Price Measurement Using Scanner Data

Time-Product Dummy Versus Time Dummy Hedonic Indexes

de Haan, Jan; Hendriks, Rens; Scholz, Michael

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

PRICE MEASUREMENT USING SCANNER DATA: TIME-PRODUCT DUMMY VERSUS TIME DUMMY HEDONIC INDEXES

by Jan de Haan

Division of Corporate Services, IT and Methodology, Statistics Netherlands, and OTB, Faculty of Architecture and the Built Environment, Delft University of Technology

Rens Hendriks

Statistics for Development Division, Pacific Community (SPC)

and

Michael Scholz*

University of Graz

This paper compares two model-based multilateral price indexes: the time-product dummy (TPD) index and the time dummy hedonic (TDH) index, both estimated by expenditure-share weighted least squares regression. The TPD model can be viewed as the saturated version of the underlying TDH model, and we argue that the regression residuals are "distorted toward zero" due to overfitting. We decompose the ratio of the two indexes in terms of average regression residuals of the new and disappearing items. The decomposition aims to explain the conditions under which the TPD index suffers from quality-change bias or, more generally, lack-of-matching bias. An example using scanner data on packaged men's T-shirts illustrates our framework.

1. Introduction

The advent of scanner data, and other electronic "big data" such as web-scraped data, has increased the potential for accurate price measurement well beyond the traditional method of price collectors visiting outlets and collecting prices for a relatively small sample of products. Electronic data usually comprise all the products sold by a certain retailer, and in the case of scanner data, quantities sold, therefore product weights, are available. Detailed characteristics are sometimes readily available as well, and if they are not, they can often be extracted from websites.

The matched-models method for constructing price indexes is inappropriate for scanner data, especially in areas where there is a regular churn in models or where prices are changed only when new models are introduced in the market. For such product markets, the Consumer Price Index (CPI) Manual (ILO/IMF/OECD/UNECE/Eurostat/The World Bank, 2004) recommended the use of explicit quality adjustments using hedonic regression techniques that take advantage of the characteristics data. The CPI Manual, however, did not discuss the issue of drift that can occur in period-on-period chained weighted price indexes and the more recently proposed multilateral index number methods to remove the chain drift while maximizing the number of matches in the data. There is a range of multilateral methods; see Diewert and Fox (2017) and Chessa *et al.* (2017).

de Haan (2015) proposed the use of two related model-based multilateral price indexes for incorporating scanner data, both estimated by expenditure-share weighted least squares regression: the time dummy hedonic (TDH) index where information on item characteristics is available, and the time-product dummy (TPD) index when this information is lacking. The name TPD method was suggested by de Haan and Krsinich (2014) as it adapts Summers' (1973) multilateral country-product dummy (CPD) method for spatial comparisons to price comparisons across time. A potential problem with the TPD method is that the resulting price index will not be explicitly adjusted for quality change. The aim of the present paper is to examine what drives the difference between the two methods.

We build on work by Silver and Heravi (2005) and Krsinich (2016). Silver and Heravi (2005) compared TPD and TDH indexes but only in a period-on-period chained context, where the bilateral TPD index equals a matched-model index. Krsinich (2016) argued that the use of longitudinal price information makes the multilateral TPD index implicitly quality-adjusted; see also Aizcorbe *et al.* (2003). It is true that in many cases (though perhaps not in oligopolistic markets with strategic pricing aimed at particular market segments) price differences among coexisting items provide us with information about the value of quality differences. Nevertheless, as with any implicit quality-adjustment method, this does not necessarily imply proper treatment of new and disappearing items and therefore does not rule out the possibility of quality-change bias or, more generally, lack-of-matching bias.

The treatment of quality change is especially important where there is a substantial churn in items sold when new models are introduced and old ones disappear. A major reason for the difference between the TPD and TDH methods potentially arises from the likely shortfall in the implicit quality adjustments of the TPD method compared to the more robust explicit quality adjustments of the TDH method, which are based on quality characteristics. Also, the TPD method cannot deal with items that are new in the sample period (these items lie exactly on the regression surface and are "zeroed out"); the TDH method does account for these items. The magnitude of the difference between the two methods will depend on the degree of churn, the extent of the quality difference between the new and disappearing models, and the adequacy of the characteristics data used in the hedonic regression to capture price-determining quality differences. Against

this background, the paper's main contribution is to draw attention to a related problem with the TPD method: that of overfitting.

Krsinich (2016) pointed out that the TPD model can be viewed as, what we call, the saturated version of a hedonic model with only categorical characteristics: the TPD model implicitly includes all the first- and higher-order interactions along with the main effects, whereas a typical hedonic model would include only main effects. In our view, this means the TPD model has too many parameters, fits the outliers, and unduly raises R squared as compared with the true underlying hedonic model. That is, the TPD model suffers from overfitting and "distorts the regression residuals toward zero." Another way to describe the problem is that overfitting potentially leads to biased out-of-sample predictions. Because quality adjustment boils down to imputing the "missing price" of new and disappearing items, that is, to making out-of-sample predictions, the TPD index is susceptible to quality-change bias.

The remainder of the paper is structured as follows. Section 2 outlines a number of different expressions for the TDH and TPD indexes, derives a decomposition of the ratio of the two indexes in terms of the average regression residuals for unmatched new and disappearing items, explains in greater detail why the TPD model is likely to suffer from overfitting, and explores potential bias in the TPD index. Section 3 illustrates our framework using scanner data on packaged men's T-shirts sold by a major Dutch chain of department stores. Section 4 discusses our findings and concludes.

2. FORMULAS AND DECOMPOSITIONS

2.1. *Formulas for TDH and TPD Indexes*

The following notation will be used: p_i^0 and p_i^t denote the price of item i in the base period 0 and in comparison period t ($t = 1, \ldots, T$), respectively; s_i^0 and s_i^t are the item's expenditure shares. Let us consider the following log-linear hedonic regression model to be estimated on the pooled data of all periods 0, …, T:

$$(1) \qquad \ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t,$$

where z_{ik} denotes item i (quantity of) characteristic k and β_k the corresponding parameter; the β_k is constrained to be fixed across time. Note that z_{ik} does not depend on time t, that is, it is assumed that the characteristics of an item do not change over time, which is not a restrictive assumption for newly produced goods. The time dummy variable D_i^t has the value 1 if the observation pertains to period t ($t = 1, \ldots, T$) and 0 otherwise; it is assumed that the errors ε_i^t are independently distributed with zero mean. The estimated parameters are denoted by $\hat{\delta}^0$, $\hat{\delta}^t$ ($t = 1, \ldots, T$), and $\hat{\beta}_k$ ($k = 1, \ldots, K$).

Following Diewert's (2005) proposal, we assume that a weighted least squares (WLS) regression is run with the expenditure shares in each period serving as weights. The TDH index going from period 0 to period t, $P_{TDH}^{0t} = \exp(\hat{\delta}^t)$, can then be written as (de Haan and Krsinich, 2017)

(2) $\quad P^{0t}_{TDH} = \dfrac{\prod_{i \in S^t} \left(p^t_i \big/ \exp\left\{ \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right\} \right)^{s^t_i}}{\prod_{i \in S^0} \left(p^0_i \big/ \exp\left\{ \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right\} \right)^{s^0_i}} = \dfrac{\prod_{i \in S^t} \left(p^t_i \right)^{s^t_i}}{\prod_{i \in S^0} \left(p^0_i \right)^{s^0_i}} \exp\left\{ \sum_{k=1}^{K} \hat{\beta}_k (\bar{z}^0_k - \bar{z}^t_k) \right\},$

where S^0 and S^t denote the sets of items sold in periods 0 and t. Exponentiation is a nonlinear transformation, and so the time dummy price indexes are not unbiased; see Kennedy (1981) for a bias-correction term. However, if the number of observations is large enough, as is the case in the data set we use in the empirical section, the correction term will most likely be minimal and can be ignored. The first expression of equation (2) writes the index as the ratio of weighted geometric averages of estimated *quality-adjusted prices* $p^0_i \big/ \exp\left\{ \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right\}$ and $p^t_i \big/ \exp\left\{ \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right\}$. The second expression adjusts the ratio of weighted geometric average prices for the changes in the weighted average characteristics $\bar{z}^0_k = \sum_{i \in S^0} s^0_i z_{ik}$ and $\bar{z}^t_k = \sum_{i \in S^t} s^t_i z_{ik}$.

