Online Data, Fixed Effects and the Construction of High-Frequency Price Indexes

Jan de Haan* and Rens Hendriks**

* Statistics Netherlands / Delft University of Technology

** Statistics Netherlands

EMG Workshop 2013

Aims of the paper

- Explain why the multilateral Time-Product Dummy index (TPD index) differs from its chained matched-model counterpart.
- Show that the multilateral TPD or ‘fixed effects’ method does not produce quality-adjusted price indexes.
- Investigate whether the TPD method is useful for estimating high-frequency price indexes from online data (for goods where quality change is not a major concern).

Outline

- Background
- Time dummy hedonic price indexes
- Time-product dummy indexes
- Unmatched items and the time-product dummy index
- A comparison with the GEKS-Jevons index
- Issues with daily online data and daily indexes
- Empirical results
- Conclusions

Background

- Possible use by Stats Netherlands of online prices obtained through web scraping

  Efficiency reasons

  Daily observations: high-frequency price indexes possible

  However, no quantity information

- Choice of index number method

  Diewert (2004): TPD method produces a matched-model index in the bilateral (two-period) case.

  Aizcorbe, Corrado and Doms (2003): TPD produces quality-adjusted price index in the multilateral (many-period) case. This seems to good to be true.

Time dummy hedonic indexes

We only consider the log-linear hedonic model. Estimating equation on the pooled data for periods t=0,1,…,T is

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

Estimating this time dummy model by OLS regression yields

$$P_{TD}^{0t} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} (p_i^t)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p_i^0)^{\frac{1}{N^0}}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^0 - \bar{z}_k^t)\right]$$

Time dummy hedonic indexes

In words: the time dummy index can be written as the product of the ratio of geometric mean prices and a quality-adjustment factor.

This exponential factor depends on the changes over time of the average characteristics.

The time dummy index is transitive and can be written as a chain index:

$$P_{TD}^{0t} = \prod_{\tau=1}^{t} \frac{\prod_{i \in S^{\tau}} (p_i^{\tau})^{\frac{1}{N^{\tau}}}}{\prod_{i \in S^{\tau-1}} (p_i^{\tau-1})^{\frac{1}{N^{\tau-1}}}} \exp\left[\sum_{k=1}^{K} \hat{\beta}_k (\bar{z}_k^{\tau-1} - \bar{z}_k^{\tau})\right]$$

Time-product dummy (TPD) indexes

Characteristics and their parameters are assumed constant over time in the time dummy model.

No characteristics available: replace unobservable hedonic effects $\sum_{k=1}^{K} \beta_k z_{ik}$ by item-specific fixed values γ_i .

Fixed effects or time-product dummy model

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

Counterpart of Country-Product Dummy (CPD) model for cross-country comparisons

TPD indexes

TPD index can be written as

$$P_{TPD}^{0t} = \exp(\hat{\delta}^t) = \frac{\prod_{i \in S^t} (p_i^t)^{\frac{1}{N^t}}}{\prod_{i \in S^0} (p_i^0)^{\frac{1}{N^0}}} \exp\left[\bar{\hat{\gamma}}^0 - \bar{\hat{\gamma}}^t\right]$$

or, because it is transitive, in chained form as

$$P_{TPD}^{0t} = \prod_{\tau=1}^{t} \frac{\prod_{i \in S^\tau} (p_i^\tau)^{\frac{1}{N^\tau}}}{\prod_{i \in S^{\tau-1}} (p_i^{\tau-1})^{\frac{1}{N^{\tau-1}}}} \exp\left[\bar{\hat{\gamma}}^{\tau-1} - \bar{\hat{\gamma}}^{\tau}\right]$$

Unmatched items and the TPD index

How are unmatched items treated in the TPD index?

Chain link of TPD index can be written as the product of the adjacent-period matched-model Jevons index and the effects of new items and disappearing items:

$$\frac{P_{TPD}^{0t}}{P_{TPD}^{0,t-1}} = \prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}} \right)^{\frac{1}{N_M^{t-1,t}}} \left[\frac{\prod_{i \in S_N^{t-1,t}} \left(\frac{p_i^t}{\exp(\hat{\gamma}_i)} \right)^{\frac{1}{N_N^{t-1,t}}}}{\prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^t}{\exp(\hat{\gamma}_i)} \right)^{\frac{1}{N_M^{t-1,t}}}} \right]^{f_N^{t-1,t}} \left[\frac{\prod_{i \in S_D^{t-1,t}} \left(\frac{p_i^{t-1}}{\exp(\hat{\gamma}_i)} \right)^{\frac{1}{N_D^{t-1,t}}}}{\prod_{i \in S_M^{t-1,t}} \left(\frac{p_i^{t-1}}{\exp(\hat{\gamma}_i)} \right)^{\frac{1}{N_M^{t-1,t}}}} \right]^{-f_D^{t-1,t}}$$

Unmatched items and the TPD index

Take clothing, for example. Prices typically decline over time, so a chained-matched model index will have a downward trend.

