Experimental clothing indexes using Australian web scraped data

Presented by Andrew Tomadini (Prices Branch Program Manager)

Australian Bureau of Statistics

Informing Australia's important decisions

Background

- ABS in a transformation environment – seeking ways to utilise 'big data' for compilation of economic statistics
- March quarter 2014 – Transactions (scanner) data introduced into the Australian CPI
- December quarter 2017 – Expansion of transactions data and introduction of multilateral index methods

Background

- What alternative big data sources are available to obtain price information?

- Web scraping – the extraction and transformation of unstructured data from the web into structured data

- The ABS is currently expanding its use of web scraped data in the CPI
  - Progressively incorporated since March 2017

Background

- Clothing and footwear – high priority for ABS
- Challenges with clothing and footwear:
  - High collection and data editing costs
  - Competitive market environment
  - Strong seasonality

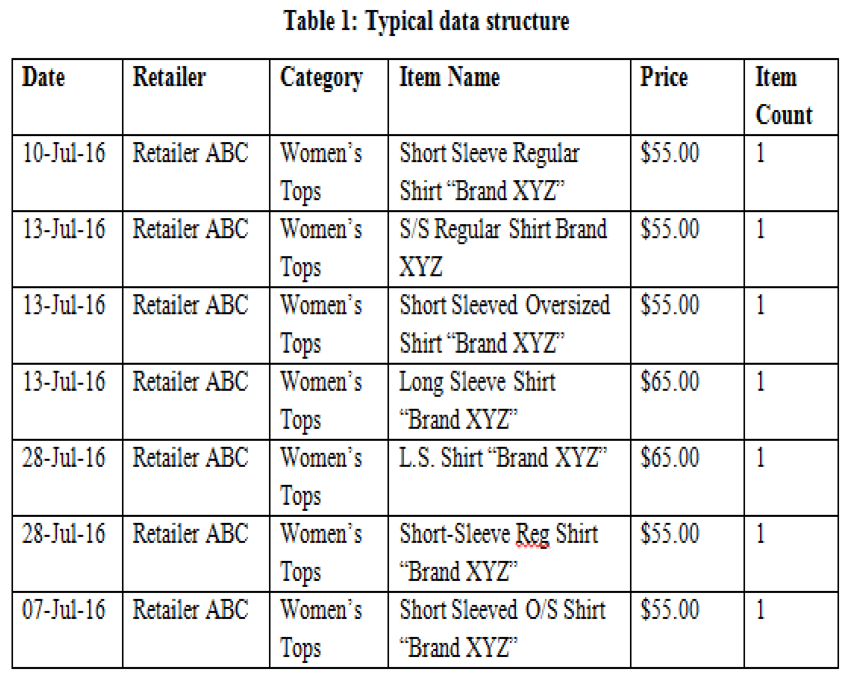


Table 1: Typical data structure

Date	Retailer	Category	Item Name	Price	Item Count
10-Jul-16	Retailer ABC	Women's Tops	Short Sleeve Regular Shirt "Brand XYZ"	$55.00	1
13-Jul-16	Retailer ABC	Women's Tops	S/S Regular Shirt Brand XYZ	$55.00	1
13-Jul-16	Retailer ABC	Women's Tops	Short Sleeved Oversized Shirt "Brand XYZ"	$55.00	1
13-Jul-16	Retailer ABC	Women's Tops	Long Sleeve Shirt "Brand XYZ"	$65.00	1
28-Jul-16	Retailer ABC	Women's Tops	L.S. Shirt "Brand XYZ"	$65.00	1
28-Jul-16	Retailer ABC	Women's Tops	Short-Sleeve Reg Shirt "Brand XYZ"	$55.00	1
07-Jul-16	Retailer ABC	Women's Tops	Short Sleeved O/S Shirt "Brand XYZ"	$55.00	1

