Experimental clothing price indexes using Australian web scraped data

By Andrew Glassock

*Views expressed in this presentation are those of the author and do not necessarily represent those of the Australian Bureau of Statistics

Australian Bureau of Statistics
Informing Australia’s important decisions
Enhancing the CPI

- ABS in a ‘transformation’ environment
  - Opportunity to expand the use of ‘big data’ in official statistics

- CPI Enhancement Project since 2015
  - Multilateral methods for transactions/scanner data (2017)
  - CPI annual re-weighting (2018)
  - Web scraping/online price collection enhancements (ongoing)
Web scraping overview

- Web scraping – an automatic collection method which extracts and converts unstructured website data into structured data

- Web scraped prices progressively incorporated into the CPI since March 2017 – direct replacement strategy currently used

- CPI Enhancing Team has been investigating methods to better utilise online price data in the CPI since April 2018
## Web scraping overview

<table>
<thead>
<tr>
<th>Transactions/Scanner Data</th>
<th>Web scraped/Online Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ‘Census’ of products collected from each retailer</td>
<td>• ‘Census’ of products collected from each retailer</td>
</tr>
<tr>
<td>• Includes weekly expenditure and quantities for each product</td>
<td>• No expenditure or quantity information provided</td>
</tr>
<tr>
<td>• Products defined by stock keeping units</td>
<td>• Stock keeping units not currently scraped</td>
</tr>
</tbody>
</table>
High priority for ABS

Competitive market structure
  – How can the ABS maintain a representative sample?

High collection and data editing costs

Product life cycle effects (Melser and Syed, 2016)
  – Seasonal products with short product life cycles and frequent ‘relaunches’
Research questions

- How can we define individual products or *homogenous* product clusters?
- Can alternative data sources be used to weight products/clusters in the absence of expenditure and quantity information?
- Which index method should be used to aggregate products/clusters to derive elementary aggregate indexes?
  - Bilateral vs multilateral indexes
Product definitions are often too detailed

- Multiple descriptions may be assigned to the same product
- Severe product churn and the ‘relaunch problem’ (Chessa, 2016)
- Distinguishes between products which are identical to consumers (e.g. black and white variants of the same t-shirt)

Clustering products provides a solution to these challenges although increases the risk of unit value (average price) bias
# Web scraping example

<table>
<thead>
<tr>
<th>Date</th>
<th>Retailer</th>
<th>Category</th>
<th>Brand</th>
<th>Type</th>
<th>Characteristics</th>
<th>Description</th>
<th>Price</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Short Sleeves</td>
<td>Short Sleeve Regular T Shirt “Brand XYZ”</td>
<td>$55.00</td>
<td>1</td>
</tr>
<tr>
<td>05-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Short Sleeves</td>
<td>S/S Regular Tee Brand XYZ</td>
<td>$55.00</td>
<td>1</td>
</tr>
<tr>
<td>05-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Short Sleeves</td>
<td>Short Sleeved Oversized T-Shirt “Brand XYZ”</td>
<td>$55.00</td>
<td>1</td>
</tr>
<tr>
<td>05-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Long Sleeves</td>
<td>Long Sleeve T.S. “Brand XYZ”</td>
<td>$65.00</td>
<td>1</td>
</tr>
<tr>
<td>28-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Long Sleeves</td>
<td>L.S. Tee Shirt “Brand XYZ”</td>
<td>$65.00</td>
<td>1</td>
</tr>
<tr>
<td>28-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Short Sleeves</td>
<td>Short-Sleeve Reg T-Shirt “Brand XYZ”</td>
<td>$55.00</td>
<td>1</td>
</tr>
<tr>
<td>28-Jan-17</td>
<td>Retailer ABC</td>
<td>Women’s Tops</td>
<td>Brand XYZ</td>
<td>T-Shirt</td>
<td>Short Sleeves</td>
<td>Short Sleeved O/S Tee “Brand XYZ”</td>
<td>$55.00</td>
<td>1</td>
</tr>
</tbody>
</table>
Aggregation structure

Expenditure Class

Elementary aggregate

Homogenous products/clusters

Product Descriptions

Garments for Women
- Womens Dresses
- Womens T-Shirts
- Retailer X Brand Y Dress
- Retailer Y Brand X Dress
- Retailer Y Brand X T-Shirts
- Retailer X Brand Y Black Dress
- Retailer X Brand Y Blue Dress
How can we aggregate products in the absence of expenditure and quantity information?

