Discussion Paper

Decomposition price indices

Léon Willenborg

January 27, 2020

In this paper we explore several indices that are based on the idea of decomposing prices or volumes and quantities, into a time component and a goods component. Our interest is in the time component which is viewed as a price index; the goods component is a nuisance parameter. The indices we consider can be viewed as derivatives of an idealized price index of the Geary-Khamis type. Among the derivatives there is the time product dummy index. There are several others which also may be of interest. In particular we look at an incremental price index, which could be used for a flash HICP, using both past and recent data. We also consider conditions for the existence of a GK index.

1 Introduction

The aim of the present paper[1)] is to study an class of price indices which have in common that they are based on a decomposition of prices in a time component and a product component. The time component can be interpreted as a price index.

The motivation for writing this paper is twofold. In the first place it want to present an interesting and important class of price indices. The GK[2)] index, an important member of this class, is being used for the CPI[3)] and HICP[4)] at CBS, for an increasing number of goods.[5)] The second reason for writing this paper is to provide a motivation for the choice of a time series model in case of a flash HICP, which is a short term predictor of the HICP (see [2]). To model a price index series one can simply postulate a time series model that is fashionable and seems appropriate. But this approach is somewhat ad hoc. The incremental approach to the GK index, yielding a motivated choice for a time series model based on a price index. Of course, for a price index different from the GK index, the incremental approach would result in a different time series model.

Apart from this, the incremental approach has the advantage that the past is summarized in a few parameters that can be used to estimate the price index for the next period, which can be attractive if past data are never revised,[6)] e.g. as a policy principle.

The price index we start with is an idealized version of the GK-index. It is an idealized version because it requires future knowledge, and knowledge (deep) into the past, which one can never have. So this idealized index can never be used. It is useless as such, but it serves as a starting point for deriving other, realistic price indices. These indices form a family of GK-type indices. As it turns out, TPD[7)] indices can be viewed as members of this family. A variant of this index is defined using a structural time series model. To this family of indices also belong incremental indices, that are updated every month as soon as new estimates for the last month become

1) The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author is grateful to Frank Pijpers and Sander Scholtus for reviewing it.
2) Named after R. Geary (cf. [4]) and S. Khamis (cf.[5]).
3) Consumer Price Index.
4) Harmonized Index of Consumer Prices.
5) A. Chessa introduced the QU-method (see e.g. [3]), where QU stands for 'Quality adjusted Unit value' . This refers to a class of price indices that includes the GK-index. The QU-method is applied to a temporal setting, whereas the GK-index was developed in the context of international comparisons. These differences, however, do not affect the formal properties of the index, and only involve interpretation.
6) Except to correct gross mistakes.
7) Time Product Dummy.

available. Some of these indices are of use for an investigation for an improved flash CPI (see [2]), using a time series approach.

The remainder of this paper is organized as follows.

In Section 2 we describe the setting to which the indices that we consider in this paper are supposed to apply to: a dynamic population of goods.[8] Typical for such a population is that goods appear, are available for some time in the market, and disappear. Possibly they are replaced by other, similar products. But it is typically not known if this is the case, and which item replaces which disappearing item. Therefore we consider the items as independent and the only way to link them is via GTIN[9] or by using common characteristics. For the goods, the assumption is that the prices are available, as well as some metadata and GTIN number, which can be used for linking items. In case of scannerdata we have information on turnover and quantities sold. This information can be used to derive prices of (aggregates of) goods for a time interval (which also might be an aggregate), and also to weigh prices. For webdata we only have metadata and prices, but no information on turnover.[10]

In Section 3 we consider an idealization of the GK index. It is useless in practice but very well suited to derive various price indices in the GK family of price indices. The situation we are dealing with is comparable to that in survival analysis, which is about handling censored data about life spans of persons (unobserved births, deaths, etc.). But in our case we deal with items that are introduced at one point in time to a market and after some time disappear from it, for instance because they are not produced (goods) or delivered (services) anymore.

Price observations of the goods in the population we consider start at some point in time and our observations start at a later time. This presents us initially with a left-censoring problem, as we do not know when some goods (the censored ones) entered the market. Supposing that we follow the population of goods in which we are interested for long enough, we see that items disappear form the market (they 'die') or new ones are introduced to it (they are 'born'). Sometimes the ones that have died re-emerge, differently packaged, possibly in a bigger or smaller bottle, box, etc. but essentially the same product, but at a higher (unit) price. This holds for most (if not all) populations of goods: they are dynamic and items come and go constantly. Each good has a finite lifespan. At the end of our observation period, we see the goods that are on the market but we do not always know when they disappear from it ('they die').[11] This is a right-censoring problem. These censoring problems concern genuine problems because they cannot be solved as they concern 'unknowable' information. This implies that, in practice, the exact price indices can never be known; they can only be approximated.

8) In the sequel we shall talk about goods, in which we are mainly interested, and do not mention services. But this does not mean that they are necessarily excluded from the theory. They are just not explicitly mentioned. But 'goods' could also include certain services as well.

9) Global Trade Item Number, which is an identifier for trade items. Former name: EAN = European Article Number.

10) This is comparable with the information obtained in traditional price taking by price takers that visit brick-and-mortar shops. Apart from being less labour intensive than price taking, web scraping can yield (much) more data, more frequently. This is a mixed blessing. On the one hand one has a lot of price information. On the other hand, this information is not always useful and it is better to discard some of it. But how to separate the wheat from the chaff? The present paper, however, is not the appropriate place to elaborate this issue.

11) Sometimes goods disappear for a short period from the market and reappear. So when a good disappears from the market one cannot always know whether it is absent temporarily or forever. Only after some time it is clear whichever is the case. Theoretically speaking one can never be sure that an item that is absent for some time will be reintroduced. In practice, however, at some point it is highly unlikely that it will appear on the market again.

In Section 4 we consider existence issues of the GK-index: under which conditions does such an index exist? We show that the question can be transformed into an existence problem in a linear algebra setting, which also can be used to provide an answer.

In Section 5 an incremental version of the GK index is considered. This index follows quite straightforwardly from the GK index. It is to be seen as a step in the direction of a price index based on time series, which does not seem to be typical for this area. For instance in [1] not a single index is considered that is derived from a time series model. It is not the case that such series are not suited for such models.

In Section 6 GK-type indices are derived using optimization models. They follow from replacing the equalities that are used to define the GK-index by a target function that is minimized. In this case the requirements are rather approximate equality than exact equality. In fact several object functions are defined, an unweighted and a weighted one. Inspired by these optimization models, other such models can be defined. They can be chosen in such a way that we move in the way of a TPD index.

In Section 7 the TPD-type indices are considered. They also decompose prices into a time and a product component, but in a different way than the GK index. Unweighted as well as weighted variants of TPD are considered. The unweighted variant is the standard variant. Also a variant is discussed where the time component is split into a year and a month component, which yields a seasonal model. In fact, this type of model can be compared to a structural time series model with a trend (year component), a season (monthly component) and noise (what remains).

In Section 8 the main findings of this paper are briefly presented, with some hints about topics that could be elaborated in future work.

The paper is completed with a list of references.

2 Setting: Dynamic population of goods

In the present paper we assume that we are dealing with a dynamic population of goods. The goods are assumed to have been divided into homogeneous classes. The lowest level is the GTIN level. From this level onwards classes of goods can be formed by aggregation. How these classes are actually formed is not of importance here. The classification used is assumed to be fixed. As the population of GTINs is dynamic, the composition of the classes can be dynamic. This should not be of concern, unless there are some fundamental changes to a class due to introduction of essentially new items. Although this may happen in practice, we consider it to be an exception. In the paper we concentrate on less dramatic changes, that do not require special measures. We denote individual GTINs as well as classes of GTINs collectively as goods.

