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

Australian Bureau of Statistics

- 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

Transactions/Scanner Data	Web scraped/Online Data
• 'Census' of products collected from each retailer • Includes weekly expenditure and quantities for each product • Products defined by stock keeping units	• 'Census' of products collected from each retailer • No expenditure or quantity information provided • Stock keeping units not currently scraped

Clothing and footwear

- 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’

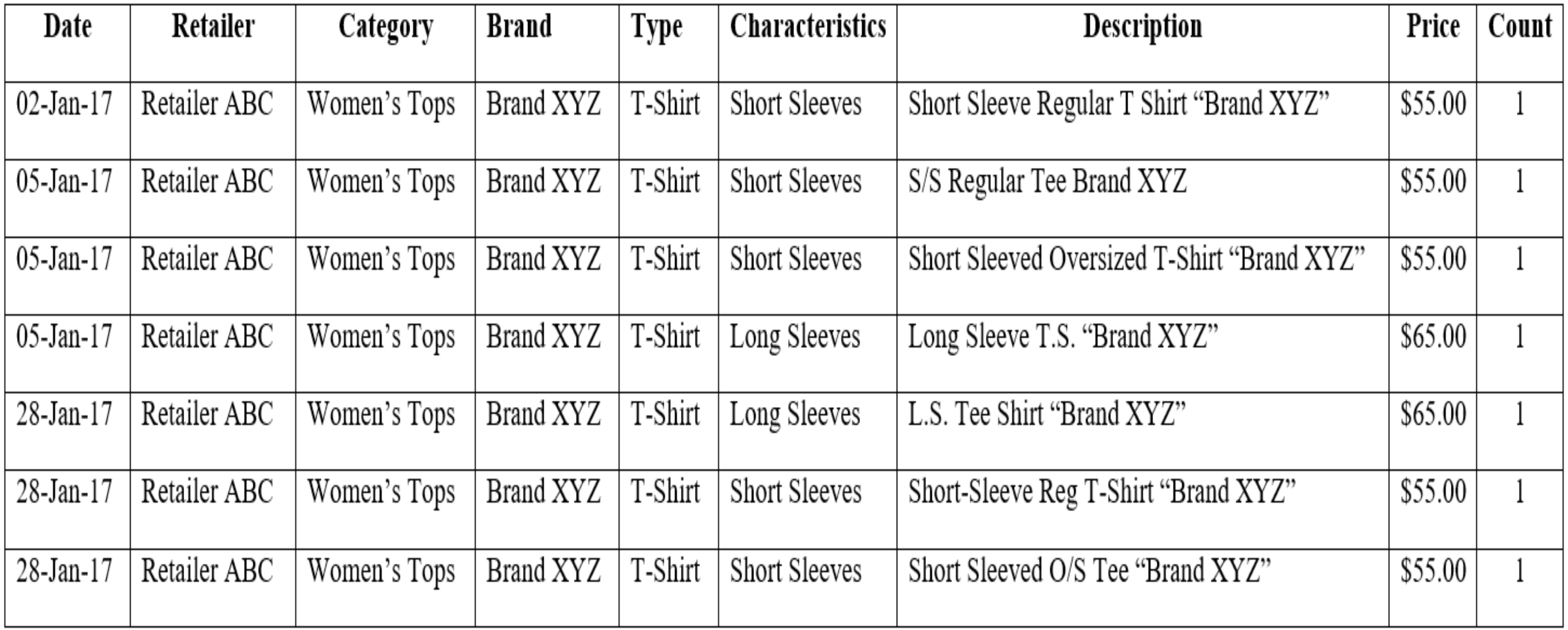
Research questions

- 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 definition

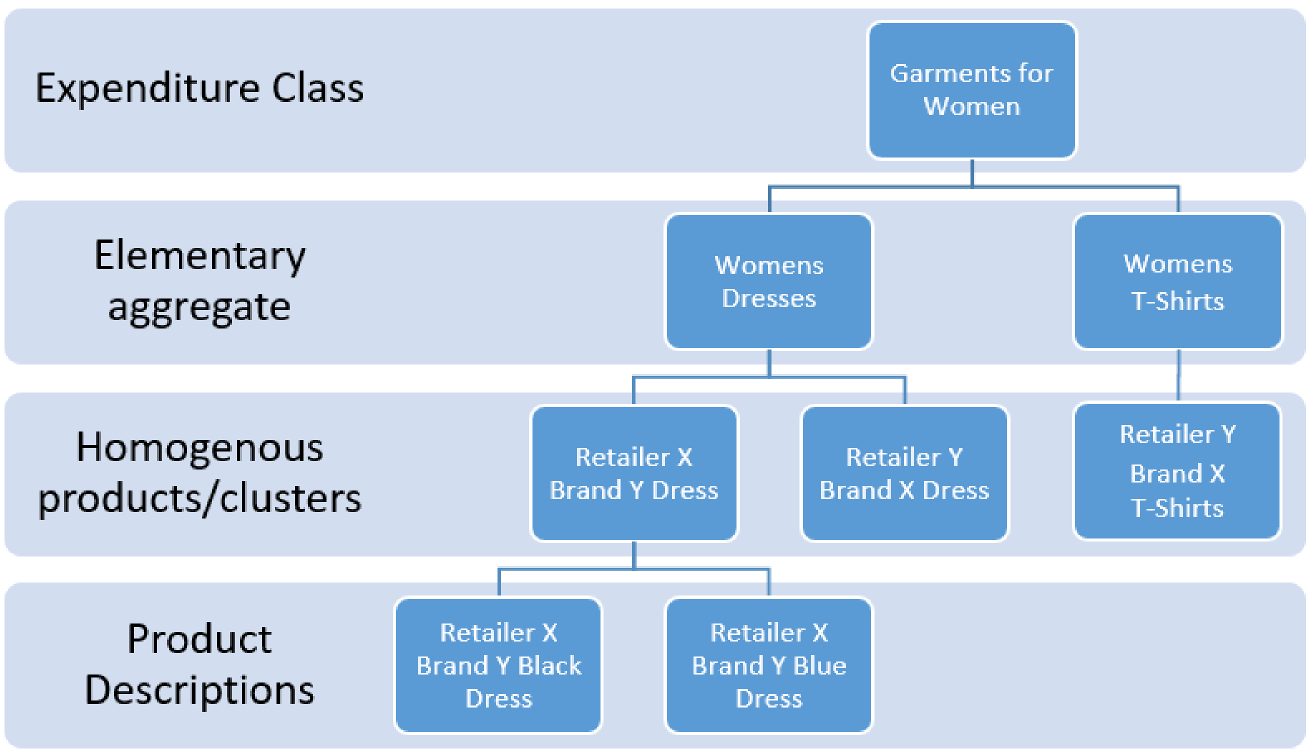
- Product descriptions are often too detailed
  - Multiple descriptions may be assigned to the same product
  - Severe product churn and the ‘relaunch problem’ (Chessa, 2016)
  - Distinguishes between products which are identical to consumers (e.g. black and white variants of the same t-shirt)
- Clustering products provides a solution to these challenges although increases the risk of unit value (average price) bias