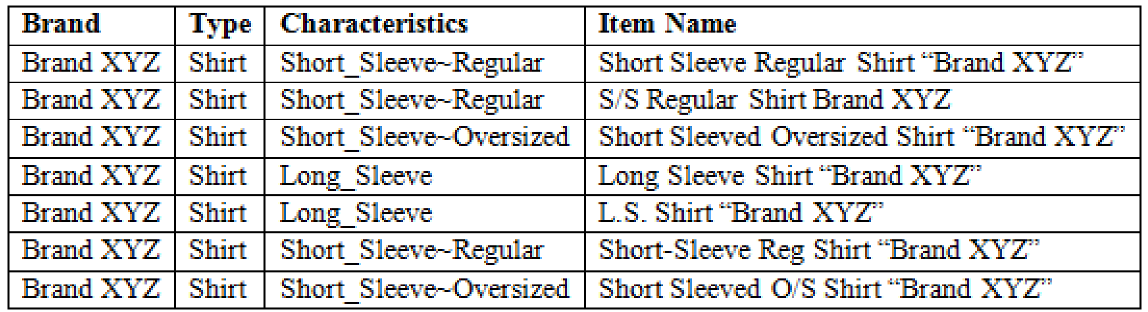
Product Definition

- Matched model indexes (e.g. Jevons, Törnqvist) rely on the ability of price analysts to identify which items are identical (i.e. homogenous) from the consumer's perspective
- Broader product definitions improve product matching over time but increase the risk of average price bias
- 3 alternative product definitions considered:
  - Item Name
  - Brand + Product Type
  - Brand + Product Type + Product Characteristics

Product Extraction

- A keyword approach was used to extract potentially important product information from item name strings

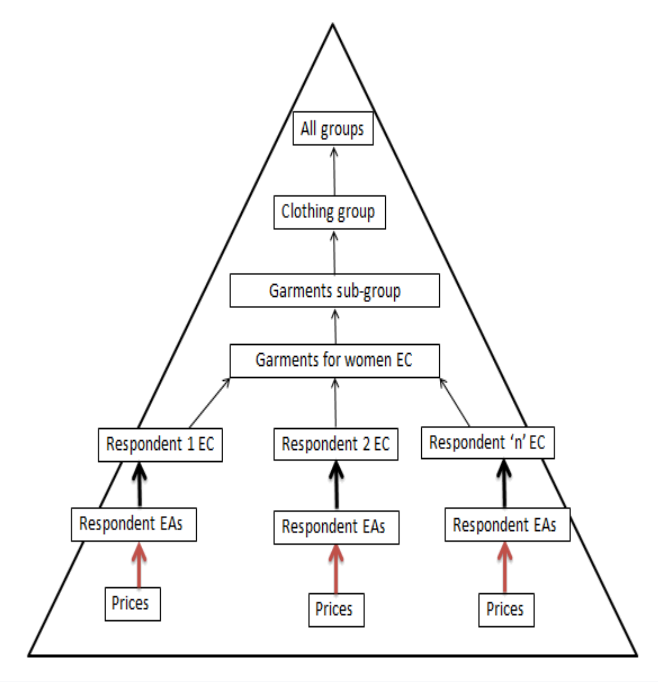
- Product information extracted included:
  - Brand
  - Product Type (e.g. t-shirt, dress, shorts)
  - Product characteristics (e.g. sleeve length, material, length)



Brand	Type	Characteristics	Item Name
Brand XYZ	Shirt	Short_Sleeve~Regular	Short Sleeve Regular Shirt “Brand XYZ”
Brand XYZ	Shirt	Short_Sleeve~Regular	S/S Regular Shirt Brand XYZ
Brand XYZ	Shirt	Short_Sleeve~Oversized	Short Sleeved Oversized Shirt “Brand XYZ”
Brand XYZ	Shirt	Long_Sleeve	Long Sleeve Shirt “Brand XYZ”
Brand XYZ	Shirt	Long_Sleeve	L.S. Shirt “Brand XYZ”
Brand XYZ	Shirt	Short_Sleeve~Regular	Short-Sleeve Reg Shirt “Brand XYZ”
Brand XYZ	Shirt	Short_Sleeve~Oversized	Short Sleeved O/S Shirt “Brand XYZ”

Aggregation Structure

- ABS currently aggregates clothing and footwear products across retailers to derive elementary aggregates (EAs)
- This presentation instead aggregates products to EAs within each retailer
- Aggregation across retailers is carried out at the Expenditure Class (EC) level