If TPD method would work, i.e. if fixed effects approximate hedonic effects well, then the unmatched items are likely to counter this downward trend – average quality-adjusted prices of new (disappearing) items likely above (below) average quality-adjusted prices of matched items.

But does the TPD method really account for new and disappearing items?

Unmatched items and the TPD index

No, it doesn't.

- Items which are observed only once during the whole sample period – are zeroed out: they are effectively dropped from the estimation.
- Thus, it still is a matched-model approach and does not adjust for quality change, even though ….
- …. theTPD index differs from the chained matched model Jevons as items which are 'new' or 'disappearing' in period-on-period comparisons are often observed multiple times during the sample period.

A comparison with the GEKS-Jevons index

Ivancic, Diewert and Fox (2011) and others adapted the GEKS method for making transitive price comparisons across countries to price comparisons across time.

$$P_{GEKS}^{0t} = \prod_{l=0}^{T} \left(P^{0l} \times P^{lt} \right)^{\frac{1}{T+1}}$$

P^{0l} and P^{lt} are bilateral price indexes between 0 and *l,* and *l* and *t*; *l* (*l*=0,…,*T)* is the link period.

Online data: no quantity information. Use of bilateral Jevons indexes (rather than Fisher indexes).

A comparison with the GEKS-Jevons index

Some findings:

- If some (unknown) time dummy hedonic model describes the data well, then TPD is a (smoothed) approximation of the matched-model GEKS-Jevons – the two methods essentially aim at the same index number formula.
- Not surprising: both methods use the exact same information, i.e. the prices all matches across the sample period or window 0,…,*T*.
- Trends may differ if e.g. the ‘true’ characteristics parameters change over time.
- TPD method probably easier to estimate.

Issues with daily online data and daily indexes

Rolling window approach can overcome revisions problem.

Window length: no longer than maximum period items are offered for sale. Depends on

- type of product;
- market circumstances;
- policy of assigning and changing item identifiers.

In practice: items identified by article numbers (EANs in scanner data) or web IDs (online data).

These identifiers may be too detailed – similar items having different IDs.

Issues with daily online data and daily indexes

Potential problems:

- item churn overestimated;
- matched-model indexes based on fewer matches than desirable;
- matched-model methods, including TPD (and GEKS), miss hidden price changes.

Issues with web scraping data

- online prices different from transaction prices;
- representativity of online data;
- changes made to website;

Issues with daily online data and daily indexes

- treatment of sales versus regular prices - daily ‘trajectory’ in offer prices does not necessarily reflect correct trend from the average consumer’s point of view due to promotional sales;
- volatility of daily price indexes;
- monthly unit values not possible with online data.

Note: scanner data might not be an ideal source for online purchases, particularly on clothing.

Potential problem: registration of goods which are returned by customers.