Web scraping example



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

Aggregation structure



Expenditure Class

Garments for Women

Elementary aggregate

Womens Dresses

Womens T-Shirts

Homogenous products/clusters

Retailer X Brand Y Dress

Retailer Y Brand X Dress

Retailer Y Brand X T-Shirts

Product Descriptions

Retailer X Brand Y Black Dress

Retailer X Brand Y Blue Dress

Aggregation weights

- 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

Aggregation weights

- 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

- Bilateral methods compare prices between two periods

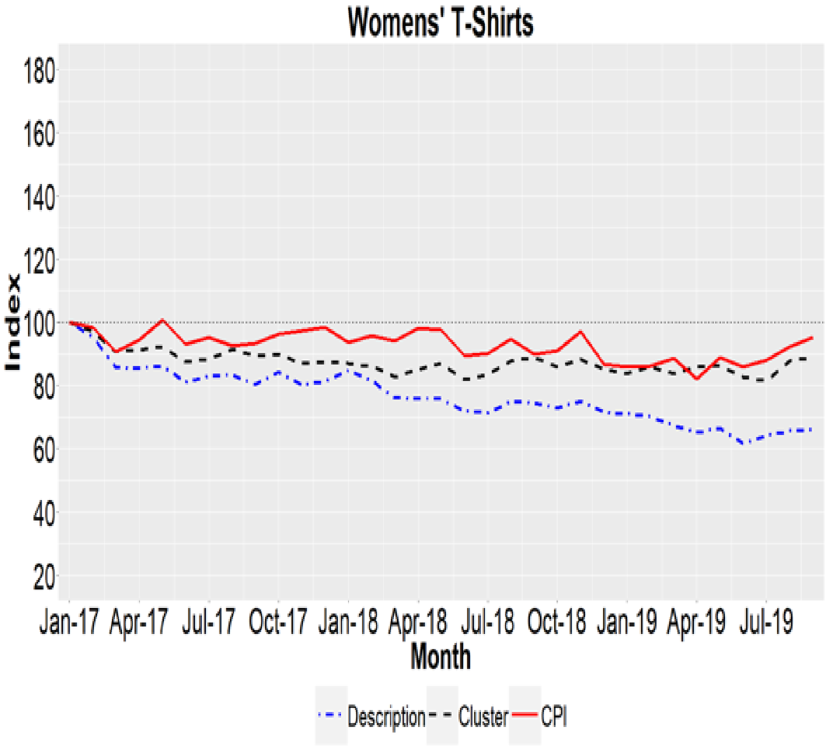
- Fixed (direct) index:

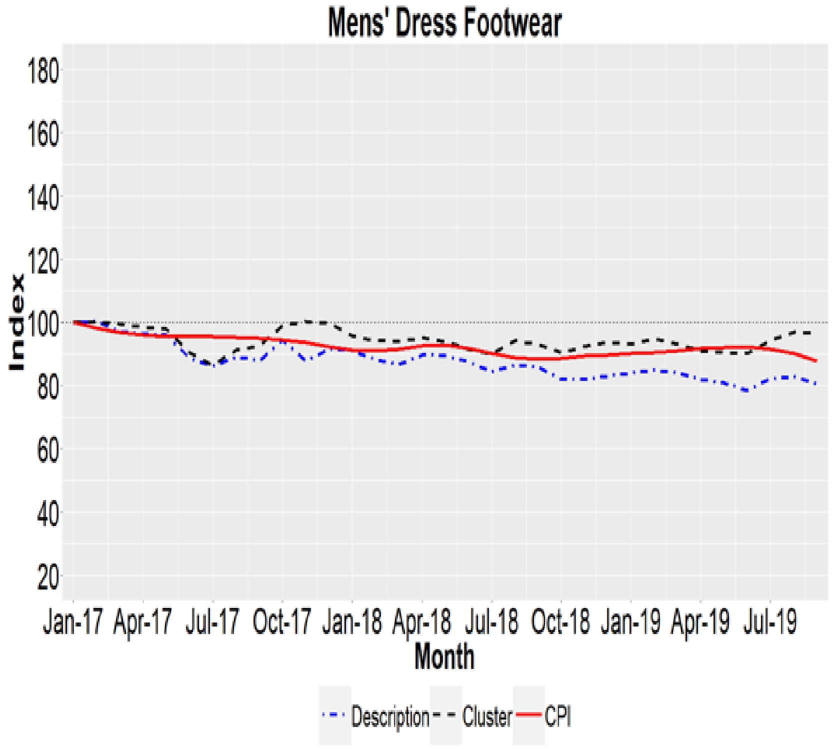
$$P_{0,t} = \prod_{i \in S_M} \left(\frac{p_{i,t}}{p_{i,0}} \right)^{\frac{w_{i,0} + w_{i,t}}{2}} \qquad (1)$$

- Period-on-period chained (indirect) index:

