Discussion Paper

Exploring the GK price index

Leon Willenborg

22 January 2024

1 Introduction

The price index that is central to this paper[1)] was introduced by Geary (see [6]) and was further explored by Khamis (see [8]). The present paper intends to explore this price index a bit more and to consider some variants. This exploration was primarily driven by the interests of the author. The goal was simply to better understand this price index. For the present author this price index has great appeal because it is based on a reasoning that seems quite natural. When one starts looking more closely, one becomes aware that refinements are possible; some of them suggest themselves, for instance by using alternative ways to average results. In the GK price index certain (obvious) ways to do this were chosen by Geary, but if one realizes that other choices are possible, one might become curious to find out what price indices can be obtained, how they would perform and how the results compare to those based on the GK index itself. It is a deep conviction of the present author that there is no best price index.[2)] At best there is a class of 'good' price indices. One should focus on finding such a class of estimators rather than on a single 'best' estimator. The advantage of having several candidates is that one can apply them to the same data, compare the results and observe differences and similarities in behavior in different situations.

The paper is organized as follows. In Section 2 the GK price index is introduced, and defined in terms of a system of equations. These equations are not symmetric and also not linear in the parameters, which are vectors of prices of goods and a vector with price index values. By using different parameters it is possible to obtain defining equations that are both symmetric in the parameters and linear. This allows us to rewrite the system of equations in such a way that two eigenvalue equations emerge. It turns out to be more profitable to consider these equations as fixed point equations.

The GK price index is not transitive, as it is based on the Paasche index. However, we can define the price index values for a subset of all ordered pairs of base periods (months) and use a method, called transitive continuation, to produce a price index for all pairs of base periods that is transitive. This is discussed in Section 3.

In Section 4 we consider an incremental definition of the GK price index. This definition also allows the set of commodities considered to be variable. Not only is this incremental approach more flexible, it also fits very well with the practice of making price indices. It is also fairly easy to adhere to a constraint that is often applied in practice, namely that index values that have been published cannot be changed.[3)]

In Section 5 we consider the problem of chaining. This is about combining short (i.e. yearly) price index series into a long one. We assume that, as is usual, around the year change modifications are made with respect to the set of commodities used: as they change over time (new items are introduced to the market, existing items are taken from the market, either temporarily or definitively. They may be replaced by other items, or not). These changes, as well due to the fact

1) The author is grateful to Sander Scholtus for his remarks on a nearly final draft of the present paper, which resulted in various improvements of the text.

2) Or more generally, a best estimator. Despite the talk of optimal estimators, closer inspection of the circumstances always shows that such estimators are the result of a host of assumptions, sometimes tacit ones.

3) Except when gross errors that have been made in the past, due to using faulty or inadequate data, come to light.

that a new sample of shops are drawn into the sample for the new year, may introduce 'discontinuities' that we want to diminish, so the price index series develops as smoothly as possible. The chaining problem is independent from the choice of index. However, we intend to suggest a chaining approach that is geared at the GK price index. We suggest some approaches but we do not deliver them here, simply due to lack of time.

In Section 6 we derive some variants of the GK price index, in particular by applying some alternative averaging methods from the one that is applied in the GK price index (that is, the arithmetical mean). This also yields alternative price indices, which happen to be weightless, that is, where all weights are equal and can assumed to be 1. One may perhaps wonder what the purpose is of considering all these variants of the GK price index. The reason is a deeply seated conviction of the present author that a single superior price index does not exist; at best an entire class of such price indices. This means that one should consider good, rivalling price indices, and apply them all to the same data and study how they perform: to which phenomena do they react the same and for which phenomena do they show different behavior? It is very much like the use of different clocks to 'estimate' local time. With a single clock one cannot detect deviations, only when using several clocks. Of course, the clocks should be of good quality to be of any value.

So far we have treated the GK price index as a non-random variable. But in practice it is actually a random variable, simply because it is applied to a sample of commodities. In Section 7 we consider some consequences of the explicit assumption that the GK price index is a random variable. In particular it opens the door to methods like bootstrapping, which can be used to give insight into the variability of the estimates, of the price index values and the 'international' prices[4)] of commodities. Apart from bootstrapping two other methods are suggested, one using time series analysis.

In Section 8 the main results of the present paper are summarized and discussed. Also some suggestions for possible future work are made. A list of references and a few appendices conclude the paper.

2 GK price index

The GK price index was originally (very briefly) defined by Geary ([6]) in the context of comparisons of the purchasing power of different countries, and more extensively investigated by Khamis ([8]), who also pointed out the possibility to use the concept in a national context as a price index. But it can also be viewed as a price index in a single country. The countries in the original representation are then replaced by base periods (say months).

The GK price index is defined by the following system of equations

4) In case the GK index is defined in a setting of multiple countries with their own currencies.

$$\alpha_i = \frac{\sum_{j\in W} v_{ij}/\pi_j}{\sum_{j\in W} q_{ij}}, \text{ for } i \in G,$$
$$\pi_j = \frac{\sum_{i\in G} v_{ij}}{\sum_{i\in G} \alpha_i q_{ij}}, \text{ for } j \in W, \tag{1}$$

where

$$v_{ij} = p_{ij} q_{ij}, \tag{2}$$

is the volume (or turnover) of commodity i sold in month j, with p_{ij} the price of commodity i in month j and q_{ij} the quantity of commodity i sold in month j.

As to the interpretation of the system equations in (1), the following. Perhaps intuitively it is easiest to assume that the months correspond to countries, each with their own currencies.[5)] The first set of equations of (1) defines the international price of each commodity i: its total value per country j, expressed in national prices, divided by the purchasing power parity (PPP) for this country summed over all countries and divided by the total quantity of this good (that is, in all countries). The second set of equations defines, for each country j, the PPP as as its Paasche price index relative to the country with the 'international prices'. (See [1], p. 45.) In our context (with commodities and their prices over time), the purchasing power parities (the π_j) can be interpreted as price index values, which are named after Geary and Khamis. We may assume that the first country (using the phrasing of the original GK setting) is the country with the 'international prices', so that we may put

$$\alpha_1 = 1,$$
$$\pi_1 = 1. \tag{3}$$

We can rewrite (1) as follows:

$$\alpha_i = \frac{\sum_{j\in W} v_{ij}/\pi_j}{\sum_{j\in W} q_{ij}}, \text{ for } i \in G,$$
$$1/\pi_j = \frac{\sum_{i\in G} \alpha_i q_{ij}}{\sum_{i\in G} v_{ij}}, \text{ for } j \in W, \tag{4}$$

and put

$$\beta_j = 1/\pi_j, \tag{5}$$

5) This is precisely the context in which the original GK price index was defined.

for $j \in W$, so that (4) can be written as

$$\alpha_i = \frac{\sum_{j \in W} v_{ij}\beta_j}{\sum_{j \in W} q_{ij}}, \text{ for } i \in G,$$
$$\beta_j = \frac{\sum_{i \in G} q_{ij}\alpha_i}{\sum_{i \in G} v_{ij}}, \text{ for } j \in W, \tag{6}$$

with

$$\alpha_1 = 1,$$
$$\beta_1 = 1. \tag{7}$$

Note the similarity of the expressions in both sets of equations. We can simplify (6) as follows

$$\alpha_i = \sum_{j \in W} r_{ij}\beta_j, \text{ for } i \in G,$$
$$\beta_j = \sum_{i \in G} s_{ij}\alpha_i, \text{ for } j \in W, \tag{8}$$

where

$$r_{ij} = \frac{v_{ij}}{\sum_{j \in W} q_{ij}},$$
$$s_{ij} = \frac{q_{ij}}{\sum_{i \in G} v_{ij}}. \tag{9}$$

The systems of equations in (6) are linear. In matrix form they can be written as

$$\alpha = R\beta,$$
$$\beta = S'\alpha \tag{10}$$

where $\alpha = (\alpha_1, \dots, \alpha_m)'$, $\beta = (\beta_1, \dots, \beta_n)'$ and where $R = (r_{ij})$ and $S = (s_{ij})$ are $m \times n$ matrices. Combining the equations in (10) we obtain

$$\alpha = RS'\alpha$$
$$\beta = S'R\beta. \tag{11}$$

For conditions that guarantee existence of solutions of the equations in (11) see Appendix A.

Obviously, if β has been determined then the $\pi_j = 1/\beta_j$ are known. π_j is the price index value comparing the base month 1 with reference month j. If these values are known then, assuming transitivity, one can compute the price index values π_{kl} for $k, l = 1, \ldots, n$. This topic is considered in Section 3.

3 Transitive continuation of a GK price index ansatz

The GK price index is based on the Paasche index, which is an intransitive price index. This, however, only becomes manifest if the Price Index DiGraph (PIDG)[6] contains cycles. If this is not the case, there is no problem with transitivity. In other words, if the PIDG is a spanning directed tree (ditree, for short) there is no violation of the transitivity property. In fact if the ditree is a spanning tree, it can be used to extend the price index in such a way that the extension is complete (that is, involves every pair of months), is a transitive price index and, being an extension, it agrees with the price index values defined for the arcs in the spanning ditree (that is, the pairs of months corresponding to its arcs). We consider the PIDG as an ansatz[7] for a price index. This method, called transitive continuaton, in fact hinges on the requirement that the extended index is transitive. The extension method yields a unique solution. In case of a ditree being of the kind typical for the Paasche price index, it is straightforward to derive, as will be shown below. However, a similar transitive extension is possible for a general price index spanning ditree as is shown in Appendix C, under the name of transitive continuation.

As remarked before if the PIDG of an intransitive price index, such as a Paasche or GK price index, is a ditree there is no violation of the transitivity property. In fact, it is possible to define a transitive price index that holds for every pair of months for which the original price index was observed. This is achieved by requiring that for any oriented cycle, the product of the price index values associated with the (oriented) cycle multiply to 1. Like in case of the cycle method[8] it is sufficient to consider the elementary cycles in the PIDG, which in this case is a complete digraph, meaning that for every pair a, b of nodes either (a, b) or (b, a) is an arc. For instance, in case of the Paasche price index the first month of a year is the base month and the price comparisons of ever month in that year are compared with those of the first month. This means that the price index values given are π_{1i} for $i = 1, \ldots, 12$. Given these values we can compute the values of π_{ij} for $i, j = 1, \ldots, 12$, assuming transitivity to hold. If we define π_{ij} as

$$\pi_{ij} = \frac{\pi_{1j}}{\pi_{1i}} = \frac{\beta_i}{\beta_j} \tag{12}$$

6) This is a directed graph where each arc (a, b) corresponds to a pair of periods being compared price wise, with a the base period and b the reference period. In the present paper these periods are in fact months. The months appearing in the arcs for a window W, typically of an uninterrupted sequence of months. W, viewed as a set, not as a sequence, is the set of nodes of the PIDG.

7) Ansatz is a German word meaning something like 'beginning', 'attempt' or 'departure'.

8) The cycle method can be used to turn a set of intransitive price index numbers into a transitive set of such numbers, trying to stay as close to the original price indices as possible. See [11] for more information on this method.

and

$$\pi_{i1} = 1/\pi_{1i}$$
$$\pi_{11} = 1 \tag{13}$$

then we have for all months i, j, k in that year that holds

$$\pi_{ij} = \pi_{ik}\pi_{kj}. \tag{14}$$

For if $i = 1$ and $j \neq 1$ we have $\pi_{1k}\pi_{kj} = \pi_{1k}\frac{\pi_{1j}}{\pi_{1k}} = \pi_{1j}$ and if $i, j \neq 1$ we have $\pi_{ik}\pi_{kj} = \frac{\pi_{1k}}{\pi_{1i}}\frac{\pi_{1j}}{\pi_{1k}} = \frac{\pi_{1j}}{\pi_{1i}} = \pi_{ij}$.

In case of the GK price index, $\pi_j = 1/\beta_j$) is in fact about comparison of month j with month 1. As shown above, transitive continuation can be applied to obtain price indices for a month i as base month and month j as comparison month, for $i, j \in W$. It is also possible to link two years and apply transitive continuation to this two year period. See Appendix C for examples.

4 Incremental GK price index

The sets of commodities that are sold and bought change over time. New commodities are introduced to the market, older commodities may be withdrawn, and may possibly be replaced by similar ones, but repackaged, rebranded and with a higher price. Which commodities are substitutes of which other commodities is a question that is sometimes difficult to answer. Here we shall take a simpler approach. We start our observation period by a set of commodities and if new commodities are introduced to the market, in due course, they will simply be added. No commodities will be eliminated. This is not needed as commodities taken from the market will also not be sold and bought, implying that the corresponding q_{ij} and v_{ij} are both equal to 0. Not discarding any commodities from the (dynamic) set of commodities has the advantage one does not have to distinguish between commodities taken permanently from the market and commodities that were temporarily unavailable. Making such a distinction is usually impossible for a statistical office, due to lack of information.

So let the months in the observation window be numbered consecutively $1, 2, 3, \ldots$. We obtain consecutive windows $W_1 = \{1\}, W_2 = \{1, 2\}$, etc., recursively defined as $W_{n+1} = W_n \cup \{n + 1\}$. We have that the consecutive time windows form a monotone, increasing 'tower' of sets $W_1 \subset W_2 \subset W_3 \subset \ldots$. We also have the sets of commodities form a tower of monotone nondecreasing sets: $G_1 \subseteq G_2 \subseteq G_3 \subseteq \ldots$. The difference between the 'growth' of the W_j on the one hand and of the G_j on the other is predictability: the W_j increase in size by 1 every month, whereas the G_i grow more randomly every month, if at all.