We assume that prices of goods are observed at the GTIN level, either by using scanner data or by using price information collected from the internet by special bots, or web scrapers. These modes of observation typically involve large amounts of data: many items and daily observations

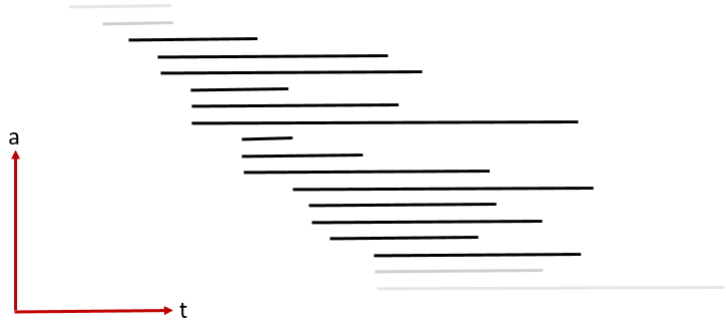

Figure 2.1 Lifespans M_i as line segments of an entire population of goods i over all time. t denotes the time dimension and a the goods dimension.

over the time window considered. This allows that price indices can be computed at finer levels of detail, both for the goods and the time units.[12] We assume that the observed prices are used to compute averages per class and per month.[13] In case scanner are used information about turnover of goods is available. This information is lacking in case of web scraped data or data obtained via price takers.[15] In this paper we sometimes assume that turnover is available to weight the prices and sometimes we do not.

We use the following notation in the present paper. For good i the set M_i denotes the period (in months) in which good i is on the market, that is, the lifespan of good i.[16] A_j denotes the goods that are on the market in month j. For good i and month j v_{ij} denotes turnover, q_{ij} quantity and p_{ij} price, with $v_{ij} = p_{ij}q_{ij}$.[17] Such information is available in scanner data, which contain (aggregates of) items that can be said to belong to class i sold in a given month j, along with descriptive (or meta-)information about these goods. This information is used to classify the goods.

The setting sketched so far assumes that time extends arbitrarily far back into the past as well as into the future. See Figure 2.1. That is not a realistic assumption. In practice we have a time window W with a limited number of months. A_W denotes the set of goods that are on the market in period W.[18] See Figures 2.2 and 2.3. Figure 2.2 shows the information that can be known (inside W) and that can not be known (inside the grey areas), and that is in fact missed when working with items alive in W. Figure 2.3 only shows the information that is known in W. The items that 'live' entirely outside W, either in its past or its future, are left out. For the items that can be observed in W only the portion inside W is shown. This illustrates that if one wants

12) This is in sharp contrast to the traditional mode of observation by price takers visiting brick-and-mortar shops: the observations are once every month and involving a relatively small number of goods. Besides, the price taking process is not error-free.

13) In the future the CPI may be based on finer time units than months. The data[14] would allow it. But there must be a demand for such data. For the moment we stick with the month as time units for the CPI.

15) In these cases one either proceeds by computing unweighted price indices, such as the Jevons price index, or one uses proxy weights from another source.

16) If a good is temporarily out of the market we discard this. This yields a consecutive sequence of months when the good was not unavailable. In the complementary months in W the good was not on the market, which is comparable to being dead. Temporarily not on the market is like being dormant.

17) All these quantities are averages, say of all items belonging to that type of good i sold in month j.

18) Starting with a window W and looking at the set of item that are alive in W, i.e. for which $W \cap M_i \neq \emptyset$, is quite natural. But one could just as well start with a set of objects Ω and consider the smallest period in which they all 'live', i.e. $\cup_{i\in\Omega} M_i$. But this is not an approach that we take in the present paper. Here, the time window W is leading.

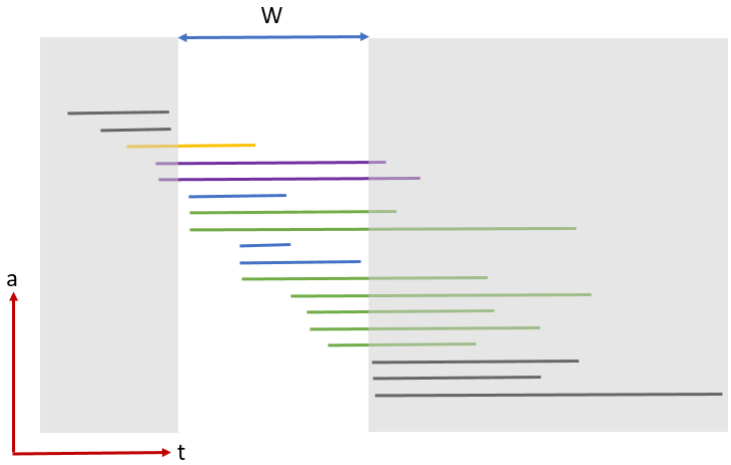

Figure 2.2 Lifespans of a population of goods and time window W. The grey blocks symbolize the past (before W) and future (after W), which are unknown. The data that can be known (in W) and that are unknown (in the grey area) are shown. Indicated are: data that cannot be observed in W (black), uncensored = completely observed in W (blue). Some data are censored with respect to W: left-censored (yellow), right-censored (green) or doubly censored (purple).

to compute the average lifespan of items in the population studied, one faces the problem of censored information. As to the goods present in W, they can be divided into three subgroups:

1. left-censored,
2. right-censored,
3. left- and right-censored.

We put

$$|W| = w \tag{1a}$$

$$a_W = |A_W|, \tag{1b}$$

where $|\cdot|$ denotes the count function, which yields the number of elements of a (finite) set.

In the sequel parameters α_i and π_j are used in different models. Hence the meaning of these parameters depends of the model in which they appear.

To illustrate the presence or absence of goods on the market consider Figure 2.4 concerning items in a web shop on a daily basis. It is clear that most items shown in Figure 2.4 are not available every day in the time period shown. This picture also illustrates that it is not so easy to determine, on the basis of the scraped information only, whether an item is temporarily not available or permanently, that is, is taken from the market. This can only be judged after some time has elapsed.[19)]

[19)] The same problem exists in case one has to decide whether a company has temporarily low activity or has stopped its activities altogether. Or in case of a person, that can be temporarily unemployed (e.g. being between jobs) or for a longer period, in which case unemployment benefit may be required as a substitute for an income from a job.

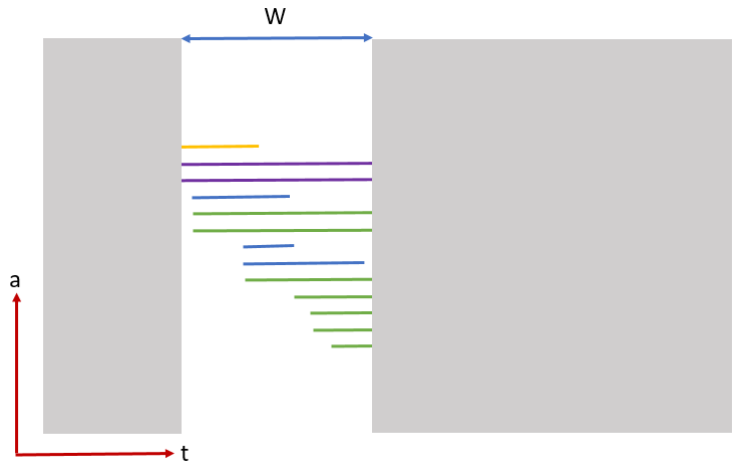

Figure 2.3 **For time window W the items in A_W are shown as well as the portions of the information known about them. For good $i \in A_W$ the observed part of the lifespan is $W \cap M_i$.**