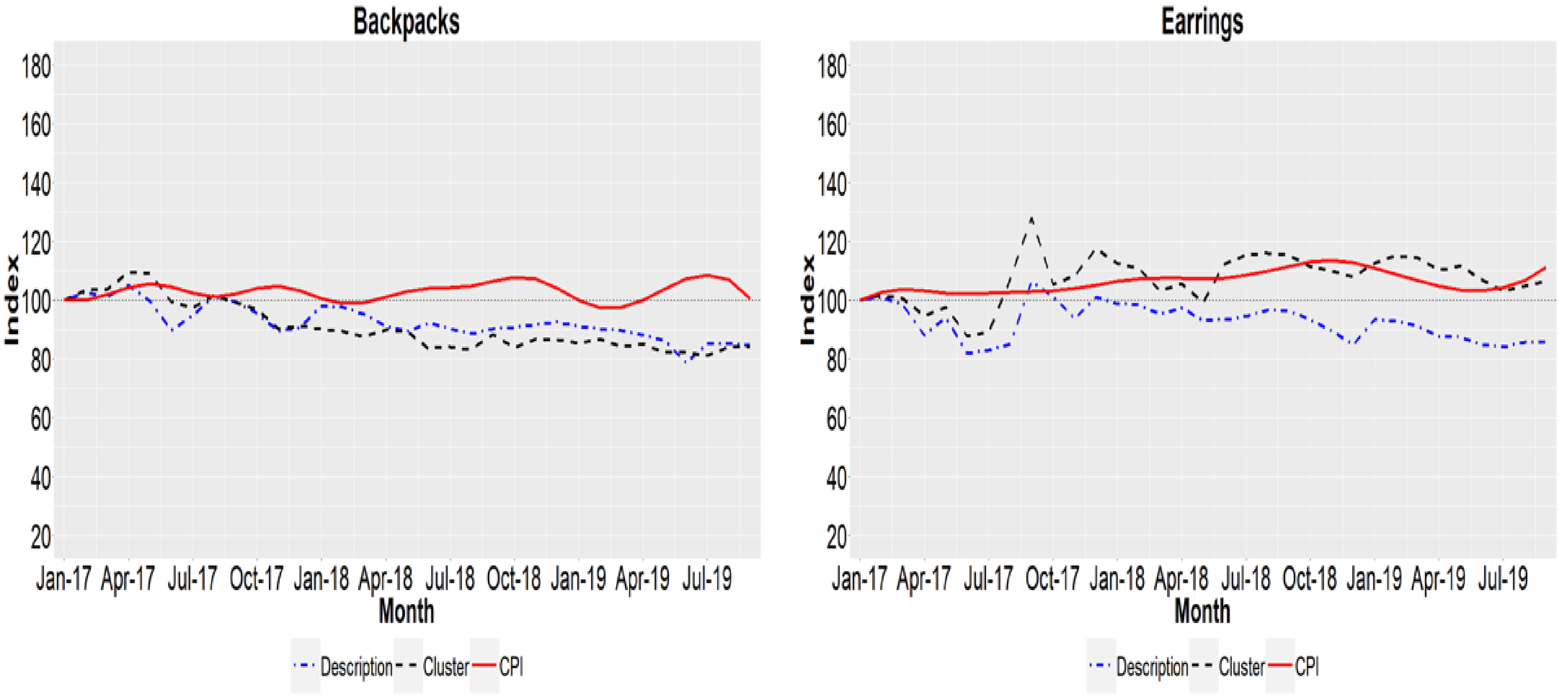
$$P_{t-1,t} = \prod_{i \in S_M} \left(\frac{p_{i,t}}{p_{i,t-1}} \right)^{\frac{w_{i,t-1} + w_{i,t}}{2}} \qquad (2)$$

