Accounting for Clearance Sales in Scanner Data: Non-Hedonic Imputations in the GEKS-Törnqvist Price Index

Jan de Haan[a]

29 April 2022

Abstract: Clearance sales, where products are purchased in relatively large quantities at unusually low prices, can lead to downward bias in matched-model (maximum overlap) price indexes from scanner data. This paper proposes a simple non-hedonic imputation approach where the "missing prices" in the multilateral GEKS-Törnqvist price index are imputed using predicted values from another multilateral index method, weighted Time Product Dummy, to mitigate bias due to clearance sales.

Keywords: missing prices, multilateral index number methods, Törnqvist price index, transactions data.

JEL Classification: C43, E31.

[a] Statistics Netherlands and Delft University of Technology; jandehaan1@gmail.com.

The views expressed in this paper are those of the author and do not necessarily reflect the views of Statistics Netherlands.

1. Introduction

It has been more than ten years since Ivancic, Diewert and Fox (2009; 2011) proposed using multilateral methods for dealing with scanner data in the Consumer Price Index (CPI). Several countries, including Australia, Belgium, Luxembourg, Netherlands, New Zealand and Norway, have now implemented multilateral methods in CPI production. Most of them chose the GEKS-Törnqvist method, to be explained in detail later, where bilateral matched-model Törnqvist price indexes are inputs. Matching products across time ensures that like is compared with like.

Clearance sales are common in supermarkets and many other retailers. Clearance prices are typically unusually low and the corresponding quantities sold relatively large. Since prices do not return to regular levels after the fast decline, matched-model price indexes potentially suffer from downward bias (though the magnitude of the bias is not necessarily the same for all types of indexes). This is true for bilateral and multilateral price indexes.

There are different ways of dealing with the issue. For example, the Australian Bureau of Statistics removes transactions that are deemed clearance sales. I propose an alternative method where the "missing prices" for the unmatched new and disappearing products are imputed in the GEKS-Törnqvist index using predicted values from another multilateral index method. The basic idea is to impute prices at regular levels to mitigate downward bias.

Section 2 defines the imputation Törnqvist price index and shows, following De Haan and Krsinich (2014), how this index can be decomposed into the matched-model Törnqvist price index and components for the unmatched products. Section 3 describes the idea behind the (non-hedonic) imputation method – it is essentially a modification of carrying forward/backward observable prices and adjusting them for inflation. Section 4 then proposes using the regression-based multilateral weighted Time Product Dummy method as imputation method because this yields useful imputed prices in a simple way. Section 5 outlines GEKS-Törnqvist with the proposed imputations. Section 6 discusses a few situations where these imputations may not be appropriate. Section 7 summarizes and points to further work.[1]

[1] Frances Krsinich (Statistics New Zealand) and I are hoping to collaborate, revise the paper and add an empirical section using supermarket scanner data for New Zealand to illustrate the method and compare the results with those based on the Australian approach.

2. The imputation Törnqvist price index

Suppose we want to measure aggregate price change for a set of products U which is fixed across the sample period $0,...,T$. The prices of products $i \in U$ in the base period 0 and in comparison periods t $(0 < t \leq T)$ are denoted by p_i^0 and p_i^t, and the expenditure shares by s_i^0 and s_i^t. A useful measure of aggregate price change between 0 and t would be the bilateral Törnqvist price index

$$P_T^{0t} = \prod_{i \in U} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} . \tag{1}$$

In reality, the set of products is unlikely to be constant; substantial churn is often found in scanner data. Let U^0 and U^t denote the set of products purchased in periods 0 and t, respectively. The (matched) set of products purchased in both period 0 and period t is $U_M^{0t} = U^0 \cap U^t$. The (disappearing) subset of U^0 that is not purchased in period t is denoted by U_D^{0t}, and the (new) subset of U^t that is not purchased in period 0 is denoted by U_N^{0t}. So, we have $U_M^{0t} \cup U_D^{0t} = U^0$ and $U_M^{0t} \cup U_N^{0t} = U^t$.