We consider here the incremental method where at each step one month (the newest) is added to the window at hand, so that W_{n+1} replaces W_n and the current set of commodities G_n is

replaced by G_{n+1}. Actually this is not standard practice when price indices are compiled at a statistical office. This results from the organization of the update work rather than a result of some deliberate methodological reasoning. However, the advantage of applying the monthly incremental method is that it is almost 'continuous' and avoids problems with chaining of short index sequences (covering single years). See Section 5 for a discussion of this topic.

The following equations define the GK price index values for the extended time window W_{n+1} and the extended set of commodities G_{n+1}:

$$\alpha_i^{n+1} = \frac{\sum_{j \in W_{n+1}} v_{ij} \beta_j^{n+1}}{\sum_{j \in W_{n+1}} q_{ij}}, \text{ for } i \in G_{n+1},$$
$$\beta_j^{n+1} = \frac{\sum_{i \in G_{n+1}} q_{ij} \alpha_i^{n+1}}{\sum_{i \in G_{n+1}} v_{ij}}, \text{ for } j \in W_{n+1} \quad (15)$$

where v_{ij} and q_{ij} are as defined in (2) but now for W_{n+1} and G_{n+1}. As before, we can solve this and obtain a solution $\bar{\beta} = (\bar{\beta}_1, \dots, \bar{\beta}_{n+1})$. We only use $\bar{\beta}_{n+1}$ from $\bar{\beta}$. For $j \in W_n$ we have the parameters $\beta_1, \dots, \beta_n$. Therefore, for W_{n+1} we have the parameters $\beta_1, \dots, \beta_n, \bar{\beta}_{n+1}$. Hence we have preserved the current price index values and used the most recent information to compute $\bar{\beta}_{n+1}$, without any restrictions on the values for the months $j \in W_n$.

This approach is very useful in practice, as it computes the most recent price index values month-by-month, as soon as the price information for the newest month has been received and processed by the statistical office. The fact that $\bar{\beta} = (\bar{\beta}_1, \dots, \bar{\beta}_{n+1})$ is also available is a bonus, as it can be used to compare the results with those of the method proposed.

The incremental definition used in the present section, is an attractive alternative to the more static one presented in Section 2. It is in line with the actual application of price index computations in practice. Note that it is possible that the set of commodities may vary over time. Note that in case of the original definition in (1) the implicit assumption was that the set of commodities is equal during the entire window W_n. This makes the incremental definition of the GK index also more flexible than the original one.

A drawback, however, of the incremental method presented here is that the set of commodities as well as the time windows only increases. That is, in each month the entire past, since the start of the observations, is carried along. The question arises whether information from a too distant past and from 'old' commodities is beneficial or harmful for computing current price indices. This topic is central to Section 5, which considers both W_n and G_n as parameters to 'play with'. How can different choices for these parameters be used to obtain better estimates for the parameters α_i and β_j? We consider the problem in the context of chaining short yearly price index sequences into a long sequence over several years.

5 Chaining

So far we have taken the roles of the time window W and the set of commodities G for granted when estimating the parameters $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\beta = (\beta_1, \dots, \beta_n)$, with $m = |G|$ and

$n = |W|$. So actually the values for α and β are dependent on the choices of these W and G, and to make this dependence explicit we should write $\alpha^{W,G}$ and $\beta^{W,G}$ instead of just α and β.

In the present section we describe another approach, different from that in Section 4, but closer to practice. We consider two consecutive years, Y and $Y+1$. We assume that the set of commodities in year Y is G_Y and in year $Y+1$ it is G_{Y+1}. These sets are assumed constant for each month in Y and $Y+1$, respectively. We could treat each year separately and come up with estimates for α and β for Y and $Y+1$, say $\alpha^Y, \beta^Y, \alpha^{Y+1}$ and β^{Y+1}. But then we have no control about the jump from β_{12}^Y to β_1^{Y+1}, and hence π_{12}^Y to π_1^{Y+1}. By using an overlap, consisting of at least one month to several months (2 or 3), one can better control the continuation of the price index series for Y to that for $Y+1$. For instance, by extending the method for year $Y+1$ to month 12 of year Y, we can make another estimate for this month: $\beta_{Y,12}^{Y+1}$. We can use the factor

$$\rho_Y = \frac{\beta_{Y,12}^{Y+1}}{\beta_{12}^Y} = \frac{\pi_{12}^Y}{\pi_{Y,12}^{Y+1}} \tag{16}$$

to adjust the sequence of price index values in $Y+1$ to obtain the following sequence of price index values for years Y and $Y+1$:

$$(\pi_1^Y, \dots, \pi_{12}^Y, \rho_Y \pi_1^{Y+1}, \dots, \rho_Y \pi_{12}^{Y+1}). \tag{17}$$

Note that (17) keeps the index values for the year Y untouched. Only those of year $Y+1$ are adjusted.

Another possibility is to extend the method of year Y to the first month of year $Y+1$. We then obtain the value $\beta_{Y+1,1}^Y$. We now define the factor

$$\sigma_Y = \frac{\beta_1^{Y+1}}{\beta_{Y+1,1}^Y} = \frac{\pi_{Y+1,1}^Y}{\pi_1^{Y+1}}. \tag{18}$$

Using this factor σ_Y we then obtain the following sequence of price index values for years Y and $Y+1$:

$$(\pi_1^Y, \dots, \pi_{12}^Y, \sigma_Y \pi_1^{Y+1}, \dots, \sigma_Y \pi_{12}^{Y+1}). \tag{19}$$

Note that for months i, j in year Y:

$$\pi_{ij}^Y = \frac{\pi_{1j}^Y}{\pi_{1i}^Y}, \tag{20}$$

and for months i, j in year $Y+1$:

$$\pi_{ij}^{Y+1} = \frac{\sigma_Y \pi_{1j}^{Y+1}}{\sigma_Y \pi_{1i}^{Y+1}} = \frac{\pi_{1j}^{Y+1}}{\pi_{1i}^{Y+1}}. \tag{21}$$

Furthermore, for month i in year Y and month j in year $Y+1$:

$$\pi_{ij}^{Y,Y+1} = \frac{\sigma_Y \pi^{Y+1}}{\pi_{1i}^{Y}} = \sigma_Y \frac{\pi^{Y+1}}{\pi_{1i}^{Y}}. \tag{22}$$

Although these methods with one month overlap yield a long sequence of price index values, with a smoother transition at the beginning of year $Y+1$, a drawback of these methods is that for every change of year a new factor (like ρ_Y or σ_Y) is needed. These factors are multiplied each time when the short sequence for a new year is chained to the current long sequence. Consider for instance the following example of a long sequence obtained by chaining four short sequences, corresponding to the years Y, $Y+1$, $Y+2$ and $Y+3$, using for example method (18):

$$(\pi_1^Y, \dots, \pi_{12}^Y, \sigma_Y \pi_1^{Y+1}, \dots, \sigma_Y \pi_{12}^{Y+1}, \sigma_{Y+1}\sigma_Y \pi_1^{Y+2}, \dots, \sigma_{Y+1}\sigma_Y \pi_{12}^{Y+2},$$
$$\sigma_{Y+2}\sigma_{Y+1}\sigma_Y \pi_1^{Y+3}, \dots, \sigma_{Y+2}\sigma_{Y+1}\sigma_Y \pi_{12}^{Y+3}). \tag{23}$$

So the question is how this could be avoided. One way to achieve this by using a longer overlap period, say of 2 or 3 (or even more months) to create a smooth transition from the sequence of price index values from year Y to that of year $Y+1$, keeping those in year Y fixed. In this overlap period the results of the method used in year Y and that of year $Y+1$ are combined by using convex combinations. The weights used in the overlap period change to give more weight to the results of the new sequence of year $Y+1$ for more recent months. For a three months overlap period, we have the input as in Tables 5.1 or 5.2. For Table 5.1 the overlap months are $(Y,10)$, $(Y,11)$ and $(Y,12)$. For Table 5.2 the overlap months are $(Y+1,1)$, $(Y+1,2)$ and $(Y+1,3)$. The columns in both tables correspond to months in year Y or $Y+1$, whereas the rows present parts of the price index sequences in that year; the bottom row gives part of the combined sequence. $\eta_k \geq 0$ and $\zeta_k \geq 0$ are weights such that $\eta_k + \zeta_k = 1$ for $k = 10, 11, 12$. Likewise for the weights $\mu_k \geq 0$ and $\nu_k \geq 0$ and $\mu_k + \nu_k = 1$ for $k = 1, 2, 3$. A value like π_{10}^Y is computed for month 10 in year Y using the method for the year Y, in particular the set of commodities for that year, that is, G_Y. However, a value like $\pi_{Y,10}^{Y+1}$ is about month 10 in year Y but computed using the method of year $Y+1$, in particular, using the set of commodities for that year, that is G_{Y+1}. Likewise for the other months in the overlap period.

$(Y,9)$	$(Y,10)$	$(Y,11)$	$(Y,12)$	$(Y+1,1)$
π_9^Y	π_{10}^Y	π_{11}^Y	π_{12}^Y	–
–	$\pi_{Y,10}^{Y+1}$	$\pi_{Y,11}^{Y+1}$	$\pi_{Y,12}^{Y+1}$	π_1^{Y+1}
π_9^Y	$\eta_{10}\pi_{10}^Y + \zeta_{10}\pi_{Y,10}^{Y+1}$	$\eta_{11}\pi_{11}^Y + \zeta_{11}\pi_{Y,11}^{Y+1}$	$\eta_{12}\pi_{12}^Y + \zeta_{12}\pi_{Y,12}^{Y+1}$	π_1^{Y+1}
π_9^Y	$(\pi_{10}^Y)^{\eta_{10}}(\pi_{Y,10}^{Y+1})^{\zeta_{10}}$	$(\pi_{11}^Y)^{\eta_{11}}(\pi_{Y,11}^{Y+1})^{\zeta_{11}}$	$(\pi_{12}^Y)^{\eta_{12}}(\pi_{Y,12}^{Y+1})^{\zeta_{12}}$	π_1^{Y+1}

Table 5.1 Two ways to chain two short price index sequences (for the years Y and $Y+1$) to a long sequence. The overlap period consists of the months $(Y,10)$, $(Y,11)$, $(Y,12)$. For the exponents holds: $\eta_i + \zeta_i = 1$ with $\eta_i, \zeta_i \geq 0$ for $i = 10, 11, 12$. Also: $\eta_{10} > \eta_{11} > \eta_{12}$.

$(Y,12)$	$(Y+1,1)$	$(Y+1,2)$	$(Y+1,3)$	$(Y+1,4)$
π_{12}^{Y}	$\pi_{Y+1,1}^{Y}$	$\pi_{Y+1,2}^{Y}$	$\pi_{Y+1,3}^{Y}$	–
–	π_{1}^{Y+1}	π_{2}^{Y+1}	π_{3}^{Y+1}	π_{4}^{Y+1}
π_{12}^{Y}	$\mu_1\pi_{Y+1,1}^{Y}+\nu_1\pi_{1}^{Y+1}$	$\mu_2\pi_{Y+1,2}^{Y}+\nu_2\pi_{2}^{Y+1}$	$\mu_3\pi_{Y+1,3}^{Y}+\nu_3\pi_{3}^{Y+1}$	π_{4}^{Y+1}
π_{12}^{Y}	$(\pi_{Y+1,1}^{Y})^{\mu_1}(\pi_{1}^{Y+1})^{\nu_1}$	$(\pi_{Y+1,2}^{Y})^{\mu_2}(\pi_{2}^{Y+1})^{\nu_2}$	$(\pi_{Y+1,3}^{Y})^{\mu_3}(\pi_{3}^{Y+1})^{\nu_3}$	π_{4}^{Y+1}

Table 5.2 Two ways to chain two short price index sequences (for the years Y and $Y+1$) to a long sequence. The overlap period consists of the months $(Y+1,1)$, $(Y+1,2)$, $(Y+1,3)$. For the exponents holds: $\mu_i+\nu_i=1$ with $\mu_i,\nu_i\geq 0$ for $i=1,2,3$. Also: $\mu_1>\mu_2>\mu_3$.

In Tables 5.1 and 5.2 two options for chaining two short price index sequences are presented: an addditive one, presented in the third row (indicated by 'a'), using convex combinations of weights, and a multiplicative one, presented in the fourth row (indicated by 'm'), using exponents from convex combinations of weights.

So far we have not said anything about the choice of the weights (η_k,ζ_k) for $k=10,11,12$ and (μ_k,ν_k) for $k=1,2,3$. The idea is that the influence of the 'old' index sequence decreases and the influence of the 'new' sequence increases as time proceeds. So a possible choice of weights could be $(\eta_{10},\zeta_{10})=(\frac{3}{4},\frac{1}{4})$, $(\eta_{11},\zeta_{11})=(\frac{1}{2},\frac{1}{2})$ and $(\eta_{12},\zeta_{12})=(\frac{1}{4},\frac{3}{4})$. Likewise for the (μ_k,ν_k), for $k=1,2,3$. Of course, this is not the only possible choice, nor necessarily the best one, but certainly an easy and obvious one. In fact, a closer study of the development of the sequences could bring to light that another choice for the weights would actually yield a smoother transition from the 'old' to the 'new' short sequence.