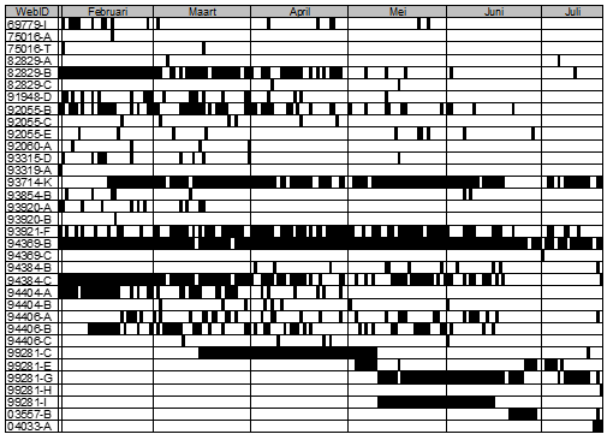

Figure 2.4 **Items of a web shop present (black rectangle) or absent (white rectangle) in a period of time (half a year). The data was collected by a web scraper that daily scraped the contents of the shop's website.**

3 GK-type price indices

3.1 GK price index

We start our investigation with the following 'idealized' GK price index:

$$\alpha_i = \frac{\sum_{j \in M_i} v_{ij}/\pi_j}{\sum_{j \in M_i} q_{ij}} \tag{2a}$$

$$\pi_j = \frac{\sum_{i \in A_j} v_{ij}}{\sum_{i \in A_j} \alpha_i q_{ij}} \tag{2b}$$

$$\pi_1 = 1 \tag{2c}$$

The first remark is that this is a price index that can never be evaluated, as it assumes knowledge of all products that currently exist, have ever existed or will exist in the future. It is impossible to possess such knowledge. It is only possible to have information about products on the market in a limited (finite) time window W.

The first of the equations in (2) defines an average price for good i over its lifespan M_i. One cannot simply add the monthly turnovers. Instead, one should discount each month with the appropriate monetary value (π_j for month j), with reference to a fixed month, in our case month 1 (with $\pi_1 = 1$).[20] The second equation defines a price index for month j as the ratio of the turnovers in terms of real money and the time-averaged prices of each of the goods.

Equations (2) describe an unrealistic situation assuming complete knowledge about past and future price information about goods. However, from (2) we can produce the following approximate price index that actually can be computed:

$$\alpha_i^{W,A_W} = \frac{\sum_{j\in W} v_{ij}/\pi_j^{W,A_W}}{\sum_{j\in W} q_{ij}} \tag{3a}$$

$$\pi_j^{W,A_W} = \frac{\sum_{i\in A_W} v_{ij}}{\sum_{i\in A_W} \alpha_i^{W,A_W} q_{ij}} \tag{3b}$$

$$\pi_1^{W,A_W} = 1, \tag{3c}$$

where we have stressed the dependence of the parameters α and π on the sets W and A_W. Normally we will not stress these dependencies to avoid overburdening the notation. We then simply write

$$\alpha_i = \frac{\sum_{j\in W} v_{ij}/\pi_j}{\sum_{j\in W} q_{ij}} \tag{4a}$$

$$\pi_j = \frac{\sum_{i\in A_W} v_{ij}}{\sum_{i\in A_W} \alpha_i q_{ij}} \tag{4b}$$

$$\pi_1 = 1. \tag{4c}$$

The dependence of the parameters on the choice of W and A_W should not be forgotten, however, even if not explicitly indicated. The fact that the dependence of A_W on W is not suppressed, is to stress that the set of goods selected depends on the choice of W.

The system of equations (4) yields a price index that can actually be computed, in the sense that the information that is needed is realistic. In fact, it remains to be shown that (4) has a solution under certain conditions, and which they are.

Note that (4) is invariant under scaling $q_{ij} \mapsto \xi q_{ij}$ and $v_{ij} \mapsto \xi v_{ij}$ for $\xi > 0$. This scale invariance property of the qs and vs is shared by all variants considered in the present paper.

[20] The π_js are comparable to exchange rates, not between different countries, but between different months in a single country.

Typically, equations like (4) are solved iteratively. In practice, convergence is typically rapid, provided a solution exists. [21)]

The indices computed in (4) are in fact with respect to reference month (= base month) $j = 1$. By using transitive closure we can extend the definition to any pair (j, k) with $j, k \in W$, as follows. π_j in (4) is in fact π_{1j}, that is the index with base month 1 and current month j. Let k be another month in W, so that index $\pi_k = \pi_{1k}$. Assuming transitivity, we define the index π_{jk} using:

$$\pi_j \pi_{jk} = \pi_{1j} \pi_{jk} = \pi_{1k} = \pi_k. \tag{5}$$

Hence, using (4) and (5), we find:

$$\pi_{jk} = \frac{\pi_k}{\pi_j} = \frac{\sum_{i \in A_W} v_{ik}}{\sum_{i \in A_W} v_{ij}} \Big/ \frac{\sum_{i \in A_W} \alpha_i q_{ik}}{\sum_{i \in A_W} \alpha_i q_{ij}}. \tag{6}$$

(6) is the ratio of a value index and a (modified) quantity index, in the sense that the quantities are weighted with the αs as specified in (4).

3.2 An approximate GK-type price index

Denominator and numerator in both equations in (4) can be viewed as ratios of arithmetic averages. Instead of such averages we can take geometric averages in numerator and denominator. But now we should be careful. We should only consider those goods-months combinations (i, j) for which $q_{ij} > 0$ for $i \in A_W$ and $j \in W$. We define

$$M^i = \{j \in W | q_{ij} > 0\} \subseteq M_i, \text{ for } i \in A_W, \tag{7}$$

and

$$A_W^j = \{i \in A_W |\ q_{ij} > 0\} \subseteq A_W, \text{ for } j \in W. \tag{8}$$

We put

[21)] A solution need not exist. For instance in (the extreme) case that there is no turnover of any of the goods considered, in some month. In more common situations a solution is likely to exist, however.

$$m^i = |M^i|, \tag{9a}$$

$$\tilde{m} = \sum_{i \in A_W^j} m^i, \tag{9b}$$

$$a_W^j = |A_W^j|, \tag{9c}$$

$$a_W = \sum_{j \in M^i} a_W^j. \tag{9d}$$

We then arrive at the following set of equations, using the notation introduced in Section 2 and in (9):

$$\alpha_i = \prod_{j \in M^i} \left(\frac{p_{ij}}{\pi_j}\right)^{m^i/\tilde{m}}, \tag{10a}$$

$$\pi_j = \prod_{i \in A_W^j} \left(\frac{p_{ij}}{\alpha_i}\right)^{a_W^j/a_W}. \tag{10b}$$

We can linearize (10a) and (10b) by taking (natural) logarithms. We then obtain:

$$\log \alpha_i = \frac{m^i}{\tilde{m}} \sum_{j \in M^i} \left(\log p_{ij} - \log \pi_j\right), \tag{11a}$$

$$\log \pi_j = \frac{a_W^j}{a_W} \sum_{i \in A_W^j} \left(\log p_{ij} - \log \alpha_i\right). \tag{11b}$$

From (11a) and (11b) the following equations for the πs and the αs can be obtained:

$$\log \alpha_i = \frac{m^i}{\tilde{m}} \sum_{j \in M^i} \log p_{ij} - \frac{m^i}{\tilde{m}} \sum_{j \in M^i} \frac{a_W^j}{a_W} \left(\sum_{i \in A_W^j} \log p_{ij} - \sum_{i \in A_W^j} \log \alpha_i \right), \tag{12a}$$

$$\log \pi_j = \frac{a_W^j}{a_W} \sum_{i \in A_W^j} \log p_{ij} - \frac{a_W^j}{a_W} \sum_{i \in A_W^j} \frac{m^i}{\tilde{m}} \left(\sum_{j \in M^i} \log p_{ij} - \sum_{j \in M^i} \log \pi_j \right). \tag{12b}$$

Solutions to equations (12a) and (12b) can be obtained iteratively (provided they exist). We shall not go into this matter here, as it leads us too far away from the main theme of the paper. Anybody interested in this subject can take it from here.