Each product purchased in period 0 and/or period t should be included in a price comparison between periods 0 and t. That is, a proper bilateral price index is defined on the union $U^0 \cup U^t = U_M^{0t} \cup U_D^{0t} \cup U_N^{0t}$ of the product sets. However, period t prices for $i \in U_D^{0t}$ and period 0 prices for $i \in U_N^{0t}$ are "missing" and must be imputed. I denote the imputed values by $\hat{p}_i^t$ and $\hat{p}_i^0$. By definition, we have $s_i^t = 0$ for $i \in U_D^{0t}$ and $s_i^0 = 0$ for $i \in U_N^{0t}$, and the *imputation Törnqvist price index* is thus defined as[2]

$$P_{IT}^{0t} = \prod_{i \in U_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} \prod_{i \in U_D^{0t}} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in U_N^{0t}} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}} . \tag{2}$$

De Haan and Krsinich (2014) showed that this index can be decomposed as[3]

$$P_{IT}^{0t} = P_{MT}^{0t} \left[\frac{P_{DIGL}^{0t}}{P_{MGL}^{0t}} \right]^{\frac{s_{D(0t)}^0}{2}} \left[\frac{P_{NIGP}^{0t}}{P_{MGP}^{0t}} \right]^{\frac{s_{N(0t)}^t}{2}} , \tag{3}$$

[2] This has also been referred to as a *single* imputation (Törnqvist) price index. A double imputation index in addition replaces the observable period 0 and period t prices for the disappearing and new products, respectively, by estimated values; see De Haan (2004) and Hill and Melser (2008).

[3] Diewert, Fox and Schreyer (2018) derived the same decomposition in a slightly different way. De Haan (2002) derived a similar decomposition for the (single) imputation Fisher price index.

where $s^0_{D(0t)} = \sum_{i \in U^{0t}_D} s^0_i$ is the aggregate period 0 expenditure share for the disappearing products and $s^t_{N(0t)} = \sum_{i \in U^{0t}_N} s^t_i$ is the aggregate period *t* expenditure share for the new products. The price indexes in the bracketed terms of (3) are: the Törnqvist index for the matched products, $P^{0t}_{MT} = \prod_{i \in U^{0t}_M} (p^t_i / p^0_i)^{(s^0_{iM(0t)} + s^t_{iM(0t)})/2}$; the geometric Laspeyres for the matched products, $P^{0t}_{MGL} = \prod_{i \in U^{0t}_M} (p^t_i / p^0_i)^{s^0_{iM(0t)}}$; the geometric Paasche for the matched products, $P^{0t}_{MGP} = \prod_{i \in U^{0t}_M} (p^t_i / p^0_i)^{s^t_{iM(0t)}}$; the imputation geometric Laspeyres index for the disappearing products, $P^{0t}_{DIGL} = \prod_{i \in U^{0t}_D} (\hat{p}^t_i / p^0_i)^{s^0_{iD(0t)}}$; the imputation geometric Paasche price index for the new products, $P^{0t}_{NIGP} = \prod_{i \in U^{0t}_N} (p^t_i / \hat{p}^0_i)^{s^t_{iN(0t)}}$. Note that the expenditure shares in these price indexes are normalized so that $\sum_{i \in U^{0t}_M} s^0_{iM(0t)} = 1$, $\sum_{i \in U^{0t}_M} s^t_{iM(0t)} = 1$, $\sum_{i \in U^{0t}_D} s^0_{iD(0t)} = 1$, and $\sum_{i \in U^{0t}_N} s^t_{iN(0t)} = 1$.

But how should we estimate the imputed prices in the imputation price indexes P^{0t}_{DIGL} and P^{0t}_{NIGP}? If information on product characteristics is available to the statistical agency, hedonic regression can be used; for applications to consumer electronics goods (in the context of the multilateral GEKS-Törnqvist method), see De Haan and Krsinich (2014) and De Haan and Daalmans (2019). When characteristics information is lacking, or if the agency is reluctant to use hedonics, non-hedonic imputation methods might be an option, depending on the circumstances. Section 3 outlines a non-hedonic imputation method which aims to deal with clearance sales.