Remark Another take on the choice of weights is to make sure that the long series does not show a clear 'break' in its development around the new year, as a result of chaining two short series. This could be achieved by using a time series model for the index value series. And it can only be done for the entire time period, and not being bounded by the restriction that old values (that are supposed to have been published) can be changed afterwards. So as an action 'behind the scenes' this is an option, with the goal to find out how the published price index values fare compared to the ones determined without such restrictions, driven by publication policy. □

6 Variants of the GK price index

In the present section we have a formal look at the GK price index, that is, look at its definition, observe some characteristics and tamper with them. This yields new price indices, which are viewed as variants of the GK price index. We do not claim that all the resulting variants have special merits. They are chiefly intended to be contrasted with the GK price index, and among themselves, applied to the same real data.

Two ideas have been used in the present section to obtain these variants. The first idea is about the averaging methods applied. In the GK price index arithmetical averaging is applied. From a mathematical point of view there is no compelling reason why this is the only possible choice. We can pick different averaging operations just as well and see which price indices emerge. The second idea is based on looking at the formulas and see where parameters occur. These, or similar, parameters could be inserted in different places, at least to make sense (or rather, not to be nonsensical, a priori). A somewhat irritating characteristic of the GK price index is that the

roles played by the π_js and the α_is are not symmetric. By introducing the β_j instead of π_j this 'defect' can be rectified. But it turns out that it is also quite easy to change the defining formulas a bit so that the desired symmetry between the price index and its 'commodity's counterpart' is obtained (see Sections 6.2.1 and 6.2.2).

Considering alternative price indices in the ways described is not merely a theoretical exercise. In practice it is never the case that there is a single best price index.[9] It is a good practice not to focus on a single candidate but to look at several promising ones. They can be used to compare the results when they are applied to the same data. It is like using several (well-made clocks) instead of a single one to determine local time. Only with several clocks one gets an insight about the 'correct' local time. In case of using a single clock this is impossible: it may be running ahead or behind or on time. One does not have a clue. In case of price indices[10] one can use several competing ones to see if they show a similar qualitative or quantitative behavior when applied to the same sets of data.

6.1 Using alternative averaging methods

6.1.1 Geometric mean

We can write the defining equations (1) of the GK price index as follows

$$\begin{aligned}\alpha_i &= \frac{\frac{1}{n}\sum_{j\in W} v_{ij}/\pi_j}{\frac{1}{n}\sum_{j\in W} q_{ij}}, \text{ for } i \in G,\\ \pi_j &= \frac{\frac{1}{m}\sum_{i\in G} v_{ij}}{\frac{1}{m}\sum_{i\in G} \alpha_i q_{ij}}, \text{ for } j \in W,\end{aligned} \tag{24}$$

where $n = |W|$ and $m = |G|$. Note that (24) uses arithmetical averages in enumerators and denominators instead of totals.

An alternative to the GK-index is obtained if we replace the arithmetic averages in (24) by geometric averages. From an algebraic perspective this seems a better choice, as only multiplication and its inverse operation, division, are used in the defining equations. We then obtain:

$$\begin{aligned}\alpha_i^{\triangle} &= \frac{\sqrt[n]{\Pi_{j\in W} v_{ij}/\pi_j^{\triangle}}}{\sqrt[n]{\Pi_{j\in W} q_{ij}}} = \Pi_{j\in W}\sqrt[n]{p_{ij}/\pi_j^{\triangle}}, \text{ for } i \in G,\\ \pi_j^{\triangle} &= \frac{\sqrt[m]{\Pi_{i\in G} v_{ij}}}{\sqrt[m]{\Pi_{i\in G}\alpha_i^{\triangle} q_{ij}}} = \Pi_{i\in G}\sqrt[m]{p_{ij}/\alpha_i^{\triangle}}, \text{ for } j \in W.\end{aligned} \tag{25}$$

[9] Or operationaliizaton of any other concept in whatever science. Concepts like distance, weight, length are not God-given. Man has to choose them, and there are typically a range of possible choices. The choice made is usually guided by matters suchs as mathematical convenience, computational feasibility, stability of the results, sensitivity for outliers, etc.

[10] As 'instruments' for measuring inflation.

Note that the quantities q_{ij} have disappeared from the defining equations (25), which is quite remarkable. It is worth noting that the price index $\pi_j^{\triangle}$ was in fact already proposed by Gerardi in the setting of international comparisons, as mentioned in [1], p. 250 (formula (7.57)).

Combining the sets of equations in (25) yields the following identities:

$$\alpha_i^{\triangle} = \frac{\sqrt[n]{\prod_{j\in W} p_{ij}}}{\sqrt[mn]{\prod_{i\in G}\prod_{j\in W} p_{ij}}}, \text{ for } i \in G,$$
$$\pi_j^{\triangle} = \frac{\sqrt[m]{\prod_{i\in G} p_{ij}}}{\sqrt[mn]{\prod_{i\in G}\prod_{j\in W} p_{ij}}}, \text{ for } j \in W. \tag{26}$$

(26) shows that in the case of a geometric mean we find explicit solutions for the prices ($\alpha_i^{\triangle}$) and the price index values ($\pi_j^{\triangle}$). As we may assume that each price $p_{ij} > 0$, the expressions for $\alpha_i^{\triangle}$ and $\pi_j^{\triangle}$ are indeed well-defined. Remarkably about (26) is that the expressions for $\alpha_i^{\triangle}$ and $\pi_j^{\triangle}$ do not depend on the quantities q_{ij}. Therefore (26) can be used in situations where the quantities q_{ij} (or, equivalently, the volumes v_{ij}) are unknown, such as in case of web scraped data. They can also be useful in case initial values are needed to start an iteration involving the computation of a GK (type) price index.

Instead of (26) it is more convenient to consider $\log \alpha_i^{\triangle}$ and $\log \pi_j^{\triangle}$ for computational purposes. Using (26) one can also derive an incremental definition of $\log \alpha_i^{\triangle}$ and $\log \pi_j^{\triangle}$. Another possibility is to use (26) to find α_i and π_j using arithmetic averages in the enumerators and denominators instead of the geometric ones that have been used. Both suggestions will not be elaborated as they are straightforward. A final remark is that to produce more robust variants of (26), one could trim extreme (log) prices first, or apply a method which dampens the influence of outliers on the estimation of $\log \alpha_i^{\triangle}$ and $\log \pi_j^{\triangle}$.

6.1.2 Harmonic mean

Let $x_1, \ldots, x_n > 0$ be n real numbers. The harmonic mean $\hat{x}_H$ of these numbers is defined as follows

$$\frac{1}{\hat{x}_H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}. \tag{27}$$

Applying (27) to (24) amounts to replacing arithmetical means in the denominators and enumerators by harmonic means, we find the resulting defining equations for the prices α_i^H and price index values π_j^H of the harmonic GK price index:

$$\alpha_i^H = \frac{\sum_{j\in W} 1/q_{ij}}{\sum_{j\in W} \pi_j^H/v_{ij}} = \frac{\sum_{j\in W} 1/q_{ij}}{\sum_{j\in W}(\pi_j^H/p_{ij})(1/q_{ij})}, \text{ for } i \in G,$$
$$\pi_j^H = \frac{\sum_{i\in G} 1/(\alpha_i^H q_{ij})}{\sum_{i\in G} 1/v_{ij}} = \frac{\sum_{i\in G}(p_{ij}/\alpha_i^H)(1/v_{ij})}{\sum_{i\in G} 1/v_{ij}}, \text{ for } j \in W. \tag{28}$$

Concentrating on the structure of (28) we can write

$$1/\alpha_i^H = \sum_{j\in W} \phi_{ij}\pi_j^H, \text{ for } i \in G,$$
$$\pi_j^H = \sum_{i\in G} \theta_{ij}/\alpha_i^H, \text{ for } j \in W. \tag{29}$$

If we put $\epsilon_i^H = 1/\alpha_i^H$ then (29) transforms into the following system of linear equations

$$\epsilon_i^H = \sum_{j\in W} \varphi_{ij}\pi_j^H, \text{ for } i \in G,$$
$$\pi_j^H = \sum_{i\in G} \vartheta_{ij}\epsilon_i^H, \text{ for } j \in W. \tag{30}$$

where

$$\phi_{ij} = \frac{1/v_{ij}}{\sum_{j\in W} 1/q_{ij}},$$
$$\vartheta_{ij} = \frac{1/q_{ij}}{\sum_{i\in G} 1/v_{ij}}. \tag{31}$$

It should be noted that (28) is somewhat restricted in its use: it can only be applied in case the quantities q_{ij} (or the volumes v_{ij}) are all > 0.

In case $q_{ij} = 0$ (or $v_{ij} = 0$) for some $(i,j) \in G \times W$, it is tempting to use ϕ'_{ij} and ϑ'_{ij}, satisfying

$$\phi'_{ij} = \frac{1/v_{ij}}{\sum_{j\in W'_i} 1/q_{ij}},$$
$$\vartheta'_{ij} = \frac{1/q_{ij}}{\sum_{i\in G'_j} 1/v_{ij}}, \tag{32}$$

where

$$W'_{i\cdot} = \{j \in W | q_{ij} > 0\} = \{j \in W | v_{ij} > 0\},$$
$$G'_{\cdot j} = \{i \in G | q_{ij} > 0\} = \{i \in G | v_{ij} > 0\}. \tag{33}$$

It is difficult to state if and when (32) has a solution for the set $\Xi = \{(i,j) \in G \times W \mid q_{ij} > 0\} = \{(i,j) \in G \times W \mid v_{ij} > 0\}$. We do not even try to speculate for which sets Ξ a solution exists. We feel that this problem is very similar to describing when 2D tables admit a solution for a rounding problem or for an IPF problem, if these tables contain 0 cells in their interior.

One can also apply (27) to (6), where the totals in the denominators and enumerators are thought to be replaced by arithmetical averages. Replacing these by corresponding harmonic averages we obtain

$$1/\alpha_i^h = \sum_{j\in W} \phi_{ij}/\beta_j^h, \text{ for } i \in G,$$
$$1/\beta_j^h = \sum_{i\in G} \chi_{ij}/\alpha_j^h, \text{ for } j \in W. \tag{34}$$

where

$$\phi_{ij} = \frac{p_{ij}/q_{ij}}{\sum_{j\in W} 1/q_{ij}},$$
$$\chi_{ij} = \frac{p_{ij}/q_{ij}}{\sum_{i\in G} 1/q_{ij}}. \tag{35}$$

Now if we define

$$\gamma_i^h = 1/\alpha_i^h,$$
$$\delta_j = 1/\beta_j. \tag{36}$$

we can write (34) as

$$\gamma_i^h = \sum_{j\in W} \phi_{ij}\delta_j^h, \text{ for } i \in G,$$
$$\delta_j^h = \sum_{i\in G} \chi_{ij}\gamma_i^h, \text{ for } j \in W, \tag{37}$$

6.1.3 Averaging operators in general

In the sections above we considered specific averaging operations associated with well-known means as arithmetic, geometric and harmonic. But there are many more examples. In the present section we abstract from the precise choice of such an operation. Instead we introduce an averaging operator 'av' and use this to define a general expression for a generalized GK index. We can apply 'av' to all definitions of the GK price index or its variants. Of course, there are so many options, we only consider a few examples. In particular it is interesting what happens when a robust averaging operator is used, or a discontinuous one like the mode.

We can generalize (24) to a price index that uses some averaging operator 'av':

$$\alpha_i^{av} = \frac{\mathrm{av}_{j\in W}\{v_{ij}/\pi_j^{av}\}}{\mathrm{av}_{j\in W}\{q_{ij}\}},\ \text{for } i \in G,$$
$$\pi_j^{av} = \frac{\mathrm{av}_{i\in G}\{v_{ij}\}}{\mathrm{av}_{i\in G}\{q_{ij}\alpha_i^{av}\}},\ \text{for } j \in W. \tag{38}$$

In fact it is even possible to use different averaging operators in the first and the second equation. It is even possible to choose four different averaging operators. But as there is no particular reason for doing so, we shall leave these options out of our discussion.

Instead of the original definition of the GK price index (1) we can also start with (6) and replace arithmetical averages by general ones. We then obtain

$$\alpha_i^{av} = \frac{\mathrm{av}_{j\in W}\{v_{ij}\beta_j^{av}\}}{\mathrm{av}_{j\in W}\{q_{ij}\}},\ \text{for } i \in G,$$
$$\beta_j^{av} = \frac{\mathrm{av}_{i\in G}\{q_{ij}\alpha_i^{av}\}}{\mathrm{av}_{i\in G}\{v_{ij}\}},\ \text{for } j \in W. \tag{39}$$

For ‘av’ we take the mode, or robust averages (that remove outliers, defined in a suitable way). The question for each choice of av is whether the system (39) has a solution and if so, if it is unique.

6.2 Tinkering with the defining equations

6.2.1 Symmetric variant of the GK price index

The following defining equations are derived from (1) by moving the π_j in the first set of equations from enumerator to denominator, resulting in:

$$\alpha_i^s = \frac{\sum_{j\in W} v_{ij}}{\sum_{j\in W} q_{ij}\pi_j^s},\ \text{for } i \in G,$$
$$\pi_j^s = \frac{\sum_{i\in G} v_{ij}}{\sum_{i\in G} q_{ij}\alpha_i^s},\ \text{for } j \in W. \tag{40}$$

where the parameters have been superscripted with an ‘s’ (from symmetric) to distinguish them from those appearing in the original sets of equations (1). Contrary to (1) the parameters in (40) play a symmetric role in the defining equations, which is appealing.

