Australian Bureau of Statistics

International Working Group on Price Indices

Paper for plenary presentation

United Nations Statistical Commission Ottawa Group

8 – 10 May 2019, Rio De Janeiro, Brazil

Prepared by: Andrew Glassock and Michael Holt

Experimental clothing indexes using Australian web scraped data

Key words: Web scraping, text mining, bilateral indexes, multilateral indexes

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Australian Bureau of Statistics.

1. Background

The Australian Bureau of Statistics (ABS) uses several modes of collection to obtain prices for the Australian CPI. These include personal visits, online, telephone, and administrative data, including transactions data. More recently, with the growth of online retailing, pricing information is able to be obtained from websites. Advances in technology and automated scraping software have enabled large scale data collection from the web.

Web scraping is a technique employed to extract large amounts of data from websites. Prices can be collected as frequently as desired for all products using purpose-built programs that scan the websites of retailers, find the relevant information and store the information in a time series. The process can be run automatically and as frequently as desired (e.g daily or weekly), providing high frequency information. As such it is deemed to be a rich source of pricing information to statistical agencies.

The ABS has been utilising new technologies to collect pricing information since May 2016. Web scraping allows approximately 500,000 prices to be collected per week compared to 1,000 prices with traditional in-field collection, providing an opportunity to significantly enhance the sample of products and prices collected. From the June quarter 2017 the ABS implemented web scraped data into the CPI, using an average calculated price of an item over a given period – essentially replacing field collected data with a web scraped price.

While this has enhanced the CPI, it is recognised that more can be done with web scraped data. The ABS is investigating new methods that will facilitate greater use of web scraped data in the compilation of the Australian CPI. This paper presents some initial findings for clothing and footwear commodities which focus on the elementary aggregation of web scraped data and attempts to contribute to conceptual/empirical discussions around options to define an individual product and the choice between bilateral and multilateral index formula. The ABS has prioritised the expansion and enhancement of web scraped data for the clothing and footwear group for the following reasons:

- The clothing and footwear group of the Australian CPI uses a large sample with high collection and data editing costs;
- Clothing and footwear markets are highly competitive, with a large number of retailers selling differentiated clothing and footwear products; and
- Clothing and footwear products are highly seasonal in nature making it difficult to calculate bilateral price indexes based on matching products between two periods.

The remainder of this paper is set out as follows. Section 2 describes the current processes used to scrape information from Australian websites. In Section 3 we discuss the methods used to derive clothing and footwear indexes at the elementary aggregate and expenditure class levels. A summary of our findings from 19 clothing and footwear retailers is presented in Section 4. Section 5 concludes and provides several areas identified for future research.

2. Process of scraping

The ABS process of web scraping information from the internet can be broken down into three main steps:

1) Confirm the website allows scraping;
2) Scraping the website; and
3) Cleaning the collected data.

Firstly, programmers setting up web scrapers should check the terms and conditions section of a website for “conditions of use” and the “robots.txt” file located within the website’s root directory to determine whether a website is eligible for scraping (and also what information can be scraped). For those websites which allow scraping, a scraper is set up to identify the website’s category structure and the relevant categories to be scraped. For example, a website may have all the individual item categories, then additional categories such as “new in” or “all items” which can be excluded since they duplicate the items in the individual categories. The scraper then proceeds to download all items and prices from the internet by pulling in data using the most suitable method given the website structure. In some instances, additional data cleaning may be required post-scraping so that only the set of items and prices remains.

Table 1 below provides a fictional example of the typical information currently scraped by the ABS. A web scraped data set consists of the date the scraping occurred, name of the retailer, website category where the item was found, a text string describing the item (item name), prices, and an item count used to flag scraped items with identical text descriptions (item counts greater than one indicate identical items on the day of the scrape).

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

3. Methods

Product Definition

Fundamentally, the basic information required to compile a price index includes prices, expenditure information (or reasonable assumptions about consumer substitution), and product classifications such as individual items or groups of homogenous items. In this respect, there are a number of challenges when constructing price indexes from web scraped data including a lack of expenditure information, selecting an appropriate item definition and index number formula. For the purposes of this paper, we focus our attention on the latter two considerations and follow the convention of using equally weighted price indexes as expenditure information is unavailable.

Before elementary aggregation can be performed, a price statistician needs to decide on an appropriate product definition. A natural choice when using large datasets is to define products using an exact matching key like a barcode or stock keeping unit (SKU) – but given the absence of these matching keys in our datasets the item name text is a reasonable starting point. There are however a number of practical data quality concerns that may impact the quality of price indexes – namely spelling variations and ordering of text (see Table 1 for several examples) that represent identical (or near identical) items. As a result, we are likely to run into several known issues when compiling matched model price indexes using the item name variable including:

- The use of a product definition that is too granular and does not accurately represent the concept of a homogenous product. This will overstate true product churn and produce price indexes that suffer from a lack of matching.
- Granular product definitions usually have distinct pricing patterns associated with their life cycle (i.e. enter at a relatively high price and exit at relatively low price). The combination of life cycle effects, product churn and matched model methods (with no linking of comparable items) produces indexes that are usually downwardly biased. This has commonly been referred to as a 'relaunch' problem (Chessa, 2016) in various scanner and web scraped data studies.

Matching products over time is an important consideration when using web scraped data since traditional approaches (such as analyst inspection) are no longer feasible from a resources perspective. To overcome this problem, several NSOs have proposed practical strategies that focus on extracting information (e.g. brand, shirt type) from text strings to form broader (clustered) product definitions. Some examples of such experimental work with large datasets include:

- The use characteristics (e.g. text string, price or price similarity, and retailer) and unsupervised learning algorithms to define products for a variety of commodity classes (Bhardwaj et al, 2017, ONS, 2017).

- The use characteristics and supervised learning algorithms to define products for home appliances and electronic products (Abe and Shinozaki, 2018).
- The use of characteristics and auxiliary retailer category information to define products for footwear (Van Loon and Roels, 2018), sports equipment (Hov and Johannessen, 2018) and department store items (Chessa et al, 2017).

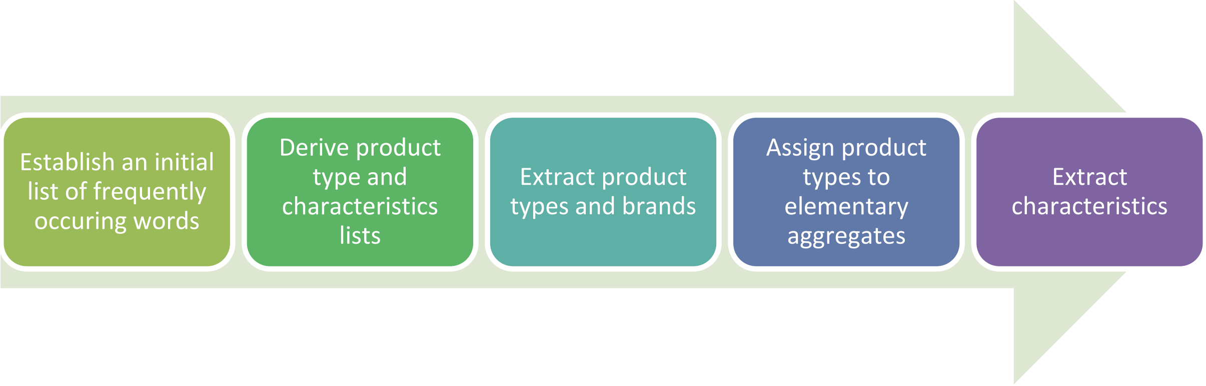
The choice of a suitable product definition however involves a trade-off. While a broader product definition than the item name is required to mitigate the relaunch issue, linking multiple item names together increases the risk of an average price bias. Item names do not necessarily capture all important price determining characteristics of the item, increasing the likelihood of heterogeneous product groupings based solely on observable characteristics. If the individual items within a product group exhibit different price behaviour, an average price bias can arise since the average price movement for the group is not representative of the average movement of the individual items.

This paper is the first ABS attempt at such experimental work on web scraping and considers three alternative product definitions based on different levels of item aggregation. The item name approach is the most granular product definition and is presented to further illustrate the problems created by severe product churn. The two remaining definitions rely on the ability to extract information from the item name about the brand, product type (e.g. whether the product is a shirt or dress) and any additional product characteristics which are likely to affect the price of the product (e.g. sleeve lengths and materials for shirts). A discussion of the process used to identify this information is provided in the following section.