Next, suppose there are N different items sold in one or more periods across the whole sample period and consider the following TPD model for the pooled data:

(3) $\qquad\qquad \ln p^t_i = \alpha + \sum_{t=1}^{T} \delta^t D^t_i + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon^t_i,$

where D^t_i is the time dummy defined earlier. D_i is a dummy variable that has the value of 1 if the observation relates to item i and 0 otherwise; the dummy for an arbitrary item N is excluded to identify the model. The parameters γ_i are item *fixed effects* ($\gamma_N = 0$). The WLS TPD index between period 0 and period t, $P^{0t}_{TPD} = \exp(\hat{\delta}^t)$, with fixed effects estimates $\hat{\gamma}_i$ ($\hat{\gamma}_N = 0$), can be written as (de Haan and Hendriks, 2013)

(4) $\qquad\qquad P^{0t}_{TPD} = \dfrac{\prod_{i \in S^t} \left(p^t_i / \exp\left\{ \hat{\gamma}_i \right\} \right)^{s^t_i}}{\prod_{i \in S^0} \left(p^0_i / \exp\left\{ \hat{\gamma}_i \right\} \right)^{s^0_i}} = \dfrac{\prod_{i \in S^t} \left(p^t_i \right)^{s^t_i}}{\prod_{i \in S^0} \left(p^0_i \right)^{s^0_i}} \exp\left\{ \bar{\hat{\gamma}}^0 - \bar{\hat{\gamma}}^t \right\},$

where $\bar{\hat{\gamma}}^0 = \sum_{i \in S^0} s^0_i \hat{\gamma}_i$ and $\bar{\hat{\gamma}}^t = \sum_{i \in S^t} s^t_i \hat{\gamma}_i$.

The TPD model (3) can be seen as a special case of the TDH model (1) in which the hedonic price effects $\sum_{k=1}^{K} \beta_k z_{ik}$ are approximated by fixed effects γ_i, up to an additive scalar. Consider the following linear model for the relation between the fixed-effects estimates $\hat{\gamma}_i$ and the estimated hedonic effects $\sum_{k=1}^{K} \hat{\beta}_k z_{ik}$:

(5) $\qquad\qquad \hat{\gamma}_i = a^t + b^t \sum_{k=1}^{K} \hat{\beta}_k z_{ik} + e^t_i,$

where e^t_i is an error term with a zero mean. If the fixed-effects estimates $\hat{\gamma}_i$ are "good approximations" of the hedonic effects $\sum_{k=1}^{K} \hat{\beta}_k z_{ik}$, we would expect parameter a^t to be close to 0, parameter b^t to be close to 1, and the variance of the errors e^t_i to be small. Note that items with the same hedonic price effects will usually have different fixed effects. The coefficients obtained from expenditure-share weighted

regressions of model (5) separately for each time period t ($t = 0, \ldots, T$) are denoted by $\tilde{a}^t$ and $\tilde{b}^t$. Since the weighted residuals sum to zero, we have

$$(6) \qquad \bar{\gamma}^t = \tilde{a}^t + \tilde{b}^t \sum_{k=1}^{K} \hat{\beta}_k \bar{z}_k^t.$$

Substituting equation (6) for periods 0 and t ($t = 1, \ldots, T$) into equation (4), dividing the result by equation (2) and some rearranging yields a decomposition of the TPD to TDH ratio in terms of changes in the intercept estimates, changes in the slope coefficients, and changes in the weighted average characteristics:

$$(7) \qquad \frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \exp\left\{\tilde{a}^0 - \tilde{a}^t\right\} \exp\left\{(\tilde{b}^0 - \tilde{b}^t) \sum_{k=1}^{K} \hat{\beta}_k \bar{z}_k^0\right\} \exp\left\{(\tilde{b}^t - 1) \sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^0 - \bar{z}_k^t)\right\}.$$

When the intercept estimates are the same in periods 0 and t ($\tilde{a}^0 = \tilde{a}^t$) and the slope coefficients equal 1 ($\tilde{b}^0 = \tilde{b}^t = 1$), the TPD index will be equal to the TDH index. These conditions are unlikely to hold in practice. First, WLS regression can produce unstable coefficients if the errors in model (5) are homoscedastic. Second, and more importantly, systematic changes in the coefficients (or the average characteristics) may occur. For example, if $\tilde{a}^t$ increases over time and everything else remains the same, the first component of equation (7) becomes increasingly smaller than 1, causing downward bias in the TPD index relative to the TDH index.

2.2. *A Decomposition in Terms of Regression Residuals*

The prices predicted by any method are denoted by $\hat{p}_i^0$ and $\hat{p}_i^t$, where we implicitly condition on the item's characteristics because they do not change over time. Using the least-squares property that the weighted sum of the residuals $u_i^0 = \ln(p_i^0) - \ln(\hat{p}_i^0)$ and $u_i^t = \ln(p_i^t) - \ln(\hat{p}_i^t)$ from the (WLS) TDH and TPD regressions is equal to zero in each period, we have

$$(8) \qquad \prod_{i \in S^0} \left(\frac{\hat{p}_i^0}{p_i^0}\right)^{s_i^0} = 1;$$

$$(9) \qquad \prod_{i \in S^t} \left(\frac{\hat{p}_i^t}{p_i^t}\right)^{s_i^t} = 1.$$

Multiplying both sides of equation (8) by $\prod_{i \in S^0} (\hat{p}_i^t / \hat{p}_i^0)^{s_i^0}$ gives

$$(10) \qquad \prod_{i \in S^0} \left(\frac{\hat{p}_i^t}{p_i^0}\right)^{s_i^0} = \prod_{i \in S^0} \left(\frac{\hat{p}_i^t}{\hat{p}_i^0}\right)^{s_i^0} = P^{0t};$$

because $P^{0t} = \hat{p}_i^t / \hat{p}_i^0$ for all i. Multiplying both sides of equation (9) by $\prod_{i \in S^t} (\hat{p}_i^0 / \hat{p}_i^t)^{s_i^t}$

yields

(11)
$$\prod_{i \in S^t} \left(\frac{\hat{p}_i^0}{p_i^t} \right)^{s_i^t} = \prod_{i \in S^t} \left(\frac{\hat{p}_i^0}{\hat{p}_i^t} \right)^{s_i^t} = \frac{1}{P^{0t}}.$$

Thus, initial expressions for the TDH and TPD indexes are

(12)
$$P^{0t} = \prod_{i \in S^0} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{s_i^0} = \prod_{i \in S^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{s_i^t} = \prod_{i \in S^0} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in S^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}}.$$

The first expression of equation (12) is a geometric Laspeyres-type price index where the period t prices for all $i \in S^0$ are predicted values from the regression. The second expression is a geometric Paasche-type price index with predicted period 0 prices for all $i \in S^t$. The two indexes are constrained to be the same, therefore equal to their geometric mean given by the third expression of equation (12), which is a Törnqvist-type price index defined on a *dynamic universe*.

We now subdivide S^0 and S^t into matched and unmatched items: $S_M^{0t} = S^0 \cap S^t$ is the set of matched items between periods 0 and t; S_D^0 is the subset of S^0 consisting of disappearing items that are not sold in period t ($S_D^0 \cup S_M^{0t} = S^0$); S_N^t is the subset of S^t consisting of new items that were not yet sold in period 0 ($S_N^t \cup S_M^{0t} = S^t$). The last expression of equation (12) then becomes

(13)
$$P^{0t} = \prod_{i \in S_M^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} \prod_{i \in S_D^0} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in S_N^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}} \left(\frac{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_i^t}{p_i^t} \right)^{s_i^0}}{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_i^0}{p_i^0} \right)^{s_i^t}} \right)^{\frac{1}{2}}.$$

Equation (13) provides some underpinning for the use of expenditure-share weighted regression to estimate the TDH and TPD models. The first three terms on the right-hand side define a *single imputation* Törnqvist price index, P_{SIT}^{0t}, where the "missing prices," that is, the period t prices for $i \in S_D^0$ and the period 0 prices for $i \in S_N^t$ are imputed. Single imputation price indexes typically apply predicted values based on regressions for each time period separately, but P_{SIT}^{0t} is based on predicted values from a pooled regression. For a comparison of time dummy hedonic and hedonic imputation indexes, see Diewert *et al.* (2009) and de Haan (2010). P_{SIT}^{0t} is not transitive, therefore dependent on the choice of the base period; the fourth term of (13) turns P_{SIT}^{0t} into the transitive (TDH or TPD) index P^{0t}.