Fixed indexes

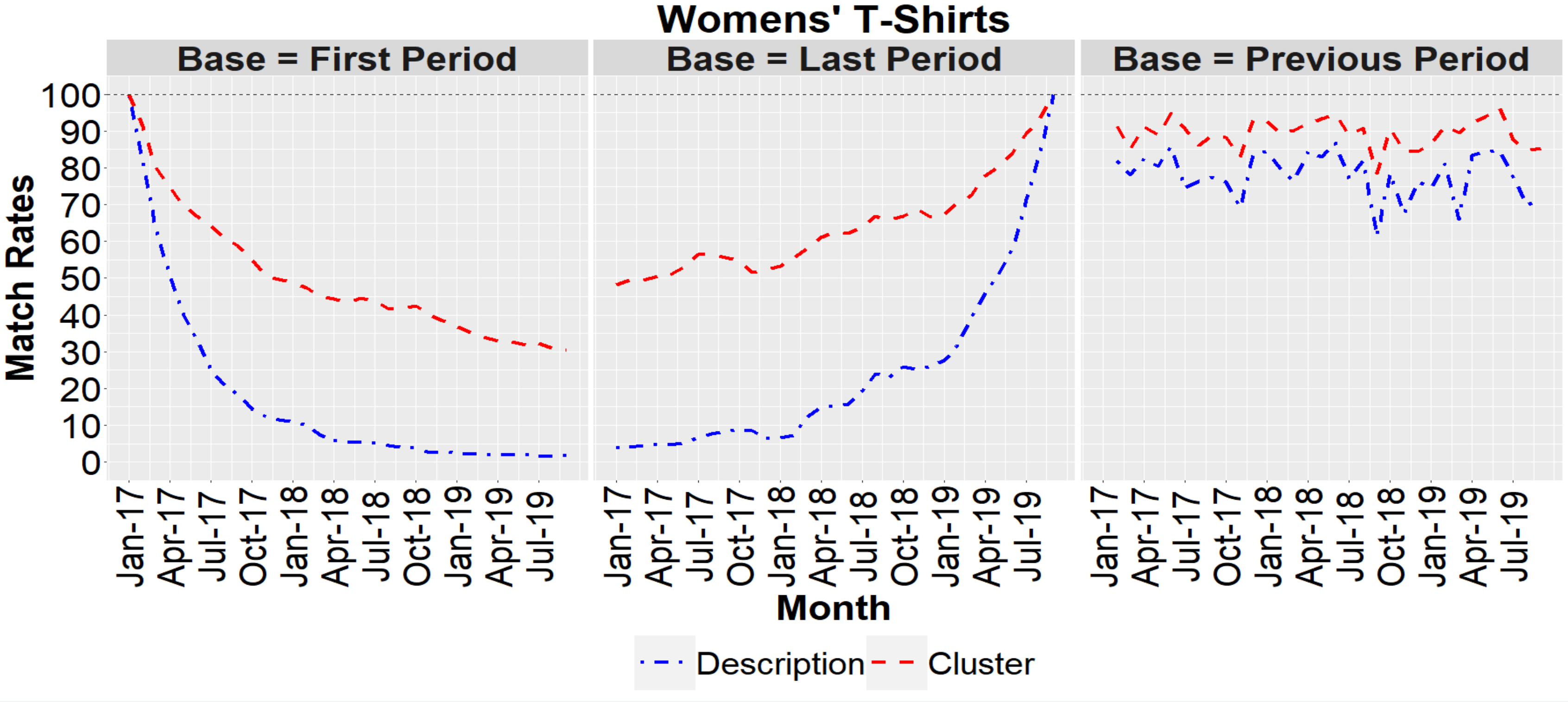




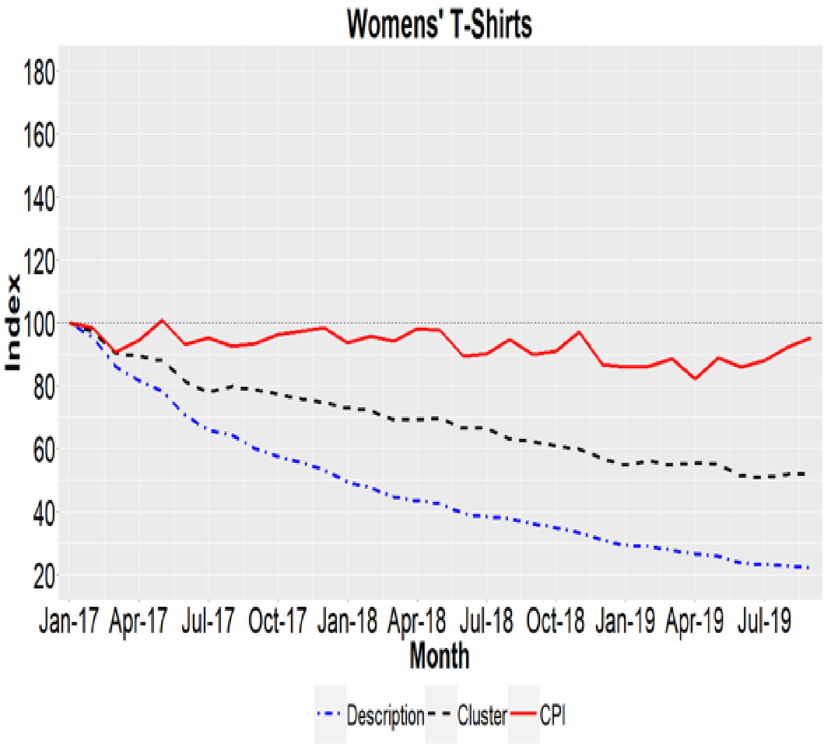
Fixed indexes

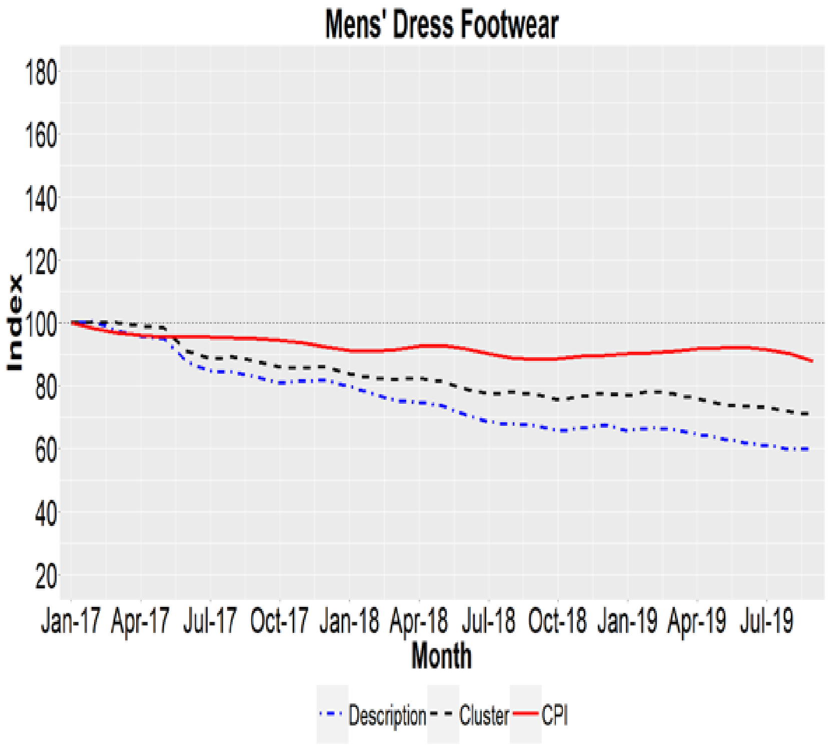


Product churn problem

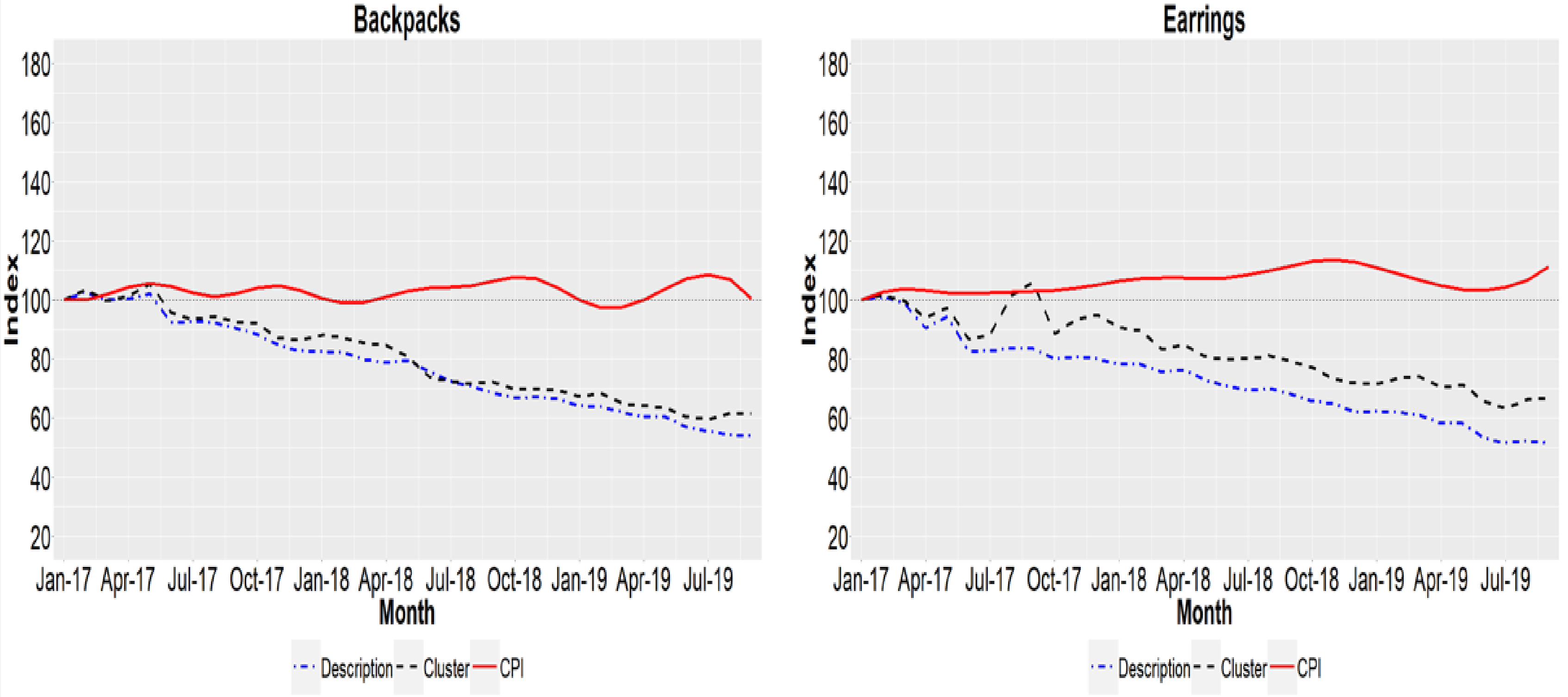


Chained indexes





Chained indexes



Multilateral methods

- Multilateral methods compare prices between three or more periods

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