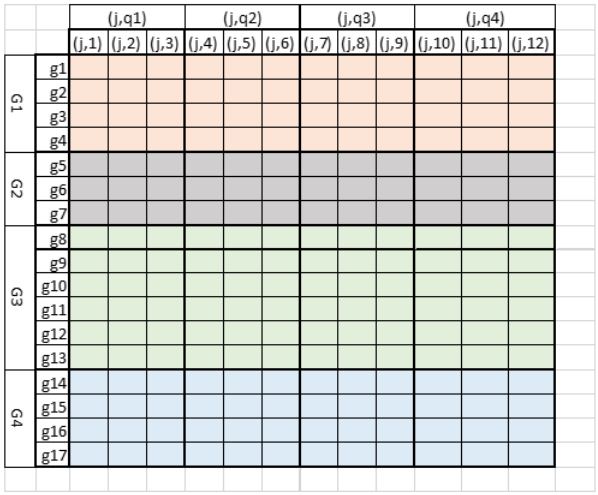

Figure 3.1 Aggregating goods over time (from months to quarters) and into broader product categories.

3.3 Aggregating the GK index

The GK price index is suitable for aggregating both in the time and in the goods dimension. This is the case because the index is defined in terms of the (observed) variables 'turnover' and 'quantity', which are both additive variables[22)23)] The variable 'price' (being the ratio of 'turnover' and 'quantity') does not have this property. Most price indices are defined in terms of 'price' and 'quantity', and so do not behave so nicely under aggregation over time as well as over goods as the GK price index does.[24)] In Figure 3.1 a situation is depicted where items are aggregated over time and over product groups. Shown are months grouped into quarters and goods lumped into broader categories. Taken together, they results in coarser period–goods combinations. The formulas 3 can be applied, suitably adapted to the new situation. Turnover and quantities can be obtained by aggregation over the newly formed periods–goods strata.

4 Existence of a GK price index

The existence of a GK price index is by no means certain. In the present section we want to find conditions that guarantee the existence of a GK price index. We present a method based on linear algebra that leads to formal criteria for the existence of a GK price index.

We start our analysis with the observation that (4) can be written as a set of linear equations:

22) Additive measures, actually.

23) In [1], Section 7.4.1 the GK price index is presented as part of a class of strongly additive methods for computing price indices.

24) Examples of such price indices are those named after Laspeyres, Paasche, Fisher and Törnqvist, to mention but a few well-known ones. In these indices prices at different periods are weighted with observed and chosen quantities. Observed prices times observed quantities yield (what could be called) observed turnovers; if these had actually been observed, they could have been used. What obstructs additivity of the price index formula is that observed prices times chosen quantities (from the reference period) are used to yield synthetic turnovers, that cannot be observed.

$$\alpha_i q_{i\cdot} - \sum_{j \in W} \kappa_j v_{ij} = 0, \tag{13a}$$

$$\sum_{i \in A_W} \alpha_i q_{ij} - \kappa_j v_{\cdot j} = 0. \tag{13b}$$

where $\kappa_j \triangleq 1/\pi_j$ for $j = 1, \ldots, |W| = w$ and $\kappa_1 \triangleq 1$, $q_{i\cdot} \triangleq \sum_{j \in W} q_{ij}$ and $v_{\cdot j} \triangleq \sum_{i \in A_W} v_{ij}$ for $i = 1, \ldots, |A_W| = a_W$.

In matrix form (13) reads:

$$\mathcal{A}\zeta = 0, \tag{14}$$

where

$$\mathcal{A} = \left(\begin{array}{cccc|cccc} q_{1\cdot} & 0 & \cdots & 0 & -v_{1,1} & -v_{1,2} & \cdots & -v_{1,w} \\ 0 & q_{2\cdot} & \cdots & 0 & -v_{2,1} & -v_{2,2} & \cdots & -v_{2,w} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & q_{a_W\cdot} & -v_{a_W,1} & -v_{a_W,2} & \cdots & -v_{a_W,w} \\ \hline q_{1,1} & q_{2,1} & \cdots & q_{a_W,1} & -v_{\cdot 1} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ q_{1,w} & q_{2,w} & \cdots & q_{a_W,w} & 0 & 0 & \cdots & -v_{\cdot a_W,w} \end{array}\right) \tag{15}$$

and

$$\zeta = \left(\begin{array}{c} \alpha_1 \\ \vdots \\ \alpha_{a_W} \\ \hline \kappa_1 \\ \vdots \\ \kappa_w \end{array}\right). \tag{16}$$

Note that $\mathcal{A}$ in (15) is a square block matrix, with in the upper right corner $-V$, where V is the matrix of turnover, and in the bottom lower corner Q, the matrix of quantities sold. If we delete the column in $\mathcal{A}$ corresponding to $\kappa_1 = 1/\pi_1 = 1$ as well as the final row in $\mathcal{A}$[25)], we can write (13) in matrix form as

$$\mathcal{B}\zeta^0 = \eta \tag{17}$$

25) In (15) the entries in this column and row have been colored red.

where

$$\zeta^0 \triangleq \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{a_W} \\ \hline \kappa_2 \\ \vdots \\ \kappa_w \end{pmatrix} \tag{18}$$

is a vector of parameters to be estimated. $\mathcal{B}$ is a square matrix of order $a_W + w - 1$. The vector η is a known $(a_W + w - 1)$- column vector, which is minus the column in $\mathcal{A}$ that corresponds to parameter κ_1 (the $(w+1)$-st column). If we assume that the turnover in each month is nonzero (and in particular, in month 1) η is a nonzero vector. We have

$$\eta \triangleq \begin{pmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{a_W,1} \\ \hline v_{\cdot,1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{19}$$

We may assume that the (rather weak) condition

$$v_{\cdot,1} > 0 \tag{20}$$

holds. This is the case iff[26)]

$$v_{i,1} > 0, \text{ for at least one } i. \tag{21}$$

Note that $v_{\cdot,1} > 0$ holds iff $q_{\cdot,1} > 0$. Condition (20) (or condition 21) implies that $\eta \neq 0$. They also imply that $\mathcal{B}$ is invertible[27)] and $\eta \neq 0$—as we shall assume—we can solve (17) for ζ^0:

$$\zeta^0 = \mathcal{B}^{-1}\eta. \tag{22}$$

Condition (20) (or condition (21)) also determines when a GK price index (4) exists.

[26)] If and only if.

[27)] The columns in $\mathcal{A}$ and hence in $\mathcal{B}$ are independent. However, the rows in $\mathcal{A}$ are dependent: adding those in the top blocks yield the same result as adding those in the bottom blocks. However removing one row implies that the remaining rows in $\mathcal{A}$ are independent, which also holds for those in $\mathcal{B}$.

5 Incremental GK index

5.1 Window updating and bi-temporality

In the approach so far the time window W was fixed and all parameters in the index formula used are estimated simultaneously, in one run. But this is not how it is usually done in practice. Instead the index would be updated every month when new price information becomes available. This means that the index is computed incrementally, in a monthly recurring iteration.

There are several ways in which the windows used can be updated. Let $W = \{c, \dots, d\}$ with $c, d \in \mathbb{N}$ and $c \leq d$. We present some examples to illustrate this window updating process. The last month of the previous year serves as the reference month for the first month of the next year.

- Incremental window: $W = \{c, \dots, d\} \rightarrow \{c, \dots, d+1\}$.
- Sliding window: $W = \{c, \dots, d\} \rightarrow \{c+1, \dots, d+1\}$.
- Yearly incremental window: Within each year the window is incremented every month until the final month is reached. In the next year the window is built from scratch. The sequence of windows is $\{(j,1)\}$, $\{(j,1),(j,2)\}$, …, $\{(j,1), \dots, (j,12)\}$, $\{(j,12),(j+1,1)\}$, …, $\{(j+1,1), \dots, (j+1,12)\}$, etc.