To gain more insight into what drives the difference between the TPD and TDH indexes, we decompose the ratio of the TPD and TDH indexes, estimated on the same data set, in terms of the average regression residuals for the new and disappearing items. From equation (8) it follows that

$$\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TPD)}^0}{p_i^0} \right)^{s_i^0} \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TPD)}^t} \right)^{s_i^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^t \right)^{s_i^0}$$

$$= \prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TDH)}^0}{p_i^0} \right)^{s_i^0} \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TDH)}^0}{\hat{p}_{i(TDH)}^t} \right)^{s_i^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^t \right)^{s_i^0},$$

which, by defining $s_{iD}^0 = s_i^0/s_D^0$, $s_{iM}^0 = s_i^0/s_M^0$, $s_D^0 = \sum_{i \in S_D^0} s_i^0$, and $s_M^0 = \sum_{i \in S_M^0} s_i^0$, and after some manipulation, yields

(14)
$$\frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \left[\frac{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TPD)}^0}{p_i^0} \right)^{s_{iD}^0}}{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TDH)}^0}{p_i^0} \right)^{s_{iD}^0}} \right]^{\frac{s_D^0}{s_M^0}} \frac{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^t \right)^{s_{iM}^0}}{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^t \right)^{s_{iM}^0}}.$$

Similarly, from equation (9) it follows that

$$\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TPD)}^t}{p_i^t} \right)^{s_i^t} \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^t}{\hat{p}_{i(TPD)}^0} \right)^{s_i^t} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^0 \right)^{s_i^t}$$

$$= \prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TDH)}^t}{p_i^t} \right)^{s_i^t} \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TDH)}^t}{\hat{p}_{i(TDH)}^0} \right)^{s_i^t} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^0 \right)^{s_i^t},$$

which, using $s_{iN}^t = s_i^t/\sum_{i \in S_N^t} s_i^t$, $s_{iM}^t = s_i^t/\sum_{i \in S_M^{0t}} s_i^t$, $s_N^t = \sum_{i \in S_N^t} s_i^t$, and $s_M^t = \sum_{i \in S_M^{0t}} s_i^t$, and again after some manipulation leads to

(15)
$$\frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \left[\frac{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TDH)}^t}{p_i^t} \right)^{s_{iN}^t}}{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TPD)}^t}{p_i^t} \right)^{s_{iN}^t}} \right]^{\frac{s_N^t}{s_M^t}} \frac{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^0 \right)^{s_{iM}^t}}{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^0 \right)^{s_{iM}^t}}.$$

Taking the geometric mean of (14) and (15) yields

(16)
$$\frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \left[\frac{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TPD)}^0}{p_i^0} \right)^{s_{iD}^0}}{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TDH)}^0}{p_i^0} \right)^{s_{iD}^0}} \right]^{\frac{s_D^0}{2s_M^0}} \left[\frac{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TDH)}^t}{p_i^t} \right)^{s_{iN}^t}}{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TPD)}^t}{p_i^t} \right)^{s_{iN}^t}} \right]^{\frac{s_N^t}{2s_M^t}}$$

$$\left[\frac{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^t \right)^{s_{iM}^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^0 \right)^{s_{iM}^t}}{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^t \right)^{s_{iM}^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^0 \right)^{s_{iM}^t}} \right]^{\frac{1}{2}}.$$

The third term of decomposition (16) can be written as

$$\left(\frac{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^t \right)^{s_{iM}^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^0 \right)^{s_{iM}^t}}{\prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TDH)}^t \right)^{s_{iM}^0} \prod_{i \in S_M^{0t}} \left(\hat{p}_{i(TPD)}^0 \right)^{s_{iM}^t}} \right)^{\frac{1}{2}} = \left(\frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} \right)^{\frac{1}{2}} \left(\frac{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^0}}{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^t}} \right)^{\frac{1}{2}}.$$

Substituting this result into equation (16) and solving for $P_{TPD}^{0t}/P_{TDH}^{0t}$ gives

$$(17) \quad \frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \left(\frac{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TPD)}^0}{p_i^0} \right)^{s_{iD}^0}}{\prod_{i \in S_D^0} \left(\frac{\hat{p}_{i(TDH)}^0}{p_i^0} \right)^{s_{iD}^0}} \right)^{\frac{s_D^0}{s_M^0}} \left(\frac{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TDH)}^t}{p_i^t} \right)^{s_{iN}^t}}{\prod_{i \in S_N^t} \left(\frac{\hat{p}_{i(TPD)}^t}{p_i^t} \right)^{s_{iN}^t}} \right)^{\frac{s_N^t}{s_M^t}} \frac{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^0}}{\prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^t}}.$$

Equation (17) can be written in terms of average regression residuals, as follows:

$$(18) \quad \frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \exp \left\{ \frac{s_D^0}{s_M^0} \left(\bar{u}_{D(TDH)}^0 - \bar{u}_{D(TPD)}^0 \right) \right\}$$

$$\exp \left\{ \frac{s_N^t}{s_M^t} \left(\bar{u}_{N(TPD)}^t - \bar{u}_{N(TDH)}^t \right) \right\} \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^0 - s_{iM}^t},$$

where $\bar{u}_{D(TPD)}^0 = \sum_{i \in S_D^0} s_{iD}^0 u_{i(TPD)}^0$, $\bar{u}_{D(TDH)}^0 = \sum_{i \in S_D^0} s_{iD}^0 u_{i(TDH)}^0$, $\bar{u}_{N(TPD)}^t = \sum_{i \in S_N^t} s_{iN}^t u_{i(TPD)}^t$ and $\bar{u}_{N(TDH)}^t = \sum_{i \in S_N^t} s_{iN}^t u_{i(TDH)}^t$. Equation (18) decomposes the ratio of the TPD index and the TDH index into three components. The first and second components are driven by the differences in the weighted average residuals for the disappearing and new items from the TPD and TDH regressions. The magnitude of these components also depends on the (relative) aggregate expenditure shares of the matched and unmatched items, s_M^0, s_M^t, $s_D^0 = 1 - s_M^0$ and $s_N^t = 1 - s_M^t$.

The third component of equation (18) can be written in terms of the period 0 residuals for the matched items, for example as

$$(19) \quad \prod_{i \in S_M^{0t}} \left(\frac{\hat{p}_{i(TPD)}^0}{\hat{p}_{i(TDH)}^0} \right)^{s_{iM}^0 - s_{iM}^t} = \exp \left\{ (\bar{u}_{M(TPD)}^{0(t)} - \bar{u}_{M(TPD)}^0) - (\bar{u}_{M(TDH)}^{0(t)} - \bar{u}_{M(TDH)}^0) \right\},$$

with $\bar{u}_{M(TPD)}^0 = \sum_{i \in S_M^{0t}} s_{iM}^0 u_{i(TPD)}^0$, $\bar{u}_{M(TPD)}^{0(t)} = \sum_{i \in S_M^{0t}} s_{iM}^t u_{i(TPD)}^0$, $\bar{u}_{M(TDH)}^0 = \sum_{i \in S_M^{0t}} s_{iM}^0 u_{i(TDH)}^0$ and $\bar{u}_{M(TDH)}^{0(t)} = \sum_{i \in S_M^{0t}} s_{iM}^t u_{i(TDH)}^0$. This term thus depends on the normalized expenditure shares of the matched items. It generally differs from 1, even without any new or disappearing items, because WLS time dummy results are model dependent. This third term may be large when the matched-items' expenditures shares in periods 0 and t differ significantly and will be equal to 1 in the unlikely event that the shares remain constant over time. Note that this term would also be equal to 1 if unweighted (OLS) regressions had been run instead of weighted regressions.

2.3. *A Priori Expectations*

With a dynamic universe, the TPD and TDH indexes are most likely to differ. Two issues are at stake: variability (variance) and systematic difference (bias).

Suppose the average residuals of the new and disappearing items in equation (18) fluctuate randomly around 0 across time, for both the TPD and TDH regressions, and the third component fluctuates around 1. While the two indexes may differ in each time period, they are expected to exhibit equal trends.

According to Krsinich (2016), the TPD index is implicitly quality-adjusted due to the use of longitudinal price information. But when quality change is important, there is no a priori reason to expect that the average residuals from the TPD regression for the new and disappearing items will be approximately equal to, or show the same trend as, those from the TDH regression. More specifically, the TPD residuals for the unmatched items tend to be "distorted toward zero" as compared with the TDH residuals. Next, we will explain why this is the case.

