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**'.

European Commission

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.
- Harvey, A.C. (1989), *Forecasting, structural time series Models and the Kalman filter*, Cambridge University Press: **local level (plus drift) model**.

2. Signal-noise ratio

- The model **controls for sale periods** (δ) and **allows for deterministic trends** (β):

$$\ln P_{0,t} = y_t = \mu_t + \delta x_t + \varepsilon_t$$

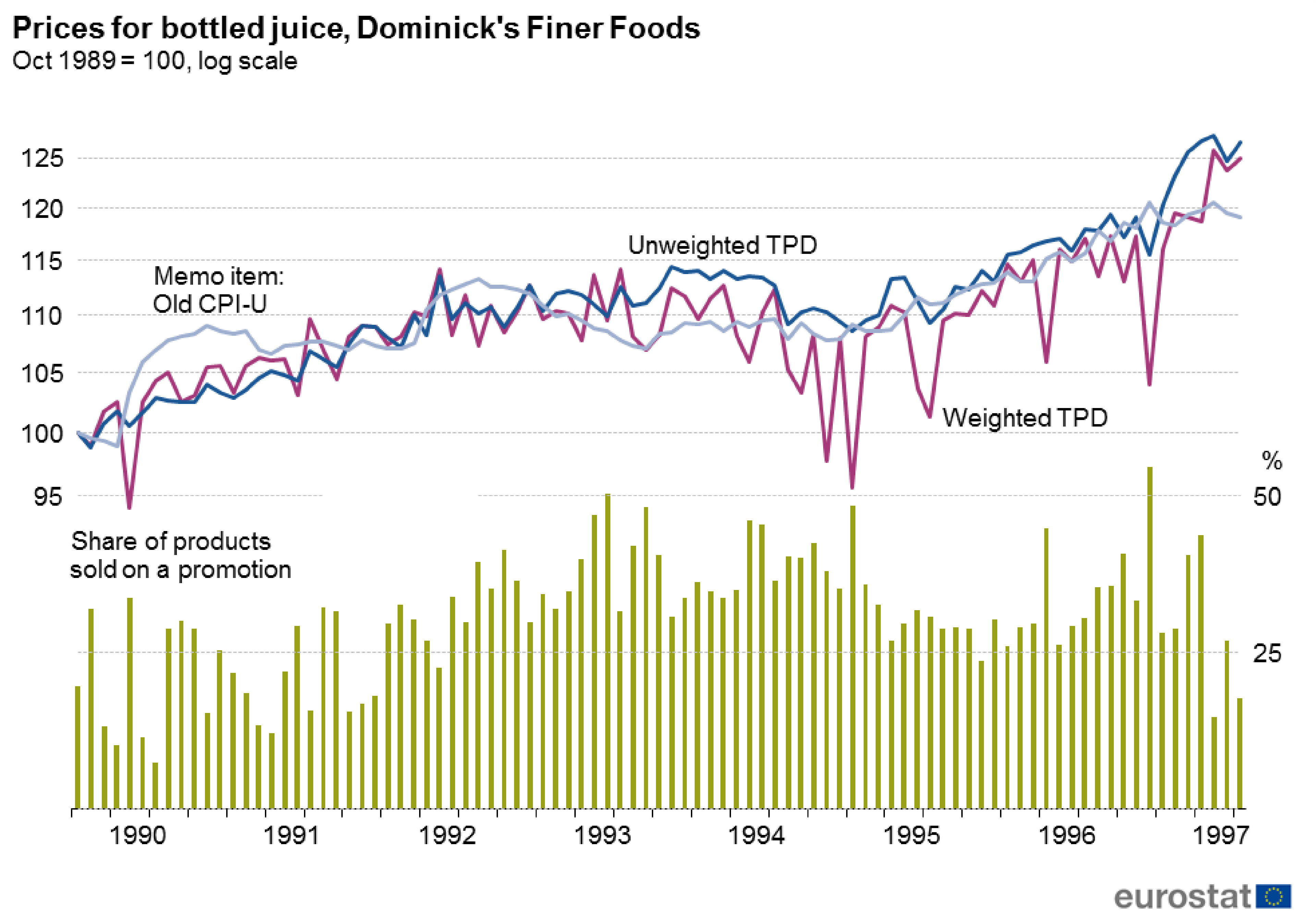
$$\mu_t = \mu_{t-1} + \beta + \eta_t$$

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

European Commission

2. Signal-noise ratio

- The **signal-noise ratio** is $q = \sigma_\eta^2/\sigma_\varepsilon^2$ 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
125
120
115
110
105
100
95
Memo item:
Old CPI-U
Unweighted TPD
Weighted TPD
%
50
25
Share of products
sold on a promotion
1990
1991
1992
1993
1994
1995
1996
1997
eurostat

2. Signal-noise ratio

	Weighted TPD				Unweighted TPD				Old CPI-U			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
δ			−0.27 (0.000)	−0.27 (0.000)			−0.08 (0.000)	−0.08 (0.000)			0.02 (0.046)	0.02 (0.041)
β		0.23 (0.054)		0.25 (0.053)		0.26 (0.009)		0.26 (0.012)		0.19 (0.051)		0.20 (0.045)
σ_{ε}^{2}	8.22 (0.000)	8.45 (0.000)	4.16 (0.000)	4.49 (0.000)	0.59 (0.003)	0.73 (0.001)	0.22 (0.085)	0.34 (0.025)	0.00	0.00	0.00	0.00
σ_{η}^{2}	1.43 (0.013)	1.01 (0.024)	1.67 (0.008)	1.17 (0.017)	1.07 (0.001)	0.74 (0.003)	1.12 (0.000)	0.82 (0.002)	0.92 (0.000)	0.88 (0.000)	0.88 (0.000)	0.84 (0.000)
q	0.17 (0.030)	0.12 (0.043)	0.40 (0.035)	0.26 (0.045)	1.83 (0.049)	1.01 (0.048)	5.00 (0.148)	2.40 (0.105)	∞	∞	∞	∞