3. Clearance sales and non-hedonic imputations

Clearance sales are characterized by unusually low prices. Like any matched (maximum overlap) price index, the matched Törnqvist potentially suffers from downward bias due to clearance sales. This will also be the case for an imputation Törnqvist index where the clearance price is carried forward and adjusted for inflation. Inflation adjustment is required to prevent further downward bias when prices are generally increasing. I will propose a non-hedonic imputation method where inflation-adjusted *regular prices* serve as imputations. Admittedly, this is not a novel idea; several statistical offices have been using a similar approach for a long time, albeit in the traditional context without scanner data.[4]

[4] The Australian Bureau of Statistics is one of them. They are using scanner data from supermarkets in the CPI, but the price indexes are matched ones, with no imputations. As mentioned earlier, transactions that are deemed clearance sales are removed to prevent downward bias.

As before, I assume that prices and quantities from scanner data are available for the entire sample period $0,...,T$. Product *i* is purchased in periods $t = 0,...,t^*$ $(t^* < T)$ and has disappeared in period $t^* + 1$. The last observed price, $p_i^{t^*}$, is a clearance price. Now how can we calculate a regular price for this product that serves as an input for the imputation of the "missing prices" for $t^* + 1,...,T$? A straightforward option is to use the price in the period prior to the clearance sale, $p_i^{t^*-1}$, and multiply it by an appropriate price index going from $t^* - 1$ to t $(t = t^* + 1,...,T)$, say $P^{t^*-1,t}$, to obtain imputed values $\hat{p}_i^t = p_i^{t^*-1} P^{t^*-1,t}$. In this case, the last observable non-clearance price is carried forward and adjusted for inflation.

There are some issues. We must be able to identify clearance sales. This can be difficult, in particular if we want our method to work at scale. So it seems worthwhile having an imputation method that does not rely on the identification of clearance sales. Furthermore, there seems to be no reason to just pick a single price, $p_i^{t^*-1}$, to indicate a regular price if more observable prices for this product are available. But then we need to decide how to combine all observations into a regular price. There is also the issue of choice of price index to adjust for inflation.

The non-hedonic method I am proposing to impute the "missing prices" consists of the following three steps: *i*) all (or most) of the observable prices of the disappearing product are deflated to obtain estimates of the base period price; *ii*) an average of these estimated base period prices is taken; *iii*) the average value is adjusted for inflation. As will be seen, this method is inspired by a multilateral index number method.

Let S_i denote the set of time periods with observable prices for product *i* which are used to estimate a regular base period price. The weighted arithmetic average of the estimated base period prices is

$$\hat{p}_i^0(A) = \sum_{t \in S_i} w_i^t \left(\frac{p_i^t}{P^{0t}} \right), \tag{4}$$

with weights w_i^t $(\sum_{t \in S_i} w_i^t = 1)$. Of course, we could use equal weights, hence take the unweighted average. The geometric counterpart to (4) is

$$\hat{p}_i^0(G) = \prod_{t \in S_i} \left(\frac{p_i^t}{P^{0t}} \right)^{w_i^t}. \tag{5}$$

In equations (4) and (5) a generic deflator P^{0t} is used. In practice, we have to choose a particular index, and the choice in (4) and (5) need not be the same. It does seem useful though to use the same price index for deflation and for subsequent “price updating” in each of the variants. In that case the imputed period *t* price for a disappearing product *i* becomes

$$\hat{p}_i^t(A) = \sum_{\tau \in S_i} w_i^\tau \left(\frac{p_i^\tau}{P^{0\tau}} \right) P^{0t} = \sum_{\tau \in S_i} w_i^\tau p_i^\tau \left(\frac{P^{0t}}{P^{0\tau}} \right); \tag{6}$$

$$\hat{p}_i^t(G) = \prod_{\tau \in S_i} \left(\frac{p_i^\tau}{P^{0\tau}} \right)^{w_i^\tau} P^{0t} = \prod_{\tau \in S_i} \left(p_i^\tau \frac{P^{0t}}{P^{0\tau}} \right)^{w_i^\tau}. \tag{7}$$

A few things are worth noting. The set S_i includes the clearance sales period t^*. This means there may still be some downward bias left due to clearance sales but taking an average of the deflated prices across multiple periods reduces the bias, hopefully to a large extent. Notice further that $P^{0t} / P^{0\tau}$ in (6) and (7) is a measure of aggregate price change between period τ and period *t*. If we want this measure to be based on the same index number formula as P^{0t} and $P^{0\tau}$, which sounds like a good idea, then we require a *transitive* price index method. Finally, given that the Törnqvist price index is geometric, I am in favor of using (7) rather than (6).