Multilateral Methods

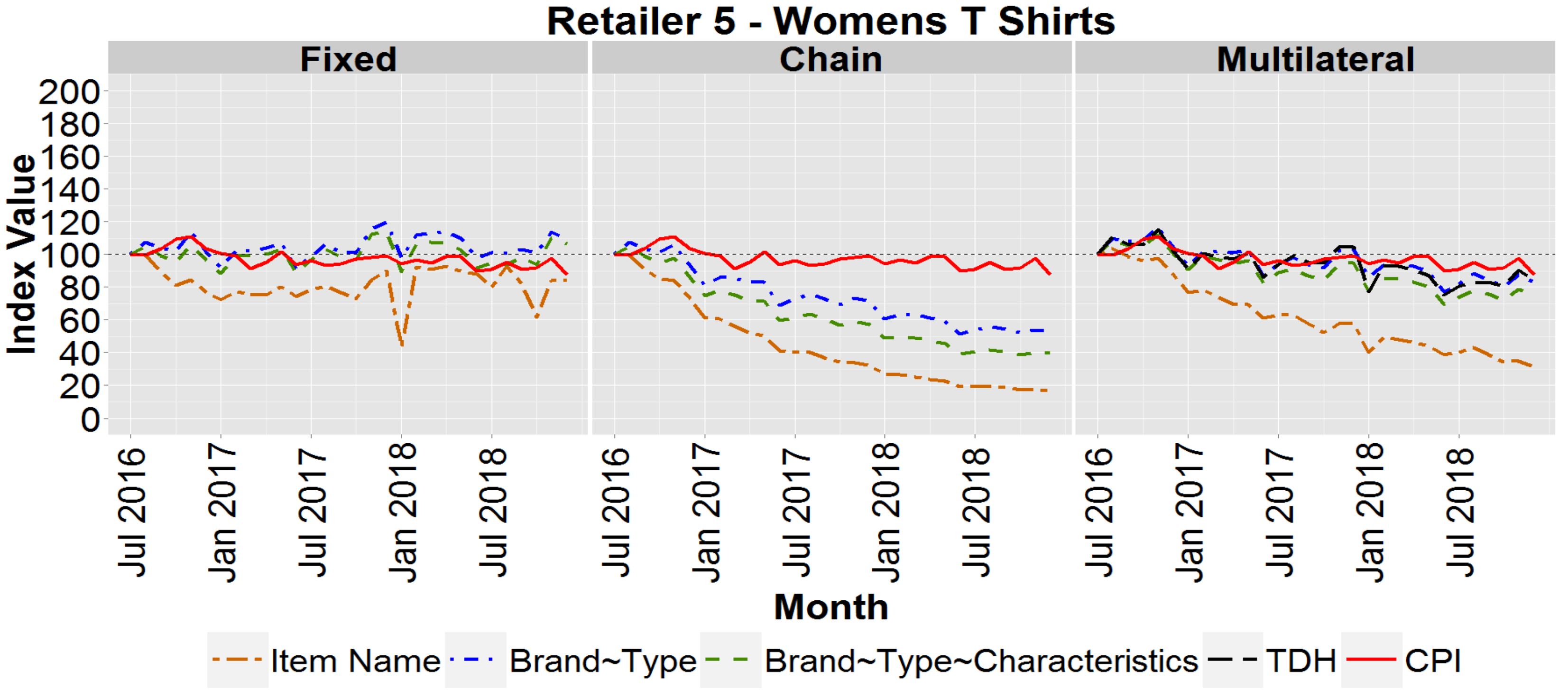
- Unweighted index methods are required since web scraped data does not contain expenditure or quantity information

- Multilateral index methods can be used to match products across multiple time periods and resolve the “chain drift” problem with chained indexes