Because the information is updated every month, the incremental approach has, in principle, the possibility to use new information about previously used data: it is possible to correct errors in data obtained in previous months. Or data that were missing before have become available and could be used, in principle. However, it is not certain that this updating can actually be done in practice.[28] Often the statistical office responsible for the compilation of the CPI (and HICP and flash HICP) has a policy that an index value once published is never replaced by an updated version, unless a serious error was made, which, obviously, has to be rectified. But this is an exception and therefore we shall not consider it here.

Without this publication restriction we need two time intervals: one time interval to indicate when a fact was true in the real world (e.g. when a certain item had a certain price, or a certain attribute) and another time interval to indicate when this was known. In the language of temporal databases the first time interval is called the 'valid time' and the second time interval is the 'transaction time'[29]

So far we have only used 'valid time'. If we want to use 'transaction time' (= 'decision time') we need to introduce a new parameter. We will write it as a superscript. However, in the discussion below we shall keep it simple: we shall not assume that previous knowledge has to be revised.[30] The important point is that the formalism can cope with such revisions.

So the incremental approach not only implies another way to compute the index, it also allows one to make use of updated information, such as corrections of previously provided figures or

[28] That is updating published figures. But one can update internal data that are not published. They could be used to see to which extent the results based on rectified and updated data differ from the published results.

[29] In the world of temporal databases there is also a third time period: the 'decision time'. This is the time period during which a fact stored in the database was decided to be valid. But we assume that 'transaction time' and 'decision time' coincide.

[30] Because this would require to actually invent an incident that has to be rectified.

new information that was unknown before.[31] It is a policy decision to use this information to correct previously published figures, or not. If publication policy forbids one to do so, one could use this information in the background, for comparison with published index values. The incremental approach makes it easy to incorporate updates, we said. This does not mean that the 'one go' approach cannot cope with them. It can, of course, by recomputing the entire index series for the updated information. It is a computational issue rather than a principled one.

5.2 The index using incremental windows

We assume that we are dealing with incremental windows. This option is chosen mainly for convenience: the formulas are less complicated to present. In practice one would probably choose for a sliding window, or a yearly incremental window. The approach then is similar to the one given below. But the formulas are a bit more complicated. We also assume that the set of goods A remains the same over time. This choice was also made to make the formulas less complicated. In practice this is only a good approximation if the time period considered is not too big. In general it is not a realistic assumption. But assuming a dynamic population would again increase the complexity of the formulas.

We then can write the equations in (4) as:

$$\alpha_i^w = \frac{\sum_{j\in W^w} v_{ij}^w/\pi_j^w}{\sum_{j\in W^w} q_{ij}^w}, \tag{23a}$$

$$\pi_j^w = \frac{\sum_{i\in A_W} v_{ij}^w}{\sum_{i\in A_W} \alpha_i^w q_{ij}^w}, \tag{23b}$$

$$\pi_1^1 = 1. \tag{23c}$$

For the $(w+1)$-st month we have the following system of equations:

$$\alpha_i^{w+1} = \frac{\sum_{j\in W^{w+1}} v_{ij}^{w+1}/\pi_j^{w+1}}{\sum_{j\in W^{w+1}} q_{ij}^{w+1}}, \tag{24a}$$

$$\pi_j^{w+1} = \frac{\sum_{i\in A_W} v_{ij}^{w+1}}{\sum_{i\in A_W} \alpha_i^{w+1} q_{ij}^{w+1}}, \tag{24b}$$

$$\pi_1^1 = 1. \tag{24c}$$

Separating terms with old (up to and including month w) and those with new results (concerning month $w+1$), we can write the equations in (24) as follows:

[31] Which is used to replace missing values by regular values.

$$\alpha_i^{w+1} = \frac{\sum_{j\in W^w} v_{ij}^{w+1}/\pi_j^{w+1} + v_{i,w+1}^{w+1}/\pi_{w+1}^{w+1}}{\sum_{j\in W^w} q_{ij}^{w} + q_{i,w+1}^{w+1}}, \tag{25a}$$

$$\pi_j^{w+1} = \frac{\sum_{i\in A_W} v_{ij}^{w+1}}{\sum_{i\in A_W} \alpha_i^{w+1} q_{ij}^{w+1}}, \tag{25b}$$

$$\pi_{w+1}^{w+1} = \frac{\sum_{i\in A_W} v_{i,w+1}^{w+1}}{\sum_{i\in A_W} \alpha_i^{w+1} q_{i,w+1}^{w+1}} \tag{25c}$$

$$\pi_1^1 = 1. \tag{25d}$$

If we assume that previous results do not need to be corrected at time $w+1$ (so that, in particular, $\pi_j^{w+1} = \pi_j^w$) we can write (25) as

$$\alpha_i^{w+1} = \frac{\sum_{j\in W^w} v_{ij}^{w}/\pi_j^{w} + v_{i,w+1}^{w+1}/\pi_{w+1}^{w+1}}{\sum_{j\in W^w} q_{ij}^{w} + q_{i,w+1}^{w+1}}, \tag{26a}$$

$$\pi_j^{w+1} = \pi_j^w, \tag{26b}$$

$$\pi_{w+1}^{w+1} = \frac{\sum_{i\in A_W} v_{i,w+1}^{w+1}}{\sum_{i\in A_W} \alpha_i^{w+1} q_{i,w+1}^{w+1}}, \tag{26c}$$

$$\pi_1^1 = 1. \tag{26d}$$

We can rewrite α_i^{w+1} as follows

$$\alpha_i^{w+1} = \phi_i^{w+1} \frac{\sum_{j\in W^w} v_{ij}^{w}/\pi_j^{w}}{\sum_{j\in W^w} q_{ij}^{w}} + \left(1-\phi_i^{w+1}\right) \frac{v_{i,w+1}^{w+1}/\pi_{w+1}^{w+1}}{q_{i,w+1}^{w+1}}, \tag{27}$$

where

$$\phi_i^{w+1} \triangleq \frac{\sum_{j\in W^w} q_{ij}^{w}}{\sum_{j\in W^w} q_{ij}^{w} + q_{i,w+1}^{w+1}}. \tag{28}$$

Because $0 \le \phi_i^{w+1} \le 1$ we have that (28) is a convex combination of a component depending on past data (up to and including month w) and a component with data from month $w+1$, the current month:

$$p_{i,w+1}^{w+1} = \frac{v_{i,w+1}^{w+1}}{q_{i,w+1}^{w+1}}, \tag{29}$$

for $i \in A_W$. In (27) this price is divided by the price index value π_{w+1}^{w+1}, which is unknown during month $w+1$ but that we can estimate on the basis of the (partial) information available. We take a closer look at π_{w+1}^{w+1} as defined in (26) and express it in 'past' and 'present' components:

$$\pi_{w+1}^{w+1} = \frac{\sum_{i\in A_W} v_{i,w+1}^{w+1}}{\sum_{i\in A_W} \left[\phi_i^{w+1}\alpha_i^{w} + \left(1-\phi_i^{w+1}\right) p_{i,w+1}^{w+1}/\pi_{w+1}^{w+1}\right] q_{i,w+1}^{w+1}}. \tag{30}$$

We can produce estimators for these parameters α and π, in particular for π_{w+1}^{w+1}, derived from the recursions.

The idea is that we try to predict estimates for month $w+1$ while the data for that month are being collected, using the partial information available at the moment of forecasting (or rather, nowcasting), as follows. We start with estimating the weights:

$$\hat{\phi}_i^{w+1} = \frac{\sum_{j\in W^w} q_{ij}^{w}}{\sum_{j\in W^w} q_{ij}^{w} + \hat{q}_{i,w+1}^{w+1}}. \tag{31}$$

These estimates use 'old' information' (up to and including month w) as well as new one (month $w+1$, yielding an estimate $\hat{q}_{i,w+1}^{w+1}$ of $q_{i,w+1}^{w+1}$, the amount sold of goods in subgroup i in month $w+1$ as far as this month has 'unfolded'. So this is a preliminary forecast based on the data for month $w+1$ that have been collected so far. As the month progresses this estimate gets better and at the end of the month will yield the figure that is produced in the regular CPI production.