R_{U}^{2}	0.36 (0.000)	0.38 (0.000)	0.60 (0.000)	0.61 (0.000)	0.08 (0.005)	0.15 (0.001)	0.33 (0.000)	0.37 (0.000)	0.00 (1.000)	0.04 (0.149)	0.04 (0.137)	0.08 (0.047)

Note: p-values in parentheses.

2. Signal-noise ratio

- **Modelling troughs in sale periods** (δ) greatly reduces noise (weighted TPD: −49%) and increases signal (+17%) as well as R_U^2 (+68%).
- **Adding deterministic trends** (β) 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** (σ_ε^2) in weighted TPD is **18-fold** that in unweighted TPD; it is not identifiable in the old CPI-U.
- **Signal** (σ_η^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** $(\delta_0 = \gamma_N = 0)$:

$$\ln p_{i,t} = \alpha + \underbrace{\delta_t}_{t=1,\dots,T} + \underbrace{\gamma_i}_{i=1,\dots,N-1} + \varepsilon_{i,t}$$

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

$$P_{0,t} = \exp \hat{\delta}_t = \frac{\prod_{i \in S_t} \left(p_{i,t} / \exp \hat{\gamma}_i \right)^{s_{i,t}}}{\prod_{i \in S_0} \left(p_{i,0} / \exp \hat{\gamma}_i \right)^{s_{i,0}}}$$

3. Dynamic factor models

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

$$\ln \mathbf{p}_t = \boldsymbol{\iota}_N \delta_t + \underbrace{\tilde{\boldsymbol{\gamma}}}_{\tilde{\gamma}_i = \alpha + \gamma_i} + \boldsymbol{\varepsilon}_t$$

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

$$\underbrace{\mathbf{y}_t}_{[N\times 1]} = \underbrace{\boldsymbol{\Theta}}_{[N\times K]} \underbrace{\boldsymbol{\mu}_t}_{[K\times 1]} + \underbrace{\boldsymbol{\mu}_0}_{[N\times 1]} + \underbrace{\boldsymbol{\varepsilon}_t}_{[N\times 1]}$$