We can rewrite (40) as

$$1/\alpha_i^s = \sum_{j\in W} \rho_{ij}\pi_j^s,\ \text{for } i \in G,$$
$$1/\pi_j^s = \sum_{i\in G} s_{ij}\alpha_i^s,\ \text{for } j \in W. \tag{41}$$

where

$$\rho_{ij} = \frac{q_{ij}}{\sum_{j\in W} v_{ij}},$$
$$s_{ij} = \frac{q_{ij}}{\sum_{i\in G} v_{ij}}. \tag{42}$$

s_{ij} has been defined (see (9)) and used before.

6.2.2 Another symmetric variant of the GK price index

The second symmetric variant derived from the defining equations of the GK index (1) is obtained by moving the α_i in the second set of equations from denominator to enumerator, making sure that it plays a similar role as π_j in the first set of equations. We then obtain the following sets of defining equations:

$$\alpha_i^\sigma = \frac{\sum_{j\in W} v_{ij}/\pi_j^\sigma}{\sum_{j\in W} q_{ij}}, \text{ for } i \in G,$$
$$\pi_j^\sigma = \frac{\sum_{i\in G} v_{ij}/\alpha_i^\sigma}{\sum_{i\in G} q_{ij}}, \text{ for } j \in W. \tag{43}$$

where the parameters have been superscripted with 'σ' to distinguish them from their counterparts in (1) and (40).

If we put

$$\omega_{ij}^W = \frac{q_{ij}}{\sum_{j\in W} q_{ij}}$$
$$\omega_{ij}^G = \frac{q_{ij}}{\sum_{i\in G} q_{ij}} \tag{44}$$

we can rewrite (43) as

$$\alpha_i^\sigma = \sum_{j\in W} \frac{p_{ij}}{\pi_j^\sigma}\omega_{ij}^W, \text{ for } i \in G,$$
$$\pi_j^\sigma = \sum_{i\in G} \frac{p_{ij}}{\alpha_i^\sigma}\omega_{ij}^G, \text{ for } j \in W. \tag{45}$$

which shows clearly that the α_i^σ and the π_j^σ depend on the relative quantities ω_{ij}^W and ω_{ij}^G, $(i,j) \in G \times W$. (45) clearly shows how prices p_{ij} are expressed relative to month 1 (in case of π_j) and commodity 1 (in case of α_i). Further reducing the number of parameters, we can rewrite (45) as

$$\alpha_i^\sigma = \sum_{j \in W} \frac{r_{ij}}{\pi_j^\sigma}, \text{ for } i \in G,$$
$$\pi_j^\sigma = \sum_{i \in G} \frac{\sigma_{ij}}{\alpha_i^\sigma}, \text{ for } j \in W. \tag{46}$$

where

$$r_{ij} = p_{ij}\omega_{ij}^W,$$
$$\sigma_{ij} = p_{ij}\omega_{ij}^G. \tag{47}$$

The r_{ij} have also been defined before, namely in (9).

6.2.3 Simplified symmetric GK price index

Tinkering with (1) we can obtain the following set of defining equations:

$$\breve{\alpha}_i = \sum_{j \in W} \frac{p_{ij}}{\breve{\pi}_j}, \text{ for } i \in G,$$
$$\breve{\pi}_j = \sum_{i \in G} \frac{p_{ij}}{\breve{\alpha}_i}, \text{ for } j \in W. \tag{48}$$

One can compare (48) with (46) to see the kinship of the defining equations of these GK price index variants. One can view (48) as a kind of minimalistic variant of the GK price index, using the same ideas but with a minimum of data needed. Note that this price index depends only on the prices of the commodities in the time window. The continuous time variant of this price index is defined by the set of equations in (60).

6.2.4 Symmetric version of a harmonic GK price index

The same idea as used in Section 6.2.1 can be applied to the harmonic GK index (29). It results in the following set of equations

$$\alpha_i^\dagger = \frac{\sum_{j \in W} 1/q_{ij}}{\sum_{j \in W} \pi_j^\dagger / v_{ij}}, \text{ for } i \in G,$$
$$\pi_j^\dagger = \frac{\sum_{i \in G} 1/q_{ij}}{\sum_{i \in G} \alpha_i^\dagger / v_{ij}}, \text{ for } j \in W. \tag{49}$$

Of course, (49) can only be applied as long as $q_{ij} > 0$ (if and only if $v_{ij} > 0$, assuming that all prices $p_{ij} > 0$). We can rewrite (49) as

$$
\begin{aligned}
1/\alpha_i^\dagger &= \sum_{j\in W} \eta_{ij}\pi_j^\dagger, \text{ for } i \in G,\\
1/\pi_j^\dagger &= \sum_{i\in G} \theta_{ij}\alpha_i^\dagger, \text{ for } j \in W.
\end{aligned} \tag{50}
$$

where

$$
\begin{aligned}
\eta_{ij} &= \frac{1/v_{ij}}{\sum_{j\in W} 1/q_{ij}},\\
\theta_{ij} &= \frac{1/v_{ij}}{\sum_{i\in G} 1/q_{ij}}.
\end{aligned} \tag{51}
$$

Compare (50) with (41).

6.2.5 Another symmetric harmonic GK price index

When we apply the same idea as was used in Section 6.2.2 to obtain symmetric defining equations for the GK price index to the harmonic GK index (28) we obtain the following set of defining equations

$$
\begin{aligned}
\alpha_i^\ddagger &= \frac{\sum_{j\in W} 1/(\pi_j^\ddagger q_{ij})}{\sum_{j\in W} 1/v_{ij}}, \text{ for } i \in G,\\
\pi_j^\ddagger &= \frac{\sum_{i\in G} 1/(\alpha_i^\ddagger q_{ij})}{\sum_{i\in G} 1/v_{ij}}, \text{ for } j \in W.
\end{aligned} \tag{52}
$$

We can rewrite (52) in such a way that

$$
\begin{aligned}
\alpha_i^\ddagger &= \sum_{j\in W} \kappa_{ij}/\pi_j^\ddagger, \text{ for } i \in G,\\
\pi_j^\ddagger &= \sum_{i\in G} \lambda_{ij}/\alpha_i^\ddagger, \text{ for } j \in W.
\end{aligned} \tag{53}
$$

where

$$
\begin{aligned}
\kappa_{ij} &= \frac{1/q_{ij}}{\sum_{j\in W} 1/v_{ij}},\\
\lambda_{ij} &= \frac{1/q_{ij}}{\sum_{i\in G} 1/v_{ij}}.
\end{aligned} \tag{54}
$$

6.3 GK price index with continuous time parameter

So far we have only considered variants of the GK price index with a discrete time parameter, like the GK price index has. In the present section we want to consider a variant with a continuous time, inspired by the Divisia and Montgomery indices (see [1], Sections 6.2.1 and 6.2.2). The volumes, quantities and prices are supposed to be functions of a continuous time parameter. The set of commodities is still discrete.

The variant of (1) is then presented by the following set of defining equations

$$\begin{aligned} \alpha_i &= \frac{\int_W v_i(t)/\pi(t)\,dt}{\int_W q_i(t)\,dt}, \text{ for } i \in G, \\ \pi(t) &= \frac{\sum_{i\in G} v_i(t)}{\sum_{i\in G} \alpha_i q_i(t)}, \text{ for } t \in W, \end{aligned} \tag{55}$$

which can also be written as

$$\begin{aligned} \alpha_i &= \int_W \frac{p_i(t)}{\pi(t)} \xi_i(t)\,dt, \text{ for } i \in G, \\ \pi(t) &= \sum_{i\in G} \frac{p_i(t)}{\alpha_i} \lambda_i(t), \text{ for } t \in W, \end{aligned} \tag{56}$$

where

$$\begin{aligned} \xi_i(t) &= \frac{q_i(t)}{\int_W q_i(u)\,du}, \text{ for } i \in G, t \in W, \\ \lambda_i(t) &= \frac{\alpha_i q_i(t)}{\sum_{k\in G} \alpha_k q_k(t)}, \text{ for } i \in G, t \in W. \end{aligned} \tag{57}$$

We have that the ξ_i and λ_i are densities:

$$\begin{aligned} \xi_i(t) &\geq 0, \text{ for } i \in G, t \in W, \\ \lambda_i(t) &\geq 0, \text{ for } i \in G, t \in W, \end{aligned} \tag{58}$$

and

$$\begin{aligned} \int_W \xi_i(t)\,dt &= 1, \text{ for } i \in G, \\ \sum_{i\in G} \lambda_i(t) &= 1, \text{ for } t \in W. \end{aligned} \tag{59}$$

Note that ξ_i is only dependent on an observable q_i, whereas λ_i depends on observables (q_k for $k \in G$) as well as on computed quantity α_i, which in turn depends on computed quantity $\pi(t)$.

The following set of defining equations can be obtained from (48) by replacing the discrete time parameter by a continuous one:

$$\begin{aligned} \tilde{\alpha}_i &= \int_W \frac{p_i(t)}{\tilde{\pi}(t)}\, dt, \text{ for } i \in G, \\ \tilde{\pi}(t) &= \sum_{i \in G} \frac{p_i(t)}{\tilde{\alpha}_i}, \text{ for } t \in W, \end{aligned} \tag{60}$$

Obviously (60) can also be seen as a special case of (56). Note that (60) only depends on the price developments of the various commodities.

7 GK price index as a random variable

Usually price indices are presented as nonstochastic quantities. But in fact they are random variables, as they are based on samples of commodities. Typically only a selection of shops are drawn into the sample. Also the commodities sold by a shop are not necessarily completely observed, but only a subset. Besides observations may be limited to one or a few days in the base period, which we assume to be one month long. In case scannerdata from a shop are used the observation may be complete, although they typically are aggregated in order to reduce the size of the data sets.

Usually the data for CPI[11)] compilation are not collected using a neat sampling design and a sampling frame. So in order to simulate the randomness in the data due to their sampling we can use bootstrapping to simulate it (see [5]).

Another approach (see e.g. [10], Chapter 3) is based on modelling the commodity prices (and derived quantities) with parametric models, using least squares or maximum likelihood for parameter estimation.

7.1 Bootstrapping

The idea is to estimate β and hence π from (11), for matrices S and R generated for each bootstrap sample. Hence, for each bootstrap sample s we have to compute β^s by solving

$$\beta^s = ((S^s)' R^s)\, \beta^s, \tag{61}$$

11) Consumer Price Index.

where

$$R^s = (r_{ij}^s) = \left(\frac{v_{ij}^s}{\sum_{j\in W} q_{ij}^s}\right) = \left(\frac{v_{ij}^s}{q_{i\cdot}^s}\right),$$
$$S^s = (s_{ij}^s) = \left(\frac{q_{ij}^s}{\sum_{i\in G} v_{ij}^s}\right) = \left(\frac{q_{ij}^s}{v_{\cdot j}^s}\right), \tag{62}$$

with the superscripts 's' referring to bootstrap sample s. For details on the computations involved see Appendix B.

7.2 Directly generating random matrices R and S

Another way to generate random matrices R and S is more direct, in the sense that it is carried out at the aggregate level of these matrices. It does not start at the level of the commodities as in case of the bootstrap approach. Apart from the generation of these matrices, the bootstrap method and the direct method have in common that they both solve the fixed point problems (11) for α and β. The computational details of the direct method can be found in Appendix B, Section B.3.

7.3 Using time series for quantities and volumes

The approach in the present section is to consider the entries in the matrices $Q = (q_{ij})$ and $V = (v_{ij})$, row-wise, so that for each commodity i we have a time series $q_{i1}, \dots, q_{in}$ and $v_{i1}, \dots, v_{in}$. If the window W is long enough time series models can be used, for each i, to model the data. Then the fitted models can be used to generate new series, by adding random noise, also estimated from the data. We then obtain series $\hat{q}_{i1}, \dots, \hat{q}_{in}$ and $\hat{v}_{i1}, \dots, \hat{v}_{in}$. We can use these to compute $\hat{R} = (\hat{r}_{ij})$ and $\hat{S} = (\hat{s}_{ij})$ where

$$\hat{r}_{ij} = \frac{\hat{v}_{ij}}{\sum_{j\in W} \hat{q}_{ij}},$$
$$\hat{s}_{ij} = \frac{\hat{q}_{ij}}{\sum_{i\in G} \hat{v}_{ij}}. \tag{63}$$

We can repeat this procedure with 'new noise' used for both time series, for the quantities and for the volumes for each commodity i. Applying these newly generated time series yields a series of estimates for both R and S. They, in turn, yield a series of estimates for α and β, as described in Section 7.2.

If this is the case, one may consider the following procedure. Consider the time series for $v_{\cdot j}$ for $j \in W$. We may safely assume that there are no missing or zero values in this aggregated values series. So a time series model can be estimated without a problem. Because the values of this aggregated values series are nonnegative it is attractive to model $(\log v_{\cdot j})$ rather than $(v_{\cdot j})$ for $j \in W$.