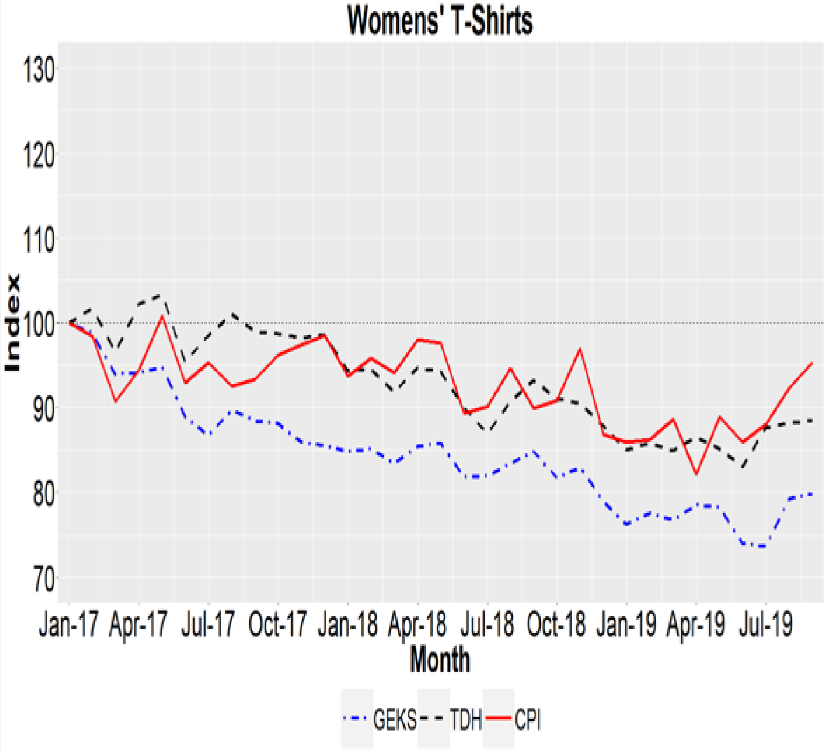
$$P_{0,t}^{GEKS} = \prod_{l=0}^{T}\left[\frac{P^{l,t}}{P^{l,0}}\right]^{\frac{1}{T+1}} = \prod_{l=0}^{T}\left[\frac{P^{0,l}}{P^{t,l}}\right]^{\frac{1}{T+1}} = \prod_{l=0}^{T}\left[P^{0,l} \times P^{l,t}\right]^{\frac{1}{T+1}} \qquad (3)$$

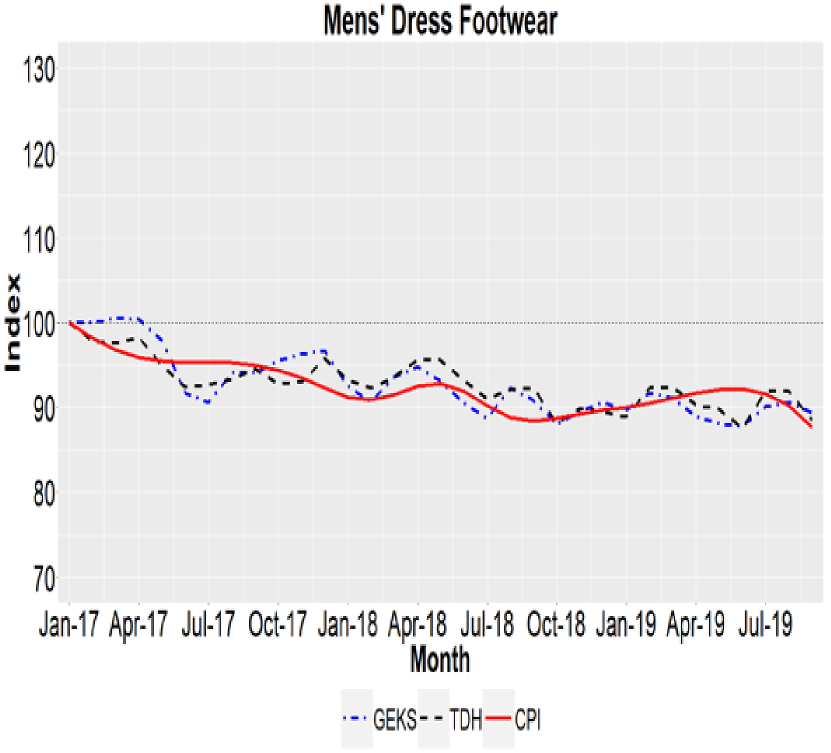
- Time dummy hedonic (TDH) index:

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k \, z_{i,k} + \epsilon_i^t \qquad (4)$$