Pre-Analysis Cleaning

In order to extract product characteristics, an iterative process was established to identify keywords contained within item names similar to the framework described in Willenborg (2017). The purpose of this process was to maximise the number of items mapped to product types, subject to minimising the amount of misclassification. The flowchart in Figure 1 illustrates this process.

Figure 1: Process flow of extracting characteristics


Establish an initial list of frequently occuring words
Derive product type and characteristics lists
Extract product types and brands
Assign product types to elementary aggregates
Extract characteristics

In the initial stage of this process, a text mining procedure was established to search through item names across all retailers to compile a list of the most frequently occurring words or phrases. This list was narrowed down to derive new lists of keywords to identify product types and characteristics, along with lists of synonyms associated with each keyword. A second text mining process was then carried out separately for each data set to map items to product types based on the ordering of the list (so that item names containing more than one product type keyword or synonym were mapped to the first type found). Unmapped items were then inspected to identify any additional keywords or synonyms and determine whether category information could be used to allocate these items. A similar check was introduced to detect misclassified items, with keywords and synonyms added, removed, or reordered accordingly. These steps were repeated on several occasions until the majority of items were successfully mapped.

For retailers offering a variety of brands, brand information was also extracted by searching for regular expressions surrounding the brand in the item name or using a list approach similar to obtaining product types by first deriving a list of brands through inspecting the retailer's website. Once the brand and product type mappings were complete, items were then linked to elementary aggregates and expenditure classes based on the assigned product type. Characteristics were then assigned to the relevant elementary aggregates and another text mining step was carried out within each elementary aggregate to allocate characteristics found in the item name (allowing multiple characteristics to be mapped to a single item). Table 2 provides an example of a subset of characteristics extracted through the iterative text mining process outlined. The full list of possible characteristics extracted from our web scraped datasets is included in Appendix 1.

Table 2: Extracting Product Information

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"

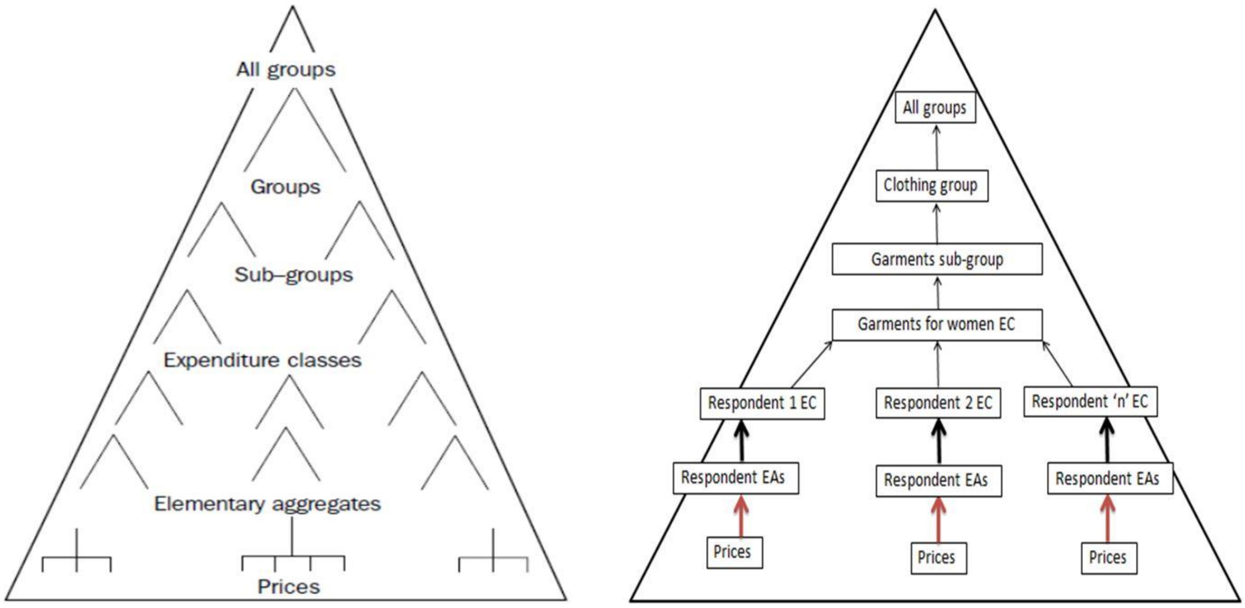
An additional pre-analysis cleaning consideration is whether any global cleaning of atypical prices observations should occur. Atypical prices could occur in web scraped data for a number of reasons including legitimate sale (or return from sale) prices, website errors or scraping errors. The approach chosen in this paper is to apply the following global rule: if a particular monthly "item name" observation differs from the median monthly clustered product definition by a factor of four, it is excluded prior to price index compilation.

Aggregation Structure

While the majority of the paper focuses on selecting a suitable product definition and price index formula, another important consideration for National Statistical Organisations (NSOs) is the choice of an appropriate aggregation structure. Traditionally, individual prices collected by the ABS are aggregated across products and retailers to produce elementary aggregate indexes. Closely related elementary aggregate indexes can then be aggregated to the expenditure class level (the lowest level published by ABS), which in turn can then be aggregated to derive the subgroup, group and total CPI.

The ABS now applies a modified aggregation structure when using scanner data below the expenditure class level in order to capture possible retailer-specific effects (ABS, 2017). Instead of using a single elementary aggregate, prices collected through scanner data are aggregated to retailer or respondent elementary aggregates and expenditure classes. Respondent expenditure classes are then weighted by the retailer's market share to obtain the expenditure class indexes. The results presented in the next section use the aggregation structure for scanner data, with a view to future implementation for web scraped data in the Australian CPI. Figure 2 compares the traditional structure (left panel) with an example of the proposed structure for the garments for women expenditure class (right panel).

Figure 2: Current and Proposed Aggregation Structures



Bilateral Indexes

The traditional bilateral approach to calculating price indexes is to directly compare the prices of products in the current period relative to their prices in a reference period. In this paper, we perform elementary aggregation using both the fixed and chained (period on period) forms of the Jevons price index between any two periods (j, k), which can be expressed as:

$$P_{Jevons}^{j,k} = \prod_{i \in S_M^{j,k}} \left(\frac{p_i^k}{p_i^j} \right)^{\frac{1}{N_M^{j,k}}}, \qquad (1)$$

Multilateral Indexes

Assuming we have an appropriate product definition, both fixed and chained bilateral indexes described above may have practical limitations. Fixed bilateral indexes compare prices and quantities from the current period relative to an earlier base period and have the problem of item attrition (i.e. product entries and exits) decreasing the amount of matched products overtime. Additionally, the period chosen as the base period is given special importance and will exclude some items (e.g. seasonal items) that are not available in the base period. While the item attrition issue observed with direct comparisons can be addressed through chaining, indirect indexes may suffer from "chain drift" due to factors including product life cycle effects (e.g. intermittent sales or products falling in price over time), the relaunch problem, and price and quantity bouncing[1].

As a compromise to some of these limitations with bilateral indexes, some web scraped studies have investigated multilateral index methods (e.g. Bhardwaj et al, 2017, ONS, 2017). Multilateral indexes allow for price comparisons across three or more time periods, and are becoming increasingly used to compile temporal indexes from scanner data to resolve the "chain drift" problem associated with promotional sales. In the context of web scraped (unweighted) data, it may be advantageous to use all the product matches across a window of data and produce transitive price comparisons across time[2]. This paper considers two well-known multilateral index methods: the GEKS approach and the time dummy hedonic (TDH) model. These methods have historically been used for spatial comparisons between countries, although in recent years have been adapted to allow for intertemporal price comparisons.

The GEKS method was first adapted for temporal price comparisons by Ivancic, Diewert and Fox (2009, 2011) to ensure transitivity in a temporal setting. To see this, consider the following GEKS index for the $T + 1$ periods from 0 to t:

$$P_{GEKS}^{0,t} = \prod_{l=0}^{T} \left[\frac{P^{l,t}}{P^{l,0}} \right]^{\frac{1}{T+1}} = \prod_{l=0}^{T} \left[\frac{P^{0,l}}{P^{t,l}} \right]^{\frac{1}{T+1}} = \prod_{l=0}^{T} \left[P^{0,l} \times P^{l,t} \right]^{\frac{1}{T+1}} \qquad (2)$$

[1] Price and quantity bouncing has frequently been cited as a major reason why chain drift has been frequently observed in scanner data. Since scanner datasets contain quantity or expenditure information, the assumption of price and quantity bouncing as a cause of chain drift can be easily verified when chain drift is observed. However, this is not the case with web scraped data which does not contain quantity or expenditure information. Therefore, where chain drift is observed in indexes constructed from web scraped data, we consider instead product life cycle and relaunch effects as alternative explanations for why this phenomenon has occurred (since these are still observed).