The next quantity we consider is

$$\hat{\alpha}_i^{w+1} = \hat{\phi}_i^{w+1}\alpha_i^{w} + \left(1-\hat{\phi}_i^{w+1}\right)\frac{\hat{p}_{i,w+1}^{w+1}}{\hat{\pi}_{w+1}^{w+1}}, \tag{32}$$

where $\hat{\pi}_{w+1}^{w+1}$ is an estimate of the price of goods in subgroup i and in month $w+1$ in progress. Most of the components it contains are known, except $\hat{\pi}_{w+1}^{w+1}$. We compute this quantity later. For now, we substitute (32) into (30) and obtain

$$\hat{\pi}_{w+1}^{w+1} = \frac{\sum_{i\in A_W} \hat{v}_{i,w+1}^{w+1}}{\sum_{i\in A_W} \left[\hat{\phi}_i^{w+1}\alpha_i^{w} + \left(1-\hat{\phi}_i^{w+1}\right] \hat{p}_{i,w+1}^{w+1}/\hat{\pi}_{w+1}^{w+1}\right) \hat{q}_{i,w+1}^{w+1}} \tag{33a}$$

$$= \frac{A_W^{w+1}}{B_W^{w+1} + C_W^{w+1}/\hat{\pi}_{w+1}^{w+1}}, \tag{33b}$$

where

$$A_W^{w+1} = \sum_{i\in A_W} \hat{v}_{i,w+1}^{w+1}, \tag{34a}$$

$$B_W^{w+1} = \sum_{i\in A_W} \hat{\phi}_i^{w+1}\alpha_i^{w}\hat{q}_{i,w+1}^{w+1}, \tag{34b}$$

$$C_W^{w+1} = \sum_{i\in A_W} \left(1-\hat{\phi}_i^{w+1}\right)\hat{p}_{i,w+1}^{w+1}\hat{q}_{i,w+1}^{w+1}. \tag{34c}$$

Solving (33) for $\hat{\pi}_{w+1}^{w+1}$ we obtain:

$$\hat{\pi}_{w+1}^{w+1} = \frac{A_W^{w+1} - C_W^{w+1}}{B_W^{w+1}} \tag{35a}$$

$$= \frac{\sum_{i=1}^{a_W} \hat{\phi}_i^{w+1} \hat{p}_{i,w+1}^{w+1} \hat{q}_{i,w+1}^{w+1}}{\sum_{i=1}^{a_W} \hat{\phi}_i^{w+1} \alpha_i^w \hat{q}_{i,w+1}^{w+1}}. \tag{35b}$$

The reasoning above shows that the definition of the GK-index in fact includes a starter to a time series approach, by considering an incremental approach were the current observation period is extended by one month. In the approach above, we did not correct the observation period at the back, where the oldest observations are. This might in fact be preferable, as one does not want to use too much historic information when computing the index. In this case we would have obtained a sliding window instead of an incrementally growing window. Also there is the issue of the length of the window: if it is too long, the past is weighing too heavily; if it is too short, more recent information gets too much weight and relevant older information is not used at all. So there is a trade-off to be found: a balance between old and new information. The problem is well-known in the context of time series analysis. A similar problem in density estimation exists: Using a wide window around a point p is likely to catch a sizable number of observations, but is not very well suited for measuring the density at p. Taking a smaller window increases the 'locality' of the density estimate at p, but has the drawback that it possibly contains too few observations, resulting in an inaccurate density estimate at p.

6 Optimization models inspired by the GK-index

In the present section we want to explore another kind of generalization of the GK index, namely one based on optimization models. This generalization is of interest as it leads naturally to the TPD index, which is a well-known index. It is discussed in Section 7.

In the remainder of this section we want to consider several object functions for optimization models, inspired by the GK-index (4). Optimizing them leads to price indices that can be viewed as intermediaries of the GK-index and the TPD-index. Several models are presented, also depending on the information that is available (volumes, quantities or prices). In case of scanner data turnover as well as volume of the sales are available (from which prices can be derived for different aggregations of goods and time periods), whereas in case of web scraping only prices are known, for which aggregations are restricted to unweighted cases. These are only approximations to the weighted cases, where the importance of goods based on sales is taken into account, which is the proper way to weight them.

The first object function we consider is the following: [32)]

[32)] For all the models implied by the object functions in the present section a constraint that holds is that $\pi_1 = 1$. This is tacitly assumed.

$$\sum_{i\in A_W}\left(\frac{\sum_{j\in W[i]} v_{ij}/\pi_j}{\sum_{j\in W[i]} q_{ij}} - \alpha_i\right)^2, \tag{36}$$

where

$$W[i] = \{j \in W \mid v_{ij} > 0\}. \tag{37}$$

This is to be minimized for the αs and πs. The meaning of the parameters is as in the GK-index case: the αs are time averaged prices and the πs are price indices. It is to be understood that π_1=1. In case of (36) the goods have the same weights. This is a bit odd if the goods are sold in different quantities. But this is easy to remedy by introducing appropriate weights ξ_i for $i \in A$. For instance, we can take ξ_i proportional to

$$v_{i\cdot} = \sum_{j\in W[i]} v_{ij}, \tag{38}$$

so that

$$\xi_i = \frac{v_{i\cdot}}{v_{\cdot\cdot}}, \tag{39}$$

where

$$v_{\cdot\cdot} = \sum_{i\in A_W, j\in W[i]} v_{ij}. \tag{40}$$

The following object function can then be used instead of (36):

$$\sum_{i\in A_W}\xi_i\left(\frac{\sum_{j\in W[i]} v_{ij}/\pi_j}{\sum_{j\in W[i]} q_{ij}} - \alpha_i\right)^2. \tag{41}$$

Incidentally, in this model the months are considered equally important. This seems to be a reasonable assumption if the turnover of the various goods considered is comparable for each of the months in W.

This assumption also leads us to the following object function:

$$\sum_{j\in W}\left(\frac{\sum_{i\in A_W[j]} v_{ij}}{\sum_{i\in A_W[j]} \alpha_i q_{ij}} - \pi_j\right)^2, \tag{42}$$

where

$$A_W[j] = \{i \in A_W \mid v_{ij} > 0\}, \tag{43}$$

the set of goods in A that were actually sold in month j.

However, in some cases some months may actually be more important than other months because they had a larger turnover. As above we could introduce a weight based on turnover. We could take such a weight, λ_j for month j, proportional to $v_{\cdot j} = \sum_{i\in A[j]} v_{ij}$:

$$\lambda_j = \frac{v_{\cdot j}}{v_{\cdot\cdot}}, \tag{44}$$

where $v_{\cdot\cdot}$ is as in (40) and use as the following object function:

$$\sum_{j\in W}\lambda_j\left(\frac{\sum_{i\in A[j]} v_{ij}}{\sum_{i\in A[j]} \alpha_i q_{ij}} - \pi_j\right)^2. \tag{45}$$

Formulas (36), (41), (42) and (45) can be applied in case volumes and quantities are available. In case only prices are known (e.g. in webscraped data) we cannot use these object functions. Instead we could use:

$$\sum_{(i,j)\in\mathcal{P}}\left(\frac{p_{ij}}{\pi_j} - \alpha_i\right)^2, \tag{46}$$

where

$$\mathcal{P} = \{(i,j) \in A_W \times W \mid p_{ij} > 0\}, \tag{47}$$

which is the set of goods in A_W on the web that were on the web site at least once in period W. [33)]

[33)] Whether it is a good idea to compute a price index only using prices and no weights based on turnover, is another matter. Our interest here is on the possibility to compute quantities that are at least lookalikes of serious price indices.