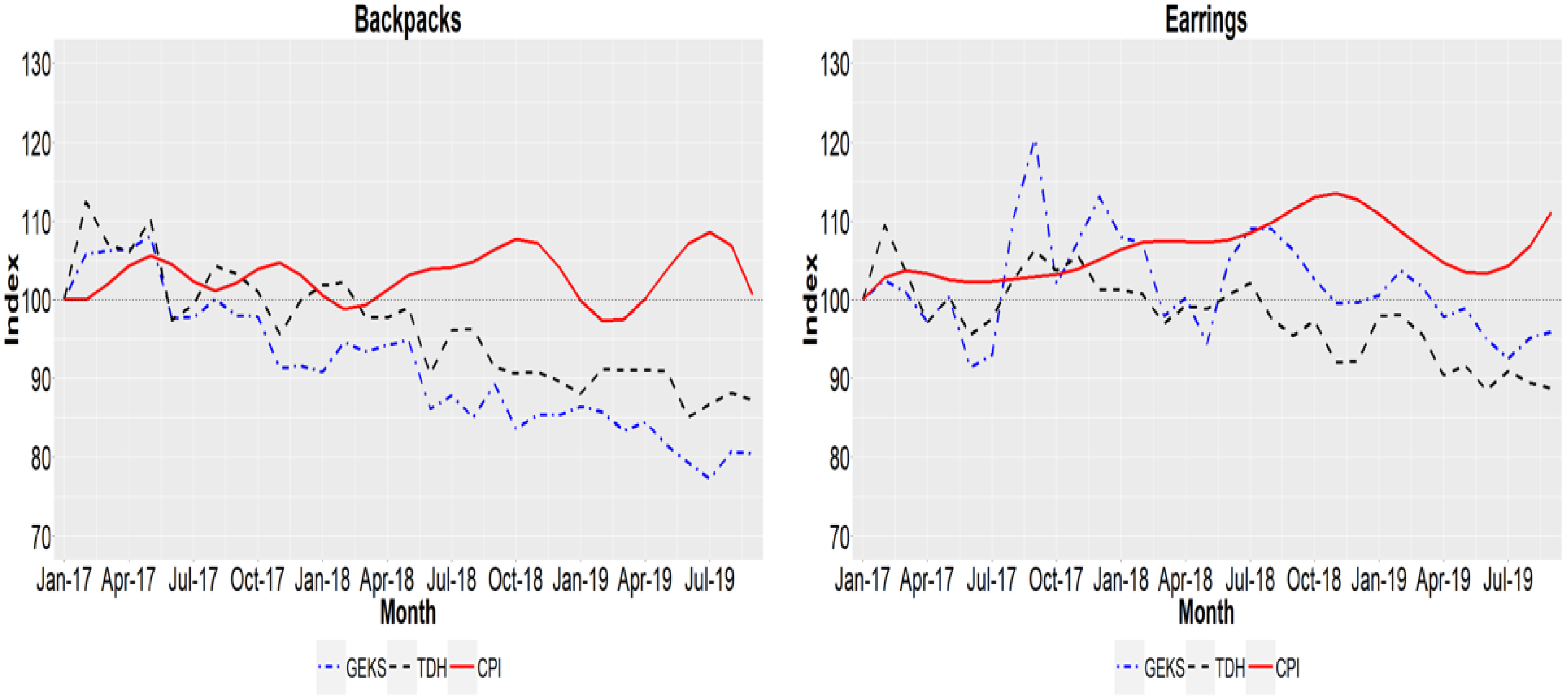
- Mean splicing is used to extend the series once new periods become available