Let such a generated aggregated value series be denoted as $\hat{v}_{\cdot j}$ We then use it to estimate a time series for each good i, by distributing its values proportional to the v_{ij}:

$$\hat{\hat{v}}_{ij} = \hat{v}_{\cdot j} \frac{v_{ij}}{v_{\cdot j}} = v_{ij} \frac{\hat{v}_{\cdot j}}{v_{\cdot j}}. \tag{64}$$

One might be tempted to define the following estimated quantities, using (64) and $v_{ij} = p_{ij} q_{ij}$:

$$\hat{\hat{q}}_{ij} = \frac{v_{ij}}{p_{ij}} \frac{\hat{v}_{\cdot j}}{v_{\cdot j}} = q_{ij} \frac{\hat{v}_{\cdot j}}{v_{\cdot j}}. \tag{65}$$

However, this choice leads to

$$\begin{aligned} \hat{\hat{r}}_{ij} &= \frac{\hat{\hat{v}}_{ij}}{\sum_{j \in W} \hat{\hat{q}}_{ij}}, \\ \hat{\hat{s}}_{ij} &= \frac{\hat{\hat{q}}_{ij}}{\sum_{i \in G} \hat{\hat{v}}_{ij}} = \frac{\hat{\hat{q}}_{ij}}{\hat{v}_{\cdot j}} = \frac{q_{ij}}{v_{\cdot j}} = s_{ij}, \end{aligned} \tag{66}$$

which is unacceptable, as $\hat{\hat{s}}_{ij} = s_{ij}$ is not randomized. So instead of (65) we should look for another randomized estimate of q_{ij}. It should be pointed out that it is not possible to consider $q_{\cdot j}$, as this may not even be well-defined: different commodities may have different quantity measures: number of units (such as bottles, boxes, packages), length, surface area, volume, weight, etc. Aggregating quantities based on different quantity measures is nonsense. Instead, we should consider $q_{i\cdot}$ for $i \in G$, as this is well-defined. Unfortunately, this is not a time series: commodities do not have a natural, linear ordering, in contrast to time, and in particular time periods such as calender months (forming a partition of a calender year).

A possibility is offered by the price matrix $P = (p_{ij})$, in particular by considering its rows. Then we have a price series per commodity i: $p_{i1}, \dots, p_{in}$. As in case of the values we can try to model them by time series, or rather their (natural) logarithms $\log p_{i\cdot}$. This yields estimates $\hat{p}_{i1}, \dots, \hat{p}_{in}$, where noise (estimated from the data) has been added, that is different for each new use of such a randomized price sequence.

We then have a new estimate of q_{ij}, instead of (65), namely

$$\hat{\hat{q}}_{ij} = \frac{v_{ij}}{\hat{p}_{ij}} \frac{\hat{v}_{\cdot j}}{v_{\cdot j}}. \tag{67}$$

Then it follows that

$$\hat{\hat{r}}_{ij} = \frac{\hat{\hat{v}}_{ij}}{\sum_{j\in W} \hat{\hat{q}}_{ij}},$$
$$\hat{\hat{s}}_{ij} = \frac{\hat{\hat{q}}_{ij}}{\sum_{i\in G} \hat{\hat{v}}_{ij}}, \tag{68}$$

We can go a step further and consider the factor $\hat{v}_{\cdot j}/v_{\cdot j}$ as a random variable. For simplicity we assume these variables to be independent, identically distributed (iid) random variables, spreading around 1. If we denote them by ξ_j for $j \in W$ and follow the approach that led to (64) and (65) then it follows that

$$\tilde{v}_{ij} = v_{ij}\,\xi_j,$$
$$\tilde{q}_{ij} = \frac{v_{ij}}{\hat{p}_{ij}}\,\xi_j. \tag{69}$$

Using (69) we define

$$\tilde{r}_{ij} = \frac{\tilde{v}_{ij}}{\sum_{j\in W} \tilde{q}_{ij}},$$
$$\tilde{s}_{ij} = \frac{\tilde{q}_{ij}}{\sum_{i\in G} \tilde{v}_{ij}} = \frac{\tilde{q}_{ij}}{v_{\cdot j}\,\xi_j} = \frac{v_{ij}}{\hat{p}_{ij}\,v_{\cdot j}}. \tag{70}$$

Of course, alternative choices for estimating volumes v_{ij} and quantities q_{ij} are possible. We can, for instance, follow the approach suggested in Appendix B.3 to find alternative estimates for the quantities q_{ij}. In Section B.4 in Appendix B more can be found about the subject of the present section, in particular concerning the use of time series models, with a computational bent.

8 Discussion

The GK price index is at the heart of the present paper. Conceptually this index is very interesting (at least to the present author). It was first presented (by Geary) in the context of PPP, comparing price levels in various countries with their own currencies. However, reinterpreting it as a price index is easy and straightforward: interpret countries as months (or more generally unit periods) in the observation period.

The ideas used in deriving this index is that one can consider the average price of commodities over time, taking inflation into account. And also to consider the total cost of the products sold in each month relative to the average costs of the commodities. This leads to two equations that involve two sets of parameters: average price of the commodities and the price index values due to inflation.

Initially it may seem that the system of equations involved can only be solved by using an iteration. But it turns out that by changing one set of parameters (the price indices) that two sets of linear equations emerge that can be combined such that fixed point equations emerge. Using Banach's contraction theorem conditions are provided when these equations have (unique) solutions. This theorem can also be used to find the conditions for which the iterations converge to (unique) solutions.

The price index values computed for the GK index are based on comparing price levels in various months in the observation period with the price level at the first month of this period. If one would like to derive price index values comparing arbitrary months within an observation period W it is easy to derive these assuming transitivity for the extended price index. We call this process transitive continuation, as it starts with a spanning ditree with index values associated with its arcs as an ansatz, which is then extended to each pair of months being compared. Transitive continuation can be compared to the cycle method, which corrects values (if necessary) so that the resulting index is transitive. Both methods use spanning trees to obtain their respective results. Transitivity is a very desirable property for price indices, if not an absolute must, for an intransitive price index shows drift (by definition), which is an undesirable phenomenon. Because the results of a comparison of two months may be dependent on the path in the PIDG chosen to connect them.

The GK price index is defined for a fixed time window and a fixed set of commodities. It is, however, straightforward to provide an incremental definition of the GK index. This is suitable in case the observation window is not fixed but dynamic, for instance with a new index value added every month. This is usually the case when compiling a CPI. It is also possible with an incremental price index to allow the set of commodities to be dynamic, even allowing it to change every month. Typically commodity updates are less frequent, mainly for practical reasons.

The modus operandi in CPI compilation is to build short, that is yearly, price index series where the set of commodities remains the same within a year. For the next year the commodity set is updated and a new price index series is started. If one is interested in a long price index series one has to apply chaining to paste the short, yearly series together. This chaining is a process that is independent of the choice of the price index that was used to compute the short series. We consider a few such chaining methods. The general approach is to use an overlap period for two consecutive short price index series, to make a 'smooth' transition from the first short series to the next. The idea underlying one of these methods is to take the values of both short series and weigh them, in such a way that the values of the more recent series gets a higher weight as time progresses. We have taken a rather simple and straightforward approach for the weights to be used. This is not based on the GK price index particularly. But it would be interesting to investigate if a long series can be found which locally resembles the yearly series as well as possible. And also, that it is such that it satisfies the set of defining equations for a GK price index well for the extended period.[12)] This problem is left as a challenge for future research.

We can take a formal look at the definition of the GK price index and its defining equations. We then see that a particular averaging method is used, namely arithmetical averaging. It is therefore tempting to see which indices emerge if we replace this averaging method by alternative ones, such as geometric or harmonic averaging. Another possibility to obtain variants

[12)] Of course, this has to be specified: what does it mean that a long price index series satisfies the defining equations of the price index 'well enough'.

of the GK price index is by tinkering with the defining equations, by moving parameters to a different part of a defining equation or by assuming that the time parameter is not discrete but continuous, as in case of the Divisia and Montgomery price indices. Both the averaging and tinkering methods lead to new sets of defining equations that sometimes have similar shapes, but with different parameter values. Particularly appealing (at the outset) seems the price indices where the α_i and the π_j play a symmetric role (i.c. (40) and (43)). We did not have the opportunity to investigate any of these new indices empirically, or investigate conditions for the existence of solutions. The comparison of the results from rivalling price indices can be revealing and enlightening in practice. So we would suggest exploring these alternative GK price indices in more depth, apply them to the same sets of real data and compare the results.

The final topic discussed in the present paper is to explore the GK price index viewed as a random variable. In fact, in most situations price index values are computed on the basis of a sample of price observations. Not all municipalities and not all shops in a municipality are included in the sample and often only a sample of the commodities they sell are observed. So it is interesting to investigate this 'random GK price index'. We suggest three possible methods: one is on the basis of bootstrap samples. Another is based on solving certain fixed point equations using random matrices. And a third method uses time series models for the time development of values, quantities and prices of commodities. In the present paper we do not delve very deeply into these methods, but only indicate how they could be used. No computational results are available. We offer this as another possible topic for future research. Of course, these 'experiments' could also include any of the alternative price indices that have been suggested in the present paper.

References

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

[2] G. Box & G. Jenkins (1970). *Time Series Analysis - Forecasting and Control*. Holden-Day.

[3] E. van Bracht & L. Willenborg (2020). *Towards a New Flash HICP*. Discussion Paper, Statistics Netherlands.

[4] J. Dugundji (1966).*Topology*. Allyn and Bacon.

[5] B. Efron & R. Tibshirani (1998). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

[6] R. Geary (1958). A Note on the Comparison of Exchange Rates and Purchasing Power between Countries. *J. of the Royal Stat. Soc., Series A*, Vol. 121, pp. 97-99.

[7] J. Hamilton (1994). *Time Series Analysis*. Princeton Universty Press.

[8] S. Khamis (1972). System of Index Numbers for National and International Purposes, *J. of the Royal Stat. Soc., Series A*, Vol. 135, pp. 96-121.

[9] Z. Mashreghi, D. Haziza & C. Léger (2016). A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys*, Vol. 10, pp. 1-52.

[10] E. Selvanathan & D. Prasada Rao (1994). *Index Numbers: A Stochastic Approach*. Macmillan.

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

[12] L. Willenborg (2020). Decomposition Price Indices. Discussion Paper, Statistics Netherlands.

Appendix
A Existential issues

In this appendix we are interested in finding sufficient conditions for the existence of solutions of the equations (11). As we want to solve these equations in practical applications, we also want to understand under which conditions (10) can be solved iteratively. It turns out that the concept of contraction plays a key role in this matter.

A.1 Banach's contraction theorem

We view (11) from the point of view of fixed point theory,[13)] The Banach contraction theorem (see below) can be applied to guarantee the existence of solutions to (11). For convenience of the reader we repeat this theorem in the present section, explain the (topological) concepts involved and link it to the context of the present paper.

First some preparations. Let (Y, d) be a metric space, where Y is a set and $d : Y \times Y \to \mathbb{R}$ a metric, and $T : Y \to Y$ be a map. A $x_0 \in Y$ is a fixed point for T if $T(x_0) = x_0$. A map $T : Y \to Y$ is d-contractive map if there is a $\theta < 1$ such that $d(Tx, Ty) \leq \theta\, d(x, y)$ for all $(x, y) \in Y \times Y$. A sequence (x_n) in Y is d-Cauchy if for all $\epsilon > 0$ there is an $N(\epsilon) \in \mathbb{N}$ such that for all $m, n \geq N(\epsilon)$ holds $d(x_m, x_n) < \epsilon$. In other words, it is the concept of a Cauchy sequence in calculus (in $\mathbb{R}$) transposed to a metric space. If every d-Cauchy sequence in Y converges the metric space (Y, d) is said to be complete. We now can state the following theorem, which is named after the Polish mathematician Stefan Banach.

Banach's contraction theorem If Y be d-complete and $T : Y \to Y$ is d-contractive then T is continuous and has exactly one fixed point. □

For a proof of this theorem see e.g. [4], pp. 305 ff.

We apply Banach's contraction theorem to find conditions for equations (11) to have solutions for α and β. Clearly, β is used to compute a GK price index, whereas α is more of a by-product. Solutions for β and α can be computed simultaneously or independently in different, but linked, iteration schemes. In separate subsections (Sections A.3 and A.4) these matters will be dealt with. But first, in Section A.2, we take a closer look at norms, both vector norms and matrix norms (derived from vector norms).

[13)] At first sight one seems to be dealing with eigenvalue problems. The challenge then is to find conditions that guarantee the existence of a solution to these equations. Conditions such as that an irreducible Markov matrix (or a multiple thereof) does possess an invariant measure, which can be viewed as an eigenvector of the transpose of a Markov matrix with eigenvalue 1. However, these particular properties do not apply to (11). in which case we should look for alternatives.

A.2 Norms

The Banach contraction theorem in the previous section is formulated for metric spaces. In the application we have in mind in the present paper we are dealing with Euclidean spaces $\mathbb{R}^k$, which have a rich structure. One can define inner products on these spaces, from which one can derive norms, and from these, in turn, metrics. In the present section we have a closer look at norms, which can be defined on $\mathbb{R}^k$ and for linear maps $\mathbb{R}^k \to \mathbb{R}^k$, for $k \in \mathbb{N}$.

A.2.1 Vector norms

In the applications of Banach's contraction theorem in the present paper we are dealing with Euclidean spaces $\mathbb{R}^k$ for various values of $k \in \mathbb{N}$. Such spaces are equipped with a natural inner product $\langle \cdot, \cdot \rangle_k$, with

$$\langle x, y \rangle_k = \sum_{i=1}^{k} x_i \, y_i, \tag{A.1}$$

from which a norm $|| \cdot ||_k$ can be derived, with

$$||x||_k = \sqrt{\langle x, x \rangle_k} = \sqrt{\sum_{i=1}^{k} x_i^2}, \tag{A.2}$$

and a metric $d_k(\cdot, \cdot)$, with

$$d_k(x, y) = ||x - y||_k^2 = \sum_{i=1}^{k} (x_i - y_i)^2, \tag{A.3}$$

for $x = (x_1, \dots, x_k)'$ and $y = (y_1, \dots, y_k)'$ column vectors in $\mathbb{R}^k$.