The next section discusses how the multilateral, hence transitive, weighted Time Product Dummy method can be used to estimate the imputed prices $\hat{p}_i^t(G)$ given by (7) for the disappearing products. Estimation of the “missing” base period prices of the new products will also be discussed.

4. Borrowing from TPD

The Time Product Dummy (TPD) method is a regression-based multilateral method that produces transitive price indexes, i.e., price indexes which are insensitive to the choice of base period.[5] Suppose there are *N* different products across the sample period, most of which are unlikely to be purchased in each period. The price of product *i* in period *t* is now modelled as the product of a factor for time, say $\exp(\delta^t)$ and a product-specific factor, say $\exp(\gamma_i)$, thus

$$p_i^t = \exp(\delta^t)\exp(\gamma_i). \tag{8}$$

[5] The Time Product Dummy method adapts Summers’ (1973) Country Product Dummy (CPD) method to comparisons across time.

After taking logarithms of both sides of (8) and adding errors ε_i^t with mean 0, the following equation is obtained, which can be estimated by least squares regression on the pooled data of the entire sample period:

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t ; \qquad (9)$$

D_i^t is a dummy variable that has the value 1 if the observation relates to period *t* and 0 otherwise, and D_i is a dummy variable with the value 1 if the observation relates to product *i* and 0 otherwise. Since an intercept α is included, dummy variables for period 0 and product *N* are left out to identify the model. The least squares parameter estimates are $\hat{\alpha}$, $\hat{\delta}^t$ $(t = 1,...,T)$ and $\hat{\gamma}_i$ $(i = 1,...,N-1)$, with $\hat{\delta}^0 = 0$ and $\hat{\gamma}_N = 0$. The predicted prices in periods 0 and *t* equal $\hat{p}_i^0 = \exp(\hat{\alpha})\exp(\hat{\gamma}_i)$ and $\hat{p}_i^t = \exp(\hat{\alpha})\exp(\hat{\delta}^t)\exp(\hat{\gamma}_i)$. The TPD price index between periods 0 and *t* is given by $P_{TPD}^{0,t} = \exp(\hat{\delta}^t)$.

I assume that (9) is estimated by *Weighted Least Squares* (WLS) regression with expenditure shares s_i^t $(t = 0,...,T)$ as weights.[6] The dummy variable (i.e., fixed effects) specification of the model ensures that the weighted regression residuals sum to zero in each time period: $\sum_{i \in U^t} s_i^t \ln p_i^t = \sum_{i \in U^t} s_i^t \ln \hat{p}_i^t$. It is easy to verify, using $P_{TPD}^{0t} = \hat{p}_i^t / \hat{p}_i^0$, that the weighted TPD index can be written as

$$P_{TPD}^{0t} = \frac{\prod_{i \in U^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{s_i^t}}{\prod_{i \in U^0} \left(\frac{p_i^0}{\hat{p}_i^0} \right)^{s_i^0}} = \prod_{i \in U^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{s_i^t} \qquad (t = 0,...,T). \qquad (10)$$

The denominator of the first expression of (10) equals 1 because the period 0 weighted regression residuals sum to zero. The TPD index can be viewed as a geometric Paasche-type price index where all the base period prices are equal to the predicted values from the estimated TPD model.