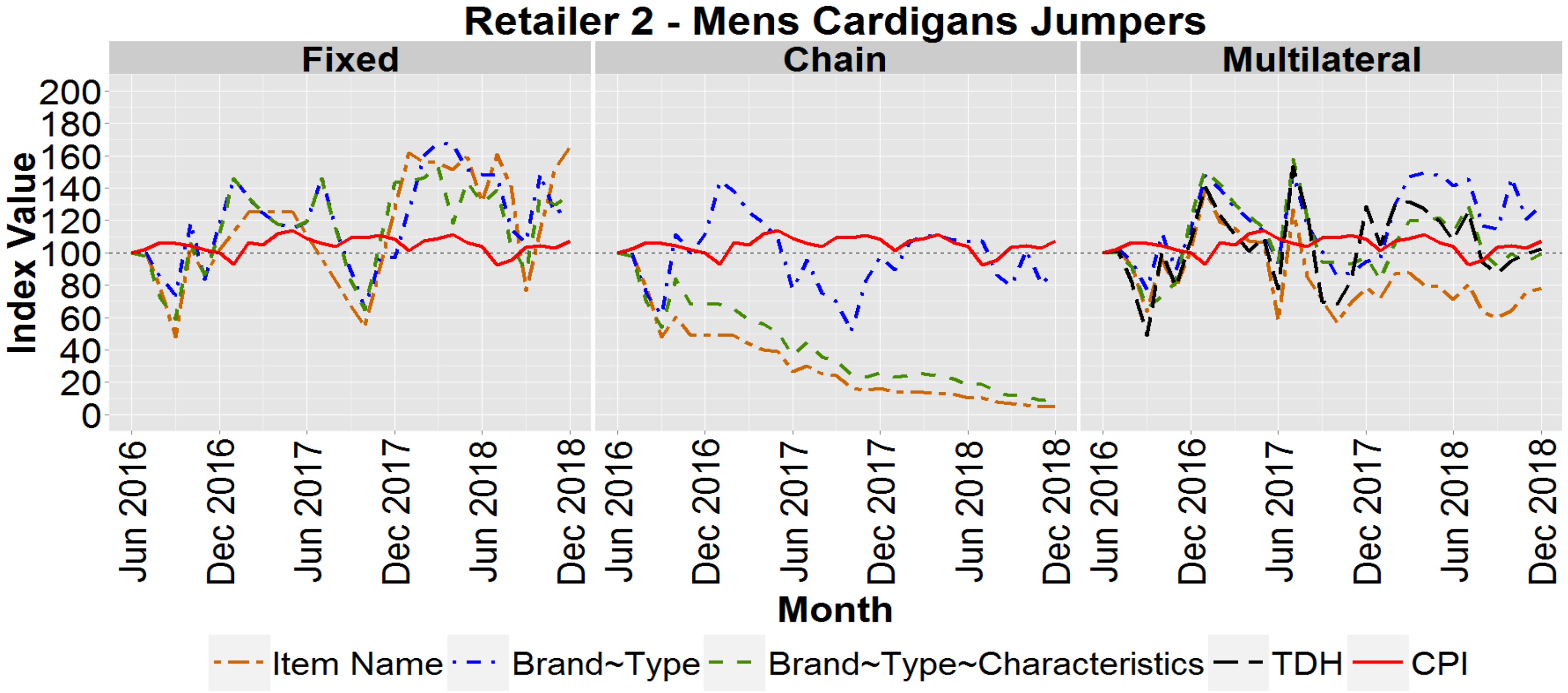
- 2 unweighted multilateral index methods are considered:
  - GEKS-Jevons (GEKS-J)
  - Time dummy hedonic (TDH) model with OLS weights

