1. Introduction

Over the past couple of years, Statistics Netherlands has been experimenting with the collection of prices from the Internet through web scraping. Online prices could perhaps replace part of the prices observed by price collectors for the compilation of the CPI.[1] Online prices might also replace data that is currently being collected from the Internet in a much less efficient way. Apart from efficiency considerations, web scraping has the advantage that prices can be monitored daily, allowing the estimation of high-frequency price indexes. In the Billion Prices Project, a research initiative at MIT that uses online data to study high-frequency price dynamics and inflation, daily price index numbers have been calculated for several countries around the world, including the Netherlands.[2] For an example on Argentina data, see Cavallo (2012)

Importantly, data on quantities purchased cannot be observed via the Internet. The lack of quantity data is problematic for the construction of price indexes, bp9( )-110.212(p)-00

In section 6 we suggest using a rolling window approach to updating the time series and discuss problems that may arise when using daily online price data, including the treatment of regular and sales prices. A related issue is whether the compilation of daily price indexes would be useful.

Section 7 provides some empirical illustrations. Our data set contains daily price observations extracted from the website of a Dutch online retailer for three products: women's T-shirts, men's watches, and kitchen appliances.

Section 8 summarizes our findings and concludes.

2. Time dummy hedonic indexes

A hedonic model explains the price of a product from its (performance) characteristics. Though other functional forms are possible, for convenience we will only consider the log-linear model

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

where p_i^t denotes the price of item i in period t; z_{ik} is the (quantity) of characteristic k for item i and β_k the corresponding parameter; δ^t is the intercept; the random errors ε_i^t have an expected value of zero, constant variance and zero covariance.

The parameters β_k in model (1) are constant across time. Pakes (2003) argues that this is a (too) restrictive assumption,[3] but it allows us to estimate the model on the pooled data of two or more periods, thus increasing efficiency. Suppose we have data for a particular product at our disposal for periods $t = 0,1,...,T$; the samples of items are denoted by $S^0, S^1,...,S^T$ and the corresponding number of items by $N^0, N^1,...,N^T$. The estimating equation for the pooled data becomes

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t, \qquad (2)$$

[3] Data permitting, this assumption can be tested. A more flexible method for estimating quality-adjusted price indexes is hedonic imputation where the characteristics parameters are allowed to change over time and the model is estimated separately in each time period. Starting from some preferred index number formula, the 'missing prices' are imputed using the predicted prices from the hedonic regressions. For a comparison of time dummy and imputation approaches, see Silver and Heravi (2007), Diewert, Heravi and Silver (2009), and de Haan (2010).

where the time dummy variable D_i^t has the value 1 if the observation pertains to period t and the value 0 otherwise; the time dummy parameters d^t shift the hedonic surface upwards or downwards as compared with the intercept term d^0. The method is usually referred to as the time dummy method.

Suppose equation (2) is estimated by Ordinary Least where $\bar{z}_k^0 = \sum_{i \in S^0} z_{ik} / N^0$ and $\bar{z}_k^t = \sum_{i \in S^t} z_{ik} / N^t$ are the unweighted sample means of characteristic k. Due to the inclusion of time dummies and an intercept into the model, the OLS residuals sum to zero in each period so that $\prod_{i \in S^0} (\hat{\varepsilon}_i^0)^{1/N^0} = \prod_{i \in S^0} (\varepsilon_i^0)^{1/N^0}$

3.

dummy method is less efficient than the hedonic time dummy method because more parameters have to be estimated. The time-product dummy method is cost efficient in that there is no need to collect information on item characteristics.

In order to derive an explicit expression for the time-product dummy index, we can follow the same steps as in section 2. For $i = 1,...,N-1$, the predicted prices in the base period 0 and the comparison periods t $(t = 1,...,T)$ are $\hat{p}_i^0 = \exp(\hat{\alpha})\exp(\hat{\gamma}_i)$ and $\hat{p}_i^t = \exp(\hat{\alpha})\exp(\hat{\delta}^t)\exp(\hat{\gamma}_i)$

We will first examine what drives the difference between the unweighted time-product dummy index and the chained matched-models index. The time-product dummy method is a special case of the time dummy method, and so the time-product dummy index (14) can be expressed as a chain index, similar to equation (9):

$$ {}^{0} \quad {}_{t\ 1} \quad \frac{\left(\ {}_{t}\ {}^{t} \right)^{\frac{1}{t}}}{\left(\ {}_{t\ 1}\ {}^{t\ 1} \right)^{\frac{1}{t\ 1}}} \exp\left[\bar{\hat{g}}^{t\ 1} - \bar{\hat{g}}^{t} \right] $$

the power of $f_D^{t-1,t} = N_D^{t-1,t} / N^{t-1}$ (the fraction of disappearing items). The factor with the average fixed effects can be written as

$$\left[\bar{\ }_{t\ 1}\quad \bar{\ }_{t}\right] \qquad \left[\frac{{}_{i\ S_N^{t\ 1,t}}\ {}_{i}^{\frac{1}{N_N^{t\ 1,t}}}}{{}_{i\ S_M^{t\ 1,t}}\ {}_{i}^{\frac{1}{N_M^{t\ 1,t}}}}\right]^{f_N^{t\ 1,t}} \left[\frac{{}_{i\ S_D^{t\ 1,t}}[\exp(\hat{\ }_i)]^{\frac{1}{N_D^{t\ 1,t}}}}{{}_{i\ S_M^{t\ 1,t}}[\exp(\hat{\ }_i)]^{\frac{1}{N_M^{t\ 1,t}}}}\right]^{f_D^{t\ 1,t}}$$

Now recall that $\hat{p}_i^t = \exp(\hat{\alpha})\exp(\hat{\delta}^t)\exp(\hat{\gamma}_i)$ or $\exp(\hat{\gamma}_i) = \hat{p}_i^t / [\exp(\hat{\alpha})\exp(\hat{\delta}^t)]$, and therefore also $\exp(\hat{\gamma}_i) = \hat{p}_i^{t-1} / [\exp(\hat{\alpha})\exp(\hat{\delta}^{t-1})]$. Substituting these results into the first factor and second factor between square brackets of (18), respectively, gives

$$\frac{P_{TPD}^{0t}}{P_{TPD}^{0,t-1}} = \prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}}\right)^{\frac{1}{N_M^{t-1,t}}} \left[\frac{\prod_{i \in S_N^{t-1,t}} \left(\frac{p_i^t}{\hat{p}_i^t}\right)^{\frac{1}{N_N^{t-1,t}}}}{\prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{\hat{p}_i^t}\right)^{\frac{1}{N_M^{t-1,t}}}}\right]^{\phi_N^{t-1,t}} \left[\frac{\prod_{i \in S_D^{t-1,t}} \left(\frac{p_i^{t-1}}{\hat{p}_i^{t-1}}\right)^{\frac{1}{N_D^{t-1,t}}}}{\prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^{t-1}}{\hat{p}_i^{t-1}}\right)^{\frac{1}{N_M^{t-1,t}}}}\right]^{-\phi_D^{t-1,t}}. \quad (19)$$

According to (19), new items will have an upward effect when their average regression residuals are greater than those of the matched items in period t, i.e., when their prices are on average unusually high. Decomposition (19) is a well-known result. It holds for any (OLS) multilateral time dummy index and can be directly derived from the fact that the regression residuals sum to zero in each period.

Equation (19) does clarify the role of items which are observed only once during the whole period $0,...,T$. By definition these are unmatched items. When using hedonic regression, they affect measured price change, as they should, but when using the time-product dummy method, they do not. To understand wh fact that, while their fixed effects can be estimated, items with a single observation are zeroed out in the two-period case, carries over to the many-period case. This does not mean that a chained matched-model Jevons index is useless, as we have seen. Items which are 'new' or 'disappearing' in comparisons of adjacent periods are typically observed multiple times during $0,...,T$ and are not zeroed out. They contain information on price change that is used in a multilateral time-product dummy regression whereas they are ignored in a chained matched-model index.

5. A comparison with the GEKS-Jevons index

The fixed effects in a time-product dummy model can be seen as item-specific hedonic price effects, assuming the parameters of the characteristics in the underlying log-linear hedonic model are constant across time. This leads Aizcorbe, Corrado and Doms (2003) and Krsinich (2013) to believe that the time-product dummy method produces a quality-adjusted price index. But measuring quality-adjusted price indexes without information on item characteristics is just not possible. This is almost trivial from a modelling point of view. In a hedonic model, the exponentiated time dummy coefficients are estimates of quality-adjusted price indexes since we control for changes in the characteristics. In the time-product dummy model, there is nothing to control for as auxiliary information on characteristics is not included.

The exponentiated time dummy coefficients in the time-product dummy method do not measure quality-adjusted price change but represent a particular type of matched-model price change. In this section, we will compare the unweighted multilateral time-product dummy method to a competing transitive approach, the unweighted multilateral

o4.41795(u)-1.5502(l)-2.02(n)32( )-405502(n)-6775(u)-1.5506(b)-1.5502(u)-1-40.8989d-2.8702(e

between periods 0 and l, periods l and t, and periods 0 and t. From section 4 it follows that

$$P^{0l}_{TPD(0,l)} = \frac{\prod_{i \in S^l} (p_i^l)^{\frac{1}{N^l}}}{\prod_{i \in S^0} (p_i^0)^{\frac{1}{N^0}}} \exp\left[\bar{\hat{g}}^0_{(0,l)} - \bar{\hat{g}}^l_{(0,l)}\right]; \qquad (24)$$

$$P^{lt}_{TPD(l,t)} = \frac{\prod_{i \in S^t} (p_i^t)^{\frac{1}{N^t}}}{\prod_{i \in S^l} (p_i^l)^{\frac{1}{N^l}}} \exp\left[\bar{\hat{g}}^l_{(l,t)} - \bar{\hat{g}}^t_{(l,t)}\right]; \qquad (25)$$

$$P^{0t}_{TPD(0,t)} = \frac{\prod_{i \in S^t} (p_0^t)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p_i^0)^{\frac{1}{N^0}}} \exp\left[\bar{\hat{g}}^0_{(0,t)} - \bar{\hat{g}}^t_{(0,t)}\right], \qquad (26)$$

Equation (27) decomposes the GEKS-Jevons price index into three factors. The first factor is the ratio of geometric mean prices in periods t and 0. The second factor is the antilog of the difference between the (arithmetic) averages of $\bar{\hat{g}}^0_{(0,l)}$ $(l = 1,...,T)$ and $\bar{\hat{g}}^t_{(l,t)}$ $(l = 0,...,T; l \neq t)$, where $\bar{\hat{g}}^0_{(0,t)}$ and $\bar{\hat{g}}^t_{(0,t)}$ count twice. The third factor is the antilog of the average of $\bar{\hat{g}}^l_{(l,t)} - \bar{\hat{g}}^l_{(0,l)}$ $(l = 1,...,T; l \neq t)$, raised to the power of $(T-1)/(T+1)$. We expect the third factor to be relatively small and fluctuate around zero over time. The GEKS-Jevons index is therefore most likely driven by the first two factors.

Let us compare decomposition (27) with decomposition (14) for the multilateral time-product dummy index. P^{0t}_{GEKS-J} and P^{0t}_{TPD} are both written as the ratio of geometric mean prices in periods t and 0, adjusted by factors based on differences in average fixed effects. The average fixed effects for period 0 and period t in (27), $\bar{\hat{g}}^0_{(0,l)}$ and $\bar{\hat{g}}^t_{(l,t)}$, can be viewed as crude approximations of $\bar{\hat{g}}^0$ and $\bar{\hat{g}}^t$ in (14) because, by assumption, they all measure the same average fixed effects, albeit estimated on different subsets of the data. Thus, the means $(\sum_{l=1}^{T} \bar{\hat{g}}^0_{(0,l)} + \bar{\hat{g}}^0_{(0,t)})/(T+1)$ and $(\sum_{l=0, l \neq t}^{T} \bar{\hat{g}}^t_{(l,t)} + \bar{\hat{g}}^t_{(0,t)})/(T+1)$ are also approximations of $\bar{\hat{g}}^0$ and $\bar{\hat{g}}^t$, but much more stable than the elements $\bar{\hat{g}}^0_{(0,l)}$ and $\bar{\hat{g}}^t_{(l,t)}$. The third factor in (27), which of course does not appear in (14), adds noise to the first two factors.

This result suggests that the unweighted time-produ67d u W n q 8.33333 0 cm BT10.7038.333

When the true characteristics parameters change over time, or if a single model is too restrictive, the basic assumption underlying the time-product dummy model will be violated. As the two methods treat the price changes of the matched items differently, a difference in trend between GEKS and time-product dummy indexes can arise. The

that regular prices stay constant over time but sales prices show an upward trend. Since promotional sales occur infrequently relative to the number of days with regular prices, the overall trend seems to be almost flat. However, if consumers mainly buy the item at times of sales,[18] then the change in sales prices would be a better indicator of the change in prices actually paid.

Partly due to promotional sales, daily price indexes may be quite volatile, at least at the product level. It is questionable whether users benefit from volatile price indexes,

products look reasonable. In Figure 3b the left scale has been adjusted in order to show that the TPD and chained Jevons indexes for kitchen appliances are also volatile, though much less so than average prices. The differences in volatility as well as in index levels between the two indexes are minor.

Figure 1: Daily price indexes of women’s T-shirts (small data set)

Figure 3a: Daily price indexes of kitchen appliances (small data set)

us that the revisions of index numbers previously estimated from the small data set are negligible in relation to the volatility of the indexes.

Figure 4: Daily TPD price indexes of women's T-shirts (large data set)

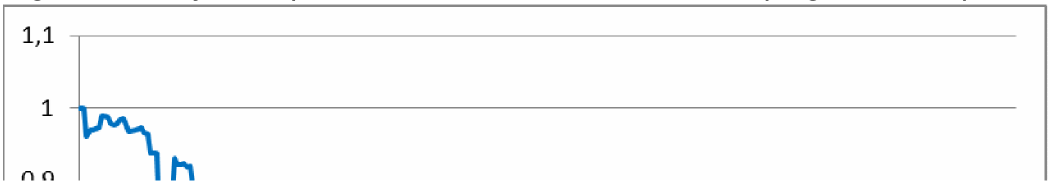
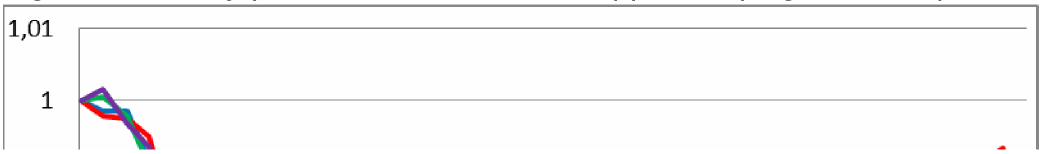


Figure 6: Daily TPD price indexes of kitchen appliances (large data set)

even though these items were most likely available for purchase. It may be worthwhile to impute temporarily 'missing prices', for example by carrying forward the latest price observations. In particular, it would be interesting to investigate how imputations affect the volatility of the daily and weekly time series.

Figure 7: Weekly price indexes of women's T-shirts (large data set)

Figure 9: Weekly price indexes of kitchen appliances (large data set)



Measuring quality-adjusted price change without data on item characteristics is just not possible. The two multilateral methods should therefore not be applied to goods where quality change is important.[21] De Haan and Krsinich (2012) show how the GEKS method can be modified to account for quality change by using hedonic rather than matched-model price indexes as input in the GEKS system.[22] For goods where quality change is of minor importance, the two methods have much to offer as compared to a period-on-period chained matched-model price index since they use all of the matches across the whole sample period. We would prefer the GEKS method because it is the most straightforward way to obtain transitive indexes and because it is a nonparametric approach whereas the time-product dummy method is model-based. Minimising model dependence seems like good advice for producing official statistics. The identification of items remains an issue. Any matched-model method breaks down when changes in item identifiers and price changes occur simultaneously.

The time-product dummy method has a practical advantage though, in particular when the aim is to construct high-frequency price index numbers using online data. If the production system can deal with very large data sets, time-product dummy indexes may be easier to estimate than GEKS indexes. Also, equations (18) and (19) provide practitioners with the opportunity to decompose the latest period-on-period price change into a matched-model index and the effects of items that are new or disappearing with respect to the previous period. The latter effects are implicitly based on the data of many earlier periods. Staff involved in production of the CPI may not like this aspect, but it is unavoidable with multilateral methods.

---

[21] This is also true for the chained matched-model Jevons method, which is how PriceStats compiles daily indexes for each product category. On their website (www.PriceStats.com/faqs) it is mentioned that "We treat all individual products [what we call items] as separate series, without making product substitutions or hedonic quality adjustments. Only consecutive price observations for exactly the same product are used to calculate price changes. So, for example, if a TV is replaced with a new, more expensive model, we do not have a price change in that category. Only when the new model starts changing its price will the index start to be affected by that product. Similarly, when a product disappears from the sample, we assume it is temporarily out of stock for a set amount of time. After that period, the product is discontinued from the index." We think their approach can give rise to upward bias for high-technology goods (due to a lack of quality adjustment) and to downward bias for clothing (due to a combination of high-frequency chaining and the use of too-detailed item identifiers).

[22] As mentioned in footnote 6, it is not possible to incorporate characteristics into a time-product dummy model; the product dummies must be left out to identify the model, turning it into a time dummy hedonic model.

matched-model Törnqvist price index $\prod_{i \in S_M^{t-1,t}} (p_i^t / p_i^{t-1})^{(s_{Mi}^{t-1}+s_{Mi}^{t})/2}$ and dividing again by the same index, but now written as $\prod$ $\prod$

References

Krsinich, F. (2011b), “Measuring the Price Movements of Used Cars and Residential Rents in the New Zealand Consumers Price Index”, Paper presented at the twelfth