An expression for the predicted base period prices from TPD can be obtained as follows. The dummy variable specification ensures that the weighted residuals for every product sum to zero across all of the time periods in which the product is purchased (the set S_i), i.e., $\sum_{t \in S_i} s_i^t \ln(p_i^t / \hat{p}_i^t) = 0$. Thus, we have

[6] As far as I know, weighting by expenditure share in a TPD regression was first proposed by Diewert (2004). Rao (2005) showed the equivalence of expenditure-share weighted CPD and his multilateral "Rao system".

$$\prod_{t\in S_i}\left(\frac{p_i^t}{\hat{p}_i^t}\right)^{s_i^t}=1. \tag{11}$$

Substituting $\hat{p}_i^t=\hat{p}_i^0 P_{TPD}^{0t}$ into (11) and rearranging gives

$$\prod_{t\in S_i}\left(\frac{p_i^t}{P_{TPD}^{0t}}\right)^{s_i^t}=\prod_{t\in S_i}(\hat{p}_i^0)^{s_i^t}. \tag{12}$$

Equation (12) implies that we can express the predicted base period price for product i as

$$\hat{p}_i^0=\prod_{t\in S_i}\left(\frac{p_i^t}{P_{TPD}^{0t}}\right)^{s_i^t/\sum_{t\in S_i}s_i^t} \qquad (i=1,\ldots,N). \tag{13}$$

Notice the similarity of equation (13) with (5); in (13), the TDH index is used to deflate the observable prices, and the (normalized) expenditure shares serve as weights. Since $\hat{p}_i^t=P_{TPD}^{0t}\hat{p}_i^0$, we find

$$\hat{p}_i^t=\prod_{\tau\in S_i}\left(\frac{p_i^\tau}{P_{TPD}^{0\tau}}\right)^{s_i^\tau/\sum_{\tau\in S_i}s_i^\tau} P_{TPD}^{0t} \qquad (i=1,\ldots,N; t=1,\ldots,T), \tag{14}$$

which is the TPD version of (7). I propose using these estimates as imputations for the "missing" period t prices of the disappearing products in the imputation Törnqvist price index, i.e. in the imputation geometric Laspeyres price index P_{DIGL}^{0t} defined in Section 2. Moreover, I also propose using the estimates (13) as imputations for the "missing" base period prices of the new products, i.e., in the imputation geometric Paasche price index P_{NIGP}^{0t}.

The TPD system of equations given by (10) and (13) can be solved iteratively, without running a regression. However, calculating the TPD index via the estimation of (9) with econometric software provides useful diagnostics such as measures of fit and standard errors. Also, modelling can help better understand the method. Importantly, we do not have to calculate the imputed prices using (13) and (14) as they can be computed from the TPD regression output: $\hat{p}_N^0=\exp(\hat{\alpha})$, $\hat{p}_i^0=\exp(\hat{\alpha})\exp(\hat{\gamma}_i)$ for $i=1,\ldots,N-1$, and $\hat{p}_i^t=\hat{p}_i^0\exp(\hat{\delta}^t)$ for $i=1,\ldots,N$ and $t=1,\ldots,T$.

Products purchased in a single period can be omitted from the TPD regression. Their observable prices lie on the regression surface, i.e., $\hat{p}_i^0=p_i^0$ and $\hat{p}_i^t=p_i^t$, and so they do not affect the results. This follows directly from equations (13) and (14).

5. GEKS-Törnqvist with TPD imputations

In scanner data, the number of matches in the data between the base period 0 and the comparison period *t* often diminishes quite rapidly over time. The bilateral imputation Törnqvist price index then becomes increasingly model-based. Also, it will ignore many matched products that may be available across the entire window $0,...,T$. An alternative approach is to calculate the period-on-period chained imputation Törnqvist price index. However, empirical research revealed that high-frequency chaining of matched-product price indexes, including superlative price indexes, can lead to significant drift (Feenstra and Shapiro, 2003; Ivancic, 2007). There are no reasons to believe that imputations for the "missing" prices will alleviate chain drift.

To deal with the chain drift problem, Ivancic, Diewert and Fox (2011) proposed using a multilateral method, in particular GEKS (Gini, 1931; Eltetö and Köves, 1964; Szulc, 1964). De Haan and Van der Grient (2011) followed up on their work and used bilateral matched-product Törnqvist price indexes rather than Fisher indexes as building blocks in the GEKS procedure.[7] I also focus on GEKS-Törnqvist but now using bilateral imputation Törqnvist price indexes, based on weighted TPD, instead of their matched counterparts.