Elementary Aggregate Results

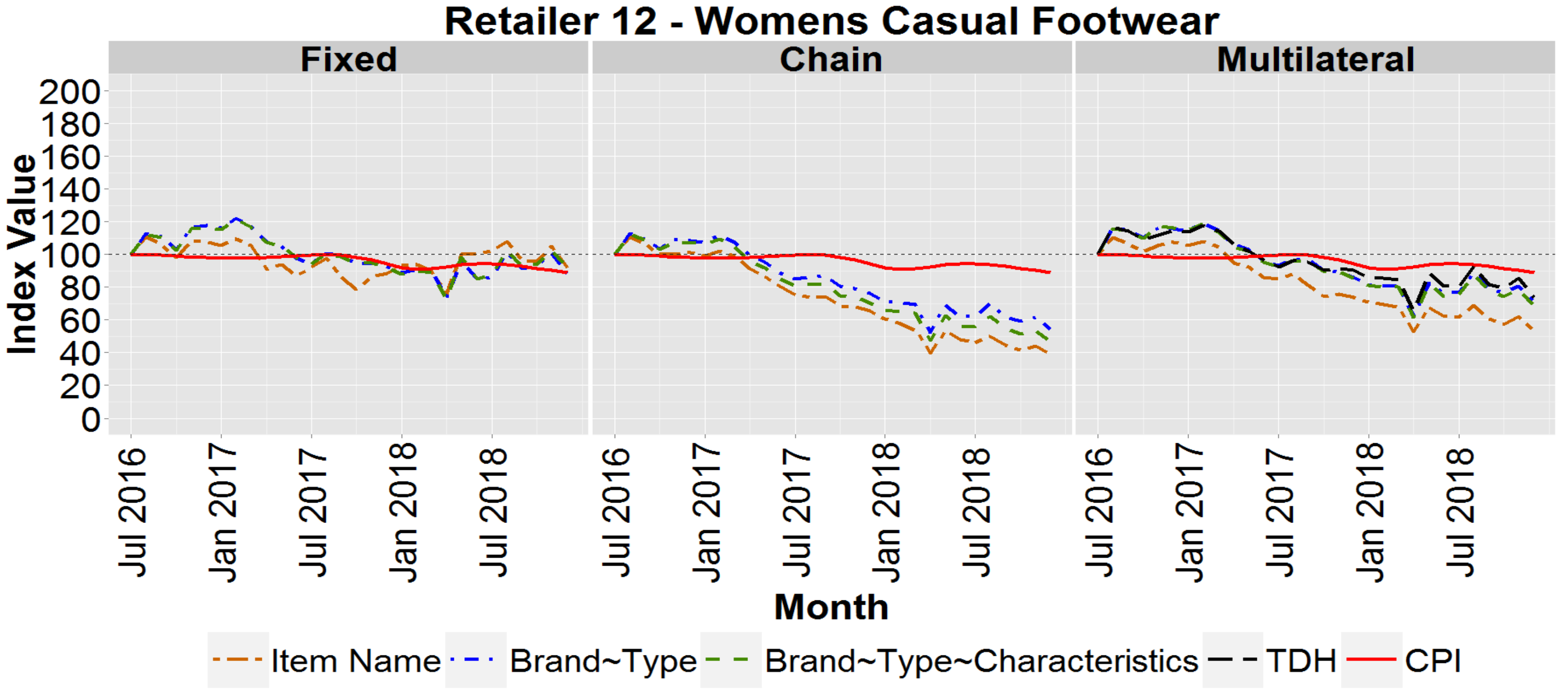
Australian Bureau of Statistics



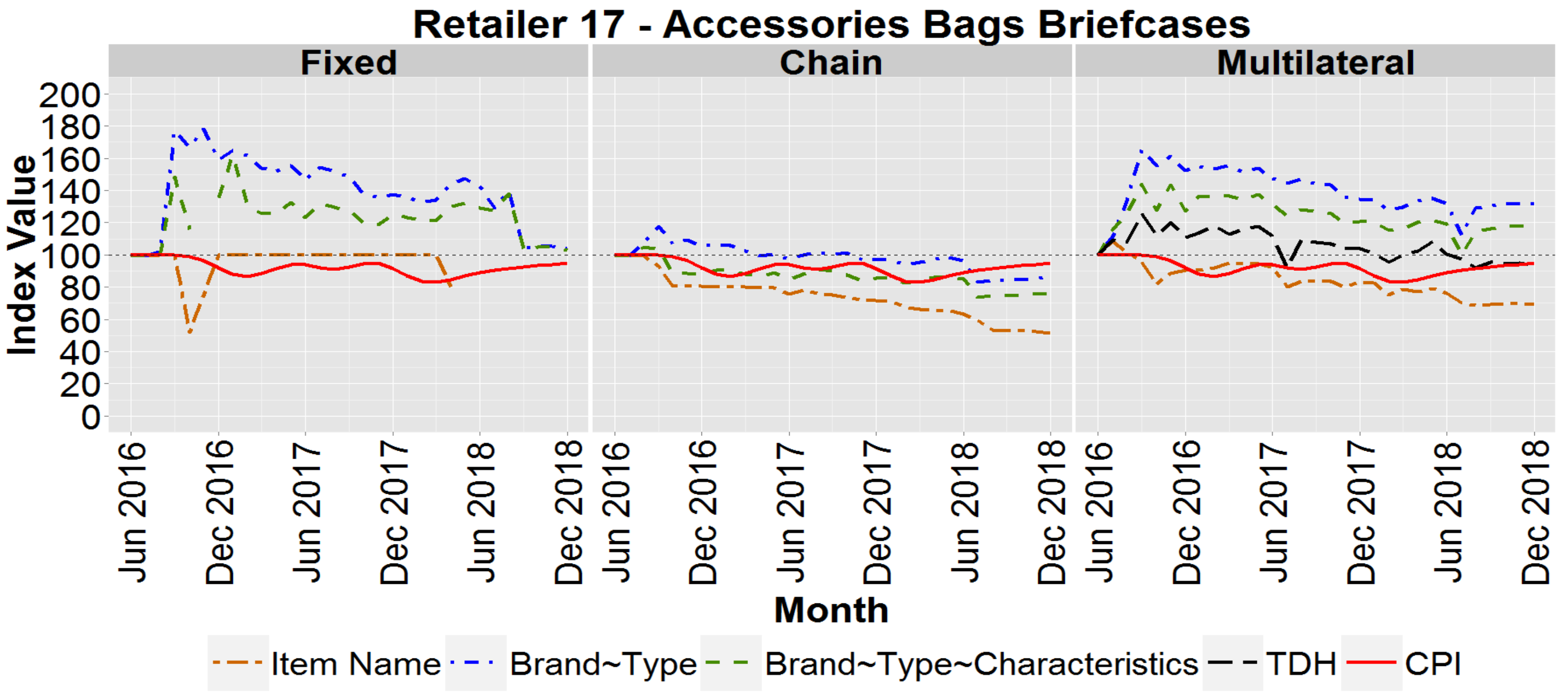
Elementary Aggregate Results



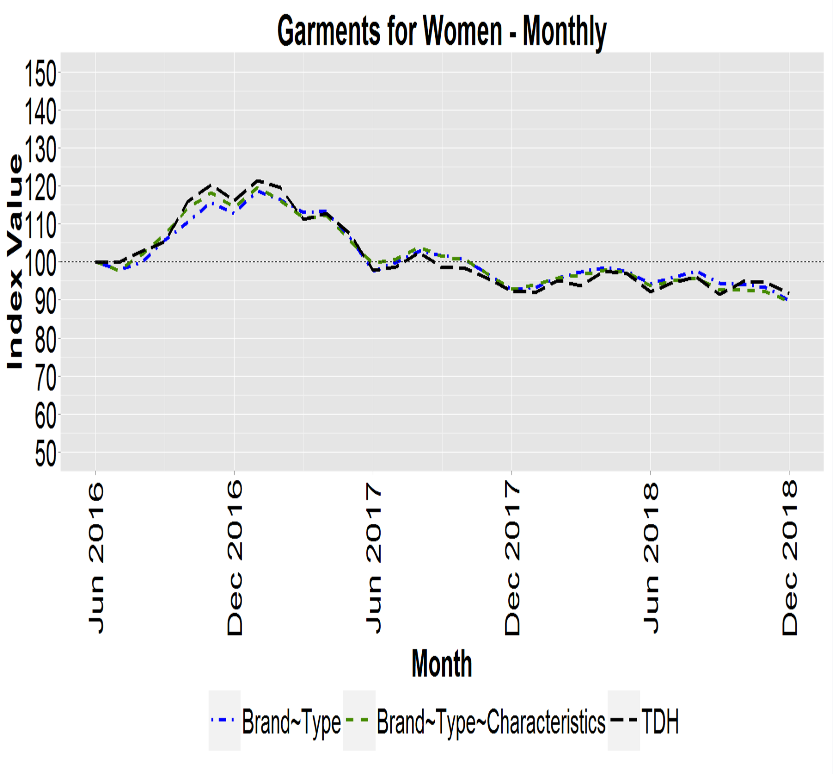
Elementary Aggregate Results

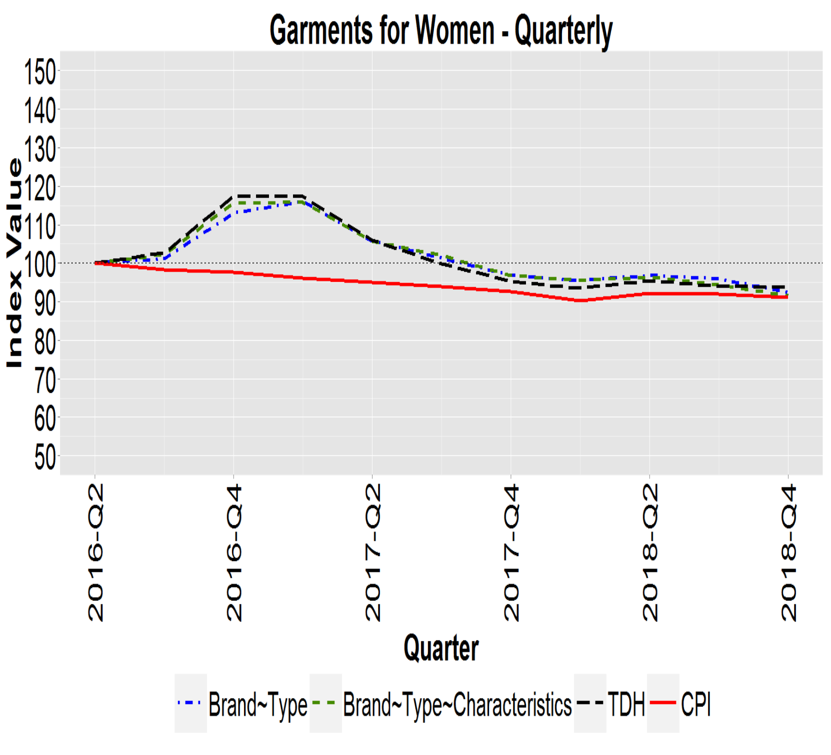


Elementary Aggregate Results

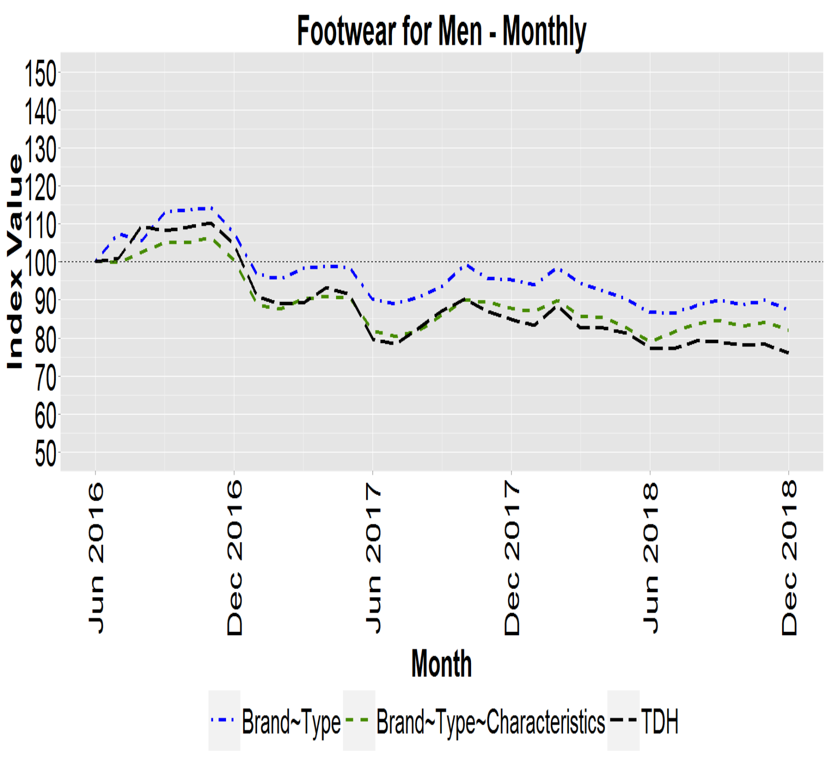


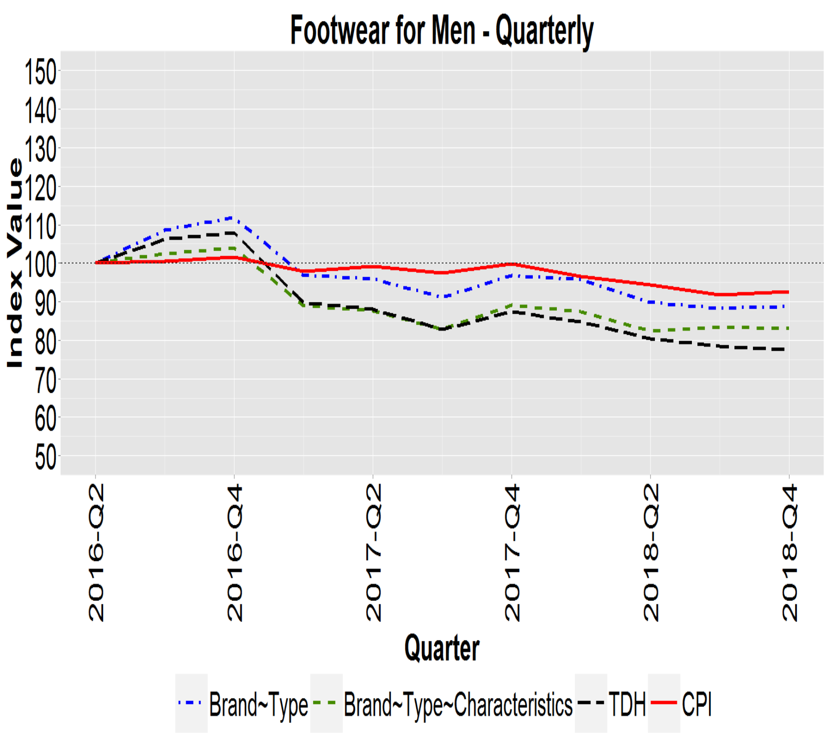
