Towards a new paradigm for scanner data price indices: applying big data techniques to big data

Jens Mehrhoff
European Commission (Eurostat)
16th Meeting of the Ottawa Group
Rio de Janeiro, 8 – 10 May 2019
Preamble

• Me at the 15th Meeting of the Ottawa Group: ‘not more data are better, **better data are better**!’
  *A ‘big data’ gaze at why electronic transactions and web-scraped data are no panacea*

• Me at the 16th Meeting of the Ottawa Group: ‘scanner and web-scraped data are **better in measurable terms** – and worse, too!’

• Also me at this meeting: ‘Panacea’s potion: **dynamic factor models**’.
1. Introduction

- **Chaining price indices** at monthly frequency, say, can lead to **significant drift**; in order to overcome chain drift, **multilateral methods** have been proposed that are **by construction drift-free**.
- These methods are borrowed from the literature on **international purchasing power parity comparisons** and may not be tailored to the problem in **intertemporal comparisons**.
1. Introduction

• The present paper proposes a shift towards a new paradigm: a model-based procedure is derived that yields figures, which do no longer possess the classical formula interpretation.

• The new index series convey a similar information content in terms of the statistical signal but come with much lower noise than the classical concepts; this is exemplified using the Dominick’s Finer Foods data set (→poster).
2. Signal-noise ratio

• **How much (more) information** is contained in price indices based on **scanner or web-scraped data** compared to traditional methods?

• **Statistical decomposition** of price indices variation in **signal and noise** using structural time series models.

2. Signal-noise ratio

- The model **controls for sale periods** \((\delta)\) and **allows for deterministic trends** \((\beta)\):

\[
\ln P_{0,t} = y_t = \mu_t + \delta x_t + \epsilon_t
\]

\[
\mu_t = \mu_{t-1} + \beta + \eta_t
\]

- The explanatory variable for sale periods \((x_t)\) is the **share of products sold on a promotion**.
2. Signal-noise ratio

• The **signal-noise ratio** is \( q = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2} \) and the **goodness-of-fit measure** is \( R_U^2 = 1 - U^2 \), where \( U \) is Theil’s inequality measure (random walk).

• Using both the **weighted and unweighted time-product dummy (TPD) approach**, price index numbers are estimated.

• The latter is less affected by quantity increases due to price decreases – very much like **web-scraped data**.
Prices for bottled juice, Dominick's Finer Foods

Oct 1989 = 100, log scale

Memo item: Old CPI-U
Unweighted TPD
Weighted TPD

Share of products sold on a promotion


% 50

25
## 2. Signal-noise ratio

<table>
<thead>
<tr>
<th></th>
<th>Weighted TPD</th>
<th>Unweighted TPD</th>
<th>Old CPI-U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$\delta$</td>
<td></td>
<td>-0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.23</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>$\sigma^2_\varepsilon$</td>
<td>8.22</td>
<td>4.16</td>
<td>0.59</td>
</tr>
<tr>
<td>$\sigma^2_\eta$</td>
<td>1.43</td>
<td>1.67</td>
<td>1.07</td>
</tr>
<tr>
<td>$q$</td>
<td>0.17</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td>$R^2_U$</td>
<td>0.36</td>
<td>0.60</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Note: $p$-values in parentheses.
2. Signal-noise ratio

• **Modelling troughs in sale periods** $(\delta)$ greatly reduces noise (weighted TPD: $-49\%$) and increases signal $(+17\%)$ as well as $R^2_U$ $(+68\%)$.

• **Adding deterministic trends** $(\beta)$ amplifies noise $(+8\%)$ and dampens signal $(-30\%)$ without significant gain in the log-likelihood function.

• **Sales periods** have more than three times the effect on weighted TPD than on unweighted TPD; they are irrelevant for the old CPI-U.
2. Signal-noise ratio

- **Noise** ($\sigma_{\xi}^2$) in weighted TPD is **18-fold** that in unweighted TPD; it is not identifiable in the old CPI-U.
- **Signal** ($\sigma_{\eta}^2$) in weighted TPD is **1.5 times** as strong as in unweighted TPD; twice compared to the old CPI-U.
- **Signal-noise ratio** ($q$) of weighted TPD is **less than a twelfth** of that of unweighted TPD; the old CPI-U is over-smoothed.
3. Dynamic factor models

• **Time-product dummy model** \((\delta_0 = \gamma_N = 0)\):

\[
\ln p_{i,t} = \alpha + \sum_{t=1}^{T} \delta_t + \sum_{i=1}^{N-1} \gamma_i + \varepsilon_{i,t}
\]