Multilateral methods

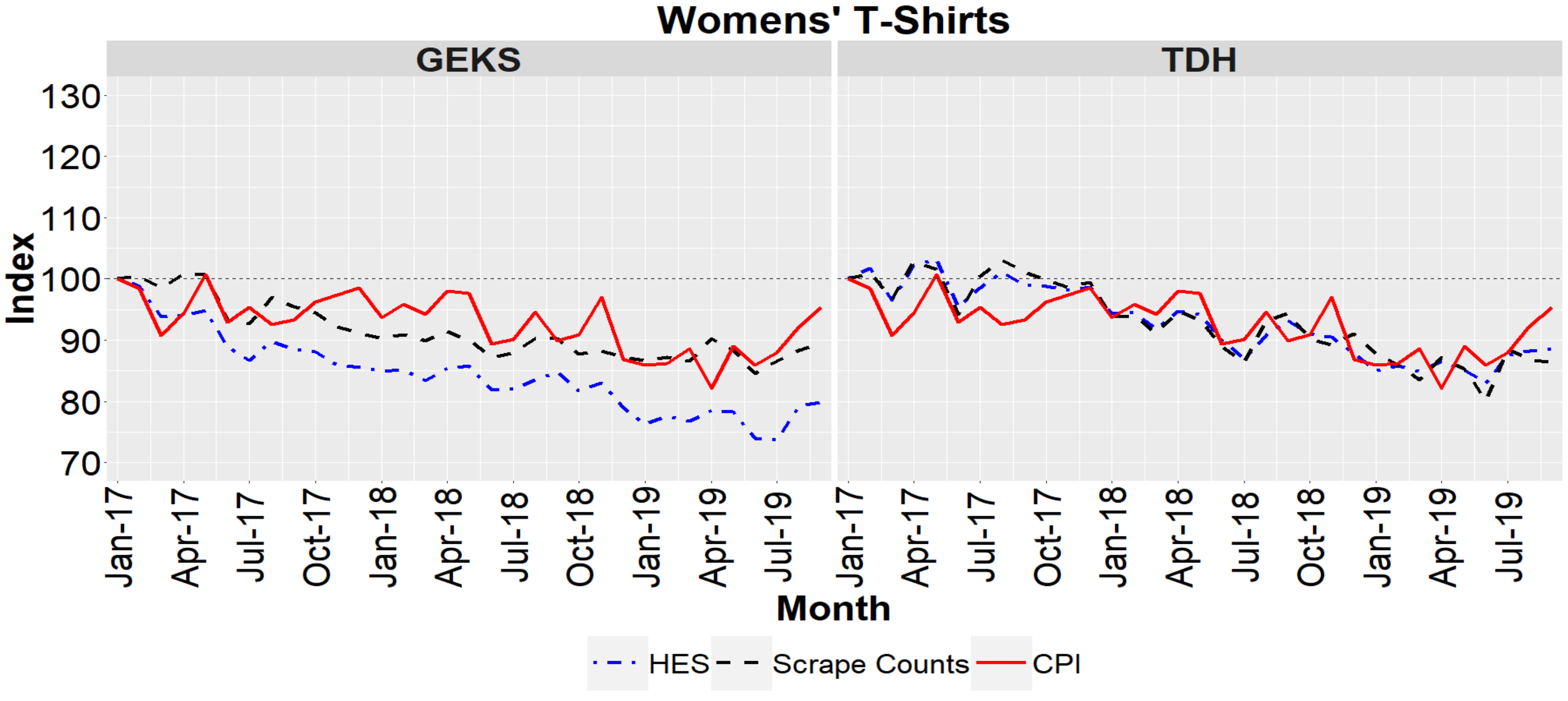




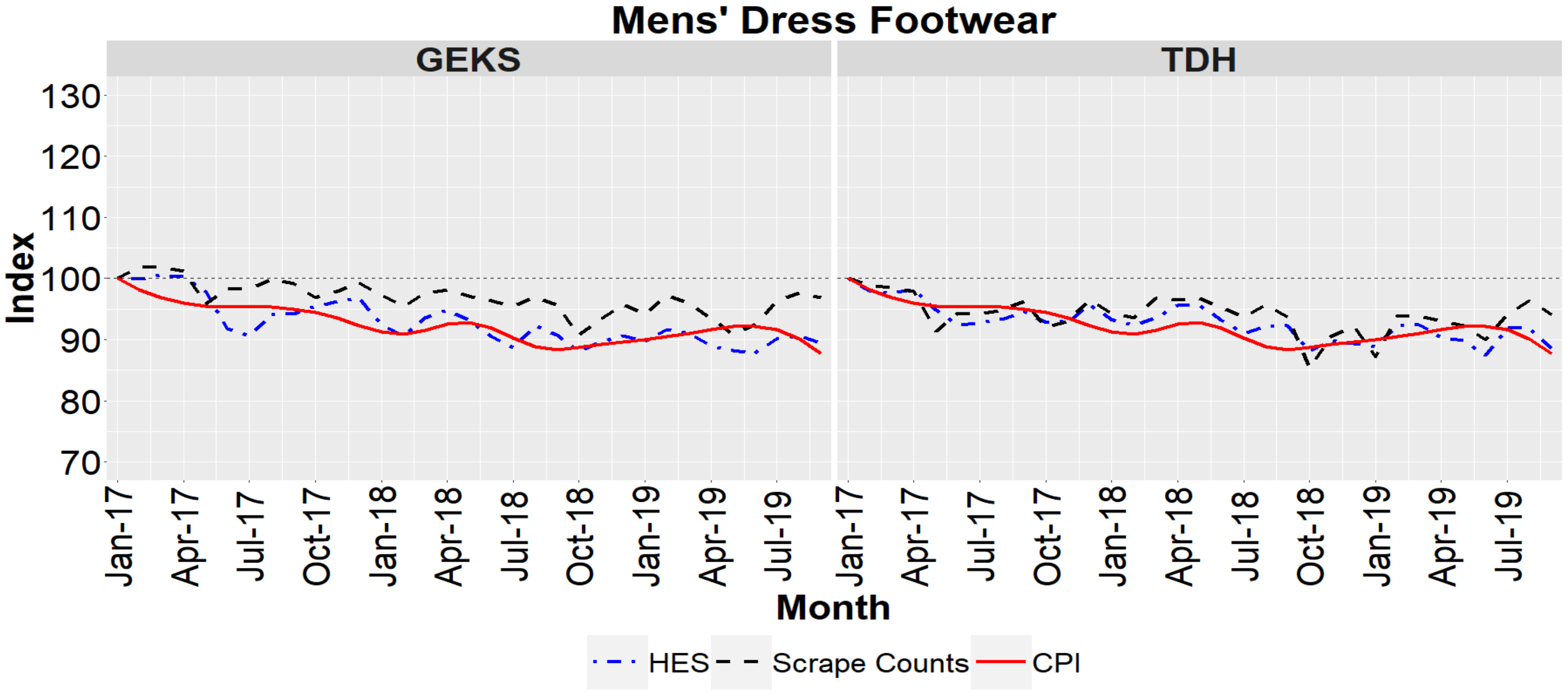
Multilateral methods



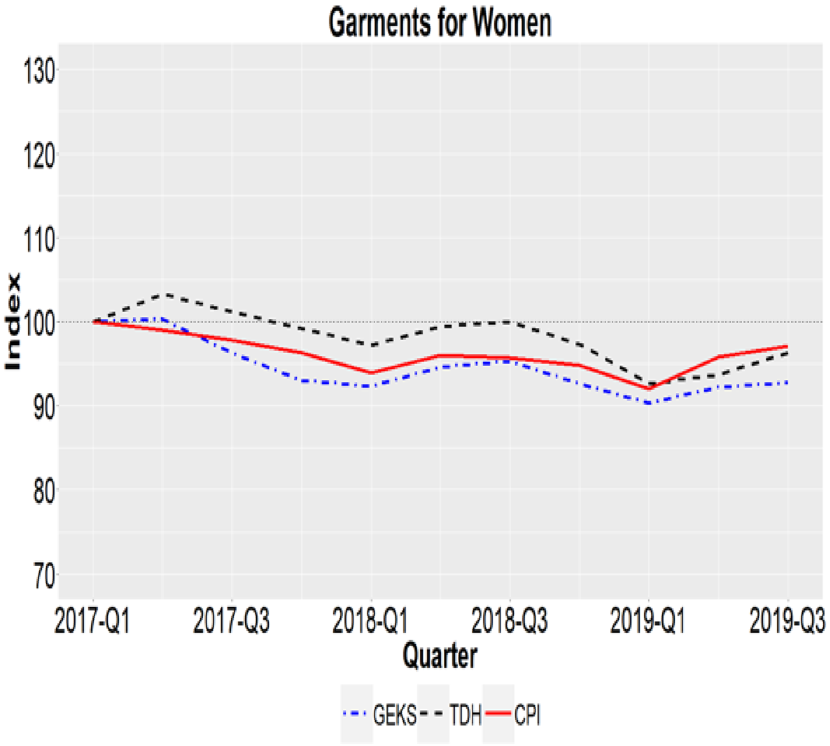
Comparison of weighting approaches

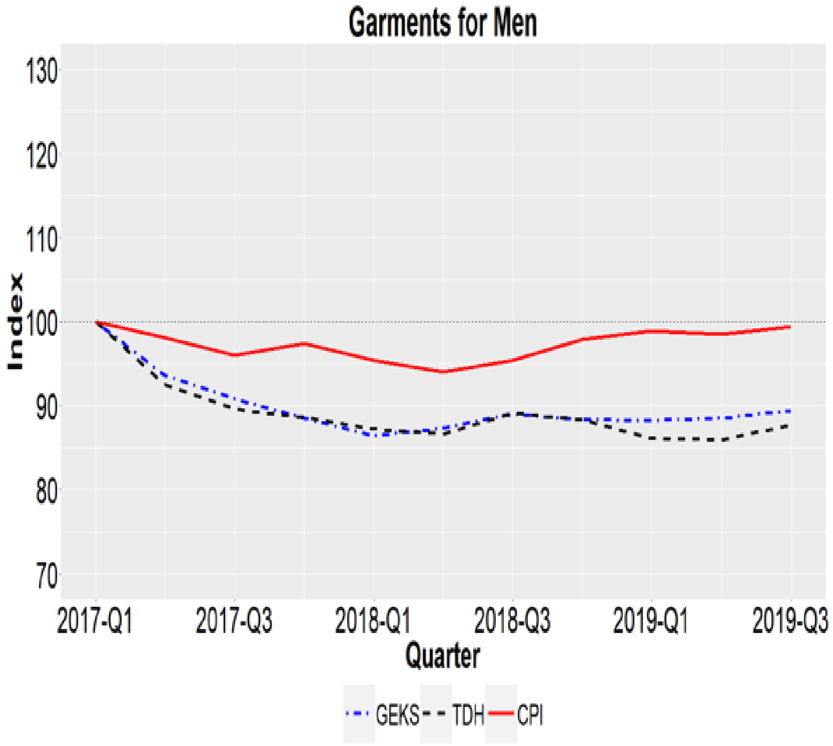


Comparison of weighting approaches