As (A.1), (A.2) and (A.3) show there is an close link between (natural) inner product, (natural) vector norm and metric in Euclidean spaces. Norm and metric are both derived from the (natural) inner product. This can also be viewed as a multiplication of a row vector $(x_1, \dots, x_k)$ and a column vector $(y_1, \dots, y_k)'$ (or vice versa):

$$\langle x, y \rangle_k = (x_1, \dots, x_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = (y_1, \dots, y_k) \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}, \tag{A.4}$$

which are also used in matrix products.

A.2.2 Matrix norms

This brings us to the concept of matrix norms, in particular those derived from vector norms. What sets matrix norms apart from vector norms is their behavior under products of matrices, including that of a matrix and a vector.

Let $A = (a_{ij})$ be a $k \times k$ matrix. A matrix norm analogous to a vector norm (A.2), is the Frobenius norm:

$$||A||_F = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} a_{i,j}^2}. \tag{A.5}$$

Another example of a matrix norm is the following supremum (sup) norm $||\cdot||_s$:

$$||A||_s = \sup_{x \neq 0} \frac{||Ax||}{||x||} = \sup_{\substack{x \\ ||x||=1}} ||Ax||, \tag{A.6}$$

where $||\cdot||$ is the vector norm defined in (A.2). In fact, the supremum norm $||\cdot||_s$ measures how much A can stretch vectors $x \in \mathbb{R}^k$.

A matrix norm has the following properties:

1. **non-negative**: $||A|| \geq 0$.
2. **definite**: $||A|| = 0 \Leftrightarrow A = 0$.
3. **absolutely homogeneous**: $||\alpha A|| = |\alpha|\,||A||$.
4. **subadditive**: $||A + B|| \leq ||A|| + ||B||$ (triangle inequality).
5. **submultiplicative**: $||AB|| \leq ||A||\,||B||$.

where A and B are matrices and x is a vector, all of the right orders, i.e. so that the multiplications shown are well-defined; $\alpha \in \mathbb{R}$ is a scalar.

In some textbooks submultiplicativity is not a requirement for a matrix norm, as any matrix norm can be rescaled to satisfy this property. Submultiplicativity is what sets matrix norms apart from vector norms, as all the other properties are shared by both types of norms.

Submultiplicativity implies as a special case: $||Ax|| \leq ||A||\,||x||$, if x is a vector. Note that here $||\cdot||$ denotes a matrix norm as well as a vector norm.

A.3 Conditions for the existence of β

For the matrix $S'R$ in (11) we have for $k, l \in W$:

$$(S'R)_{kl} = \sum_{i=1}^{m} \frac{v_{il} q_{ik}}{v_{\cdot k} q_{i\cdot}} = \sum_{i=1}^{m} \frac{p_{il}}{p_{ik}} \frac{q_{il}}{q_{i\cdot}} \frac{v_{ik}}{v_{\cdot k}}. \tag{A.7}$$

Evidently, $(S'R)_{kl} \geq 0$, for each entry (k, l), is concisely expressed as $S'R \geq 0$, which means that $S'R$ is a nonnegative matrix.

Consider the three factors in the sum of (A.7): p_{il}/p_{ik}, $q_{il}/q_{i\cdot}$ and $v_{ik}/v_{\cdot k}$, which are about the price ratio of commodity i in two months k, l in W, the relative quantity of commodity i in a month l in W and the relative volume of commodity i in month k in W. In particular, we are interested in (sharp) upperbounds for each of these factors, independently of each other. For commodity i define

$$\begin{aligned} f_i^p &= \max_{k,l \in W} \{p_{il}/p_{ik}\}, \\ f_i^q &= \max_{l \in W} \{q_{il}/q_{i\cdot}\}, \\ f_i^v &= \max_{k \in W} \{v_{ik}/v_{\cdot k}\}. \end{aligned} \tag{A.8}$$

Now define

$$\begin{aligned} f_*^p &= \max_{i \in G} \{f_i^p\}, \\ f_*^q &= \max_{i \in G} \{f_i^q\}, \\ f_*^v &= \max_{i \in G} \{f_i^v\}. \end{aligned} \tag{A.9}$$

Then it follows that for $k, l \in W$

$$(S'R)_{kl} \leq \sum_{i=1}^{m} f_*^p f_*^q f_*^v = m f_*^p f_*^q f_*^v = M_1, \tag{A.10}$$

for convenience.

Remark For f_*^q and f_*^v the folllowing inequalities apply:

$$\begin{aligned} 1/m \leq f_*^q \leq 1 \\ 1/m \leq f_*^v \leq 1 \end{aligned} \tag{A.11}$$

The lowerbounds hold, for otherwise it would follow that $\sum_{i\in G} f_i^q < 1$ and $\sum_{i\in G} f_i^v < 1$, which are obviously incorrect. □

We have

$$||S'R\,x|| \leq ||S'R||\,||x|| \leq \sqrt{mn}M_1\,||x|| \tag{A.12}$$

where we have used

$$||S'R|| \leq \sqrt{mn}\,M_1. \tag{A.13}$$

This follows from

$$\begin{aligned} ||S'R\,x|| &= \sqrt{\sum_{i=1}^{m}\langle a_i, x\rangle^2} \leq \sqrt{\sum_{i=1}^{m}||a_i||^2\,||x||^2} = \sqrt{\sum_{i=1}^{m}||a_i||^2}\,||x|| \\ &\leq \sqrt{mnM_1^2}\,||x|| = \sqrt{mn}M_1\,||x||, \end{aligned} \tag{A.14}$$

where a_i is the i-th row of $A = S'R$, for $i = 1, \ldots, m$, which is a vector of length n. The entries of A, and hence the components of the a_i, are nonnegative.

Suppose that

$$\sqrt{mn}M_1 < 1 \tag{A.15}$$

holds which implies that $S'R$ is a contraction. Then according to Banach's contraction theorem there exists a (single) fixed point β such that

$$S'R\,\beta = \beta. \tag{A.16}$$

A solution to (A.16) can be found by iteration. See Appendix B, Section B.1.

We can find another—sharper—bound of the entries of $S'R$. We can consider the factors of (A.7) in combination rather then separately and determine bounds. For commodity $i \in G$ we define

$$f_i = \max_{k,l\in W} \frac{q_{ik}v_{il}}{q_{i\cdot}v_{\cdot k}} = \frac{1}{q_{i\cdot}} \max_{k,l\in W} \frac{q_{ik}v_{il}}{v_{\cdot k}}, \tag{A.17}$$

and using (A.17) we define in turn

$$f_* = \max_{i \in G} f_i. \tag{A.18}$$

Then we have for $k, l \in W$

$$(S'R)_{kl} \leq \sum_{i=1}^{m} f_* = m f_* = M_2, \tag{A.19}$$

for convenience. Hence M_2 is an alternative, and indeed sharper, bound for the entries of $S'R$ than M_1 is. Consequently, if

$$\sqrt{mn} M_2 < 1 \tag{A.20}$$

then $S'R$ is a contraction.

Remark We can sharpen the upperbounds $\sqrt{mn}M_1$ and $\sqrt{mn}M_2$ by noting that the factor mn can in fact be replaced by ζ, which denotes the number of nonzero entries in $S'R$. This follows when refining the reasoning that led to the upperbounds just mentioned. So the sharpened upperbounds are $\sqrt{\zeta}M_1$ and $\sqrt{\zeta}M_2$, respectively. □

A.4 Conditions for the existence of α

Now we are dealing with the matrix RS' and the corresponding fixed point equation in (11) satisfied by α, that is $\alpha = RS'\alpha$. Of course, proving the existence of a fixed point α is similar to proving the existence of a fixed point β in Section A.3. Only the details differ.

For $k, l \in G$ we have:

$$(RS')_{kl} = \sum_{j=1}^{n} \frac{v_{kj} q_{lj}}{q_{k\cdot} v_{\cdot j}} = \sum_{j=1}^{n} \frac{p_{kj}}{p_{lj}} \frac{q_{kj}}{q_{k\cdot}} \frac{v_{lj}}{v_{\cdot j}}. \tag{A.21}$$

For month $j \in W$ define

$$\begin{aligned} h_j^p &= \max_{k,l \in G} \{p_{kj}/p_{lj}\}, \\ h_j^q &= \max_{k \in G} \{q_{kj}/q_{k\cdot}\}, \\ h_j^v &= \max_{l \in G} \{v_{lj}/v_{\cdot j}\}. \end{aligned} \tag{A.22}$$

Now define in turn

$$h_*^p = \max_{j \in W}\{h_j^p\},$$
$$h_*^q = \max_{j \in W}\{h_j^q\}, \tag{A.23}$$
$$h_*^v = \max_{j \in W}\{h_j^v\}.$$

Then for $k, l \in G$ the following inequality holds:

$$(RS')_{kl} \leq \sum_{j=1}^{n} h_*^p h_*^q h_*^v = n h_*^p h_*^q h_*^v = N_1, \tag{A.24}$$

for convenience.

We have

$$||RS'\, x|| \leq ||RS'||\, ||x|| \leq \sqrt{mn} N_1\, ||x||. \tag{A.25}$$

The derivation of (A.25) is similar to that of (A.14). If $\sqrt{mn} N_1 < 1$ then RS' is a contraction.

As in Section A.3, we can find another, sharper, bound of the entries of RS'. For month $j \in W$ we define

$$h_j = \max_{k,l \in G} \frac{v_{kj} q_{lj}}{q_{k\cdot} v_{\cdot j}} = \frac{1}{v_{\cdot j}} \max_{k,l \in G} \frac{v_{kj} q_{lj}}{q_{k\cdot}}. \tag{A.26}$$

Using (A.26) we define in turn:

$$h_* = \max_{j \in W} h_j. \tag{A.27}$$

From (A.21) we infer that for $k, l \in G$ holds:

$$(RS')_{kl} \leq \sum_{j=1}^{n} h_* = n h_* = N_2, \tag{A.28}$$

for convenience. Hence N_2 is an alternative, and indeed sharper, bound for the entries of RS' than N_1 is. Consequently, if $\sqrt{mn} N_2 < 1$ then RS' is a contraction.

Applying a remark at the end of Section A.3 we can sharpen the bounds $\sqrt{mn} N_1$ and $\sqrt{mn} N_2$ by using ζ as defined in that section, to $\sqrt{\zeta} N_1$ and $\sqrt{\zeta} N_2$, respectively.

A.5 Contractions and iterations

We now consider an iteration to compute α and β as given in (10). This scheme has great intuitive appeal. However, we want to stress that it only yields a (unique) result, independent of the initial value under certain conditions. These conditions involve certain linear maps to be contractions. The iteration we want to consider is defined as follows.

$$\begin{aligned} \alpha_i^{k+1} &= \sum_{j \in W} r_{ij} \beta_j^k, \text{ for } i \in G, k \in \mathbb{N}, \\ \beta_j^{k+1} &= \sum_{i \in G} s_{ij} \alpha_i^k, \text{ for } j \in W, k \in \mathbb{N}, \end{aligned} \tag{A.29}$$

where the coefficients r_{ij} and s_{ij} are as given by (9). As initial conditions we take

$$\begin{aligned} \alpha_i^0 &= 1, \text{ for } i \in G, \\ \beta_j^0 &= 1, \text{ for } j \in W. \end{aligned} \tag{A.30}$$

In matrix form (A.29) can be written as

$$\begin{aligned} \alpha^{k+1} &= R\, \beta^k \\ \beta^{k+1} &= S'\, \alpha^k, \end{aligned} \tag{A.31}$$

where $R = (r_{ij})$ and $S = (s_{ij})$ as defined in (10) and $k \in \mathbb{N}_0$. The initial conditions written in vector form are

$$\begin{aligned} \alpha^0 &= \iota_m, \\ \beta^0 &= \iota_n, \end{aligned} \tag{A.32}$$

where ι_m and ι_n are the all ones (column) vectors of length m and n, respectively.

We can even represent (A.31) more concisely as

$$\begin{pmatrix} \alpha^{k+1} \\ \beta^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & R \\ S' & 0 \end{pmatrix} \begin{pmatrix} \alpha^k \\ \beta^k \end{pmatrix}. \tag{A.33}$$

and the initial conditions (A.32) as

$$\begin{pmatrix} \alpha^0 \\ \beta^0 \end{pmatrix} = \begin{pmatrix} \iota_m \\ \iota_n \end{pmatrix} \tag{A.34}$$

If for the iteration defined by (A.33) and (A.34) the linear map defined by the matrix

$$\begin{pmatrix} 0 & R \\ S' & 0 \end{pmatrix} \tag{A.35}$$

is a contraction, then it follows that

$$||\begin{pmatrix} 0 & R \\ S' & 0 \end{pmatrix}|| = \sqrt{||R||^2 + ||S||^2} < 1. \tag{A.36}$$

(A.36) implies that the following inequalities hold:

$$\begin{aligned} &||R|| < 1, \\ &||S|| < 1. \end{aligned} \tag{A.37}$$

Reversely, if (A.37) holds, we have that

$$\begin{aligned} &||RS'|| \leq ||R||\,||S'|| = ||R||\,||S|| < 1, \\ &||S'R|| \leq ||S'||\,||R|| = ||R||\,||S|| < 1, \end{aligned} \tag{A.38}$$

so that RS' as well as $S'R$ are contractions.