Empirical results

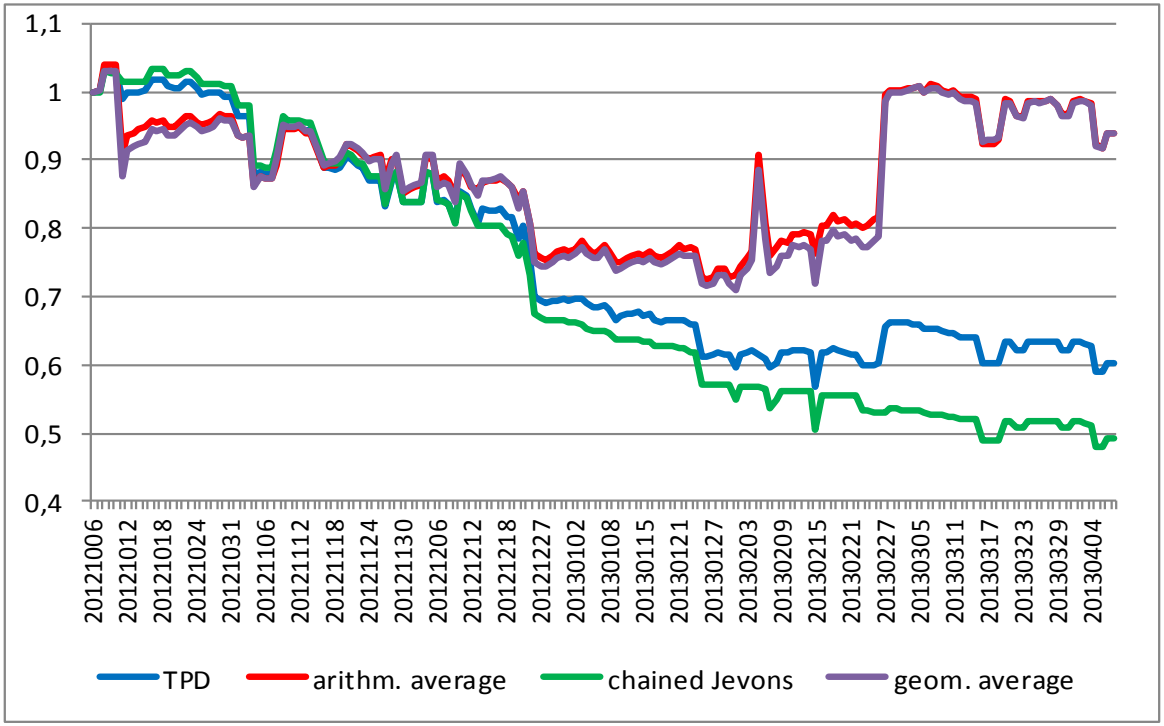
Main goal:

to illustrate that different types of indexes - TPD, chained matched-model Jevons and GEKS-Jevons - can have different trends and can be highly volatile when constructed at a daily frequency.

Data set

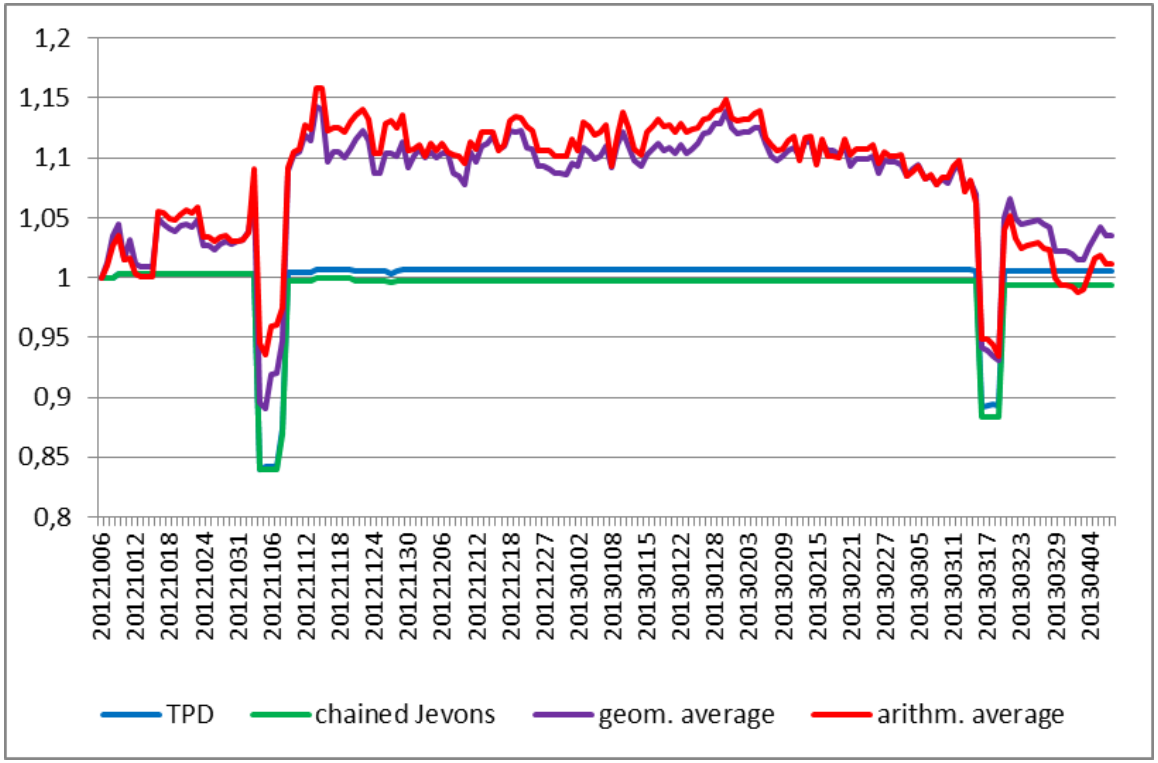
- daily prices extracted from website of Dutch online store - no physical store so only (potential) online purchases
- women's T-shirts; men's watches, kitchen appliances
- 6 October 2012 – 8 April 2013 (12 August 2013)