• **Expenditure-share weighted TPD index**:

\[
P_{0,t} = \exp \hat{\delta}_t = \frac{\prod_{i \in S_t} (p_{i,t}/\exp \hat{\gamma}_i)^{s_{i,t}}}{\prod_{i \in S_0} (p_{i,0}/\exp \hat{\gamma}_i)^{s_{i,0}}}
\]
3. Dynamic factor models

- **Products stacked** into $N$-vector:

$$\ln p_t = \mathbf{i}_N \delta_t + \tilde{\mathbf{y}} + \varepsilon_t$$

$$\tilde{y}_i = \alpha + \gamma_i$$

- **Dynamic factor model** (DFM) with $K$ common trends:

$$\begin{bmatrix} y_t \end{bmatrix}_{N \times 1} = \begin{bmatrix} \Theta \\ \mu_t \end{bmatrix}_{N \times K} + \begin{bmatrix} \mu_0 \\ \varepsilon_t \end{bmatrix}_{N \times 1}$$
3. Dynamic factor models

- If $\mu_t$ **scalar** ($K = 1$) as well as $\Theta$ **restricted** to $\iota_N$:

  $$y_t = \iota_N \mu_t + \mu_0 + \varepsilon_t$$

- Then $y_t = \ln p_t$, $\mu_t = \delta_t$ and $\mu_0 = \tilde{\gamma}$:

  $$\ln p_t = \iota_N \delta_t + \tilde{\gamma} + \varepsilon_t$$
3. Dynamic factor models

- **Key difference:** TPD model estimates $\delta_t$ as independent time dummies; DFM uses structural time series modelling instead.

\[
\ln p_t = \iota_N \mu_t + \tilde{\gamma} + \varepsilon_t
\]

\[
\mu_t = \mu_{t-1} + \eta_t
\]
3. Dynamic factor models


1. Estimation of $\delta_t$ and $\tilde{\gamma}$ as well as $\Sigma_\epsilon$ (diagonal) by means of the TPD model.
2. Estimation of $\sigma_\eta^2$ by regressing $\delta_t$ onto its lags, i.e. *conditional on TPD estimates.*
3. Dynamic factor models

- Populate the **state-space model** with the estimates of $\tilde{\gamma}$, $\Sigma_\varepsilon$ and $\sigma_\eta^2$ (but not $\delta_t$!) and compute an improved estimate of $\mu_t$ using the Kalman smoother:

\[
\ln p_t - \tilde{\gamma} = u_N \mu_t + \hat{\Sigma}_\varepsilon^{1/2} \varepsilon_t^*
\]

\[
\mu_t = \mu_{t-1} + \hat{\sigma}_\eta \eta_t^*
\]
Prices for bottled juice, Dominick's Finer Foods

Oct 1989 = 100, log scale

Unweighted TPD
Kalman smoother
Weighted TPD
### 3. Dynamic factor models

<table>
<thead>
<tr>
<th></th>
<th>Weighted TPD</th>
<th>Unweighted TPD</th>
<th>Kalman smoother</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>( \delta )</td>
<td></td>
<td>-0.27</td>
<td>-0.27</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.23</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>( \sigma^2_\epsilon )</td>
<td>8.22</td>
<td>4.16</td>
<td>0.59</td>
</tr>
<tr>
<td>( \sigma^2_\eta )</td>
<td>1.43</td>
<td>1.67</td>
<td>1.07</td>
</tr>
<tr>
<td>( q )</td>
<td>0.17</td>
<td>0.40</td>
<td>1.83</td>
</tr>
<tr>
<td>( R^2_U )</td>
<td>0.36</td>
<td>0.60</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Note: \( p \)-values in parentheses.
3. Dynamic factor models

• Kalman smoother with θ restricted to \( \mu_N \) is essentially unweighted.

• Results are about the same as regards modelling troughs in sale periods and (not) adding deterministic trends vis-à-vis unweighted TPD.

• Noise can be reduced by a factor of 5\(1/2\) compared to unweighted TPD.

• Signal still is three-fifth of that of unweighted TPD.

• Signal-noise-ratio is more than three times that of unweighted TPD.
Postscript

• Kalman smoother can produce **substantial improvements in estimates** if the signal of the common component is **persistent** (so time averaging helps) and **small** (so substantial noise remains after cross-section averaging).

• **Work in progress:**
  - Expenditure-share weighted index
  - More refined time series model for $\mu_t$
  - Real-time performance (non-revisable)
  - Maximum-likelihood estimation, etc.
Contact

New e-mail address from early-August on:

jens.mehrhoff@bundesbank.de