In practice, items can be identified by a finite number of observable attributes, the range of possible values being discrete rather than continuous. In other words, a set of categorical variables (some of which will be ordinal) for each attribute can describe the different items belonging to the product category. Suppose we cross-classify all the categorical variables and know to which cell each item belongs. Obviously, some cells can be empty in practice since not all combinations may be feasible to produce or sell. Suppose further that we specify a TDH model using additive dummy variables for the *main effects* and multiplicative dummies for all first and higher-order *interaction terms*. As was shown by Krsinich (2016), this fully interacted or, as we will call it, saturated TDH model is essentially equivalent to the TPD model.

A problem with the saturated model is that it includes many irrelevant variables. The inclusion of interaction terms—certainly higher-order terms—in a hedonic model is difficult to justify due to interpretation problems. In a typical hedonic model, one would only find main effects for categorical variables and perhaps some first-order interaction terms. Also, the TPD model implicitly includes all the variables that are incorporated into the key that identifies the items, including attributes that may not be important from the consumers' point of view. One such key, which is always available in scanner data sets received from retailers, is GTIN (Global Trade Item Number). GTIN is a unique and universal identifier, developed by GS1, to define trade items, that is, "products or services that are priced, ordered or invoiced at any point in the supply chain" (https://www.gs1.org/ standards/id-keys/gtin).

The implicit inclusion of irrelevant variables in the TPD model is likely to lead to *overfitting*, in particular as compared to a TDH model that includes only main effects for attributes that are deemed important from the consumers' perspective. That is, the TPD model fits the outliers, unduly raises R squared, and "distorts the residuals toward zero."

Note that items that are observed only once during the sample period lie on the regression surface so that their residuals are exactly equal to zero, but this is probably a minor problem. Econometrics textbooks tell us that the inclusion of irrelevant variables does not lead to bias, conditional on the sample data. If we want the exponentiated time dummy coefficient from a TPD or TDH regression to be a quality-adjusted price index, imputation of the "missing prices" for unmatched items is required; see equation (13). These imputations are *out-of-sample predictions* and can be biased.

While bias can also arise for the TDH model, the TPD model is likely to be more affected as overfitting makes out-of-sample prediction very problematic. Put differently, we expect significant differences in the imputed "missing prices" for unmatched items between the TPD and TDH methods. When the *average* imputations differ significantly, substantial differences between the TPD index and the TDH index can arise. As shown by equation (18), the ratio of the two indexes can be analyzed by comparing the average residuals for the unmatched new and disappearing items instead of the average imputed values.

Silver and Heravi (2005) argued that the difference in the average residuals for the matched and unmatched new and disappearing items is the driver of the difference between a hedonic index and the corresponding matched-model index. Using scanner data for consumer electronics products, they found generally negative average residuals (i.e., relatively low observed prices, given their characteristics) for old models, or disappearing items in our language, and positive average residuals (relatively high observed prices) for new models. They attributed this result to the prevailing *pricing strategies* of retailers and manufacturers: inventory cleaning, or dumping, for old models and price skimming for new models.

To see what can happen and to simplify matters somewhat, let us assume that the aggregate expenditure shares of the new and disappearing items in equation (18) are the same ($s_D^0 = s_N^t = s_{UM}^{0t}$) and the third component equals 1. Equation (18) then reduces to

$$(20) \qquad \frac{P_{TPD}^{0t}}{P_{TDH}^{0t}} = \exp\left\{ \frac{s_{UM}^{0t}}{1 - s_{UM}^{0t}} \left[\left(\bar{u}_{N(TPD)}^t - \bar{u}_{D(TPD)}^0 \right) - \left(\bar{u}_{N(TDH)}^t - \bar{u}_{D(TDH)}^0 \right) \right] \right\}.$$

Suppose that the above inventory cleaning and price skimming strategies apply and we have $\bar{u}_{N(TDH)}^t > 0$ and $\bar{u}_{D(TDH)}^0 < 0$, or $\bar{u}_{N(TDH)}^t - \bar{u}_{D(TDH)}^0 > 0$. Because the residuals from the TPD regression for the new and disappearing items are "distorted toward zero," we expect to find $\bar{u}_{N(TPD)}^t - \bar{u}_{D(TPD)}^0 < \bar{u}_{N(TDH)}^t - \bar{u}_{D(TDH)}^0$, therefore $P_{TPD}^{0t} < P_{TDH}^{0t}$. Thus, under these pricing strategies, the TPD index is most likely downward biased compared to the TDH index.

Dumping and price skimming are not confined to products where quality change due to technical progress is important, such as consumer electronics goods; such pricing strategies can be found for groceries as well (Melser and Syed, 2016). An extreme case, referred to by Chessa (2016) as *re-launching*, arises when items are replaced by "new" items that essentially represent the same goods, apart perhaps from minor differences in packaging, but with different GTINs. This phenomenon seems to occur regularly for many product categories in the Netherlands. The prices of the replacement items are often higher than those of the replaced items; apparently, the retailers/manufacturers have some degree of market power, and consumers are unable to substitute away from the replacements. If items are identified by GTIN, the TPD method and matched-model methods cannot pick up disguised price increases due to re-launches, in contrast to the TDH method. It has been known for a long time that the GTIN level can sometimes be too fine for index construction (Reinsdorf, 1999; de Haan, 2002). Retailers' internal product

codes or stock-keeping units are more stable and maybe a more suitable choice; see for example ABS (2016).

2.4. *Weighting or Not and Other Regression Issues*

When scanner data, or similar transactions data, are not available to the statistical agency, a weighting will not be possible. This reflects the traditional situation where only prices are collected for a sample of items (in a sample of outlets). Weighting will also not be possible when electronic data on prices and characteristics are extracted from websites. Statistical agencies are increasingly using web-scraped data in the compilation of the CPI because this is a cost-effective way to replace the traditional price collection while at the same time being able to increase the sample of items priced.

Our analyses can be readily modified to the unweighted case by running OLS instead of WLS regressions, which will produce unweighted TDH and TPD indexes. The modification is straightforward, and so there is no need to include it in the paper, but we do show the OLS results in Section 3 to compare them with the WLS results. Also, there are issues with web-scraped data, such as the best way to calculate average monthly prices from, for example, daily price observations, which have not been resolved and are beyond the scope of our paper.

WLS is typically used to correct for heteroscedasticity and to achieve more efficient estimates. A prerequisite is the knowledge of the underlying error structure or the possibility to estimate it consistently. For example, Adjibolosoo (1993) studies various hypothesized error structures and compares the estimation results in a Monte Carlo study. Note that for a homoscedastic error before the expenditure-share weighting, such a weighting will induce heteroscedasticity and inefficient estimation (Solon *et al*., 2015). In the empirical part, we applied a Breusch-Pagan test that rejected the null hypothesis of homoscedasticity for all OLS/WLS TDH/TPD regressions. We also estimated the error structure as proposed by Harvey (1976) and corrected for heteroscedasticity. The results show that the TPD index is much more affected by this exercise than the TDH index. While the first exhibits substantial drift, the second retains the overall trend and shows deviations from the uncorrected counterpart only in special periods.

We do not include a broader discussion on heteroscedasticity in this paper for the following reasons: (1) the estimated coefficients in the TDH and TPD regressions are still consistent under heteroscedasticity, and we are in a rather large sample context; (2) our contribution aims to understand the difference between methods used in practice by statistical agencies. These methods are addressed in Section 2.2 without a correction for heteroscedasticity. Nevertheless, the issues of heteroscedasticity and also of regression diagnostics to handle potential outliers are of importance and are a topic for future research (some of our results are available in the online supplement).

Diewert's (2005) WLS procedure, which is used in this paper, has been criticized as observations are also weighted by the influence they have, that is, observations with high weight and influence may receive "too much weight than merited" (Silver and Heravi, 2005, Appendix 10; Silver, 2018). However, we do not recommend (automatic) deletion of influential observations.

The true underlying hedonic model is unknown, of course. A careful selection of characteristics and limited use of interaction terms is required, but some arbitrariness cannot be avoided. This is a disadvantage as compared with the TPD method. Another potential problem is parameter fixity in the TDH and TPD models. Regularly updating the coefficients seems required.

Like all multilateral methods, TDH and TPD have a revisions problem: when data for the next period, in our analysis period $T+1$, is added to the sample and price indexes are estimated from the extended sample, previously estimated index numbers will change. The literature offers several solutions for a nonrevisable CPI constructed in real time. One solution is to use a rolling-window approach in which the latest price movement, in our case between period T and period $T+1$, is spliced onto the price index level for period T. While in Section 3 we use a fixed sample period without updating, in the Appendix, we present TDH and TPD indexes using a rolling-window extension approach.

