{1/8}(p_{11}^2)^{1/4}(p_{12}^2)^{1/4}}$

5.3.5 Group method for the situation in Table 5.3.1. The subgroup structure, although indicated, is not used in the calculations.

Subgroup	t=1	t=2
SG1	$p_{11}^1, p_{12}^1, p_{13}^1, p_{14}^1$	p_{21}^1
SG2	p_{11}^2, p_{12}^2	$p_{21}^2, p_{22}^2, p_{23}^2$
SG3	p_{11}^3, p_{12}^3	-
SG4	-	p_{21}^4, p_{22}^4
SG5	-	$p_{21}^5, p_{22}^5, p_{23}^5$
SG6	-	-
Group chain index method for group	$J_{12} = \frac{(p_{21}^1)^{1/9}(p_{21}^2)^{1/9}(p_{22}^2)^{1/9}(p_{23}^2)^{1/9}(p_{21}^4)^{1/9}(p_{22}^4)^{1/9}(p_{21}^5)^{1/9}(p_{22}^5)^{1/9}(p_{23}^5)^{1/9}}{(p_{11}^1)^{1/8}(p_{12}^1)^{1/8}(p_{13}^1)^{1/8}(p_{14}^1)^{1/8}(p_{11}^2)^{1/8}(p_{12}^2)^{1/8}(p_{11}^3)^{1/8}(p_{12}^3)^{1/8}}$	

We now compute the subgroup index for the transitions $t = 2 \rightarrow t = 3$ and $t = 3 \rightarrow t = 4$. In Tables 5.3.6 and 5.3.7 the relevant subgroups for these transitions are represented.

5.3.6 Subgroups relevant for the transition from t=2 → t=3.

Subgroup	t=2	t=3
SG1	p_{21}^1	p_{31}^1, p_{32}^1
SG4	p_{21}^4, p_{22}^4	$p_{31}^4, p_{32}^4, p_{33}^4$
SG5	$p_{21}^5, p_{22}^5, p_{23}^5$	p_{31}^5

5.3.7 Subgroups relevant for the transition from t=3 → t=4.

Subgroup	t=3	t=4
SG4	$p_{31}^4, p_{32}^4, p_{33}^4$	p_{41}^4, p_{42}^4
SG6	p_{31}^6	$p_{41}^6, p_{42}^6, p_{43}^6$

Note that for transition $t = 2 \rightarrow t = 3$ we have subgroups that are different from those in case of the transition $t = 3 \rightarrow t = 4$. The indices have been collected in Table 5.3.8, specified in terms of the observed prices. This gives an idea of the different weights that are used in various chain indices.

5.3.8 Chain indices at subgroup and group level for transitions t=2 → t=3 and t=3 → t=4.

Chain indices at subgroup and group level
$I_{23}^1 = \frac{(p_{31}^1)^{1/2}(p_{32}^1)^{1/2}}{(p_{21}^1)^1}$
$I_{23}^4 = \frac{(p_{31}^4)^{1/3}(p_{32}^4)^{1/3}(p_{33}^4)^{1/3}}{(p_{21}^4)^{1/2}(p_{22}^4)^{1/2}}$
$I_{23}^5 = \frac{(p_{31}^5)^1}{(p_{21}^5)^{1/3}(p_{22}^5)^{1/3}(p_{23}^5)^{1/3}}$
$I_{34}^4 = \frac{(p_{41}^4)^{1/2}(p_{42}^4)^{1/2}}{(p_{31}^4)^{1/3}(p_{32}^4)^{1/3}(p_{33}^4)^{1/3}}$
$I_{34}^6 = \frac{(p_{41}^6)^{1/3}(p_{42}^6)^{1/3}(p_{43}^6)^{1/3}}{(p_{31}^6)^1}$
$\sqrt[3]{I_{23}^1 I_{23}^4 I_{23}^5} = \frac{(p_{31}^1)^{1/6}(p_{32}^1)^{1/6}(p_{31}^4)^{1/9}(p_{32}^4)^{1/9}(p_{33}^4)^{1/9}(p_{31}^5)^{1/3}}{(p_{21}^1)^{1/3}(p_{21}^4)^{1/6}(p_{22}^4)^{1/6}(p_{21}^5)^{1/9}(p_{22}^5)^{1/9}(p_{23}^5)^{1/9}}$
$\sqrt{I_{34}^4 I_{34}^6} = \frac{(p_{41}^4)^{1/4}(p_{42}^4)^{1/4}(p_{41}^6)^{1/6}(p_{42}^6)^{1/6}(p_{43}^6)^{1/6}}{(p_{31}^4)^{1/6}(p_{32}^4)^{1/6}(p_{33}^4)^{1/6}(p_{31}^6)^{1/2}}$

Example: aggregating over the subgroups in a group

The present example only adds some graphical information to a situation that may arise in practice. We are dealing with a group consisting of six subgroups of a class of the internal classification of clothing.