We can in fact weaken condition (A.37) to

$$||R||\,||S|| < 1 \tag{A.39}$$

and still find that both RS' and $S'R$ are contractions, as (A.39) also implies (A.38).

We have not been able to prove (or disprove, by finding a counter-example) the following statement:

$$RS' \text{ is a contraction} \Leftrightarrow S'R \text{ is a contraction.} \tag{A.40}$$

or equivalently, whether the existence of α implies that of β and vice versa.

B Computational issues

In this appendix we consider several computational problems that one faces when applying the approach taken in the main text. We do not intend to give a detailed discussion of these problems. The aim is only to make some pertinent remarks and provide some useful pointers to the literature. In particular, we want to address the following computational issues: to compute the price index values by solving the defining equations (11) iteratively. This is considered in Section B.1 below. The next problem concerns the bootstrapping of the GK index, in particular how to compute bootstrap samples from 'base data'. They should reflect the uncertainty in the data due to the collecting of data, which is a somewhat implicit and informal form of sampling, without a sampling design and sampling frame, etc. The next problem we consider is how to sidestep bootstrapping by generating random matrices for R and S directly. They can be used to solve (11) and yield random solutions. There is one problem that could be considered in the present appendix as well due to its nature but that is not: how to compute a transitive continuation for a given spanning ditree. This problem, however, is considered in Appendix C, as it requires a bit more preparation.

In this paper various price index formulas are derived. See Appendix D for an overview. For each of these problems computational issues as discussed above arise. We will not deal with them here, as it would be rather repetitive. In the present appendix we consider only the case of the GK index itself. The remaining cases are left for future consideration.

B.1 Solving some fixed point problems iteratively

One way to compute the GK price index is by solving the system of equations (1) or the equivalent system (6). We consider the latter. A natural way of solving it is by the using the iteration (A.29), in case the linear map defined in (A.35) is a contraction, which in turn implies that both RS' and $S'R$ are contractions, as was shown in Section A.5 of Appendix A.

Instead of solving (A.29) for both α and β one can also solve the equations in (11) separately for α and β. For α we have the iteration

$$\alpha^{k+1} = RS'\,\alpha^k, \tag{B.1}$$

with initial value

$$\alpha^0 = \iota_m. \tag{B.2}$$

For β we have the iteration

$$\beta^{k+1} = S'R\,\beta^k. \tag{B.3}$$

with the initial value

$$\beta^0 = \iota_n. \tag{B.4}$$

As (A.38) shows, RS' and $S'R$ are also contractions. Therefore the iterations just presented converge to unique solutions, independent from the initial value chosen.

B.2 Bootstrapping the GK index

In this section we consider bootstrapping to 'generate' random versions of the matrices R and S. The idea is to draw a large numbers of samples from the base data (say scannerdata from various shops) and compute the matrices R and S from these base data for each sample. From these bootstrap matrices R and S we compute $S'R$ and RS' and solve the corresponding fixed point problems to compute estimates of β and α, respectively, as explained in Section 7.1. In the present section we explain the structure of base data and how bootstrapping these data can be applied.

B.2.1 Base data used for bootstrapping

We assume that the base data is a file consisting of records with the following variables

- RID: record id.
- YM: reference period (year, month).
- MID: municipality id.
- SID: shop id.
- CID: commodity id, to which corresponds a commodity description. We assume that the CID is in fact a COICOP[14] group with the greatest level of detail that is appropriate.
- v: the total value of the commodities sold in month YM, per commodity CID and shop SID combination in the sample.
- q: the total quantity of commodities sold in in month YM, per commodity CID and shop SID combination in the sample.
- p: unit price of the commodity in month YM, per commodity CID and shop SID combination in the sample. So we have only one price for a commodity sold by a shop in month YM. So if there is a price change of a commodity sold at a shop a representative price has to be picked for this commodity in this month. We have $p = v/q$, assuming $q > 0$ (and hence $v > 0$).[15]

The idea is that municipalities MID and shops SID are $1 : n$ linked (i.e. each municipality in the data has at least 1 shop), and that shops SID are $1 : n$ linked to commodities CID (i.e. each shop in the data has at least 1 commodity).

It should be pointed out that this file does not contain directly observed data, but it is a file derived from such files, with unnecessary details removed. In case of scannerdata the unit of

[14] Classification Of Individual COnsumption according to Purpose. A classification published by the United Nations Statistics Division. See https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf.

[15] The assumption is that each commodity sold in a shop has a price $p > 0$.

observed data are receipts. A receipt contains information about commodities that have been purchased together by a customer. Receipts are therefore also sources of interesting marketing information, in particular about commodities bought together by customers. However, we are only interested in the commodities bought in a shop in a given month, the prices paid and the quantities and volumes sold. So whereas the receipt information is truely basic information, the information that we are interested in is not that basic any more. It is obtained after rearrranging and aggregating the information on the receipts, with three purposes in mind. The first is to remove details from the data that are not necessary for our goals. The second is that removing unnessary details in the data enhances their safety. The third purpose is to reduce the size of the data, which is beneficial for computations and storage. Therefore individual transactions (symbolized by receipts) are aggregated to a higher level.[16)]

B.2.2 Bootstrapping using the base data

In Table B.1 an overview is presented of the various sampling steps that could be involved, each month YM, in the bootstrap case.[17)] It should be stressed that this is only an example to illustrate the method. Other choices are possible. The general idea is that the bootstrapping should mimmick the actual sampling procedure that was applied.

Step	Sample units	Key	Description
1	Municipalities	MID	Municipalities in the data file.
2	Shops	MID × SID	Shops in the data file, located in the municipalities sampled in step 1.
3	Commodities	MID × SID × CID	Commodities in the data fie, sold by the shops selected in step 2.

Table B.1 Units involved in the bootstrapping. See explanation of an example of such a procedure in the text.

We now briefly describe the bootstrapping scenario summarized in Table B.1. This scenario is applied to the records belonging to each period YM. It has to be repeated for every month YM as represented in the data file. First randomly choose the municipalities proportional to size, that is the number of inhabitants. Then, for each selected municipality randomly select shops, again proportional to size, which is measured by turnover in the previous year. For each sample we assume that the total number of shops is equal to that of the original data file.[18)] In the next, and final, step we draw for each shop in the sample, a sample from the commodities it has sold in month YM. This sample of commodities is obtained, let's say, by applying simple random sampling with replacement, and is independent of the sample drawn in any other shop. It should be noted that if a shop was drawn multiple times in the sample in the previous step, the bootstrap samples of its commodities are likely to be different, as they are obtained independently. This would also be the case for different shops.

Above it was assumed that each month shops are drawn into the sample, independently from previous months. However, it may be the case that certain shops (the bigger ones, for instance) are in the selected sample every month. For instance when we are dealing with paneldata. In this case there is no extra randomness due to the selection of these shops and this should also be

16) The drawback of this is that some uncertainty is removed in the bootstrapping, as one cannot draw a sample from the receipts associated with a shop. However, if data size is not an issue, one may choose to keep the information specified at the receipt level. Not to allow this possibility is a matter of convenience rather than of principle.

17) See Table 2, p.10 in [3] for the steps in the case of price taking in physical shops.

18) This need not be the case, however. A variable number of shops is also possible. This introduces an extra source of randomness.

left out of the bootstrap sampling. The same can be said about municipalities. Some may always in the sample, whereas others are not.

Finite population bootstrapping is the subdomain of bootstrapping that matters here. The interested reader is referred to [9] for inspiration and guidance in this area.

B.3 Generating random matrices R and S directly

We start with the matrices $Q = (q_{ij})$ and $V = (v_{ij})$ and $P = (p_{ij})$. The final matrix is strictly speaking redundant as it can be computed from the first two matrices, using $v_{ij} = p_{ij}q_{ij}$. Nevertheless it will be used in the methods we propose in the present section.

First we produce the following two matrices, on the assumption that $q_{i\cdot} > 0$ for $i \in G$ and $v_{\cdot j} > 0$ for $j \in W$:

$$M_q = (q_{ij}/q_{i\cdot}) \tag{B.5}$$
$$M_v' = (v_{ij}/v_{\cdot j}), \tag{B.6}$$

where M_q is an $m \times n$ and M_v is an $n \times m$ Markov matrix, as they have nonnegative entries and row sums equal to 1.[19)]

Now we use M_q and M_v to draw samples, row-wise, of size n, which is a number in the order of the number of commodities in the original sample. The sampling is done independently for each row of the matrix in question, and with replacement. Each row in these matrices is normalized to 1, that is, has row sum equal to 1. Multiplying the i-th row of M_q by $q_{i\cdot}$ and the j-th row of M_v by $v_{\cdot j}$ yields two matrices, $\tilde{M}_q$ and $\tilde{M}_v$ that can be viewed as empirical versions of M_q and M_v. We need to check that for the entries in M_q with $M_{q,ij} > 0$ the values $\tilde{v}_{ij}/\tilde{q}_{ij}$ do not deviate too much from p_{ij}. More formally, we should check that

$$|\tilde{v}_{ij}/\tilde{q}_{ij} - p_{ij}| < \rho, \tag{B.7}$$

for some $\rho > 0$ that has been selected as a threshold. If this condition is satisfied, we keep the values $\tilde{v}_{ij}$ and $\tilde{q}_{ij}$.

If condition (B.7) is violated, we apply the following randomized imputation procedure:

$$\begin{aligned} &\tilde{v}_{ij} \leftarrow p_{ij}\tilde{q}_{ij} \text{ with probability } 1/2, \\ &\tilde{q}_{ij} \leftarrow \tilde{v}_{ij}/p_{ij} \text{ with probability } 1/2, \end{aligned} \tag{B.8}$$

[19)] Note that (B.6) defines the transpose of M_v.

where '←' denotes imputation (or replacement): the value on the left of the arrow is replaced by the value to its right, if selected. The idea is of course that one of the options is chosen. After all entries have been processed in this way, and values have possibly been replaced by new values, the updated matrices Q, V and P are denoted as $\tilde{Q}$, $\tilde{V}$ and $\tilde{P}$, respectively. They are mutually consistent, in the sense that $\tilde{v}_{ij} = \tilde{p}_{ij}\tilde{q}_{ij}$ holds for all entries of $\tilde{Q}$, $\tilde{V}$ and $\tilde{P}$, for $i \in G$ and $j \in W$.

B.4 Time series models for values and prices

This section is an extension of Section 7.3. It mainly comments on certain issues arising in the application of time series to price index data. Our main inspiration is [3], which deals with an application of time series to a price index problem, i.c. a new kind of flash HICP.[20] What is new about this method is that it attempts to use historic data to predict the newest price index value. Traditionally, the HICP is computed using only (incomplete) current data and no historic data.[21]

The focus in [3] is on autoregressive (AR) models, as they are simple models. The idea was to see if these models work well enough. If not one can move on to consider more complex models. However, not only the standard type of AR models are used in [3] (with a maximum lag dictated by a criterion like AIC) as such models would fail[22] with seasonal data. In such a case it is profitable to use information of a year ago, that is, from the similar season in the previous year. [3] furthermore noted that, in a qualitative sense, different types of price developments exist: not only periodic price developments (reflecting a yearly, seasonal pattern) but also 'staircase-like' price developments, where prices are constant for some time and then jump to another level, where they stay for some time, until the next jump brings them to a new level.

AR models are suitable for randomization, as they can be generated from independent drawings from normal distributions $\mathcal{N}(\mu, \sigma^2)$ with expectation $\mu = 0$ and a variance σ^2 to be estimated from the data. See e.g. [7], Section 3.4. These drawings can be viewed as random shocks that generate the time series at hand. So it is straightforward to use these shocks again to generate new versions of the time series.

AR models are special cases of extended classes of models which go under the names of ARMA (= AutoRegressive Moving Average), ARIMA (= AutoRegressive Integrated Moving Average) models.[23] It seems inappropriate to consider more complicated time series models, such as state-space models or Kalman filters, for our application, prior to the exploration of the suitability of the much simpler AR models.[24]

If the commodities i are specified in great detail one can expect many entries q_{ij} in Q and v_{ij} in V to be equal to 0 or to be 'missing'. This may complicate the estimation process for the time series per commodity. This is unlikely to happen if one considers the aggregated volume series $(v_{\cdot 1}, \ldots, v_{\cdot n})$.

[20] This is a method to predict the price index value of a month (typically january of a new year) at a time when the new data are not yet completely available. HICP = Harmonized Index of Consumer Prices.

[21] A combination of both approaches, using historic data as well as current data, is another obvious choice, resulting in even better (or at least not worst) estimates. This extension has already been mentioned in [3], Section 5 (Discussion).

[22] In not being parsimonious enough

[23] For which [2] is the classic reference. For a more recent, and broader, coverage of time series analysis see [7].

[24] Also, the price index series we are dealing with, may be too short for such models. Too many parameters and too little data.

As prices, volumes and quantities are nonnegative, it is obvious to model the logarithms of these quantities instead of the quantities themselves. In case the errors one is dealing with are small enough, one can simply work with the price, volumes or quantities directly.

In case of prices (see (67)) 0 cells and missing data (no sales or no observed sales) do not present problems, in contrast to 0 volumes or missing volumes: commodities have a price for each $j \in W$ even if they are not sold. So in case of no (known) sales, we can try to impute missing prices. This can be done in various ways, for instance by using one of the following simple imputation models: repeat the last known price of a commodity. Or use a linear interpolation (or extrapolation) model, whichever is needed: in the middle or at the end of the period W. Or aggregate the commodities such that at the higher aggregation level prices are available.