3. Empirical Illustration

For an empirical illustration, we use scanner data on packaged men's T-shirts. The data run from February 2009 to March 2013 and cover all the department stores belonging to a major Dutch retail chain. In addition to prices, that is, monthly unit values across all stores, and quantities sold, we have information on six categorical attributes that have been extracted from the available product descriptions: shape of neck (O or V), fabric (basic or organic), sleeve length (short or long); number of T-shirts per package (1, 2, or 3), color (white, black, or other), and fit (normal or stretch).

3.1. *EAN as Item Identifier*

In our data set, items are identified by barcode or European Article Number (EAN), the European version of GTIN. Across the 4-year sample period, 1953 different items were sold. The item turnover rate is high: from the more than 500 items that were sold in the first month, only 10% were still sold in the last month. The total number of items sold in each month is huge. Many packages with different EANs probably contain the same physical product or can be described by the same set of attributes so that the "true" rate of product churn may be overstated.

Figure 1 plots two indexes based on EAN as item identifier: the unit value index and the monthly chained Törnqvist price index. Both indexes have their problems. The unit value index is defined as the ratio of total expenditure divided by total quantities sold in the periods compared. It is affected by compositional change, giving rise to a volatile time series and possibly also a wrong trend. Weighted price indexes, including superlative price indexes such as the Törnqvist, are prone to chain drift when consumers stock up goods during sales periods (Ivancic *et al*., 2011; de Haan and van der Grient, 2011). The chained Törnqvist index does indeed have a downward drift, especially during the first half of the sample period.

Although multilateral price indexes are free from chain drift by construction, this does not mean that they are necessarily unbiased. Figure 2 shows the expenditure-share weighted TPD and TDH indexes. The TDH index shows a plausible
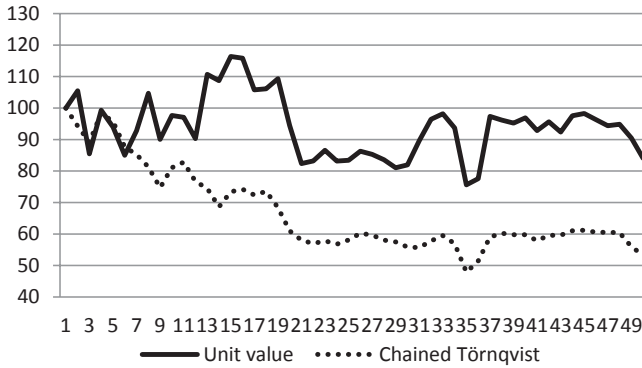
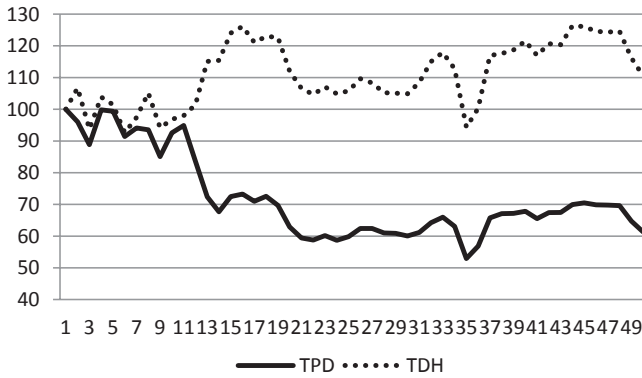Figure 1. Unit Value Index and Chained Törnqvist Index

Figure 2. TPD Index and TDH Index

trend, but the TPD index appears to be severely downward biased. The bias in the TPD index mainly arises in months 12 and 13 when organic T-shirts were introduced in the stores and largely replaced basic T-shirts. Surprisingly, the volatility of the TDH index is of the same order of magnitude as that of the unit value index, in spite of the fact that the TDH method controls for quality mix changes.

Table 1 contains the regression results for the weighted TDH model. Organic T-shirts are cheaper than basic T-shirts, other things equal. Organic T-shirts are made from materials grown in compliance with organic agricultural standards, but they need not be 100% organic to use the organic label. The negative coefficient is perhaps somewhat surprising; it suggests that consumers in this retail chain—who mostly have a low to middle income—were unwilling to pay a premium for organic T-shirts during our sample period. The signs of the other coefficients are as expected. Partly due to a large number of observations (24,797), all coefficients are highly significant, except for the attributes shape of the neck ("V") and fit ("stretch"). The R squared value from the TDH regression (0.7607) is satisfactory, but nevertheless much lower than that from the TPD regression (0.9108). This

TABLE 1
REGRESSION RESULTS FOR TDH MODEL

Attribute	Dummy	Estimate	Std. error	t-value	Signif.
Intercept		1.947536	0.009396	207.27	***
Shape of neck	V	0.006629	0.003154	2.10	*
Fabric	Organic	−0.237870	0.006555	−36.29	***
Sleeve length	Long	0.209994	0.004885	42.99	***
# T-shirts per package	2	0.489098	0.004065	120.31	***
	3	0.638112	0.005641	113.12	***
Color	White	−0.047670	0.003413	−13.97	***
	Black	−0.035280	0.005025	−7.02	***
Fit	Stretch	−0.021716	0.006636	−3.27	**

Note: Significance codes: ***0.001, **0.01, *0.05, # observations: 24,797; R squared: 0.7607; adjusted R squared: 0.7601; the results for the time dummy variables are not shown.
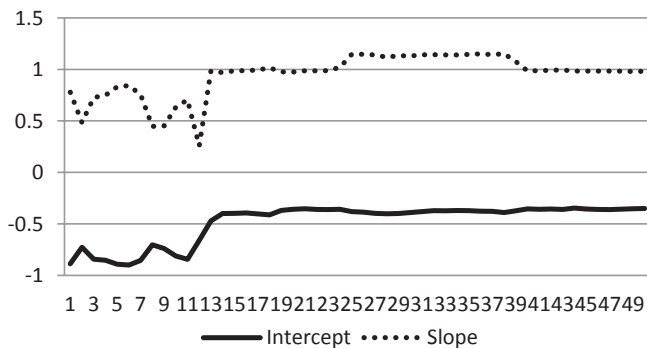

Figure 3. Coefficients from Regression Model (5)

confirms our suspicion that the TPD method unduly raises R squared as compared with the TDH method. Notice that our results provide evidence of highly nonlinear pricing. Other things equal, packages of two and three T-shirts are 1.6 and 1.9 times as expensive as a package of one T-shirt. On nonlinear pricing issues with applications to scanner data, see Fox and Melser (2014).

Decomposition (7) of the TPD index to TDH index ratio is based on a simple linear regression of the estimated TPD fixed effects against the estimated TDH hedonic price effects. Figure 3 plots the regression coefficients. As of month 14, the coefficients are quite stable, with the slope coefficient being close to the optimal value of 1. Before month 14, the regression coefficients are extremely volatile, and the TPD fixed effects are poor approximations of the hedonic price effects. The results of decomposition (7) in Figure 4 indicate that the change in the intercept term (the first component) mainly drives the change in the TPD to TDH ratio.

Decomposition (7), while instructive, says little about the causes of the change in the TPD to TDH ratio. This is where decomposition (18) comes into play. Here, the weighted average regression residuals of the new and disappearing items (with respect to the first or base month) are important drivers of the TPD to TDH ratio. As shown by Figure 5, the average TDH residuals tend to be positive for new items
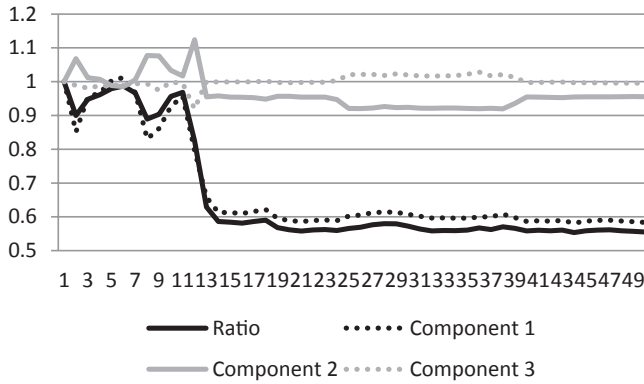
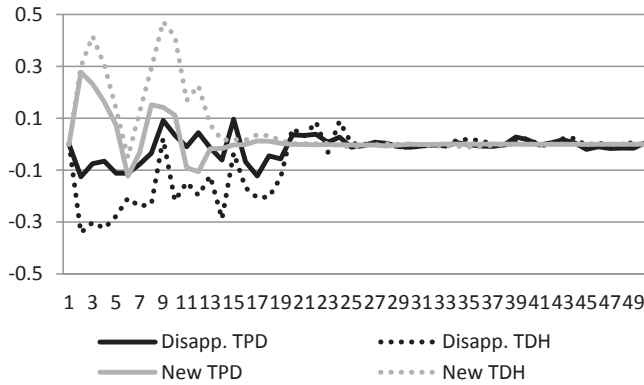Figure 4. Decomposition (7) of TPD to TDH Ratio

Figure 5. Weighted Average Regression Residuals

and negative for disappearing items during the first half of the sample period. The absolute values of the average TPD residuals are much smaller; in that sense, they are "distorted toward zero." These findings are consistent with an inventory cleaning and price skimming strategy, or re-launching of items, and with the TPD index sitting below the TDH index. In the second half of the sample period, the average residuals for the new and disappearing items from the two regressions turn out to be very small.