3. Dynamic factor models

- If $\boldsymbol{\mu}_t$ **scalar** $(K = 1)$ as well as $\boldsymbol{\Theta}$ **restricted** to $\boldsymbol{\iota}_N$:

$$\mathbf{y}_t = \boldsymbol{\iota}_N \mu_t + \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}_t$$

- Then $\mathbf{y}_t = \ln \mathbf{p}_t$, $\mu_t = \delta_t$ and $\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\gamma}}$:

$$\ln \mathbf{p}_t = \boldsymbol{\iota}_N \delta_t + \tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}_t$$

European Commission

3. Dynamic factor models

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

$$\ln \mathbf{p}_t = \boldsymbol{\iota}_N \mu_t + \tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}_t$$

$$\mu_t = \mu_{t-1} + \eta_t$$

3. Dynamic factor models

- Stock, J.H., and Watson, M.W. (2011), 'Dynamic factor models,' in: Clements, M.P., and Hendry, D.F. (eds.), *The Oxford handbook of economic forecasting*, Oxford University Press: **third generation factor estimation**.

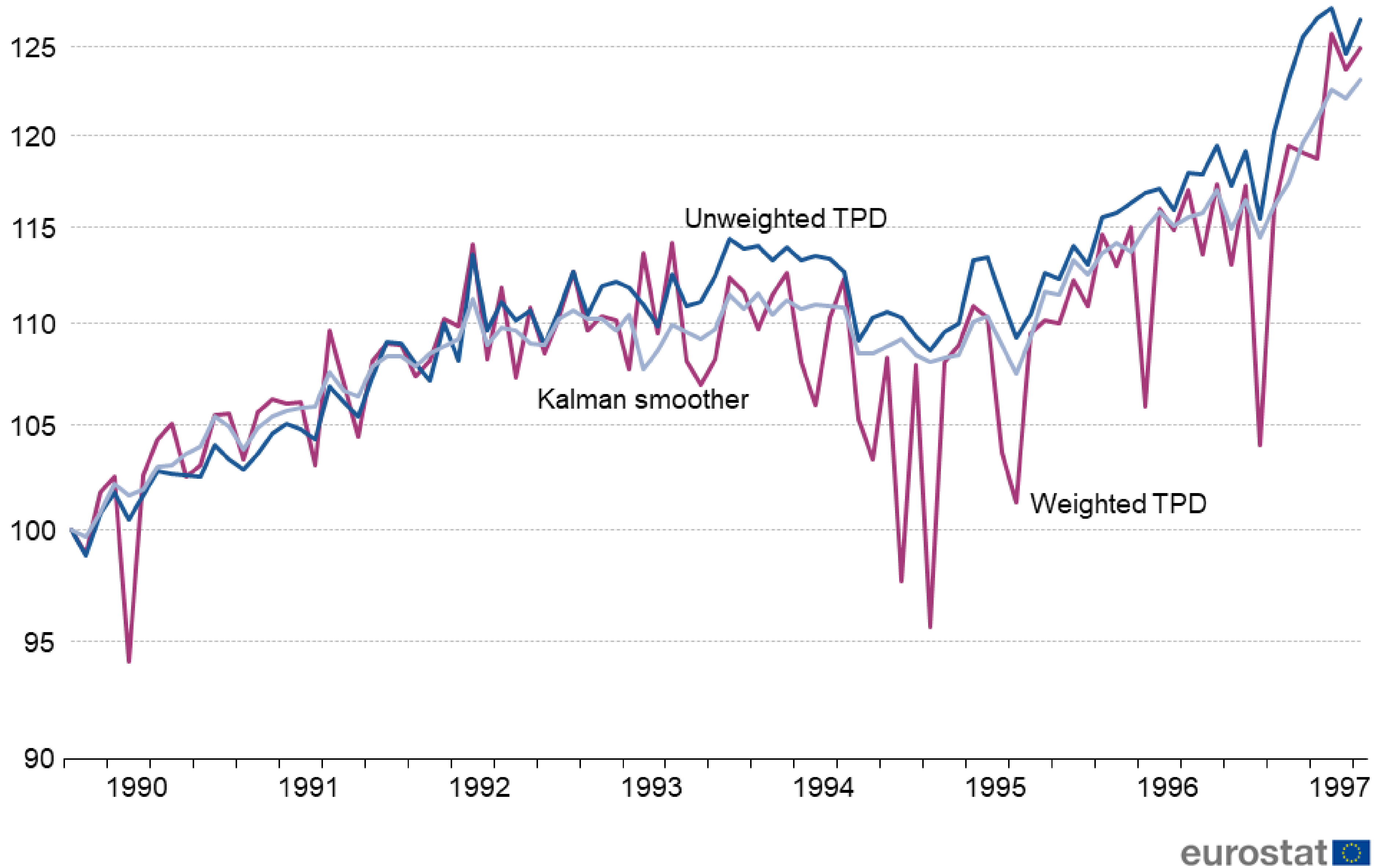
1. Estimation of δ_t and $\tilde{\boldsymbol{\gamma}}$ as well as $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ (diagonal) by means of the **TPD model**.
2. Estimation of σ_η^2 by **regressing** δ_t onto its lags, i.e. **conditional on TPD estimates**.

3. Dynamic factor models

- Populate the **state-space model** with the estimates of $\tilde{\boldsymbol{\gamma}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ and σ_η^2 (but not δ_t!) and compute an improved estimate of μ_t using the **Kalman smoother**:

$$\begin{aligned} \ln \mathbf{p}_t - \hat{\tilde{\boldsymbol{\gamma}}} &= \boldsymbol{\iota}_N \mu_t + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{1/2} \boldsymbol{\varepsilon}_t^* \\ \mu_t &= \mu_{t-1} + \hat{\sigma}_\eta \eta_t^* \end{aligned}$$

Prices for bottled juice, Dominick's Finer Foods
Oct 1989 = 100, log scale



3. Dynamic factor models

	Weighted TPD				Unweighted TPD				Kalman smoother			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
δ			−0.27 (0.000)	−0.27 (0.000)			−0.08 (0.000)	−0.08 (0.000)			−0.05 (0.000)	−0.05 (0.000)
β		0.23 (0.054)		0.25 (0.053)		0.26 (0.009)		0.26 (0.012)		0.23 (0.005)		0.23 (0.004)
σ_ε^2	8.22 (0.000)	8.45 (0.000)	4.16 (0.000)	4.49 (0.000)	0.59 (0.003)	0.73 (0.001)	0.22 (0.085)	0.34 (0.025)	0.12 (0.102)	0.19 (0.021)	0.04 (0.299)	0.11 (0.084)
σ_η^2	1.43 (0.013)	1.01 (0.024)	1.67 (0.008)	1.17 (0.017)	1.07 (0.001)	0.74 (0.003)	1.12 (0.000)	0.82 (0.002)	0.71 (0.000)	0.52 (0.001)	0.67 (0.000)	0.48 (0.000)
q	0.17 (0.030)	0.12 (0.043)	0.40 (0.035)	0.26 (0.045)	1.83 (0.049)	1.01 (0.048)	5.00 (0.148)	2.40 (0.105)	5.91 (0.159)	2.70 (0.088)	16.53 (0.317)	4.58 (0.152)
R_U^2	0.36 (0.000)	0.38 (0.000)	0.60 (0.000)	0.61 (0.000)	0.08 (0.005)	0.15 (0.001)	0.33 (0.000)	0.37 (0.000)	0.02 (0.203)	0.10 (0.009)	0.22 (0.000)	0.29 (0.000)

Note: p-values in parentheses.

3. Dynamic factor models

- **Kalman smoother** with Θ restricted to ι_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 5½** 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 μ_t
  - Real-time performance (non-revisable)
  - Maximum-likelihood estimation, etc.

Contact

New e-mail address from early-August on:

jens.mehrhoff@bundesbank.de

European Commission