Denoting the link period by l $(0 \leq l \leq T)$, the imputation GEKS-Törnqvist price index going from period 0 to period t $(t = 1,...,T)$ can be expressed as

$$P^{0t}_{IGEKS-T} = \prod_{l=0}^{T} \left[P^{0l}_{IT} P^{lt}_{IT} \right]^{\frac{1}{T+1}} . \tag{15}$$

Note that *l*, the "base period" in P^{lt}_{IT}, can be greater than *t*. Taking the mean across all possible link periods ensures transitivity. For $l = 0$ and $l = t$, we have $P^{00}_{IT} P^{0t}_{IT} = P^{0t}_{IT}$ and $P^{0t}_{IT} P^{tt}_{IT} = P^{0t}_{IT}$ – in GEKS, the direct comparison between 0 and *t* weights twice as much as each of the indirect comparisons.

Substituting (3) into (15) yields the following decomposition:

$$P^{0t}_{IGEKS-T} = P^{0t}_{MGEKS-T} \Omega^{0t} , \tag{16}$$

where $P^{0t}_{MGEKS-T} = \prod_{l=0}^{T} \left[P^{0l}_{T} P^{lt}_{T} \right]^{1/(T+1)}$ is the matched GEKS-Törnqvist price index and where Ω^{0t} is defined by

[7] The GEKS-Törnqvist index is also known as CCDI (Caves, Christensen and Diewert, 1982; Inklaar and Diewert, 2016) index.

$$\Omega^{0t} = \prod_{l=0}^{T} \left[\left[\frac{P_{DIGL}^{0l}}{P_{MGL}^{0l}} \right]^{\frac{s_{D(0l)}^{0}}{2}} \left[\frac{P_{NIGP}^{0l}}{P_{MGP}^{0l}} \right]^{\frac{s_{N(0l)}^{l}}{2}} \right]^{\frac{1}{T+1}} . \tag{17}$$

This imputation component has a rather complex structure. Of course, there is no need to estimate it independently if we want to know its value, because it can be calculated as $\Omega^{0t} = P_{IGEKS-T}^{0t} / P_{MGEKS-T}^{0t}$.

Estimation of the proposed imputation GEKS-Törnqvist price index is easy. For the window $0,...,T$ and N different products, we just need to complete the $(T+1)\times N$ matrix of observed prices with the TPD-based imputations and then calculate a GEKS-Törnqvist price index from the full matrix (and the corresponding matrix of expenditure shares).[8] Summers (1973) stressed that his CPD method was specifically designed to fill the holes in the prices matrix, called the “tableau”. Essentially, I am making use of this property.

Empirical work has indicated that the matched GEKS-Törnqvist price index (and other multilateral indexes) can be sensitive to the choice of window length (ABS, 2017). One of the causes seems to be clearance sales. Due to the imputations for the “missing prices”, imputation GEKS-Törnqvist indexes are likely to be less sensitive to clearance sales, hence to the choice of window length.

To include strongly seasonal products, the window should be at least 13 months long (or 5 quarters for a quarterly CPI). In multilateral indexes, past price movements affect measured recent price changes so that recent changes become less “characteristic” as the window length grows. To reduce the loss of characteristicity, the window should not be too long.

A drawback of multilateral methods is that when additional data is available and the indexes are estimated on the extended data set, previous estimates will be revised. Different methods have been suggested to deal with such revisions; for an overview, see Van Kints, De Haan and Webster (2019). Diewert and Fox (2022) and Fox, Levell and O’Connell (2022) recommended using a rolling window of 25-months (9-quarters) with a so-called mean splice for updating the matched GEKS-Törnqvist index. I recommend this strategy for imputation GEKS-Törnqvist too.

[8] Graham White’s R package IndexNumR includes matched GEKS-Törnqvist and weighted TPD price indexes and could be used. The package is available from CRAN and GitHub.