Expenditure Class Results





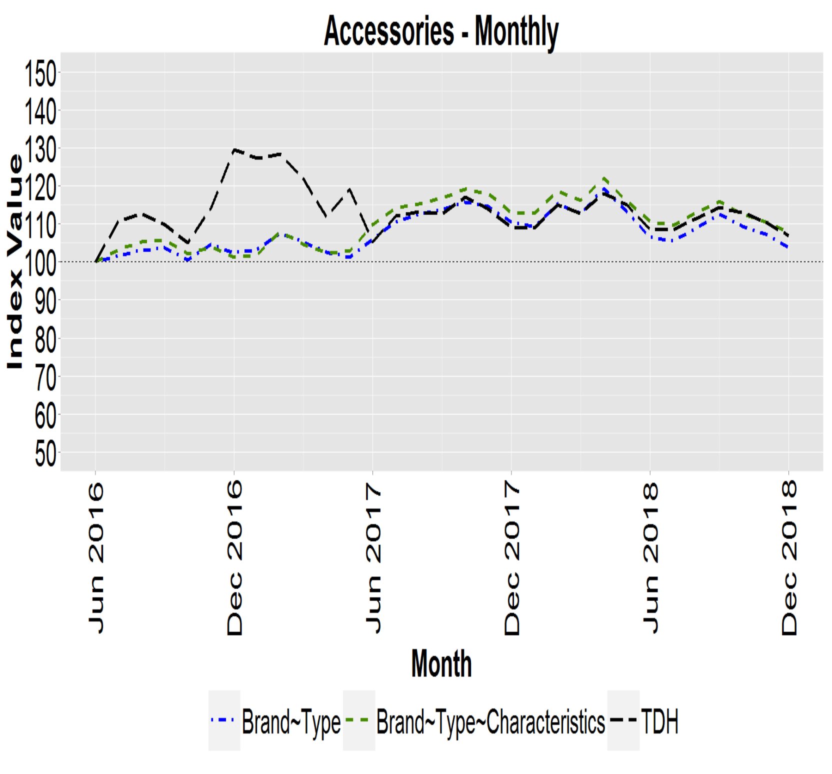
Expenditure Class Results

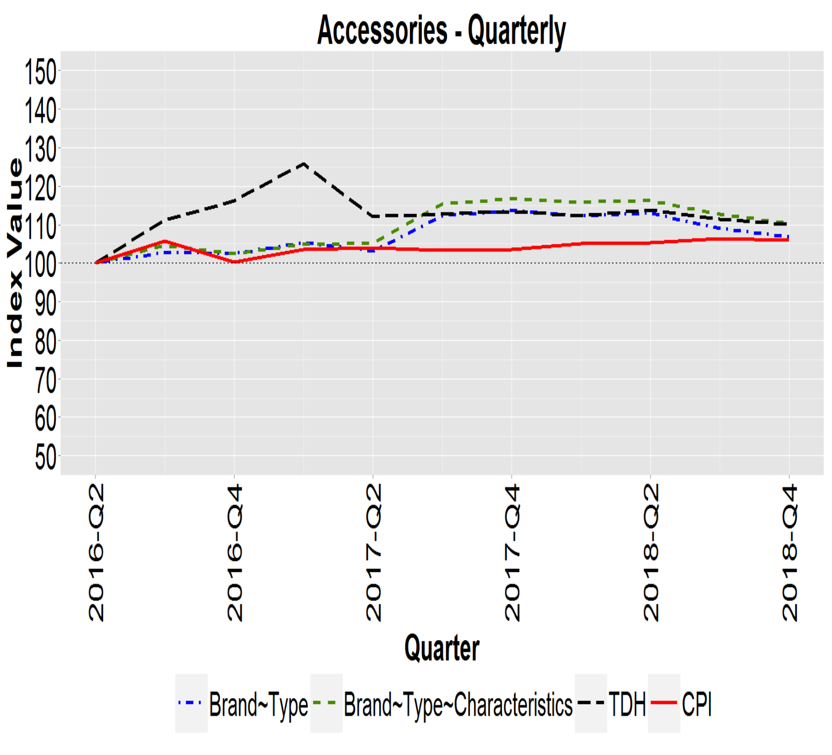




Expenditure Class Results

Australian Bureau of Statistics





Conclusions

- Pre-processing to form ‘clustered’ homogenous products is one viable strategy for NSOs to consider for ‘dynamic’ basket categories
- At the elementary level, our clothing results exhibit downward drift for chained indexes
- Fixed and multilateral indexes produced the ‘most plausible’ results with broader clothing product definitions

Conclusions

- Characteristic extraction more difficult with some footwear and accessory indexes – sparse text data means some heterogeneity still exists in our broader product definitions
- At the published level, experimental multilateral results broadly comparable with CPI equivalent

Further development

- Web scrapers maintained by ABS Prices Branch – funding attempts to expand across organisation
- Alternative strategies for forming clustered homogenous products
- Alternative strategies for respondent aggregation
- Alternative strategies for weighting individual products within clustered homogenous definitions

Australian Bureau of Statistics

Questions?