Daily indexes; women's T-shirts; small data set



- TPD above chained Jevons, as expected
- Substantial downward bias – too detailed identifiers
- Extremely volatile; trend in average prices more plausible

Daily indexes; men's watches; small data set

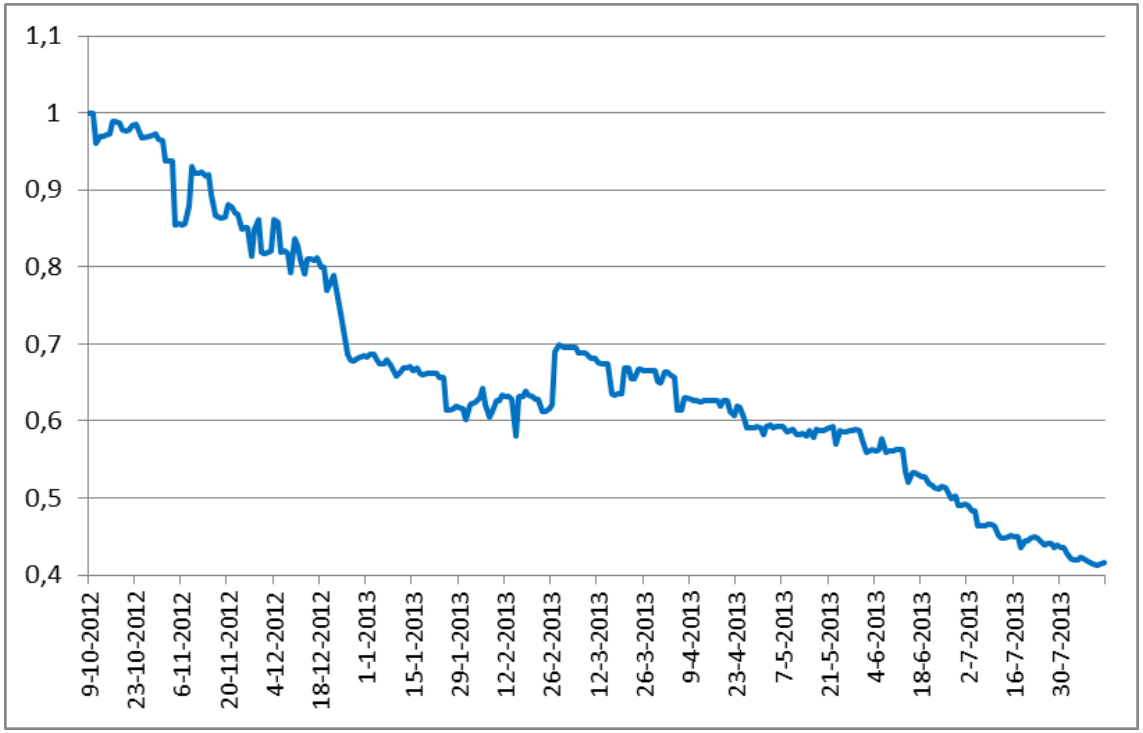


- Heterogeneity – erratic behavior average prices
- TPD and chained Jevons very similar and reasonable