[2] As described in de Haan and Hendricks (2016, p.26), provided the product definition is appropriate, multilateral methods "have much to offer as compared to a period-on-period chained matched-model price index since they use all of the matches across the whole sample period".

The first two equalities indicate that the GEKS index can be expressed as either the geometric mean of the ratio of bilateral price indexes between periods 0 and t using each comparison period l as the base, or alternatively as the geometric mean of the ratio of bilateral price indexes across all comparison periods using periods 0 and t as the base. The third equality shows that the first two equalities together imply that the GEKS index satisfies the time-reversal test required for transitivity since the GEKS index between periods t and 0 will be equal to the reciprocal of the index between periods 0 and t. Since web scraped data does not contain quantity information, the GEKS index can be derived using the Jevons index as the bilateral link; we refer to this as GEKS-J for the remainder of the paper.

Although the GEKS-J method makes greater use of available information compared to the fixed and chained Jevons indexes, this approach still relies on the ability to match products across time. In order to reduce the amount of product churn, a broader product definition can be used where clusters of items are formed based on observable information about the brand, type, and characteristics of each item. However, this approach may result in clusters containing a limited number of items and does not allow the effects of each feature on the average cluster movement to be easily identified. Furthermore, because the clustering approach considers only within group variation, possible correlations between the price movements of closely related clusters with common features (e.g. branded short sleeve shirts made from different materials or those with the same material but different sleeve lengths) are ignored.

An alternative multilateral approach which incorporates both new and disappearing products is the time dummy hedonic (TDH) regression model. The TDH model estimates (logarithmic) price movements for each product as a function of time and product characteristics. In log-linear form, the TDH model can be expressed as:

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

The parameters δ^t and β_k in Equation (1) denote the effects associated with the period t time dummy D_i^t and the product characteristics $z_{i,k}$ respectively. De Haan and Krsinich (2014) showed that when the TDH model is estimated using *weighted least squares* (WLS), the resulting price index between periods 0 and t is equal to the ratio of weighted geometric average prices in each period multiplied by an implicit quality adjustment factor for changes in weighted average characteristics. Similarly to the GEKS approach, the TDH model in this paper assumes equal weights due to the absence of expenditure information. In this case, the WLS minimisation problem reduces to estimation of the TDH model via *ordinary least squares* (OLS).

Since the GEKS and TDH methods use a window of periods to calculate the price movement between periods 0 and t, revisions to the index occur whenever the window is updated to reflect new information becoming available. This poses a challenge for the ABS and other NSOs who traditionally operate under a "no revisions" policy. To overcome this problem, calculated price movements can be linked to the existing CPI series by choosing a suitable splicing method. We follow the mean-splicing approach used by the ABS in compiling elementary aggregate indexes from scanner data (ABS, 2017) which takes the geometric average price movement between the current period and the previous $T + 1$ periods.

4. Results

In this section, we present a summary of our main findings up to the expenditure class level for the clothing and footwear group of the Australian CPI. Indexes were calculated across 19 clothing and footwear retailers, covering 74 standardised elementary aggregates from 7 expenditure classes. All elementary aggregation was performed on a monthly frequency, with all multilateral time series constructed using a rolling 13 month window (given the limited length of time series data currently available). In all experimental indexes below, monthly average prices are derived as the arithmetic average of available daily price observations.

Clothing Elementary Aggregates

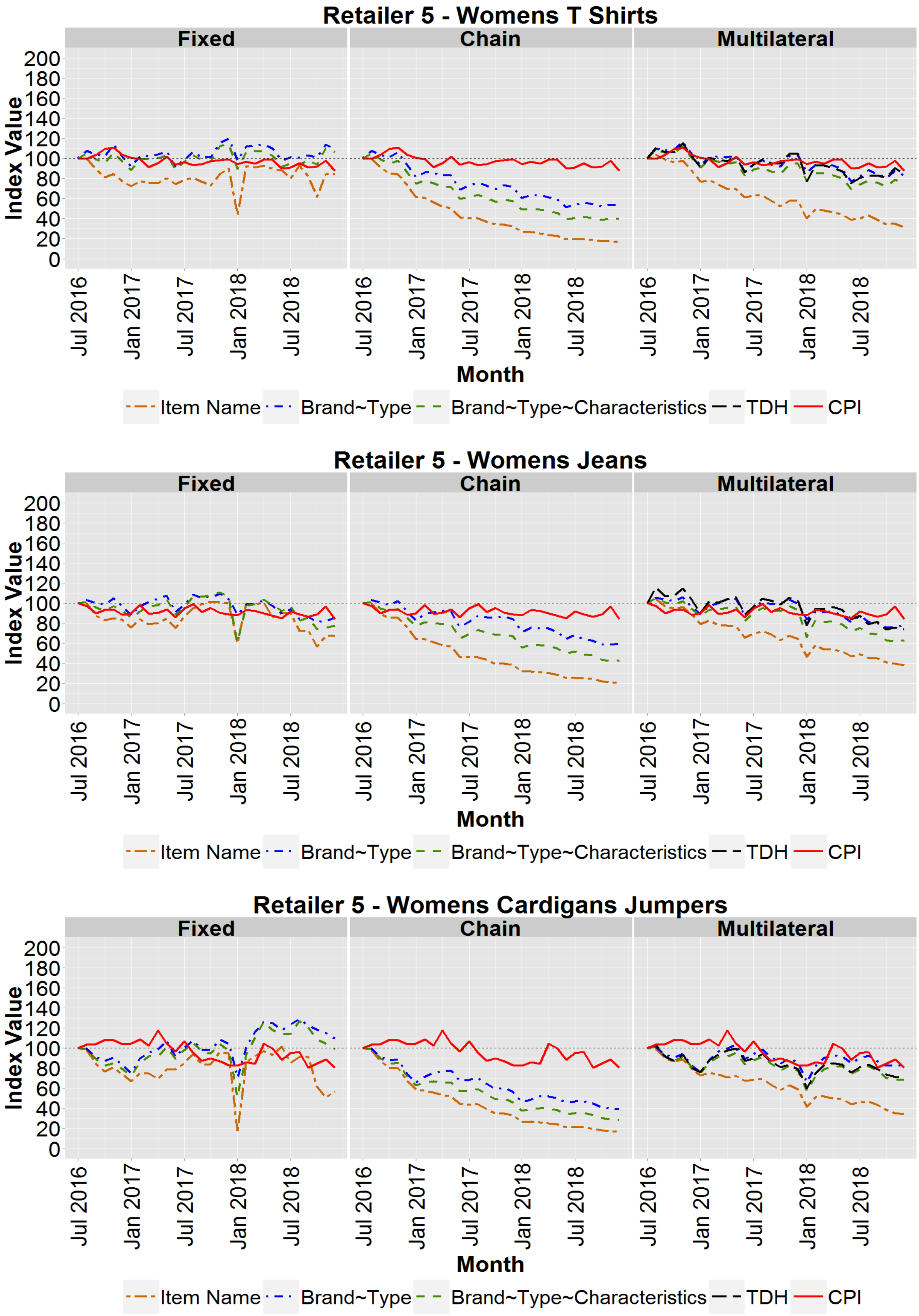
We begin our analysis with a discussion of elementary aggregates for clothing, footwear and accessories commodities which is the primary focus of this paper. Garments are divided into three expenditure classes in the Australian CPI: garments for children, garments for men, and garments for women. The following panels present a selection of results, and include comparisons to an equivalent CPI series where possible to observe where any large differences exist in the short and long term.

Figure 3 provides a comparison of elementary aggregate indexes for Retailer 5[3]. Looking at the fixed Jevons panel, the experimental series track the CPI series fairly closely for the "brand~type" and "brand~type~characteristics" product definitions while the "item name" usually diverges below the CPI. On the other hand, the chained Jevons indexes drift downwards substantially over time for all product definitions, with the most granular product definition (i.e. "item name") drifting the most severe. The downward drift observed for chained indexes is likely being driven by products leaving the market at historically low prices, and the inability to record the appropriate price rise when it re-enters the market.