Formula (46) can also be used as the starting point for object functions that can be used in case volume and quantity information is available, possibly in the form of proxy information.[34)]

Assuming such weights can be computed, the following object function can be used instead of (41):

$$\sum_{(i,j)\in\mathcal{P}} \xi_i \left(\frac{p_{ij}}{\pi_j} - \alpha_i \right)^2, \tag{48}$$

A variant of formula (48) is

$$\sum_{(i,j)\in\mathcal{P}} \xi_i \lambda_j \left(\frac{p_{ij}}{\alpha_i} - \pi_j \right)^2, \tag{49}$$

A variant of object function (49) is:

$$\sum_{(i,j)\in\mathcal{P}} \xi_i \lambda_j \left(\frac{p_{ij}}{\pi_j} - \alpha_i \right)^2, \tag{50}$$

A variant of formula (48) is

$$\sum_{(i,j)\in\mathcal{P}} \lambda_j \left(\frac{p_{ij}}{\alpha_i} - \pi_j \right)^2, \tag{51}$$

An alternative to (48) and (49) is the following object function, where the αs and πs play a symmetric role:

$$\sum_{(i,j)\in\mathcal{P}} \xi_i \lambda_j \left(\frac{p_{ij}}{\alpha_i \pi_j} - 1 \right)^2, \tag{52}$$

In case no weights are available, one can of course use

$$\sum_{(i,j)\in\mathcal{P}} \left(\frac{p_{ij}}{\alpha_i \pi_j} - 1 \right)^2, \tag{53}$$

[34)] For instance, if turnover information is available for a similar shop, or for a branch in retail.

Finally, we have the following object function, which is a variant of (52):

$$\sum_{(i,j)\in\mathcal{P}} \xi_i \lambda_j \left(p_{ij} - \alpha_i \pi_j\right)^2, \tag{54}$$

where ξ_i and λ_j are as in (39) and (44), respectively. With object function (54) we are close to the TPD index, which we shall deal with in Section 7.

For the present section we have achieved our goal, namely to show how the GK-index and the TPD-index can be linked, by presenting a series of optimization models that can be used to compute various price indices. These price indices can be viewed as intermediate forms between the GK-index and the TPD-index. We could elaborate and study each of these indices, but as this is not the goal of the present subsection. We leave the subject with this insight, and move on to the next group of price indices, the family of TPD-type indices.

7 TPD-type indices

7.1 Standard TPD index

Comparing the approach above with that for the TPD method may suggest to look at an optimization method to define an index. In case of the TPD method the model assumption is

$$p_{ij} = \gamma \alpha_i \pi_j \varepsilon_{ij}, \tag{55}$$

where $p_{ij} > 0$ is the observed price of good[35] i in month j, $\gamma > 0$ is a constant, $\pi_j > 0$ is a price component that varies over time, $\alpha_i > 0$ is a parameter that varies over the product groups (COICOPs) and can be seen as a factor to adjust the π_js. These are parameters to be estimated. $\epsilon_{ij} > 0$ is an error term. We can interpret π_j as a price index with base month 1, if we take $\pi_j = 1$ for $j = 1$.

Taking (natural) logarithms on both sides of the equality sign in (55) yields:

$$\log p_{ij} = \log \gamma + \log \alpha_i + \log \pi_j + \log \varepsilon_{ij}. \tag{56}$$

A common way to estimate the parameters in (56) is by minimizing the sum of squares of the log error terms:

[35] This can be a GTIN or a group of GTINs that form a natural unit, and that has, e.g. the property that it exists for a long time, even if the GTINs that it consists of change over time.

$$\sum_{(i,j)\in\mathcal{P}} \left(\log \varepsilon_{ij}\right)^2 = \sum_{(i,j)\in\mathcal{P}} \left(\log p_{ij} - \log \gamma - \log \alpha_i - \log \pi_j\right)^2. \tag{57}$$

In Section 7.2 a weighted variant of (57) is presented, as an alternative.

Another possibility is to use the L^1-norm, which yields

$$\sum_{(i,j)\in\mathcal{P}} \| \log \varepsilon_{ij} \| = \sum_{(i,j)\in\mathcal{P}} \| \log p_{ij} - \log \gamma - \log \alpha_i - \log \pi_j \|. \tag{58}$$

An advantage of using an L^1-norm is that the estimates tend to be more robust.

Minimizing (57) yields estimates for the parameters $\log \gamma$, the $\log \alpha_i$s and the $\log \pi_j$s and hence for γ, the α_is and the π_js.

Our interest is actually only in estimating the values of the π_js, as they are price indices. The other parameters, i.e. γ and the α_is, can be viewed as nuisance parameters in our application. We can get rid of them by considering price ratios of a good i in two different months, say j_1 and j_2, as (55) yields:

$$\frac{p_{i,j_2}}{p_{i,j_1}} = \frac{\pi_{j_2}}{\pi_{j_1}} \varepsilon_{i,j_1,j_2}, \tag{59}$$

where $\varepsilon_{ij_1j_2}$ is an error term. By taking (natural) logarithms we could still estimate the π_js using an optimization procedure, which minimizes the error term, that is

$$\sum_{i\in A_W, j_1, j_2 \in W} \left(\log \varepsilon_{ij_1j_2}\right)^2 = \sum_{i,j_1,j_2} \left(\log p_{i,j_2} - \log p_{i,j_1} + \log \pi_{j_1} - \log \pi_{j_2}\right)^2. \tag{60}$$

7.2 Weighted TPD index

The object function (57) is unweighted: each contribution to the sum has equal value. If turnover is available as a variable, it is preferable to weight each contribution. The following variant of object function (57) is a linearly weighted target function

$$\sum_{(i,j)\in\mathcal{P}} w_{ij} \left(\log \varepsilon_{ij}\right)^2 = \sum_{(i,j)\in\mathcal{P}} w_{ij} \left(\log p_{ij} - \log \gamma - \log \alpha_i - \log \pi_j\right)^2, \tag{61}$$

with suitably chosen weights w_{ij}.

A weighted version of (57) is

$$\sum_{(i,j)\in\mathcal{P}} v_{ij}\left(\log\varepsilon_{ij}\right)^2 = \sum_{(i,j)\in\mathcal{P}} v_{ij}\left(\log p_{ij} - \log\gamma - \log\alpha_i - \log\pi_j\right)^2. \tag{62}$$

A weighted variant of (58) is

$$\sum_{(i,j)\in\mathcal{P}} v_{ij}||\log\varepsilon_{ij}|| = \sum_{(i,j)\in\mathcal{P}} v_{ij}||\log p_{ij} - \log\gamma - \log\alpha_i - \log\pi_j||. \tag{63}$$

We can also take (59) as a starting point and consider a weighted version of (60), resulting in

$$\sum_{i\in A_W, j_1, j_2\in W} w_{i,j_1,j_2}\left(\log\varepsilon_{ij_1j_2}\right)^2 = \sum_{i\in A_W, j_1, j_2\in W} w_{i,j_1,j_2}\left(\log p_{ij_2} - \log p_{ij_1} + \log\pi_{j_1} - \log\pi_{j_2}\right)^2. \tag{64}$$

as a target function.

It is also possible that for a pair of months j_1, j_2 an item is not available in both months. In fact, the items that are available depend on the pair of months considered.

The idea is now to average the price ratios (59) over the set $A_{j_1j_2}$ for each pair j_1, j_2. This averaging is done by taking the weight of subgroup i as well as the distance of j_1 and j_2 into account, using, for instance, the following definition for the weights:

$$w_{j_1,j_2} = \frac{v_i}{1 + |j_1 - j_2|}, \tag{65}$$

where v_i is the turnover of good i.[36] Note that the weights defined in (65) are inversely proportional to the distance of the months in a pair and proportional to the turnover.