Figure 6 plots the aggregate expenditure shares of the unmatched and matched items. Two months witness dramatic changes with respect to the preceding months. In month 13, the expenditure share of new items rises from 0.08 to 0.84, and in month 25, the expenditure share of disappearing items rises from 0.09 to 0.62. The shares of the matched items in periods 0 and t drop accordingly. Note that the ratio of the period 0 aggregate expenditure shares for the disappearing and matched items and the ratio of the period t aggregate expenditure shares for the new and matched items cause *leveraging* in decomposition (18). The bigger these relative aggregate expenditure shares are, the more important the differences between the average residuals for the disappearing and new items from the TPD and TDH regressions become.
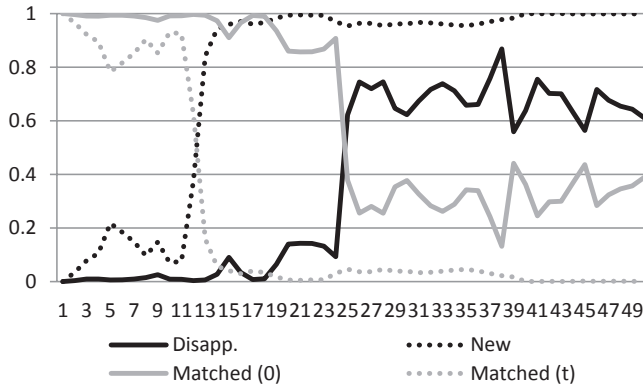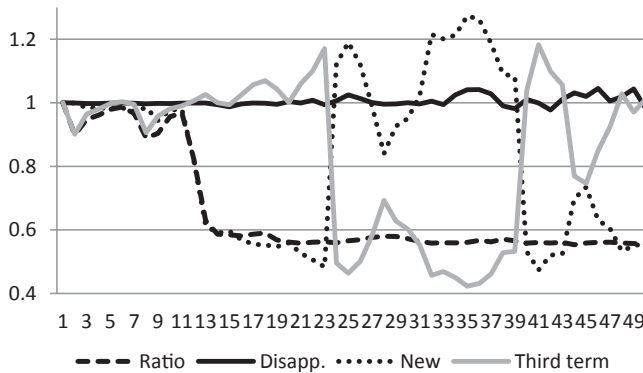
Figure 6. Aggregate Expenditure Shares

Figure 7. Decomposition (18) of TPD to TDH Ratio

Figure 7 shows the results of decomposition (18). As we already saw, the TPD to TDH index ratio remains quite stable after the introduction of organic T-shirts. This suggests that the TPD method performs well unless a structural break in the assortment takes place, in the sense of a sudden introduction of many new items with a significant expenditure share. New items contribute most to the TPD to TDH ratio, except between months 22 and 38 when the third term takes over. In that period, the average residuals of the new items, although small, are positive and have high leverage. Interestingly, the contribution of the disappearing items is negligible.

Our TPD and TDH indexes are weighted, but it is of some interest to compare them with their unweighted counterparts, obtained by running OLS rather than WLS regressions. As Figure 8 shows, the impact of weighting is significant. While the trend of the unweighted TPD index is similar to that of our weighted TPD index, the volatility of the unweighted version is much greater. For the TDH index too, the volatility of the unweighted version is much greater than that of the weighted one. Here, the trend is also affected, especially during the introduction of organic T-shirts. This example confirms how important weighting is for the construction of price indexes. It also shows why the use of scanner data, or similar
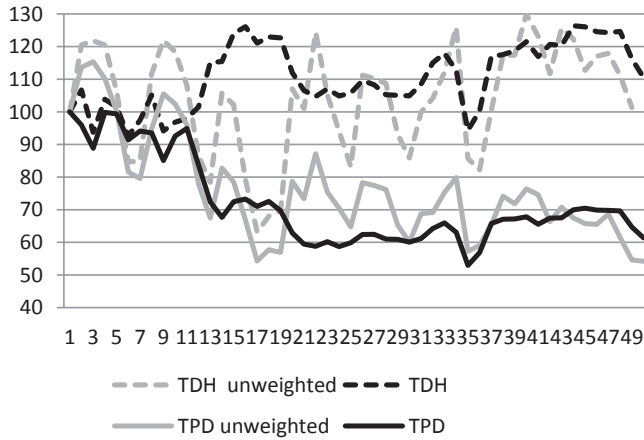
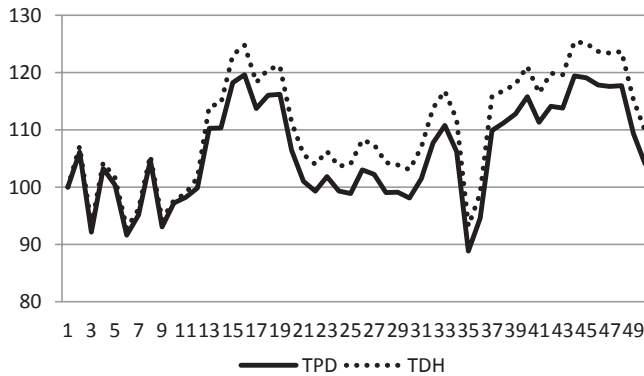Figure 8. Weighted and Unweighted TDH and TPD Indexes (EAN Level)

Figure 9. TPD Index and TDH Index, Group Level

types of transactions data, is preferred over electronic data, such as data scraped from websites, where expenditure information is not available.

3.2. *Some Results at the Group Level*

A straightforward way of getting rid of re-launches and disguised price changes is to identify items by cross classifying the (categorical) attributes rather than by barcode. In doing so, any difference between the TPD and TDH indexes can be entirely attributed to the implicit use of first-order and higher-order inter-action terms in the TPD regression. A full cross-classification of the six categorical attributes available in the scanner data yields $2 \times 2 \times 2 \times 3 \times 3 \times 2 = 144$ possible combinations or "groups," as we will call them, but only 37 of those are actually found in the data. Prices are calculated as unit values across all the EANs belonging to the respective groups.

A comparison of Figure 9 with Figure 2 reveals that the group-based TPD index differs substantially from the EAN-based TPD index. This shows how

TABLE 2
REGRESSION RESULTS FOR TDH MODEL, GROUP LEVEL

Attribute	Dummy	Estimate	Std. error	t-value	Signif.
Intercept		1.906238	0.029731	64.12	***
Shape of neck	V	0.008876	0.009980	0.89	
Fabric	Organic	−0.231588	0.020739	−11.17	***
Sleeve length	Long	0.210602	0.015455	13.04	***
# T-shirts per package	2	0.529237	0.012863	41.15	***
	3	0.678165	0.017848	38.00	***
Color	White	−0.038794	0.010799	−3.59	***
	Black	−0.024664	0.015900	−1.55	
Fit	Stretch	0.012907	0.020998	0.62	

Note: Significance codes: *** 0.001, ** 0.01, * 0.05; # observations: 1,289; R squared: 0.8664; adjusted R squared: 0.8602; the results for the time dummy variables are not shown.
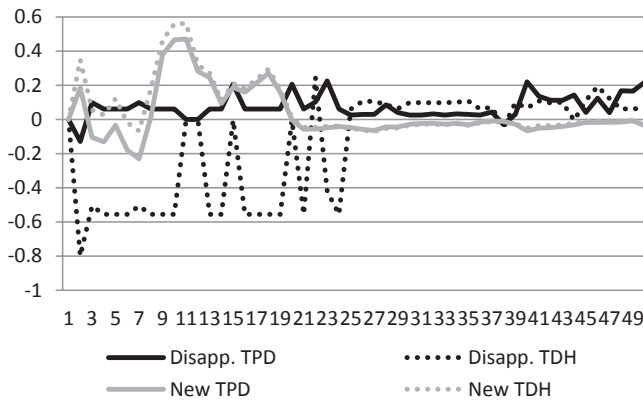

Figure 10. Weighted Average Regression Residuals, Group Level

sensitive the TPD method can be to the choice of item identifier. Assuming the groups can be viewed as homogeneous products, there must have been disguised price increases the EAN-based TPD index was unable to pick up, which is tantamount to saying there must have been a lack of matching. In accordance with our expectations, the group-based TDH index is very similar to its EAN-based counterpart. Although the TPD index comes closer to the TDH index at the group level, a gap remains.