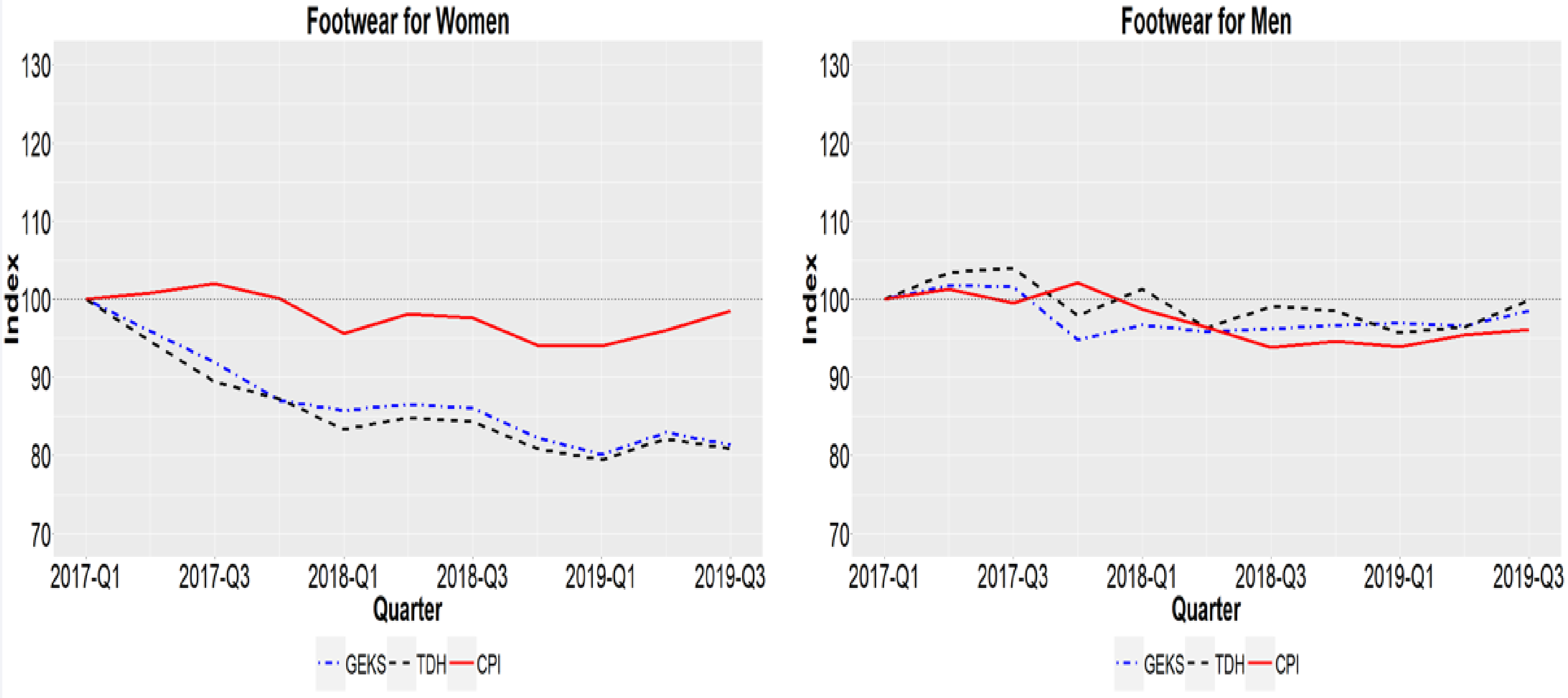
Expenditure class results



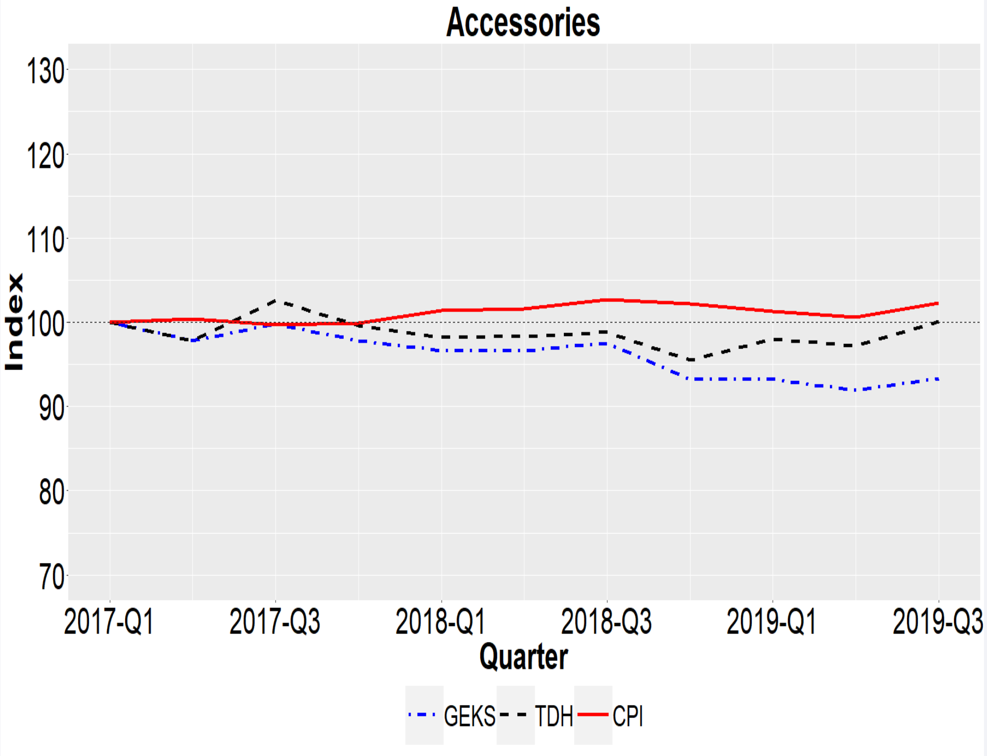


Expenditure class results

Australian Bureau of Statistics



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

Australian Bureau of Statistics

Questions?