Unweighted indexes (e.g. Jevons, OLS) are traditionally used

- Does not account for consumer substitution effects
- Evidence of stronger downward bias in the presence of life cycle effects

Weighted indexes (e.g. Tornqvist, WLS) using expenditure share proxies

- A number of studies/NSOs considering this strategy including Van Loon (2019), Antoniades (2017) and Chessa and Griffioen (2017).
Aggregation weights

- **ABS Retail Trade Survey (RTS)** - retailer sales data

- Two approaches used to disaggregate retailer sales to the product level

- **Option 1:** Household Expenditure Survey (HES) method
  - Retailer sales divided by elementary aggregate using HES
  - Elementary aggregates weights are consistent across retailers unless unavailable
  - Equal expenditure is assumed for products with the same retailer and elementary aggregate combination
Option 2: Scrape count method

- Number of products scraped used to proxy for quantities purchased
- Retailer sales split by elementary aggregate according to scrape count shares
- Scrape count shares for each retailer and elementary aggregate combination used to allow for unequal expenditure across products

Proxy weights are derived by dividing estimated product expenditure by total elementary aggregate expenditure across all retailers
Bilateral methods compare prices between two periods

- Fixed (direct) index:
  \[
P_{0,t} = \prod_{i \in S_M} \left( \frac{p_{i,t}}{p_{i,0}} \right)^{w_{i,0} + w_{i,t}}
  \]
  \( (1) \)

- Period-on-period chained (indirect) index:
  \[
P_{t-1,t} = \prod_{i \in S_M} \left( \frac{p_{i,t}}{p_{i,t-1}} \right)^{w_{i,t-1} + w_{i,t}}
  \]
  \( (2) \)
Fixed indexes

![Graphs showing indexes for Womens' T-Shirts and Mens' Dress Footwear from Jan-17 to Jul-19.](image-url)
Product churn problem

![Diagram showing match rates for different periods for Women's T-Shirts. The diagram includes three columns: Base = First Period, Base = Last Period, and Base = Previous Period. Each column shows a graph with data points for each month from January 2017 to July 2019.]
Chained indexes

Womens' T-Shirts

Mens' Dress Footwear

Index

Month

Australian Bureau of Statistics Informing Australia’s important decisions
Chained indexes

Australian Bureau of Statistics  Informing Australia’s important decisions
Multilateral methods

Multilateral methods compare prices between three or more periods.

Gini, Elteto, Koves and Szulc (GEKS) index:

\[
P_{0,t}^{GEKS} = \prod_{l=0}^{T} \left( \frac{P^{l,t}}{P^{l,0}} \right)^{1/T+1} = \prod_{l=0}^{T} \left( \frac{P^{0,l}}{P^{l,t}} \right)^{1/T+1} = \prod_{l=0}^{T} \left[ P^{0,l} \times P^{l,t} \right]^{1/T+1}
\]  

(Time dummy hedonic (TDH) index:

\[
\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{i,k} + \epsilon_i^t
\]

Mean splicing is used to extend the series once new periods become available.
Multilateral methods
Multilateral methods
Comparison of weighting approaches

The diagram compares the index values of women's T-shirts for two different approaches: GEKS and TDH. The index values are shown for each month from January 2017 to July 2019. The series are differentiated by the following lines:

- HES
- Scrape Counts
- CPI

The GEKS approach shows a more stable index value compared to the TDH approach, which exhibits more fluctuations. The CPI line is consistently higher than the HES and Scrape Counts lines, indicating a higher price index for women's T-shirts.
Comparison of weighting approaches

Mens' Dress Footwear

<table>
<thead>
<tr>
<th></th>
<th>GEKS</th>
<th>TDH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan-17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr-17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul-17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct-17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oct-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jan-19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apr-19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jul-19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Month

HES - Scrape Counts - CPI

Australian Bureau of Statistics Informing Australia's important decisions

05/12/2019
Expenditure class results

Garments for Women

![Graph showing expenditure for women over quarters]

Garments for Men

![Graph showing expenditure for men over quarters]

Index

Quarter


Australian Bureau of Statistics Informing Australia’s important decisions

05/12/2019
Expenditure class results

Footwear for Women

Footwear for Men

Index

Quarter


Index

Quarter


GEKS - TDH - CPI
Expenditure class results
Conclusions

- Pre-processing to form ‘clustered’ homogenous products is one viable strategy for NSOs to consider for ‘dynamic’ basket categories.

- Pooling data across retailers is one strategy to produce coherent and weighted aggregate price indexes.

- At the elementary level, our results exhibit downward drift for chained indexes.
Conclusions

- Annually fixed and multilateral indexes (homogenous cluster definitions) produced the most similar results to CPI indexes.
- Multilateral indexes our current preferred strategy for mitigating fixed and chained limitations.
- Future ABS work will focus on a quality framework for using web scraped data.
- ABS plan to release information paper during 2020 detailing framework for consultation.
Questions?