5.3.9 Spanning trees for six subgroups in a one year period and some price comparisons both at the subgroup and group level.

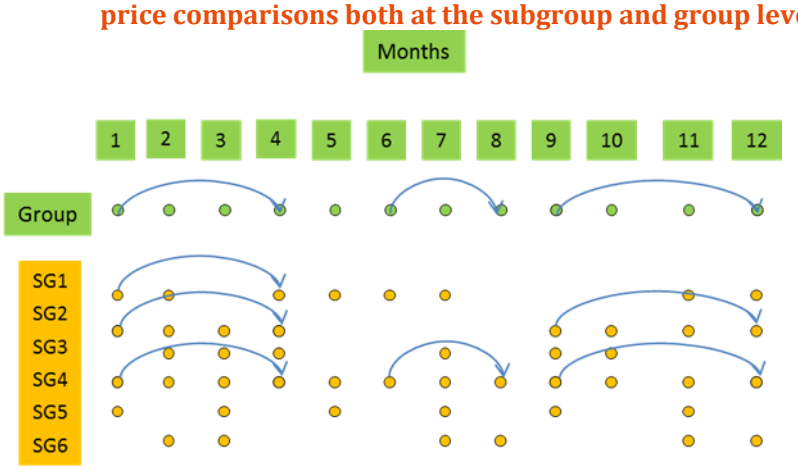

Figure 5.3.9 shows some of the arcs at the subgroup level and the corresponding ones at the group level. There are many more such arcs, but drawing them would be a major task and would result in a cluttered picture. Note that the subgroups are completely independent of each other. The class of the internal classification of clothing does not play a role whatsoever. It obviously plays a role when the indices associated with these arcs are aggregated at the group level. This produces a PIDG with 12 states (corresponding to the months in the reference year). This PIDG does not necessarily satisfy our cyclicity constraints, in particular transitivity. But we can apply the method of Willenborg (2017) to produce transitive indices from these. ■

Example with seasonal products

This example requires a period of two consecutive calendar years, the current and the previous one. This is unusual as one typically works within a single calendar year when compiling the CPI. [25]

Typical for seasonal goods is that they do not exist, or are not for sale, the entire year but only in part of the year, say in one or two seasons. Think of winter coats and summer dresses. The seasonal goods in a class of the internal classification of clothing are supposed to be distributed over one of more subgroups.

Comparison of prices of such goods within a calendar year is of limited value. We would like to compare prices of such goods also with those a year earlier, so we need to consider a period of two consecutive calendar years. We may expect that the seasonal clothing articles available in different calendar years are different. So the method used to compare prices of a particular seasonal subgroup between months in different calendar years should not be based on methods expecting an overlap in articles. However when comparing prices within a calendar year we may find an overlap, in particular in consecutive months. So we can apply a mix of comparison methods when considering price developments. So summarizing the difference of the treatment of seasonal goods compared to non-seasonal goods is:

[25] An alternative is to work with a rolling window of one year.

- The period in which the prices are compared (two calendar years instead of one)
- The methods of comparison of the prices of the subgroups, may be different for months of the same calendar year and months of different calendar years.

In practice it is often not possible to revise earlier published price index values, as a result of a publication policy that strives to publish 'definitive' figures as soon as possible. But one can use this information to calculate new price indices. For internal use one could study the effect of increased knowledge, and extra data, on the estimates. In order to comply with the publication restrictions, one can use an incremental method (see Willenborg, 2017, Section 3).

6. Summary and discussion

We started with a discussion about the various methods to calculate average monthly prices of subgroups of articles. These average prices are the raw material for elementary price indices. They, in turn, are used to calculate aggregated price indices. As the section on this issue shows, there are more alternatives to the method that has been chosen for the group method.

The group method is in fact a unit value method, although that terminology does only partly apply to the clothing data that have been the key data for which the group method has been developed (e.g. for differently sized packs of similar products, like socks). The group method is easy to apply and is able to cope with the highly dynamic population of clothing. It is based on classes of articles that are defined broadly enough so that most classes have prices each month. The group method employs the internal classification of clothing. But the group size was gauged to a particular store. Another benefit of the group method is that it yields a transitive price index. It is conceptually easy, and easy to apply, provided one is able to handle the automatic coding of garments on the basis of available metadata. This metadata is mostly alphanumeric data (text data) and handling it requires techniques from machine learning (like feature extraction).

A (possible) drawback of the group method is – by its very nature - that groups are used that are large enough to 'exist' during the entire year, and hence may be rather heterogeneous[26] in some cases. Also it may show price fluctuations due to volume effects (e.g. due to an increased offer of low priced garments). The subgroup method tries to ameliorate these defects, by looking at smaller, more homogeneous groups of items. But using subgroups, as they are smaller in size than groups, may imply that

[26] The heterogeneity meant here is that of articles, and indirectly that of prices, but not necessarily that of elementary price indexes. It may very well be that a heterogeneous group of articles and prices is actually quite homogeneous when looking at price ratios and elementary price indexes.

they often yield no average prices, as the corresponding subgroups are empty in certain months. This implies that the group method cannot be applied, and an alternative method is needed. This is what the subgroup method is intended for.