6. A few issues

It is worth emphasizing that the choice for applying TPD-based imputations was made to simplify implementation in CPI production, in particular when the statistical office is calculating TPD along GEKS-Törnqvist. It can be argued that in our context it would be more appropriate to use the matched GEKS-Törnqvist index for deflating the observable prices and adjusting the average base period price estimate for inflation. Instead of (14), we would then have

$$\hat{p}_i^t = \prod_{\tau \in S_i} \left(\frac{p_i^\tau}{P_{MGEKS-T}^{0\tau}} \right)^{s_i^\tau \big/ \sum_{\tau \in S_i} s_i^\tau} P_{MGEKS-T}^{0t} \qquad (i = 1, ..., N), \tag{18}$$

But in this case, it is not so obvious why we should weight by expenditure share. Taking the unweighted average of the deflated prices can perhaps be justified as this treats all time periods as equally important. Note also that, for measuring aggregate price change, expenditure-share weighted TPD is arguably better than unweighted TPD. Yet it will be interesting to examine how imputations from unweighted TPD compare to the proposed imputations.

A potential problem with scanner data is that barcodes may change even if the products stay the same from the consumers' perspective, for instance in case of slight changes in the type of packaging. Price changes that occur during such *relaunches* do not affect matched bilateral indexes when products are identified by barcode (Reinsdorf, 1999; De Haan, 2002), and this carries over to multilateral price indexes. The proposed non-hedonic imputation method cannot resolve the relaunch problem. Lamboray (2022) discussed a non-hedonic imputation approach (in a GEKS-Fisher context) where unit values across all the available products are used for imputing the "missing prices". This approach does take relaunches into account but at the risk of introducing unit value bias. Some statistical offices, notably those in Belgium and Australia (ABS, 2017), are using the Stock Keeping Unit (SKU) rather than the usually more detailed Global Trade Item Number (GTIN) to identify individual products. This is likely to alleviate the relaunch problem without introducing unit value bias because only very similar products tend to receive the same SKU.[9]

[9] The Netherlands use the multilateral Geary-Khamis method (Geary, 1958; Khamis, 1972) for scanner data. This method does not allow imputations. To account for relaunches, Chessa (2016) defined products at a much less detailed level, potentially leading to unit value bias.

There are other situations where the proposed imputation method is unsuitable. This method was developed for products sold in supermarkets and other retailers where quality improvements due to technological change are not very important. For consumer electronics goods and other high-tech products, *hedonic imputation methods* are needed, which explicitly adjust for quality change using information on product characteristics.[10] Hedonic imputations will also deal with the relaunch problem (De Haan and Daalmans, 2019).

Clearance sales and relaunches belong a broader phenomenon: life cycle pricing. Life cycle effects are common for consumer electronics goods and fashion goods like clothing. Not surprisingly, Greenlees and McClelland (2010) found that rolling-window GEKS did not reduce downward bias in the matched Törnqvist price index for apparel while hedonic regression did. Melser and Syed (2016) discussed a non-hedonic method to deal with life cycle effects. Their method requires estimating life cycle functions, and I am afraid this hampers implementation in CPI production.

7. Summary and future work

Clearance sales can lead to downward bias in matched price indexes, including GEKS-Törnqvist indexes. There are various ways of dealing with the issue. One way is to try and identify clearance sales and delete them. The identification of clearance sales may, however, be arbitrary. More broadly, deleting data which are essentially "correct" is not an ideal solution. This paper proposes an alternate method where the "missing prices" in the GEKS-Törnqvist price index are imputed using predicted values from the weighted Time Product Dummy.

Unfortunately, I have not yet been able to apply the proposed method to data. As mentioned earlier (in footnote 1), Frances Krisinich (Statistics New Zealand) and I are aiming to collaborate again and apply the method to New Zealand supermarket scanner data to examine how it performs and how the results compare with those found using a clearance sales filter.

[10] De Haan and Krsinich (2014) estimated rolling-window hedonic imputation GEKS-Törnqvist indexes for consumer electronics goods using New Zealand scanner data. The bilateral imputation Törnqvist price indexes that served as inputs in GEKS were found by running a specific type of weighted bilateral Time Dummy Hedonic regressions. Their ITRYGEKS method has been implemented in the New Zealand CPI (Statistics New Zealand, 2014).