Despite some disparity between methods, the GEKS-J indexes with "brand~type" and "brand~type~characteristics" product definitions and the TDH indexes generally exhibit similar patterns to the CPI. On the other hand, the GEKS-J indexes using the "item name" product definition continue to drift downwards by close to 50% by the end of the analysis period. One explanation for the persistence of chain drift with the item name GEKS-J indexes is that matching products across multiple periods does not alleviate the problem of drift when matching products beyond the first period is difficult. These findings reaffirm our support

[3] Retailer 5 is a large retailer with a variety of different brands, product types and characteristics.

Figure 3: Garments Elementary Aggregates for Retailer 5[4]



Retailer 5 - Womens T Shirts

Retailer 5 - Womens Jeans

Retailer 5 - Womens Cardigans Jumpers

[4] Note for the multilateral panel results, the different product definitions series use the GEKS-J method.

that the item name product definition is too disaggregated for the purposes of constructing price indexes.

Footwear Elementary Aggregates

Elementary aggregate indexes for casual footwear are provided in Appendix 2. Casual footwear product types include sneakers, slip-ons, espadrilles, flats, loafers, moccasins and plimsolls. Similar to the clothing results, we found substantial downward drift in all chained indexes in the vicinity of between 40-60 percent which we again consider an unrealistic representation of price trends faced by Australian households across this period. Alternatively, the TDH, fixed (with broader definitions such as "brand~type" and "brand~type~characteristics") and GEKS-J indexes (with broader definitions) provide the most comparable results to the CPI. While the extraction of product information from item name descriptions worked reasonably effectively for clothing, the process of proved to be significantly more challenging for footwear. In many instances, product characteristics could not be easily identified; resulting in product definitions based primarily on product types (and brands) such as "shoes" or "boots", which may introduce volatility or average price bias.

Accessories Elementary Aggregates

Accessories elementary aggregates can be broadly categorised into 3 groups: travel-related accessories (including bags, purses and wallets, suitcases), jewellery (earrings, necklaces, watches, rings and bangles) and clothing accessories (hats, sunglasses, scarves, gloves, ties). As the results for selected accessories elementary aggregates presented in Appendix 3 highlight, the TDH, fixed (with broader definitions) and GEKS-J indexes (with broader definitions) again show the most plausible results relative to the CPI. Within many accessories elementary aggregates such as jewellery, which contain items ranging in price from under $100 to over $10,000, prices are widely dispersed. If this within price variation is not adequately captured through observable product characteristics, average prices will not accurately reflect underlying price movements and are therefore biased. Retailers offering a homogenous brand or a limited range of brands are likely to be at greater risk of average price bias since product types and characteristics may not sufficiently capture quality differences within each accessories elementary aggregate.

Expenditure Class Indexes

In this subsection, we derive expenditure class indexes by aggregating the retailer elementary aggregate indexes to the retailer expenditure class level and averaging the movement across retailers. Only the multilateral elementary aggregate indexes (with broader definitions) are used for upper level aggregation due to their promising performance in the previous subsection. Aggregation to the retailer expenditure class and all retailers' expenditure class levels is carried out using a Laspeyres-type approach. In the first stage, expenditure class indexes for each retailer are calculated by weighting each elementary aggregate by expenditure shares obtained from the Household Expenditure Survey (HES). However, aggregation across retailers is more problematic since information about expenditure on each retailer expenditure class is required. The ABS Retail Trade Survey provides a natural starting point for obtaining expenditure weights for each retailer, although not all retailers represented in our sample are included in the survey. Expenditure information from supplementary data sources is therefore also required to attain expenditure weights for these retailers. For both simplicity and to ensure the all retailers' expenditure class captures every retailer in our sample, the expenditure class results presented in this paper assign an equal weight to each retailer.

Expenditure class indexes for garments are presented in Figure 4. Monthly indexes are presented on the left hand side of the figure, while quarterly indexes are shown on the right[5]. For all index methods, strong price rises from one sampled retailer cause each garments expenditure class to increase in the first few months of the sample. Over time prices gradually return towards their initial value, with the monthly TDH index for men's garments and all monthly indexes for children's and women's garments slightly below the reference value at the end of the sample period. Overall, the monthly indexes for garments expenditure classes are encouraging: the results do not appear to be overly volatile throughout the sample and are consistent with the hypothesis that clothing indexes are flat or falling due to competitive Australian retail conditions.

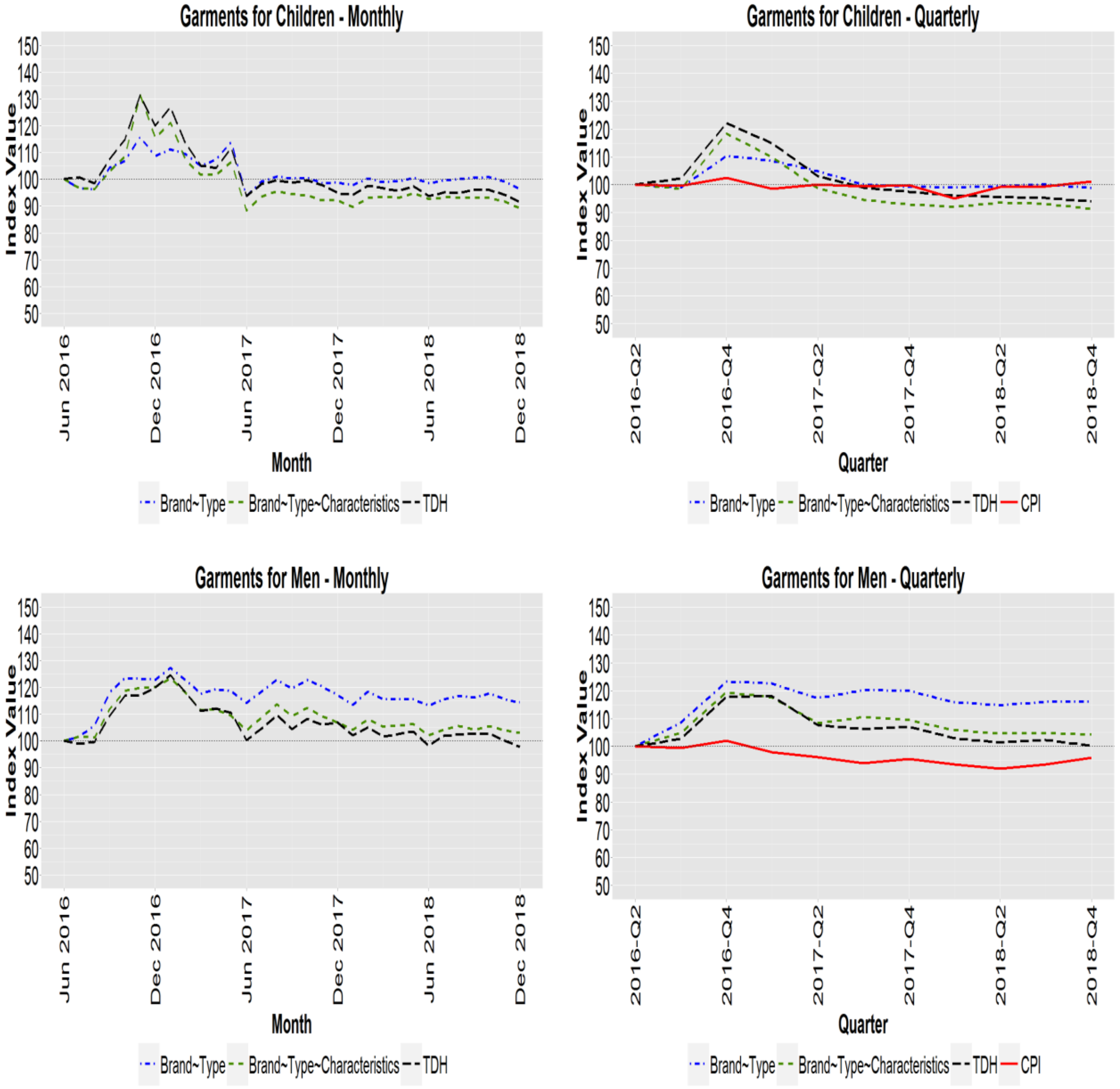
The quarterly indexes for garments expenditure classes yield similar conclusions to the monthly indexes but also highlight how each approach compares to the published CPI indexes. For children's garments, which remained relatively stable and close to the initial value throughout the sample, each of the three methods closely tracks the published index once the initial price rise has subsided. Similarly, the experimental indexes for women's garments are close to the corresponding CPI series in the second half of the sampling interval, although the expenditure class has trended slightly downwards over time. In contrast, garments for men displays a greatest divergence between the estimated indexes and the CPI series. While the CPI has fallen, each of the estimated indexes remains above the initial value throughout the entire sample. The garments for men expenditure class also shows the greatest divergence between the GEKS-J index excluding item characteristics, which remains significantly higher than the initial value at the end of the sample, and the other experimental indexes, which converge towards the reference value. Nevertheless, the quarterly results for garments expenditure classes further reaffirm our view that using observable product information including brand, product type, and additional characteristics provides a reasonable approach to constructing web scraped price indexes for clothing garments.