The subgroup method is proposed in particular in the present paper as an alternative to the group method that has been applied in earlier parts of the project to improve the data collection for the CPI. The group method is currently being used in the production of the CPI. It is based on subgroups which are refinements of the classes of the internal classification of clothing. These classes are too broad. The subgroups are more homogeneous.

The subgroup method is a refinement of the group method, which is based on a stratification of the population of clothing items. But it still uses groups of articles, that is, consisting of several EANs, to calculate prices (price averages). As in the group method, this allows similar articles to 'flow into' the respective subgroups, and others to disappear from these subgroups. In that sense both the group and subgroup method are preferable to a matched pair method. The matching of articles is on the basis of 'similarity' (common characteristics) rather than on 'identity' (equality of EANs). In terms of matching, the subgroup and group method use object characteristic matching - more commonly known as statistical, probabilistic or synthetic matching) and the matched model uses identity object matching (more commonly known as exact matching).

When aggregating the elementary price indices at the subgroup level to rough elementary indices (meaning that they may not be transitive) at the group level various methods can be used. But when geometric means are used, this is most elegant from a mathematical point of view. In this case price indices that are generalizations of Cobb-Douglas indices are obtained. The exponents that are used to weigh the various prices, reflect the sizes of the subgroups being used. So there is a form of weighing prices, but this is not very directly relatable to turnover. For internet data in general, explicit information on turnover is lacking.

When comparing 'matched model' methods applied at the item level with subgroup methods, it is clear why they fail in case of short-lived articles such as clothing items: they do not permit comparison with other items, within a year or in the previous year. The level of description, say at the EAN or web-id level, is too detailed and items which satisfy such strict descriptions only exist possibly for one season. And even within one season, they are treated as being unique, and are therefore 'beyond comparison'. The comparison is made possible in the traditional approach (where physical shops are visited by price takers) because the price takers link articles at the EAN level, and the matching used is statistical, not exact. They are able to do this 'matching by hand' because they have to handle only a relatively small number of articles. Such matching is impossible to used when big volumes of data are to be handled as in case of internet data (as well as scanner data, for that matter). Firstly, the information is lacking as to which items to link. Secondly, the matching task is too time-consuming and error-prone in view of the volume of data to be handled, day after day. Automatic coding is called for, which is, in case of the group or subgroup

method, to be carried out on the basis of characteristics of the items (meta-information).

In order to be able to produce an acceptable price index the description of the articles has to be coarser, so that this comparison sets of items (groups or subgroups) is possible. Classes of the internal classification of clothing may not be homogeneous enough. The challenge is to find subgroups of the right size, not too small so as to make comparison impossible (because they are too often without any items) and not too big (to avoid too much heterogeneity).

The situation is somewhat reminiscent of density estimation at a particular point *x* in the domain of its probability density. If an estimate is based on a small interval around *x*, then the estimate would be highly accurate for the point. But because the interval is small, there will hardly be any observations inside this interval, if any at all. So an estimate based on such an interval would be highly unreliable in the best case and non-existent in the worst case. On the other hand, if a broad interval would have been chosen around *x*, there would be plenty of observations inside this interval, but the density estimated at *x* can hardly be called representative for the density around *x*. So an interval somewhere between these extremes is called for.

The group and subgroup methods are both stratification methods, using different levels of detail. They are both relatively simple to apply, at least in case of the 'happy flow', when all preconditions are met for the methods to be applied. Automatic coding is required to apply the methods, but this method is not exclusive for these methods. It is an attractive way to cope with new articles ('relaunches'). More complicated and laborious alternatives are also possible. For instance one could link items in different months on the basis of common web-ids which are also in the data. This would take care of the products that are on offer in both months ('continuing products'). But then one would miss the items that are present in only one of a pair of months, that is, either disappearing or new articles. These articles cannot be ignored, as they can have profound influence on the course of the price indices calculated. Significant price rises or price drops may be missed. There are no lists linking new products with the ones they are supposed to replace.[27] For those items one could apply statistical matching and link items on the basis of common secondary characteristics (metadata). This method is more time-consuming than the group or subgroup method, and therefore was not chosen as a first method to apply to the data of web shop A. But it is certainly an interesting method to investigate, that, when applied for subgroups, can be seen as an elaboration of the subgroup method.

[27] They may exist in the web shop but we (CBS) did (and still do) not have them.

References

Balk, B. M. (2008). Price and quantity index numbers, Cambridge University Press.

Willenborg, L. (2010). Chain indexes and path independence. Report, CBS, The Hague.

Willenborg, L. (2017). Transitivizing elementary price indices for internet data using the cycle method. Discussion Paper, CBS, The Hague.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information