Daily indexes; kitchen appliances; small data set

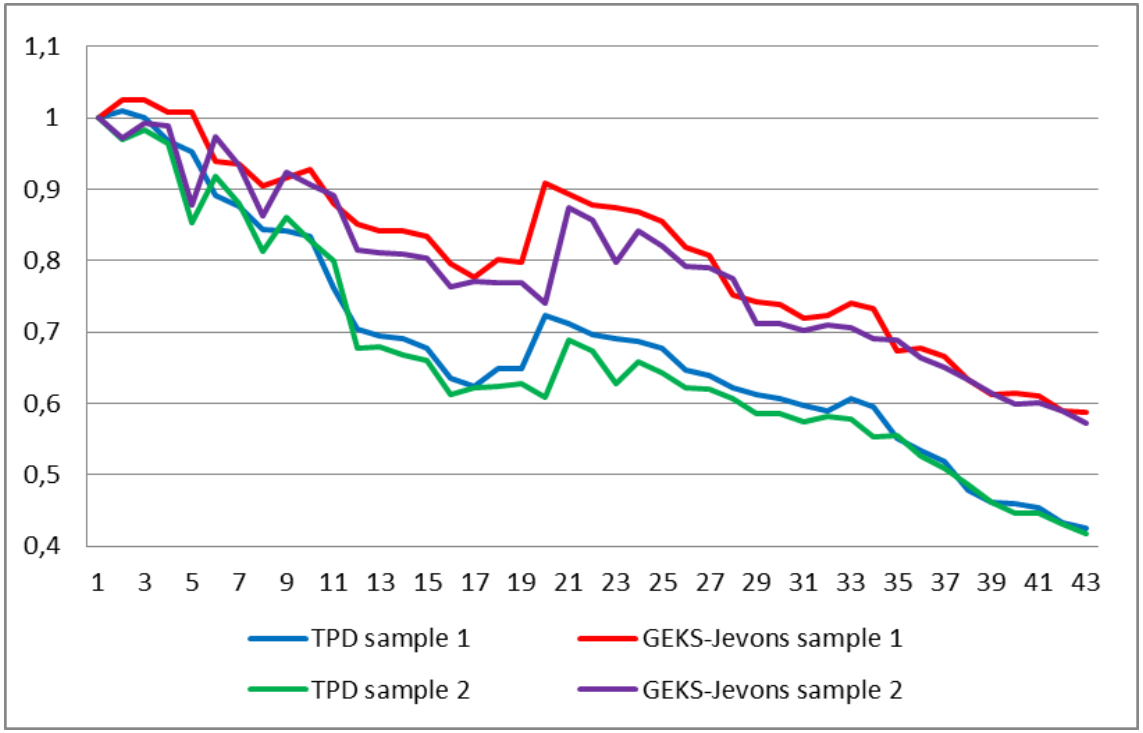
- Compositional change early November 2012

Daily TPD indexes; women's T-shirts; large data set



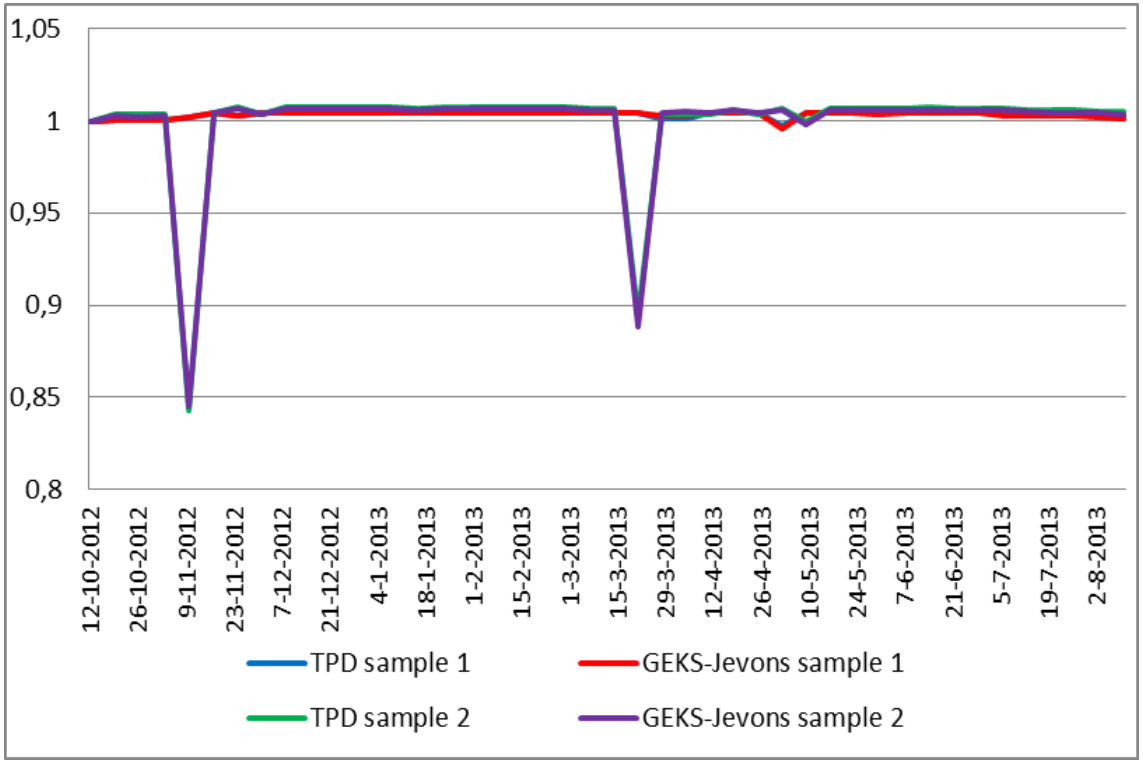
- Confirms downward bias of TPD index (decline of almost 60% within 10 months!)
- Comparison with small data set: revisions very small