References

Australian Bureau of Statistics (ABS) (2017), "An Implementation Plan to Maximise the Use of Transactions Data in the CPI", Information Paper 6401.0.60.004, ABS, Canberra, Australia.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity", *Econometrica* 50, 1393-1414.

Chessa, A.G. (2016), "A New Methodology for Processing Scanner Data in the Dutch CPI", *Eurona* 1, 49-69.

Diewert, W.E. (2004), "On the Stochastic Approach to Linking the Regions in the ICP", Discussion Paper 04-16, Department of Economics, The University of British Columbia, Vancouver, Canada.

Diewert, W.E. and K.J. Fox (2022), "Substitution Bias in Multilateral Methods for CPI Construction", *Journal of Business & Economic Statistics* 40, 355-369.

Diewert, W.E., K.J. Fox and P. Schreyer (2018), "The Digital Economy, New Products and Consumer Welfare", ESCoE Discussion Paper 2018-16, Economic Statistics Centre of Excellence, London, United Kingdom.

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

Feenstra, R.C. and M.D. Shapiro (2003), "High-Frequency Substitution and the Measurement of Price Indexes", pp. 123-150 in R.C. Feenstra and M.D. Shapiro (eds.), *Scanner Data and Price Indexes*. Chicago: University of Chicago Press.

Fox, K.J., P. Levell and M. O'Connell (2022), "Multilateral Index Number Methods for Consumer Price Statistics", ESCoE Discussion Paper 2022-08, Economic Statistics Centre of Excellence, London, United Kingdom.

Geary, R.C. (1958), "A Note on the Comparison of Exchange Rates and Purchasing Power between Countries", *Journal of the Royal Statistical Society A* 121, 97-99.

Gini C. (1931), "On the Circular Test of Index Numbers", *International Review of Statistics* 9, 3-25.

Greenlees, J. and R. McClelland (2010), "Superlative and Regression-Based Consumer Price Indexes for Apparel Using U.S. Scanner Data", Paper presented at the General Conference of the IARIW, 27 August 2010, St. Gallen, Switzerland.

de Haan, J. (2002), "Generalised Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index", *Journal of Official Statistics* 18, 61-85.

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

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

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

de Haan, J. and J. Daalmans (2019), "Scanner Data in the CPI: The Imputation CCDI Index Revisited", Paper presented at the 16th meeting of the Ottawa Group, 8-10 May 2019, Rio de Janeiro, Brazil.

Hill, R. and D. Melser (2008), "Hedonic Imputation and the Price Index Problem: An Application to Housing", *Economic Inquiry* 46, 593-609.

Inklaar, R. and W.E. Diewert (2016), "Measuring Industry Productivity and Cross Country Convergence", *Journal of Econometrics* 192, 426-433.

Ivancic, L. (2007), "Scanner Data and the Construction of Price Indices", PhD thesis, School of Economics, The University of New South Wales, Sydney.

Ivancic, L., W.E. Diewert and K.J. Fox (2009), "Scanner Data, Time Aggregation and the Construction of Price Indexes", Discussion Paper 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.

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

Khamis, S.H. (1972), "A New System of Index Numbers for National and International Purposes", *Journal of the Royal Statistical Society A* 135, 96-121.

van Kints, M., J. de Haan and M. Webster (2019), "Utilizing Big Data and Multilateral Index Methods to Produce the Australian CPI: Theory, Implementation and Empirical Results", *Statistical Journal of the IAOS* 35, 387-405.

Lamboray, C. (2022), "Matching, Grouping, Linking: What Impact Does Product Specification Have on a Fisher Price Index?", Paper presented at the 17th meeting of the Ottawa Group, 7-10 May 2022, Rome, Italy.

Melser, D. and I.A. Syed (2016), "Life Cycle Price Trends and Product Replacement: Implications for the Measurement of Inflation", *Review of Income and Wealth* 62, 509-533.

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

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

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

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

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