Table 2 contains the regression results for the expenditure-share weighted TDH model at the group level. At this level, the number of observations is considerably lower (1,289) than at the EAN level. The coefficients for "shape of neck" and "fit" have now become insignificant. The difference between the R squared values from the TPD and TDH regressions is rather small at the group level, 0.8844 versus 0.8664, which is not surprising since group churn is modest compared with EAN churn. Because aggregation of EANs into groups reduces noise in the prices data, R squared is higher at the group level than at the EAN level.

The weighted average regression residuals from the TPD and TDH regressions at the group level in Figure 10 exhibit a similar pattern as those at the EAN
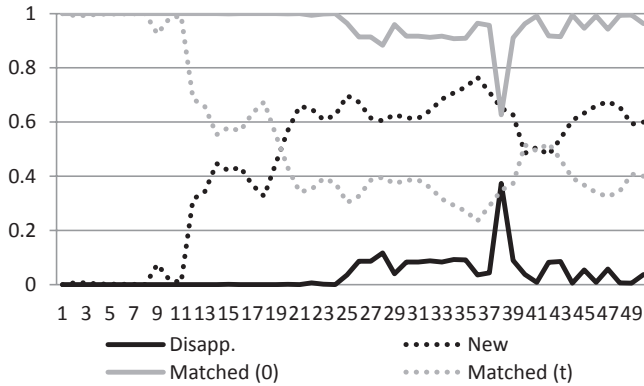
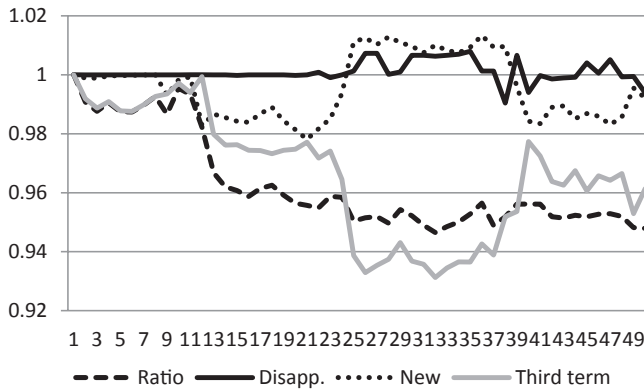Figure 11. Aggregate Expenditure Shares, Group Level

Figure 12. Decomposition (18) of TPD to TDH Ratio, Group Level

level. There are a few noticeable differences though. The average TPD residuals for the new items do not differ much any longer from the average TDH residuals. Also, the average TPD residuals for the disappearing items are now mostly positive.

As can be inferred from Figure 6, the ratio of the period *t* expenditure shares for the new and matched items after month 13 is huge when items are identified by EAN, up to more than 2400 in month 40. This is very different for the group-based items. The patterns of the group-based aggregate expenditure shares shown in Figure 11 are similar to those in Figure 6, but much less pronounced. For example, the share of new items never gets above 0.77, which implies that the ratio of the period *t* expenditure shares for new and matched items never exceeds 3.3. Put differently, the degree of leverage is much less at the group level than at the EAN level.

The results of decomposition (18) at the group level are plotted in Figure 12. The third term, which depends on the change in the matched items' normalized expenditure shares, now contributes most. This is simply because at the group level there are few new and disappearing items; if all items were matched, this term would be exactly equal to the TPD to TDH index ratio.

4. Discussion and Conclusions

The TPD model is essentially a pooled regression model with fixed effects for all items sold during the sample period and with no time-varying variables other than dummies for time. The inclusion of fixed effects in a regression model estimated on a balanced panel controls for unobservable characteristics. Scanner data sets are unbalanced panels, however; there is generally substantial turnover of items; that is, the data are characterized by many entries (new items) and exits (disappearing items). If the item universe was static rather than dynamic, the use of a conventional matched-model index would suffice, and modeling was not required at all.

In this paper, we have discussed some of the issues that arise when applying the TPD method in a dynamic-universe context to construct quality-adjusted price indexes. Our main point is that the TPD model likely suffers from overfitting because it includes all first- and higher-order interactions as compared with the "true" TDH model, which typically includes only the main effects. We derived a decomposition explaining the ratio of the TPD and TDH indexes in terms of the weighted average regression residuals for the unmatched new and disappearing items (and a component that depends on the changes in the matched items' normalized expenditure shares) and applied the decomposition to scanner data on men's T-shirts.

Our results pointed to a downward bias in the TPD index for T-shirts, especially if items are identified by EAN (barcode). Clothing is well known for its lack of matching due to seasonality. For example, summer clothes disappear in autumn and re-appear in spring but often with different EANs. These (strongly) seasonal items are re-launched and potentially exhibit disguised price changes. Seasonality was not the major cause of the downward bias of the TPD index in our example, however. While (weak) seasonal effects in sales do occur as more short-sleeve T-shirts are sold in spring and summer than in autumn and winter, the problem was rather a sudden sharp decline in the number of matches in the data due to the introduction of organic T-shirts that immediately took up a very large share of expenditure at the expense of basic T-shirts.

A potential issue in our hedonic model is *omitted variables bias*. The brand has often been used as a proxy for unobserved characteristics in hedonic models (Triplett, 2006), but this retailer only sells T-shirts under a house brand. Also, an indicator for the quality of fabric other than basic or organic is not included as we are limited by the relatively broad product descriptions in the scanner data sets. Omitted variables in hedonic regressions can lead to bias in the resulting price indexes, but the bias is not necessarily large. Suppose we left out "fabric" from the model. This is an interesting case since the introduction of organic T-shirts had a big impact on the difference between the TDH and TPD indexes, especially at the EAN level. Recall that the coefficient for organic in the original EAN-based TDH regression (Table 1) was negative and highly significant. Figure 13 shows what happens to the TDH index if we delete this variable: almost nothing.

A comparison of the new regression results in Table 3 with the old ones in Table 1 reveals that the downward effect of "organic" is now largely being picked up by the dummy for "stretch" due to a high correlation between these variables.
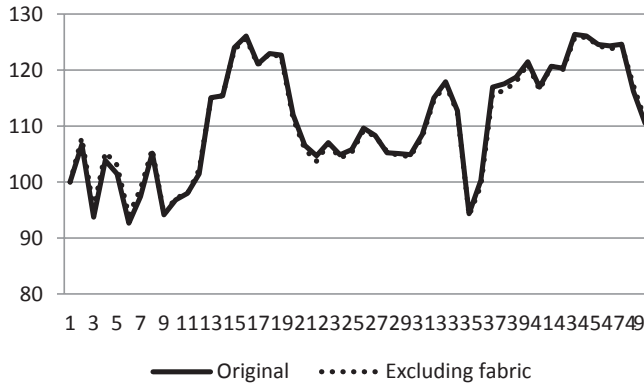
Figure 13. TDH Index and TDH Index Excluding Fabric, EAN Level

TABLE 3
REGRESSION RESULTS FOR TDH MODEL EXCLUDING FABRIC, EAN LEVEL

Attribute	Dummy	Estimate	Std. error	t-value	Signif.
Intercept		1.929611	0.009630	200.38	***
Shape of neck	V	0.015069	0.003228	4.67	***
Sleeve length	Long	0.312975	0.004080	76.71	***
# T-shirts per package	2	0.496930	0.004166	119.28	***
	3	0.664429	0.005741	115.74	***
Color	White	−0.056324	0.003494	−16.12	***
	Black	−0.066130	0.005083	−7.021	***
Fit	Stretch	−0.202044	0.004515	−44.57	***

Note: Significance codes: ***0.001, **0.01, *0.05; # observations: 24,797; *R* squared: 0.7480; adjusted *R* squared: 0.7474; the results for the time dummy variables are not shown.