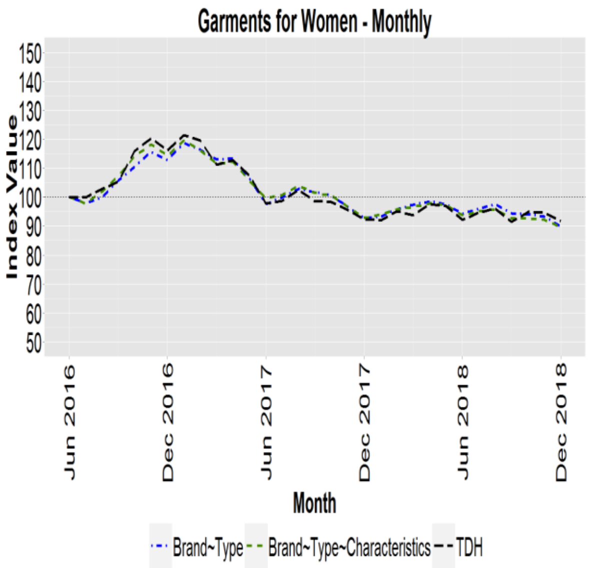
Appendix 4 presents the monthly and quarterly indexes for each footwear expenditure class. Similarly to the garments expenditure classes, footwear expenditure classes exhibit a large increase during the early months of the sample due to abnormal price rises from one retailer.

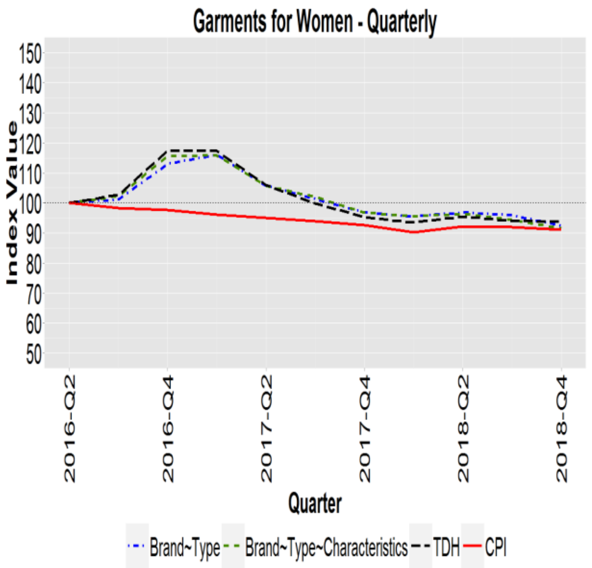
[5] The Australian CPI is currently compiled on a quarterly basis. ABS is currently undertaking a feasibility study to investigate the possibility of producing a monthly CPI.

Once this effect subsides however, there is a divergence in trends between the footwear expenditure classes. Estimated indexes for children's footwear remain above the initial value

Figure 4: Garments Expenditure Class Indexes







throughout the majority of the sample, whereas the footwear for men expenditure class displays a strong downward trend. In contrast, prices for women's footwear remain reasonably stable despite falling marginally later in the sample.

With the exception of footwear for women, the deviations of the quarterly indexes for footwear from their corresponding CPI indexes are relatively large when compared with those for garments. In the case of footwear for children, the persistence of higher prices for the experimental indexes following the initial price increase contrasts the decrease in the CPI throughout the sample. For men's footwear however, the multilateral indexes overstate the downward impact of falling prices after the rise in the index has subsided. Both these findings indicate that when large price movements cannot be explained by observable product information, multilateral indexes may potentially overstate the impact and persistence of these price changes.

Finally, the accessories expenditure class indexes presented in Appendix 5 highlight a similar feature. Both the GEKS-J and TDH indexes yield accessories expenditure class estimates which are above the CPI series throughout the sample. In particular, the TDH index deviates the furthest from the CPI due to the abnormal increase early in the sampling interval. A higher TDH index continues to persist throughout the sample, with the index only starting to return towards the initial value towards the end of the sample. For the accessories expenditure class, the GEKS-J indexes do not appear to display an increase at the start of the sample but instead begin to rise later in the interval. Although the rise in the GEKS-J indexes occurs more gradually compared to the TDH index and begins at a later point, the effects are once again prolonged due to the use of a multilateral window. Despite rising above the accessories CPI, the GEKS-J index excluding item characteristics is the only index which remains reasonably close to the CPI for the majority of the sample period. However, as discussed in the previous section, this method is subject to a greater risk of heterogeneity in calculation of an average price at the elementary aggregate level compared with the other two experimental indexes. To alleviate these concerns and ensure price movements are accurately captured, further attention is therefore required for the accessories and footwear expenditure classes.

5. Conclusion and further development work

The purpose of this paper was to provide an overview of the methods ABS is currently investigating to enhance the use of web scraped data for the purpose of constructing clothing and footwear price indexes. We investigated web scraped data from 19 clothing and footwear retailers and were able to extract brand information, product types and additional characteristics for each item which was used to derive elementary aggregate price indexes for each retailer. Given the dynamic nature of clothing, our investigations focussed on a selection of product definitions and unweighted bilateral and multilateral index methods. We conclude the following from this study:

- Some form of clustering raw web scraped data is required for an appropriate homogenous product as mentioned in other studies such as ONS (2017) and Van Loon and Roels (2018). In most cases, product definitions such as the "brand~type" and "brand~type~characteristics" performed similarly.
- Chained Jevons indexes performed poorly across all product definitions. Fixed and multilateral methods with broader product definitions performed reasonably well for clothing commodities. Our view is that multilateral methods have advantages relative to fixed indexes for elementary aggregation in their ability to match across all possible time periods in a window.
- The empirical results at the elementary aggregate level for footwear and accessories were less conclusive. A number of footwear and accessories elementary aggregates were subject to significant downward drift, high volatility, and concerns of average price bias where product information proved difficult to extract. Further attention is required for footwear and accessories in order to alleviate these concerns.
- At the published expenditure class level, garments indexes remain relatively stable and close to the corresponding quarterly CPI, particularly for the TDH model and GEKS-J index using all observable product information, whereas footwear and accessories expenditure classes show a greater divergence in some instances due to the persistence of large price movements.

Although this paper focused on the choice of a suitable product definition and multilateral index methods for constructing price indexes from web scraped data, several areas for future research were identified through the process. Areas of possible investigation include the following:

- Alternative strategies for respondent elementary aggregation (especially if retailers have small samples for a specific elementary aggregate).
- Alternative assumptions for weighting individual products for clustered product definitions.

- Alternative strategies for forming clustered homogenous products (e.g. Bhardwaj, 2017, ONS, 2017).
- Alternative strategies for cleaning atypical price observations.

6. References

Australian Bureau of Statistics (ABS) (2009), “Information Paper: An Implementation Plan to Maximise the Use of Transactions Data in the CPI”, cat. no. 6401.0.60.004. Canberra, Australia.

Abe, N. and Shinozaki, K., (2018). “Compilation of Experimental Price Indexes using Big Data and Machine Learning: A Comparative Analysis and Validity Verification”. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland.

Chessa, A.G. (2016), “A New Methodology for Processing Scanner Data in the Dutch CPI”, Eurona 1/2016, 49-69.

Chessa, A.G., Verburg, J., and Willenborg, L. (2017), “A comparison of price index methods for scanner data”. Paper presented at the 15th Meeting of the Ottawa Group on Price Indices, 10-12 May 2017, Eltville am Rhein, Germany.

Bhardwaj, H. Flower, T., Lee, P., Mayhew, M. (2017), “Research indices using web scraped price data: August 2017 update”, ONS Research Article, Office of National Statistics (ONS).

de Haan, J. and F. Krsinich (2014), “Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes”, Journal of Business & Economic Statistics 32, 341-358.

Hov, K., and Johannessen, R. (2018). “Using scanner data for sports equipment”. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland.

Ivancic, L., W.E. Diewert and K.J. Fox (2009), “Scanner Data, Time Aggregation and the Construction of Price Indexes,” Discussion Paper 09-09, Department of Economics, University of British Columbia, Vancouver, Canada.

Ivancic, L., W.E. Diewert and K.J. Fox (2011), “Scanner Data, Time Aggregation and the Construction of Price Indexes,” Journal of Econometrics 161, 24-35.

Office of National Statistics (ONS) (2017), “Research indices using web scraped price data: clothing data”, ONS Research Article, Office of National Statistics (ONS).

Van Loon, K., and Roels, D. (2018). “Integrating big data in the Belgian CPI”. Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 7-9 May 2018, Geneva, Switzerland.

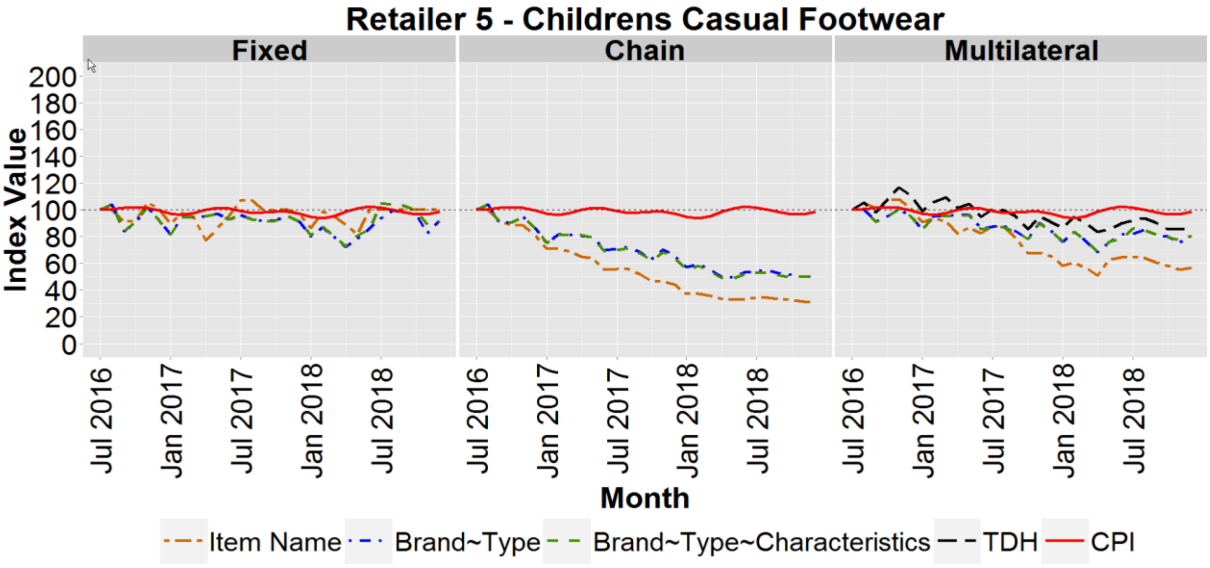
Willenborg, L. (2017). “Stratification and price index computation”. CBS Discussion paper, Statistics Netherlands.

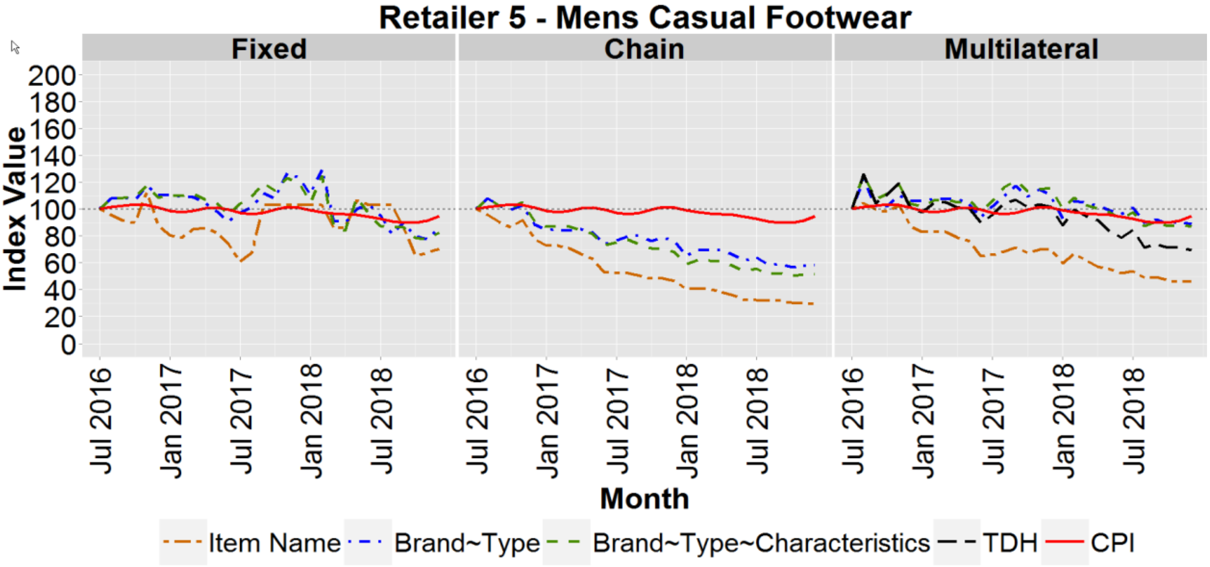
7. Appendices

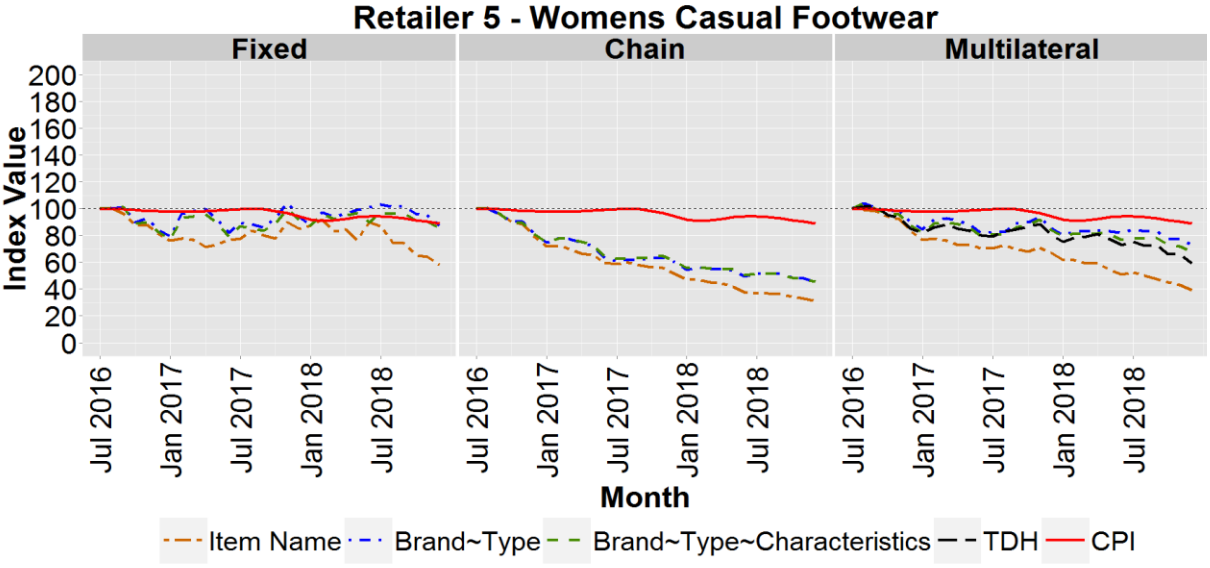
Appendix 1: Full suite of possible variables extracted

Variable	**Example of characteristics**
Brand (text)	Nike, Adidas.
Type (text)	T-shirt, Blouse, Jeans.
Pack size (numeric)	2-pack, 2-pairs.
Material (text)	Cotton, silk.
Gender (text)	Boys, girls.
Length (text)	Maxi, mini.
Metal (text)	Gold, silver.
Method (text)	Jacquard, knit.
Neck type (text)	Crew, turtleneck.
Sleeve length (text)	Short-sleeve, long-sleeve.
Style (text)	Oversized, cropped.
Textile (text)	Corduroy, denim.

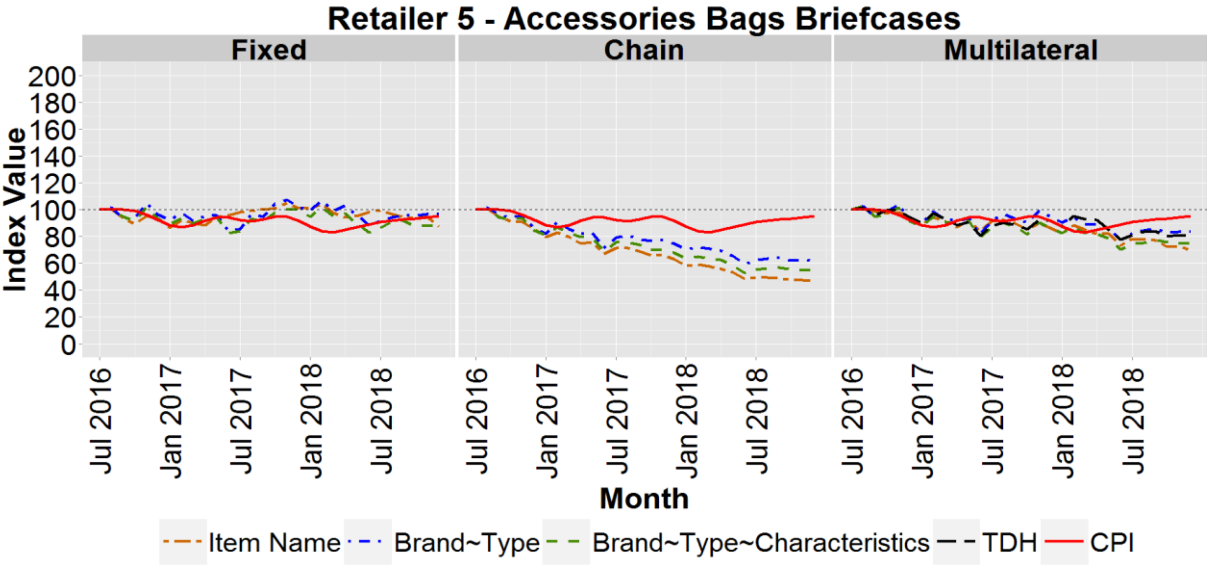
Appendix 2: Footwear Elementary Aggregates for Retailer 5

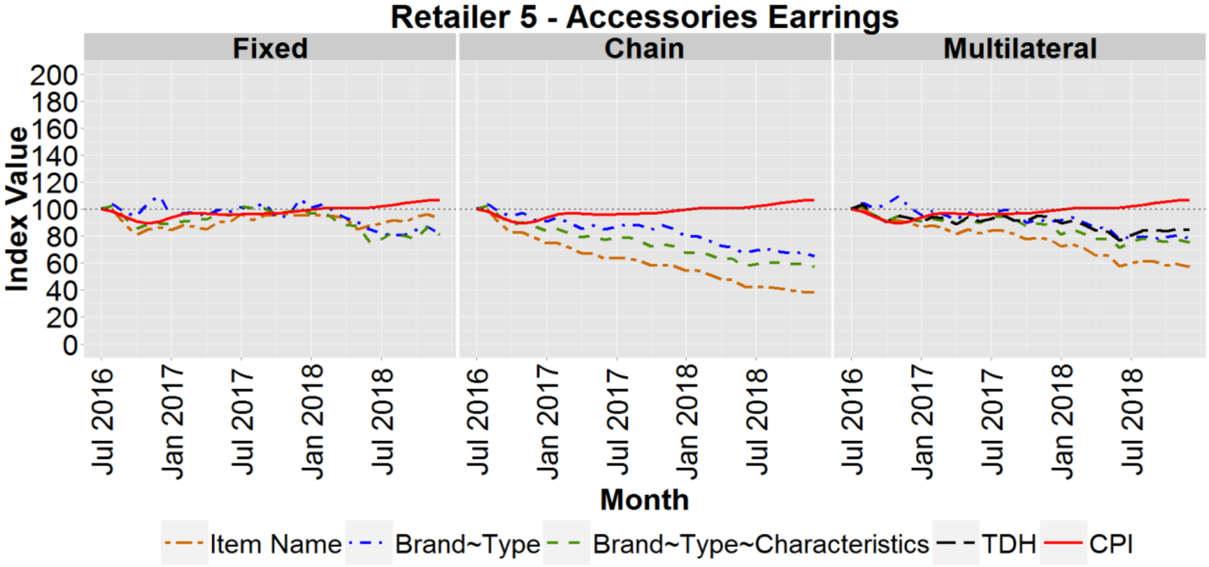


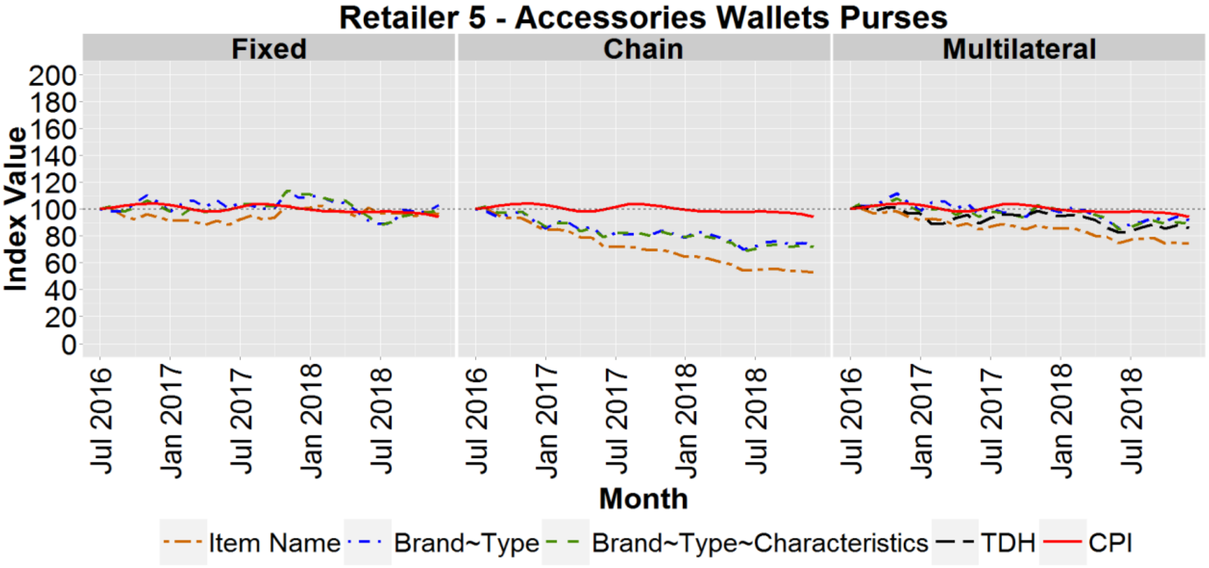




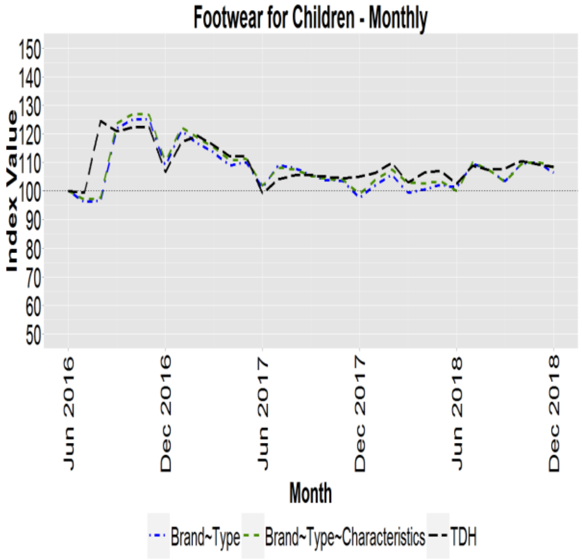
Appendix 3: Accessories Elementary Aggregates for Retailer 5