‘Weekly’ indexes; women’s T-shirts; large data set



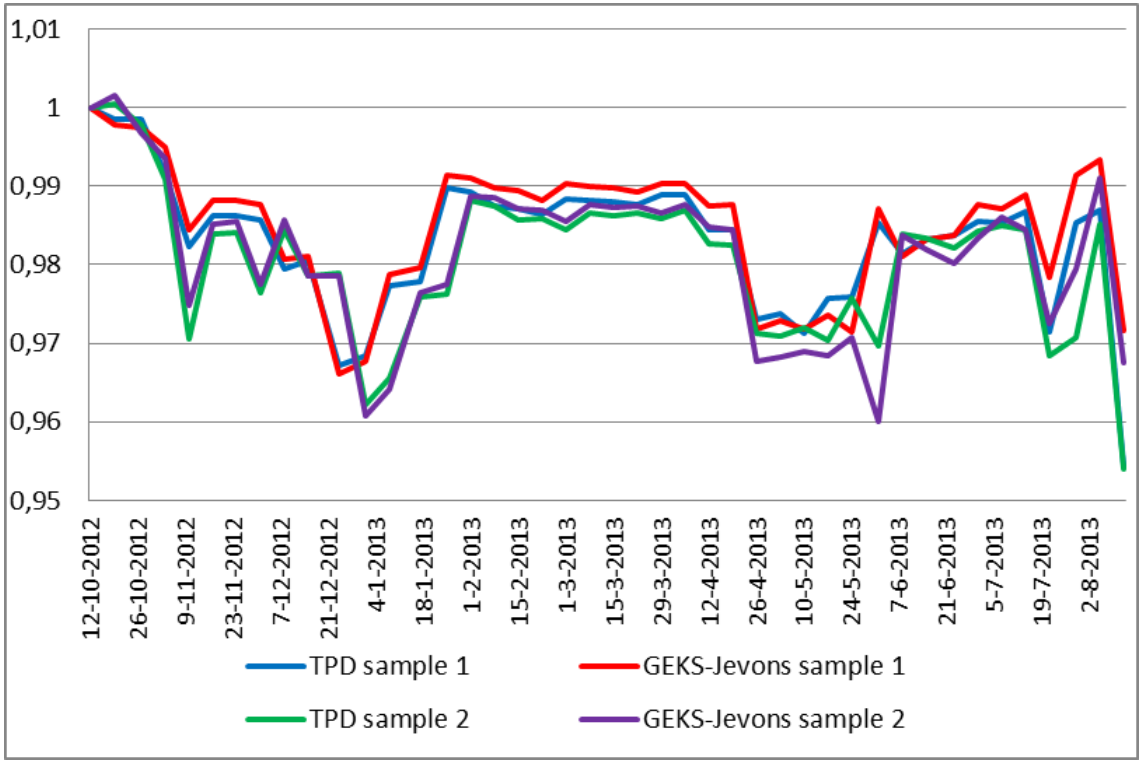
- GEKS Jevons does not fall as fast as TPD
- Only small differences between the two samples
- Drawing samples does not change the picture

‘Weekly’ indexes; men’s watches; large data set



- TPD and GEKS-Jevons very similar, as expected

‘Weekly’ indexes; kitchen appliances; large data set



- TPD and GEKS-Jevons very similar, as expected

Conclusions

- While fixed effects in TPD model can be viewed as item-specific hedonic effects, ….
- …. this does not mean that TPD produces a quality-adjusted index.
- Where quality change is unimportant: multilateral indexes (TPD, GEKS) preferred over period-on-period chained indexes.
- Regression-based TPD will be easier to estimate than GEKS.
- Potential problem: hidden price changes - identification issue.
- Weighted TPD or GEKS if quantity data is available, but ….
- …. quantity data for online purchases might be unreliable.