This example also reminds us that multicollinearity is not such a big issue when estimating TDH indexes, where we are interested in the predicted prices rather than the estimated characteristics parameters.

The multilateral TDH method is not the only way to estimate transitive quality-adjusted price indexes. de Haan and Krsinich (2014) proposed a (rolling-year) GEKS approach called ITRYGEKS, where the missing prices of the new and disappearing items in the bilateral comparisons—in their case measured by bilateral Törnqvist price indexes—are imputed using bilateral TDH regressions, based on a result derived by de Haan (2004). Statistics New Zealand implemented this method in the CPI for many consumer electronics products (Statistics New Zealand, 2014). A strong point of (ITRY)GEKS is its reliance on a superlative index number formula; it is grounded in standard index number theory. A practical disadvantage is its complexity and the fact that many models must be estimated each month.

The above is not to say that weighted TDH has no theoretical underpinning. As shown by equation (13) in Section 2.2, the weighted TDH index (and

the weighted TPD index) can be written as an imputation Törnqvist price index times a factor that induces transitivity. de Haan and Krsinich (2017) showed that the expenditure-share weighted TDH index will be an accurate approximation of a so-called *quality-adjusted unit value index*. The latter is a modified unit value index where the observed prices and quantities are replaced by quality-adjusted prices and quantities to standardize the various items. The quality-adjusted unit value approach is appealing for products existing of broadly comparable items. Irrespective of the interpretation, the TDH model should be restricted to broadly comparable items, that is, applied at a low level of aggregation, because different products typically have different sets of characteristics or different parameters for the same characteristics.

The use of the TPD method can lead to biased results when there is insufficient matching due to re-launches of identical items but with different barcodes or due to a sudden introduction of new items that account for a large share of expenditure. The first problem is essentially a data problem: the barcode/EAN may be too detailed a level to compare like with like. Data permitting, we could identify items by cross classifying the categorical attributes and apply the TPD method. This does not necessarily resolve the second problem because the TPD method wrongly treats all interaction terms as quality characteristics. If enough information on characteristics is available, the TDH method is our preferred choice. Moreover, there would be no need to form groups: the TDH method can be directly applied to scanner data at the barcode level.

If the information on characteristics is not available, the TPD model could perhaps be improved by incorporating a life cycle function. The extended model would control for time, unobserved characteristics, and "age." For an application to scanner data, see Melser and Syed (2016), Bils (2009) and Abe *et al*. (2016) estimated life cycle functions as well, albeit not in a TPD context. We doubt, however, that this approach is fit for CPI production given the complexity of such functions. Moreover, a life cycle approach is unlikely to resolve the problem of re-launches and disguised price changes as both their timing and magnitude are rather unpredictable.

References

Abe, N., Y. Ito, K. Munakata, S. Ohyama, and K. Shinozaki, "Pricing Patterns Over Product Life-Cycle and Quality Growth at Product Turnover: Empirical Evidence From Japan," *Working paper 16-E-5*. Bank of Japan, Tokyo, 2016. (Available from https://www.boj.or.jp/en/research/wps_rev/wps_2016/data/wp16e05.pdf)

Adjibolosoo, S., "Estimation of Parameters of Heteroscedastic Error Models Under Various Hypothesized Error Structures," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 42, 123–33, 1993.

Aizcorbe, A., C. Corrado, and M. Doms, "When Do Matched-Model and Hedonic Techniques Yield Similar Price Measures?" *Working Paper 2003-14*. Federal Reserve Bank of San Francisco, San Francisco, 2003. (Available from https://www.frbsf.org/economic-research/files/wp03-14bk.pdf)

Australian Bureau of Statistics (ABS), "Making Greater Use of Transactions Data to Compile the Consumer Price Index," *Information Paper 6401.0.60.003*. ABS, Canberra, 2016. (Available from https://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0.60.003)

Bils, M., "Do Higher Prices for New Goods Reflect Quality Growth or Inflation?" *Quarterly Journal of Economics*, 124, 637–75, 2009.

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

Chessa, A.G., J. Verburg, and L. Willenborg, "A Comparison of Price Index Methods for CPI Construction Using Scanner Data," Paper presented at the 15th Ottawa Group meeting, 10–12 May 2017, Eltville am Rhein, Germany, 2017. (Available from https://www.ottawagroup.org)

Diewert, W. E., "Weighted Country Product Dummy Variable Regressions and Index Number Formulae," *Review of Income and Wealth*, 51, 561–70, 2005.

Diewert, W. E., S. Heravi, and M. Silver, "Hedonic Imputation Versus Time Dummy Hedonic Indexes," in W. E. Diewert, J. Greenlees, and C. Hulten (eds), *Price Index Concepts and Measurement, Studies in Income and Wealth*, University of Chicago Press, Chicago, 70, 87–116, 2009.

Diewert, W. E. and K. J. Fox, "Substitution Bias in Multilateral Methods for CPI Construction Using Scanner Data," *Discussion Paper 17-02*, Vancouver School of Economics, The University of British Columbia, Vancouver, Canada, 2017. (Available from https://econ.sites.olt.ubc.ca/files/2017/03/pdf_paper-erwin-diewert-17-02-SubstitutionBiasetc.pdf)

Fox, K. J. and D. Melser, "Non-Linear Pricing and Price Indexes: Evidence and Implications From Scanner Data," *Review of Income and Wealth*, 60, 261–78, 2014.

de Haan, J., "Generalised Fisher Price Indexes and the Use of Scanner Data in the Consumer Price Index," *Journal of Official Statistics*, 1, 61–85, 2002.

———, "The Time Dummy Index as a Special Case of the Imputation Törnqvist Index," Paper presented at the 8th Ottawa Group meeting, 23–5 August 2004, Helsinki, Finland, 2004. (Available from https://www.ottawagroup.org)

———, "Hedonic Price Indexes: A comparison of Imputation, Time Dummy and 'Re-Pricing' Methods," *Jahrbücher fur Nationalökonomie und Statistik*, 230, 772–91, 2010.

———, "A Framework for Large Scale Use of Scanner Data in the Dutch CPI," Paper presented at the 14th Ottawa Group meeting, 20–22 May 2015, Tokyo, Japan, 2015. (Available from https://www.ottawagroup.org)

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

de Haan, J. and R. Hendriks, "Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes," Paper presented at the Economic Measurement Group Workshop, 28–29 November 2013, Sydney, Australia, 2013. (Available from https://www.business.unsw.edu.au/research-site/centreforappliedeconomicresearch-site/Documents/Jan-de-Haan-Online-Price-Indexes.pdf)

de Haan, J. and F. Krsinich, "Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes," *Journal of Business & Economic Statistics*, 32, 341–58, 2014.

———, "Time Dummy Hedonic and Quality-Adjusted Unit Value Indexes: Do They Really Differ?" *Review of Income and Wealth*, 64, 757–76, 2017.

Harvey, A., "Estimating Regression Models with Multiplicative Heteroscedasticity," *Econometrica*, 44, 461–65, 1976.

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

ILO/IMF/OECD/UNECE/Eurostat/The World Bank, *Consumer Price Index Manual: Theory and Practice*, ILO Publications, Geneva, 2004.

Kennedy, P. E., "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 71, 801, 1981.

Krsinich, F., "The FEWS Index: Fixed Effects with a Window Splice," *Journal of Official Statistics*, 32, 375–404, 2016.

Melser, D. and I. A. Syed, "Life Cycle Price Trends and Product Replacement: Implications for the Measurement of Inflation," *Review of Income and Wealth*, 62, 509–33, 2016.

Reinsdorf, M. B., "Using Scanner Data to Construct CPI Basic Component Indexes," *Journal of Business & Economic Statistics*, 17, 152–60, 1999.

Silver, M., "How to Measure Property Price Indices Better," *EURONA*, 1, 35–66, 2018.

Silver, M. and S. Heravi, "A Failure in the Measurement of Inflation: Results From a Hedonic and Matched Experiment Using Scanner Data," *Journal of Business & Economic Statistics*, 23, 269–81, 2005.

Solon, G., S. J. Haider, and J. M. Wooldridge, "What are We Weighting for?" *Journal of Human Resources*, 50, 301–316, 2015.

Statistics New Zealand, "Measuring Price Change for Consumer Electronics Using Scanner Data," 2014. (Available from https://www.stats.govt.nz/browse_for_stats/economic_indicators/CPI_inflation/cpi-rice-change-scanner-data.aspx)

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

Triplett, J.E., *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes; Special Application to Information Technology Products*, OECD, Paris, 2006.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site: