A comparison of price index methods for scanner data

15th Meeting of the Ottawa Group
Eltville am Rhein, 9-12 May 2017

Antonio Chessa (CBS, CPI department)
Outline

• Use of electronic data in Dutch CPI
• Challenges when switching to scanner data
• Study set-up: Data, index methods and choice aspects
• Results
• Analysis
• Conclusions
• **Scanner data:**
  - Covers more than 20% in terms of Coicop weights
  - For supermarkets only scanner data are used since Jan. 2013

• **Internet prices:**
  - Different data collection tools have been developed
  - Web scraping: From 1 to 8 retailers/web shops in 2015-2017
  - Web scraping only for clothing, but we have plans to extend this to other types of products
Scanner data: Challenges

• **Price index calculation:**
  - Fixed baskets → dynamic populations:
    - How to include new products?
    - Relaunches: How to handle item replacements when processing 10-100,000 GTINs per retailer?
  - Expenditures at GTIN level:
    - Allow weighting within Elementary Aggregates
    - How to control for drift when weighting and including dynamics?

• **Implications for CPI process:**
  - Large data sets require top-down analysis of results
  - Periodic maintenance: Tracking of metadata changes by retailer
  - Efficient handling of different tasks
Recent developments in Dutch CPI

- Increased use of electronic data in past three years

- **Introduction of QU/GUV method (Geary-Khamis):**
  - Mobile phones (Jan. 2016)
  - Dutch department store chain (Jan. 2017)
  - Next (July 17 - Jan. 18): DIY stores and drugstores
  - Supermarkets are under study (GK vs current method/Jevons)

- **Research programme:**
  - How does GK compare with other methods?
  - Question is part of a 4-year programme at CBS
Comparative study

• **Scanner data:**
  - Dutch chain of department stores, 4-year period
  - 4 product groups: T-shirts, pastries, office supplies, bed clothing

• **Index methods:**
  - Bilateral: Jevons, Törnqvist, etc. (chained and direct)
  - Multilateral: GK, GEKS, TPD, hedonic

• **Choice aspects:**
  - Updating method (multilateral methods)
  - Length of time window (multilateral methods)
  - Product differentiation: by GTIN vs GTIN group (common char’s)
Results: 1. Weighting (GTIN level)
2. Bilateral vs Multilateral (GTIN level)

![Graphs showing Bilateral vs Multilateral (GTIN level) for different categories such as Bed clothing, Pastries, Men's T-shirts, and Office supplies. Each graph compares TPD-index FBME, Törnqvist MoM, and Törnqvist direct.]
3. Multilateral methods (GTIN groups)

Bed clothing

Pastries

Men's T-shirts

Office supplies
4. Updating method (GTIN groups)

Bed clothing

Pastries

Men's T-shirts

Office supplies

GK-index FBME  |  GK-index WS  |  GK-index MS
Analysis: 1. Bilateral methods

• Equal weighting may severely distort indices
• Chained methods may lead to severe drift
• Direct methods miss contributions from new products
• Problematic for dynamic populations
2. Multilateral methods

- GK and TPD give practically the same results

- *Hedonic vs TPD:*
  - TPD: Model parameters for combinations of characteristics
  - Traditional hedonic models: No interactions among attributes
  - Does this explain the differences between TPD and hedonic?

- *Differences GEKS method:*
  - Only for T-shirts. Exceptional case?
  - How does the GEKS relate to other methods?
Hedonic *without* interaction terms

**Note:**

In the above cases, hedonic models are applied to the full window of 50 months. In the preceding cases, time windows of 13 months were used.
Hedonic with interaction terms

- Pairwise interactions among item attributes are included in hedonic models
- Model fits based on AIC and BIC are improved, in spite of adding parameters
Rewriting the GEKS-Törnqvist

\[ P_{0,t} = \prod_{z=0}^{T} \left( \frac{P_{0,z}}{P_{t,z}} \right)^{\frac{1}{T+1}} = \prod_{z=0}^{T} \left( \frac{\prod_{i \in G} \left( \frac{p_{i,t}}{p_{i,z}} \right)^{\frac{s_{i,t} + s_{i,z}}{2}}}{\prod_{i \in G} \left( \frac{p_{i,0}}{p_{i,z}} \right)^{\frac{s_{i,0} + s_{i,z}}{2}}} \right)^{\frac{1}{T+1}} =: \tilde{p}_t / \tilde{p}_0 \]

\[ \tilde{p}_t = \left\{ \prod_{i \in G} \left( \frac{p_{i,t}}{v_i} \right)^{s_{i,t}} \right\}^{\frac{1}{T+1}} \left\{ \prod_{i \in G} \left( \frac{p_{i,t}}{v_i''} \right)^{\frac{1}{T+1} \sum_{z=0}^{T} s_{i,z}} \right\}^{\frac{1}{2}} \]

\[ v_i' = \prod_{z=0}^{T} p_{i,z}^{\frac{1}{T+1}} \quad v_i'' = \prod_{z=0}^{T} p_{i,z}^{s_{i,z} / \sum_{z=0}^{T} s_{i,z}} \]

Leads to a **downward bias** in cases with dump prices for disappearing items!
Sensitivity GEKS to choice of base month

Men's T-shirts (GEKS)

Men's T-shirts (GK)
More evidence for downward bias GEKS

Hair care

Oral care

TPD
GK
GEKS
After setting a dump price filter...

**Hair care**

**Oral care**

- TPD
- GK
- GEKS
Concluding remarks

• Consider inclusion of weights within EAs

• *Multilateral methods:*  
  - Fixed base updating methods are free of chain drift  
  - Cannot be excluded for splice methods  
  - 13-month period + fixed base month in line with CPI practice  
  - GK, TPD similar results; GEKS, CCDI sensitive to downward bias

• *High priority:*  
  - Product differentiation (relaunch problem)  
  - Text mining, web scraping, attribute selection  
  - Large data sets $\Rightarrow$ efficient handling of monthly work in CPI