For fixed good i and months j_1 and j_2 we can consider price ratios. Let

$$A_W[j_1, j_2] \triangleq A_W[j_1] \cap A_W[j_2]. \tag{66}$$

We want to approximate price ratios p_{ij_2}/p_{ij_1} averaged over $A[j_1, j_2]$ by ratios of price indices, using (65) as weights, combined in the following model:

[36] In some suitable period of time. It does not make sense to use a lot of old turnovers if one wants to be current. On the other hand, using only fairly recent turnovers does not do justice to the average turnover of a good over time.

$$\frac{p_{ij_2}}{p_{ij_1}} = \prod_{i \in A_W[j_1,j_2]} \left(\frac{\pi_{j_2}}{\pi_{j_1}}\right)^{w_{j_1,j_2}} \eta_{i,j_1,j_2}, \tag{67}$$

where the η_{i,j_1,j_2}s are error terms. The procedure is like before: take (natural) logarithms to the left-hand and the right-hand side of (67) and look for those π_j's that minimize the object function $\sum_{i,j_1,j_2} \left(\log \eta_{ij_1j_2}\right)^2$. We save the details.

7.3 TPD-model with trend and seasonal components

In model (55) each month μ has a separate parameter p_μ. There is no notion that there are seasonal effects within a year. But often in economic data it is appropriate to assume that a phenomenon is described by a trend and a periodic or seasonal component.[37] So if month μ is represented by a pair (j, m), where j denotes the year and $m \in \{1, \ldots, 12\}$ the calendar month, we can decompose p_μ in a yearly price p_j^J and a monthly price p_m^W:

$$p_\mu = p_j^J p_m^W. \tag{68}$$

p_j^J can be viewed as a yearly trend and p_m as a monthly adjustment factor.[38] We thus obtain an adjusted TPD model with a yearly trend and a monthly cycle:[39]

$$p_{\alpha jm} = \gamma p_\alpha^G p_j^J p_m^W \varepsilon_{\alpha jm} \tag{69}$$

containing a trend, a seasonal component and a product component. If the time window contains $|W|$ months, and $|W| = |W|$ div $12 + |W|$ mod 12, then the number of parameters in model (69) is $|A_W| + |W|$ div $12 + 12 + 1 = |A_W| + |W|$ div $12 + 13$, for $|W| \geq 12$. Model (69) can be linearized by taking (natural) logarithms left and right of the equality sign. The parameters in model (69) can then be estimated by minimizing the target function:[40]

$$\sum_{\alpha,j,m} (\log \varepsilon_{\alpha jm})^2 = \sum_{\alpha,j,m} (\log p_{\alpha jm} - \log \gamma - \log p_\alpha^G - \log p_j^J - \log p_m^W)^2. \tag{70}$$

In Table 7.1 we have shown a situation with a time window W of 50 months, so there are data of four full years of prices, and for two months in year 5. The data in month 50 is assumed to be incomplete. Hence it is written in red, to distinguish it from the other months in W.

37) We consider months as the seasonal units of time.
38) It would have been possible to use 4 seasons instead of 12 months, or another subdivision of the year, provided there is enough data to estimate the parameters.
39) In state space models for economic data one often assumes that they consist of a trend component, a seasonal component and a noise component, combined in a linear model.
40) We take the unweighted variant. A weighted variant can easily be formulated.

	Jan	Feb	Mar	...	Dec
Y1	1	2	3	...	12
Y2	13	14	15	...	24
Y3	25	26	27	...	36
Y4	37	38	39	...	48
Y5	49	50	-	...	-

Table 7.1 Price data: the months $1, \dots, 49$ are complete and month 50 (final month) is incomplete.

To estimate the parameters for model (70) one can proceed in a similar fashion as in case of the TPD model, as the adjusted model is a refinement of the TPD model.

The estimates from (70) yield the trend and seasonal components of a price index $\mathcal{I}_{j,p}$ defined as follows:

$$\mathcal{I}_{j,p} \triangleq \gamma_{\mathcal{I}} p_j^J p_m^W, \tag{71}$$

for $j = 1, 2, \dots$ and $m = 1, 2, \dots, 12$ with

$$\gamma_{\mathcal{I}} \triangleq \frac{1}{p_1^J p_1^W}, \tag{72}$$

so that $\mathcal{I}_{1,1} = 1$.

Another approach to produce a price index based on optimizing (70) is the following. One first estimates the parameters as above. Then the estimate for $\log p_\alpha$, say $\widehat{\log p_\alpha}$, is used to adjust the prices $p_{\alpha jm}$ for the goods component:

$$\frac{p_{\alpha jm}}{\hat{p}_\alpha}, \tag{73}$$

where $\hat{p}_\alpha = \exp(\widehat{\log p_\alpha})$.

In the next step we use model (70) again to estimate the various components again. As the prices are (somewhat) corrected for the influence of the goods, we expect the estimates for the p_α^Gs to be closer to 1 for all goods α. Repeating this procedure yields estimates for $\log p_j^J$ ($j = 1, 2, \dots$) and $\log p_m^W$ ($m = 1, 2, \dots, 12$) as well as for $\log \gamma$. The latter estimate can also be obtained by calibrating the price index estimates, by assuming that $p_1^J = 1$ and $p_1^W = 1$. This yields estimates of a price index like (71), but with the trend and seasonal components replaced by the n^{th} iterate for p^J and p^W, for some, sufficiently high, n.

8 Discussion

In this paper we have discussed several methods to compute price indices that are all variants or generalizations of an idealized GK index. Within this family we also have the TPD-index. The main aim was to show which indices can be derived from the idealized GK-index. These indices are presented without trying to pick a favorite, as it is not clear that there is an obvious winner. But as closely related indices it is interesting to see the results they produce when applied to the same data and compare them.

It is clear that there is an abundance of indices of the decomposition type available. In practice the choice for a suitable index is limited by the circumstances: the variables that are available (volumes and quantities or only prices), the frequency of the data collection, the publication schedule adopted (in case of using incremental estimates), etc.

Several of the generalizations led to time series models: the incremental model (for nowcasting) and the seasonal TPD model. The seasonal TPD-model in Section 7.3 can be seen as a structural time series model (see [**?**]), but it is arrived at in a different way. A product parameter acts as a nuisance parameter. In a typical TPD model one would first aggregate over the various product groups to obtain average monthly prices, which are then used to model the time series and, in particular, estimate the components for trend and season.

One of the goals of this paper is to push 'standard' price index theory in the direction of time series models. The application we have in mind is the flash HICP. The current version does not use historical data. In [2] a proposal is made for a new version of this estimator using historical data exclusively or additional to current data. One can use time series models as they are typically used in all kinds of applications (such as MA, ARMA, ARIMA, Kalman-filter, etc), but which are not particularly geared to the price index application. As the present paper shows it is also possible to come up with models inspired by price index theory itself. This is easily achieved by using an incremental approach to price index computation.

When modeling a price index series as a time series, on can always use a direct method which models the prices themselves as the quantities to model. But in some cases, time series of turnover and quantity data are available from which one can compute the prices. The availability of these series is also assumed for the GK-price index. In this case one can also proceed to model value and quantity series using time series modeling, and compute series for prices from these series. In [2] this is not attempted. There is no guarantee that this approach will give better results, but as the approach is rather obvious, it is inviting to investigate it.

References

[1] B. Balk (2008). *Price and Quantity Index Numbers*. Cambridge University Press.

[2] E. van Bracht & L. Willenborg (2019). Towards a New Flash HICP. Discussion Paper, CBS, The Hague.

[3] A. Chessa (2016). The QU-Method: A New Methodology for Processing Scanner Data. Proceedings of Statistics Canada Symposium 2016.

[4] R. Geary (1958). A Note on the Comparison of Exchange Rates and Purchasing Power between Countries. Journal of the Royal Statistical Society A 121, pp. 97–99.

[5] S. Khamis (1972). A New System of Index Numbers for National and International Purposes. Journal of the Royal Statistical Society A 135, pp. 96–121.

Colophon

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

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information