C Transitive continuation

In this appendix a method, called transitive continuation (TC) is described that can be used to produce a transitive index from an intransitive one. In that sense it can be compared to the cycle method (CM), in purpose. The concept 'transitive continuation' is inspired by a similar one, namely 'analytic continuation', in complex function theory. This is a method to extend the domain of an analytic function.

The idea underlying TC is the following: Suppose we have an intransitive price index π. This means that for months $i, j, k \in W$, the time window, we have $\pi_{ij} \neq \pi_{ik}\pi_{kj}$. We now pick pairs of months $i, j \in V$, such that the set of edges $\{i, j\}$ that we obtain are the edges of a spanning tree T covering W, which means that every element in W is part of at least one edge of T. This means that for every pair of points k, l there is a path Γ connecting k and l.

The price index obtained when extending π for each pair i, j is denoted by π^*. In case $(i, j) \in DT$, where DT is a directed tree such that the underlying tree is T, we have $\{i, j\} \in T$, or equivalently $(i, j) \in T$ and $(j, i) \in T$.

For $i \in W$ we have $\pi^*_{ii} = \pi_{ii} = 1$, and for each pair $i, j \in W$ we define $\pi^*_{ji} = 1/\pi_{ij}$, assuming that π_{ij} is known, that is, has been computed from observed data.

Now let K_W be the complete graph on W. Consider the cycles of K_W and in particular the set of elementary cycles induced by DT. For each elementary cycle C there is exactly one arc e missing. Now orient the arcs in such a way that the reoriented elementary cycle $\vec{C}$ is a directed, closed path. Now we want the product of the price index values associated with arcs to equal to 1. This yields an equation from which that value for π_e can be computed: the reciprocal of the product of the price index value associated with the remaining arcs in $\vec{C}$.

$$\pi^*_e = 1/\Pi_{d\in\vec{C}'}\pi_d = \Pi_{d\in\vec{C}'}\pi_d^{-1} = \Pi_{d\in\vec{C}'}\pi_{d^{-1}}, \tag{C.1}$$

where $\vec{C}'$ is $\vec{C}$ without arc e and d^{-1} is the reoriented arc d. So because we used a spanning tree T (or rather a spanning ditree DT) the values for the missing arcs can be uniquely defined, so that the define a transitive price index π^* based on π. Of course, the result depends on the choice of DT. For another choice of directed tree we would have obtained a different continuation.

Three examples of spanning trees are now provided. The first one is typical for the Paasche price index (and hence the GK price index), where there is one base month and every other month is compared with this base month. See Figure C.1. The second one is the month-by-month comparison, where each month is compared to its immediate predecessor (if it exists). See Figure C.2. The third example is a kind of mix beteen the first two: it uses bases months that shift so that months that are too far apart are not directly compared. See Figure C.3.

The next two examples are about spanning trees for a two year period, which are linked together by using an extra arc, with a base month in the first year and a comparison month in the second year. They are shown in Figures C.4 and C.5. Figure C.4 pertains to the situation of two consecutive years where both months 1 and month 13 are base months, for year 1 and for year 2

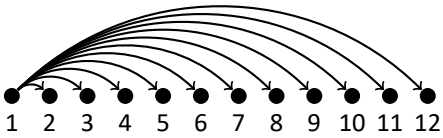

Figure C.1 Spanning tree for GK index for a one year period, comparing each month with the first month as the base month.

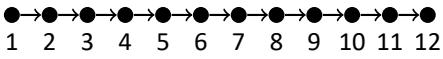

Figure C.2 Spanning tree for a one year period, based on a month-by-month comparison of successive months.

respectively. This is typical for the Paasche price index. Note that this 'concatenation' of yearly spanning trees yields a new spanning tree for the period of the combined years. The arc linking year 1 with year 2 is $(1, 13)$. In Figure C.5 we have a mix of two spanning trees: one with a single base month in year 1 and one a linear digraph corresponding to a month-by-month comparison in year 2. In this case the extra arc linking year 1 to year 2 is $(1, 13)$.

In the examples above it was the case that the spanning tree could be freely chosen. In practice it may be the case that the PIDG is given and a spanning ditree is a subtree of the PIDG. This means in particular that its arcs should be chosen from the arcs in the PIDG. Of course, the PIDG should have enough arcs so that is has a subditree.

One may wonder if such a spanning tree can be chosen in an optimal way, according to some sensible criterion. In fact this is possible if one realizes that it is best if the arcs only connect months that are closeby, so that the overlap of the sets of commodities associated with these months is substantial. So if $e = (i, j)$ is an arc, $|i - j|$ or $(i - j)^2$ could be used of measures of the closeness of the months i and j. We also denote this by $||e||$. This is an obvious choice, but other options to measure closeness of months exist, such as $||e||^2$, which is computationally more attractive. So, in words, we are looking for a subditree DT of a PIDG P, in such a way that the sum of the lengths of its arcs is minimal and such that DT is a spanning tree (so that each vertex should be represented in at least one arc). Formally, the goal is to solve the following optimization problem:

$$\min \{ \sum_{e \in DT} ||e|| \; | \text{DT} \subseteq \text{ P is a spanning tree} \}. \tag{C.2}$$

So this problem looks for a particular minimum spanning tree. Instead of $||e||$ in (C.2) one can, of course, use other arc lengths, such as $||e||^2$.

In case the PIDG is a complete digraph on V, the minimum spanning tree we are looking for is the month-by-month ditree.

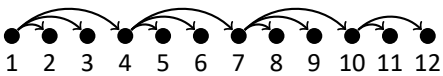

Figure C.3 Spanning tree for a one year period with shifting base months so that only months are compared that are not too far apart (at most 3 months).

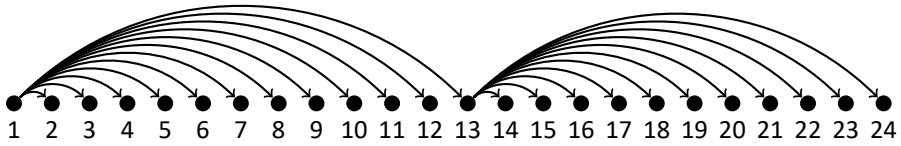

Figure C.4 Spanning tree for GK index for a two year period, with the first month of each year (months 1 and 13) as the base month for that year.

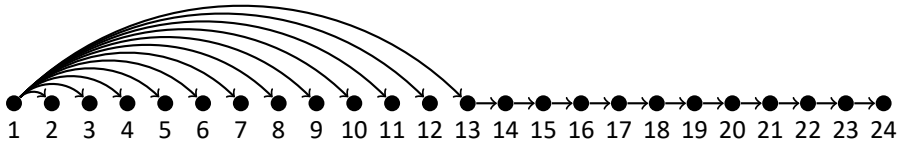

Figure C.5 Spanning tree for a two year period, with different spanning trees for year 1 and year 2.

We consider two examples of transitive continuation, one based on Figure C.4 and the other on Figure C.5. In both cases we want to find an expression for $\pi_{11,15}$. In the first case we find it to be equal to

$$\pi_{11,15}^{C.4} = \frac{\pi_{1,13}\pi_{13,15}}{\pi_{1,11}}, \tag{C.3}$$

and in the second case

$$\pi_{11,15}^{C.5} = \frac{\pi_{1,13}\pi_{13,14}\pi_{14,15}}{\pi_{1,11}}. \tag{C.4}$$

It is perhaps instructive to end this appendix with a comparison of TC and CM. Both methods are intended to produce a transitive price index from one that is either intransitive or one that is incomplete, in the sense that not all months are being compared. CM does this by looking at the given PIDG and PIG[25)], and in particular to its cycle space. It then aims at adjusting the price index values in such a way that the new ones are transitive, meaning that the product of the values associated with each arc (oriented in a suitable way) in each elementary cycle, is 1. The adjustment is done in such a way that the adjusted values are as close as possible to the original values, according to some quadratic criterion. Constrained Generalized Linear Regression is used to find the adjusted values.

TC considers a spanning tree that is a subtree T of the PIDG (under the assumption that such a spanning tree exists, which is the case if there are enough comparisons). Like CM, TC also considers elementary cycles of the PIDG. Each arc not on T can be used 'to complete' a cycle. For each elementary cycle, the missing price index value is in fact defined as the requirement is fulfilled that the product of the values associated with the arcs on each elementary cycle equals 1.

25) PIG = Price Index Graph, which is the underlying graph of a PIDG. This is the graph that is obtained by replacing each arc (a, b) of the Price Index Digraph (PIDG) by the edge $\{a, b\}$. See e.g. [11].

So CM obtains its aim by modifying observed price index values, whereas TC computes the missing values for certain arcs from conditions derived from a transitivity requirement.

Remark Given a intransitive PIDG G with cycles, one can also apply TC instead of CM to adjust it to a transitive PIDG, as follows. Draw a random sample of spanning trees from G. Complete each of these using TC. Then average the results for the completed trees. This yields a PIDG $\bar{G}$ that is transitive. This procedure has the property that no value associated with an arc of G that is not situated on a cycle, is left unchanged. □

D Systems of equations for various GK type price indices

We have collected all systems of equations for variants of the GK price index, with discrete time parameter, that have been proposed or derived in the main text in Table D.1, so that it is easy to compare them to each other. In particular the various coefficients involved are of interest. The coefficients for each variant define two matrices. For instance, $R = (r_{ij})$ and $S = (s_{ij})$ in case of variant 1, as specified in (10).

Variant	Equations	Formulas	Parameters
1	(8), (9)	$\alpha_i = \sum_{j\in W} r_{ij}\beta_j$	$r_{ij} = \frac{v_{ij}}{\sum_{j\in W} q_{ij}}$
–	–	$\beta_j = \sum_{i\in G} s_{ij}\alpha_i$	$s_{ij} = \frac{q_{ij}}{\sum_{i\in G} v_{ij}}$
2	(30), (31)	$\epsilon_i^H = \sum_{j\in W} \varphi_{ij}\pi_j^H$	$\varphi_{ij} = \frac{1/v_{ij}}{\sum_{j\in W} 1/q_{ij}}$
–	–	$\pi_j^H = \sum_{i\in G} \vartheta_{ij}\epsilon_i^H$	$\vartheta_{ij} = \frac{1/q_{ij}}{\sum_{i\in G} 1/v_{ij}}$
3	(37), (35)	$\gamma_i^h = \sum_{j\in W} \phi_{ij}\delta_j^h$	$\phi_{ij} = \frac{p_{ij}/q_{ij}}{\sum_{j\in W} 1/q_{ij}}$
–	–	$\delta_j^h = \sum_{i\in G} \chi_{ij}\gamma_i^h$	$\chi_{ij} = \frac{p_{ij}/q_{ij}}{\sum_{i\in G} 1/q_{ij}}$
4	(41), (42)	$1/\alpha_i^s = \sum_{j\in W} \rho_{ij}\pi_j^s$	$\rho_{ij} = \frac{q_{ij}}{\sum_{j\in W} v_{ij}}$
–	–	$1/\pi_j^s = \sum_{i\in G} s_{ij}\alpha_i^s$	$s_{ij} = \frac{q_{ij}}{\sum_{i\in G} v_{ij}}$
5	(46), (47)	$\alpha_i^\sigma = \sum_{j\in W} r_{ij}/\pi_j^\sigma$	$r_{ij} = \frac{v_{ij}}{\sum_{j\in W} q_{ij}}$
–	–	$\pi_j^\sigma = \sum_{i\in G} \sigma_{ij}/\alpha_i^\sigma$	$\sigma_{ij} = \frac{v_{ij}}{\sum_{i\in G} q_{ij}}$
6	(48)	$\breve{\alpha}_i = \sum_{j\in W} p_{ij}/\breve{\pi}_j$	–
–	–	$\breve{\pi}_j = \sum_{i\in G} p_{ij}/\breve{\alpha}_i$	–
7	(50), (51)	$1/\alpha_i^\dagger = \sum_{j\in W} \eta_{ij}\pi_j^\dagger$	$\eta_{ij} = \frac{1/v_{ij}}{\sum_{j\in W} 1/q_{ij}}$
–	–	$1/\pi_j^\dagger = \sum_{i\in G} \theta_{ij}\alpha_i^\dagger$	$\theta_{ij} = \frac{1/v_{ij}}{\sum_{i\in G} 1/q_{ij}}$
8	(53), (54)	$\alpha_i^\ddagger = \sum_{j\in W} \kappa_{ij}/\pi_j^\ddagger$	$\kappa_{ij} = \frac{1/q_{ij}}{\sum_{j\in W} 1/v_{ij}}$
–	–	$\pi_j^\ddagger = \sum_{i\in G} \lambda_{ij}/\alpha_i^\ddagger$	$\lambda_{ij} = \frac{1/q_{ij}}{\sum_{i\in G} 1/v_{ij}}$

Table D.1 Variants of the GK index as derived in Sections 2 and 6.

Note the rightmost column of Table D.1: the various parameters are functions of the v_{ij}, q_{ij} and p_{ij}. It is interesting to consider the corresponding matrices using real data to compute their entries.

Note also that variants 1, 2, 3 and 5 are similar in form, but with different parameters. Likewise the variants 4 and 6 are similar in form, as are the variants 5 and 7.

Note furthermore that if one of the matrices has entries that can be viewed as prices, the other has entries that can be viewed as reciprocals of prices.

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