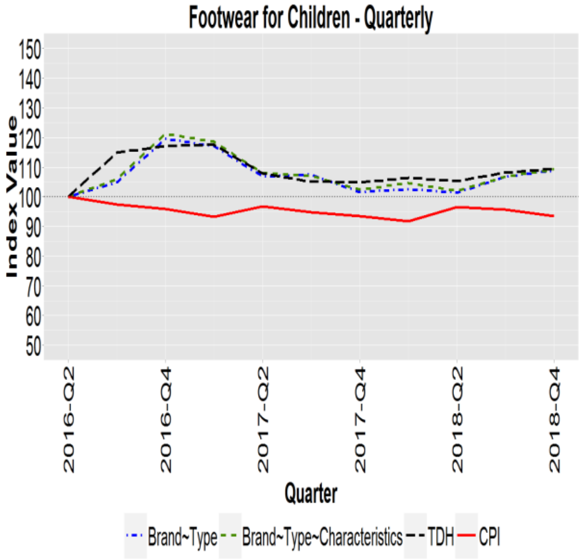


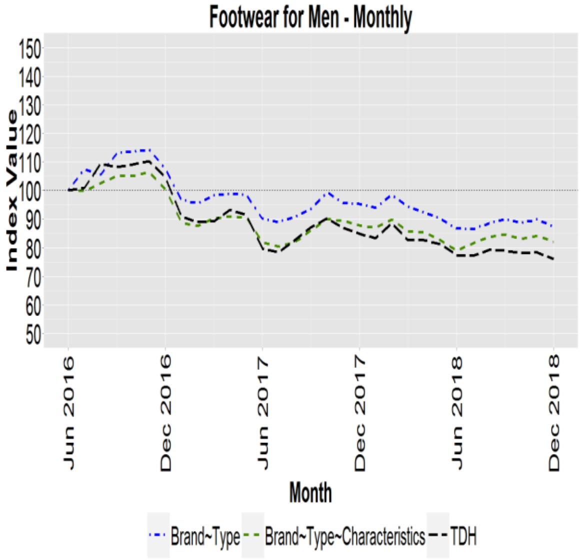


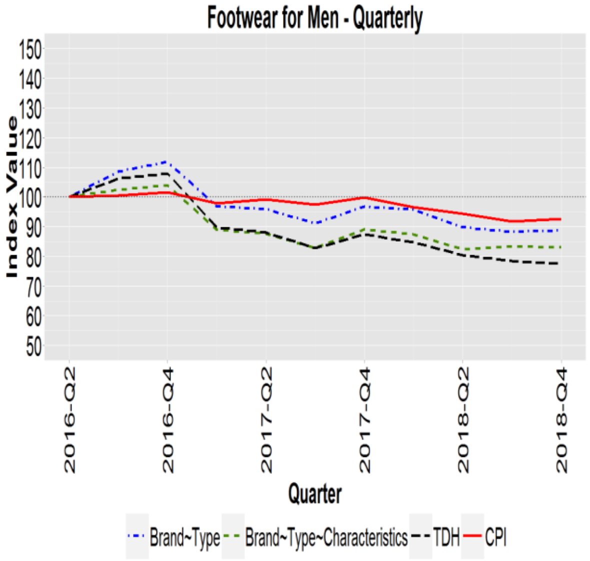


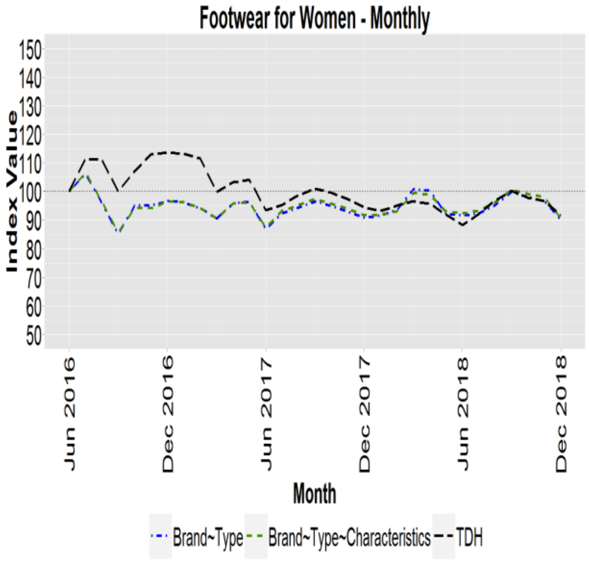
Appendix 4: Footwear Expenditure Class Indexes

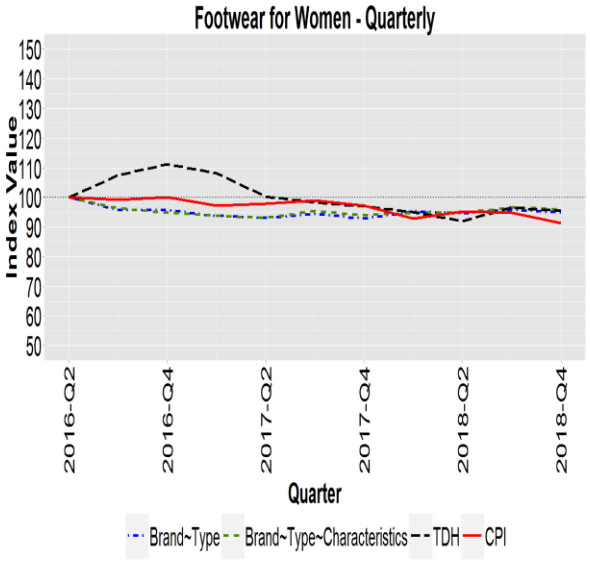




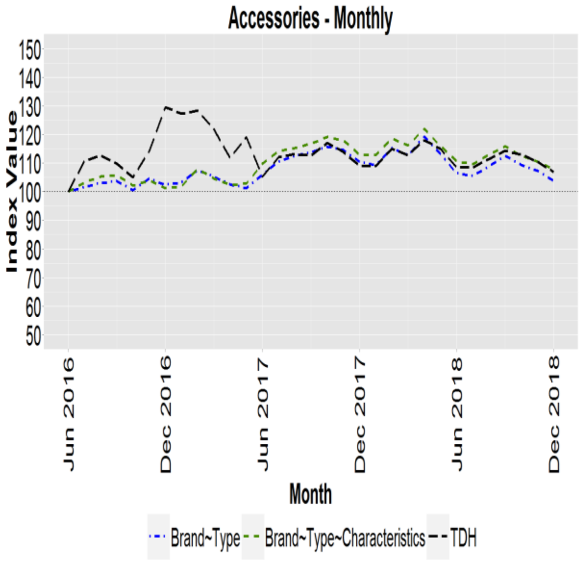


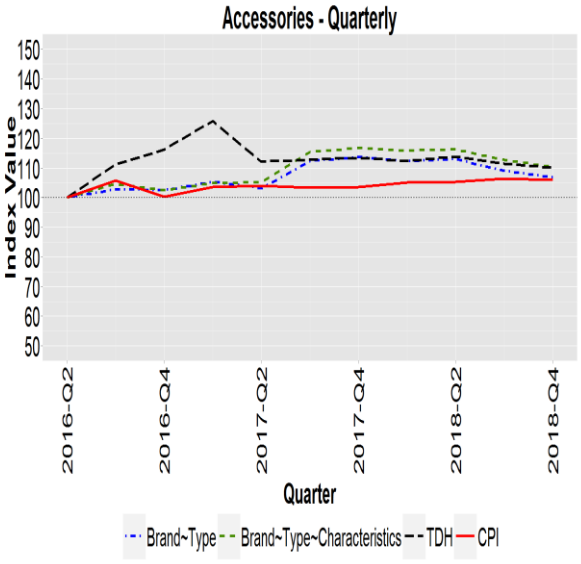






Appendix 5: Accessories Expenditure Class Indexes


Accessories - Monthly
Index Value
Jun 2016
Dec 2016
Jun 2017
Dec 2017
Jun 2018
Dec 2018
Month
Brand~Type
Brand~Type~Characteristics
TDH


Accessories - Quarterly
Index Value
2016-Q2
2016-Q4
2017-Q2
2017-Q4
2018-Q2
2018-Q4
Quarter
Brand~Type
Brand~Type~Characteristics
TDH
CPI