Office for National Statistics

Article

Research indices using web scraped price data: clothing data

Analysis into using web scraped clothing data to produce experimental price indices, including the CLIP (Clustering Large datasets Into Price indices) method which has recently been developed by ONS.

Contact:
Tanya Flower
cpi@ons.gsi.gov.uk
+44 (0)1633 455171

Release date:
31 July 2017

Next release:
To be announced

Notice

19 December 2017

An error has been found in Research indices using web scraped price data: clothing data, Research indices using web scraped price data: August 2017 update, Analysis of product turnover in web scraped clothing data and its impact on methods for compiling price indices, and ONS methodology working paper series number 12 – a comparison of index number methodology used on UK web scraped price data.

This affects the chained Jevons indices presented in Figures 4 to 9 of the first paper, Figure 6 of the second paper, Figures 9 to 12 of the third paper, and Figures 11 and 12 of the fourth paper along with accompanying commentary and analysis. This was due to an error in the way that the indices were calculated.

Please be aware of this if using this data. We will correct this error in early 2018.

We apologise for any inconvenience. Please contact Chris Payne or Tanya Flower for more information.

Table of contents

1 . Introduction

Alternative data sources such as web scraped and point of sale scanner price datasets are becoming more commonly available, providing large sources of price data from which measures of consumer inflation could potentially be calculated. Over the past few years, new methods have evolved for compiling price indices from such data sources. This article extends previous ONS research on using web scraped data to compile price indices for clothing items. These are early analyses using experimental techniques to help us develop our statistical methodology and are not comparable with headline estimates of inflation. We would strongly caution against their use in economic modelling and analysis.

Clothing items generally experience much higher rates of product churn (that is, products coming in and out of stock) compared with other expenditure categories. This is due to the fast-paced nature of the fashion industry, with high seasonality in clothing ranges and changing fashion trends. This makes it difficult to follow prices over time. For example, a new range of swimwear could be introduced at the beginning of the summer, be heavily discounted at the end of summer and then replaced entirely by winter wear clothing.

This has contributed to a number of measurement challenges when we include clothing prices in our consumer price inflation measures. For example, the clothing and footwear division was a major contributor to the divergence between the CPI (Consumer Prices Index) and RPI (Retail Prices Index) inflation measurements, this is explained in the article CPI and RPI: Increased Impact of the Formula Effect in 2010 (ONS, 2011). These difficulties mean that there is a particular interest in investigating generating clothing price indices using alternative data sources.

This article summarises analysis into using web scraped clothing data to produce experimental price indices from the article Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices (Payne, 2017). It also includes a number of additional indices, including the CLIP (Clustering Large datasets Into Price indices) method, which has recently been developed by Office for National Statistics (ONS). The CLIP method is considered of particular interest as it is based on clustering similar items together and may reduce the problem of product churn associated with calculating price indices using web scraped clothing data.

The data was provided by WGSN (World's Global Style Network), a global trend authority specialising in fashion. The structure of this article is as follows. Section 2 gives some background on alternative data sources for price collection and Section 3 gives some further information on the data used. Section 4 outlines the different methods used to produce price indices with the web scraped data. In Section 5 we present results from the different methods, these indices are then compared with a special aggregate of the CPIH index. In Section 6 we summarise the findings and present areas for future research. Charts for each of the web scraped items and aggregate indices are presented in the "Data" section of this release.

2 . Move to alternative data sources

Alternative data sources such as scanner data and web scraped data have been enabled by technological developments in recent years. Scanner data are datasets collected by retailers as products are scanned through the till. Average prices can then be derived from this transactional data. However, scanner data may not be available for smaller retailers. We have also to date been unable to obtain scanner data from large retailers. We have therefore focused our research on investigating how web scraped data can be used.

Web scraping collects the price information directly from the retailer's websites. A web scraper is a tool that reads the HTMLs on the website and extracts the data needed. The methods required to store, process and analyse web scraped data will also inform our use of scanner data in future.

In January 2014, we began a research project to use web scrapers to collect grocery prices from three online retailers as part of the ONS Big Data project. Since the pilot was launched, we have published a number of updates on research into using web scraped data to produce experimental price indices, including methodologies that differ from the more traditional fixed base indices such as the CPIH (a measure of consumer price inflation that includes owner occupiers' housing costs). The most recent article was Research indices using web scraped data: May 2016 update. A further article was published in November 2016 Research indices using web scraped price data: clustering large datasets into price indices (CLIP), which looked to overcome the problem of high levels of product churn associated with web scraped data by clustering groups of products together.

This work identified a number of benefits in using web scraped data over the traditional method of data collection but it also highlighted a number of limitations. Some of the main issues with using web scraped data to calculate price indices are summarised in this section for reference.

In the traditional method of price collection, price collectors use their market knowledge to select products which are a representative sample of goods and services, while in theory web scraping collects all prices from the website. This greatly increases the coverage of goods and services available, but it also increases the difficulty of forming a representative basket as there is no expenditure information on what products are actually purchased by consumers, unless weights are available from another source. This means that all products will have the same weight.

Studies by other national statistical agencies have shown that this may lead to downward bias in indices that are calculated from web scraped data, compared with indices calculated from other sources such as scanner data (Chessa and Griffioen, 2016). This is a particular problem with online retailers (such as Amazon) who sell a large number of product lines, as shelf space is not a limiting factor.

The large number of prices collected using web scraping also leads to greater product churn (the data contains a higher number of products that move into and out of the market over time). This is less of an issue for the traditional collection as collectors are able to replace products that go out of stock by using comparable replacements if available. Collectors will also use their market knowledge to choose products that they expect to remain in stock and be representative of what consumers purchase (these products are more likely to remain in stock as retailers would keep stock levels high).

However, high product churn can cause difficulties in matching items over time. The issue is compounded when looking at the clothing industry, as it is also a sector which experiences particularly high levels of product churn due to high seasonality and changing fashion trends. These limitations affect the methodologies that can be used to calculate price indices from the web scraped clothing data.

3 . WGSN clothing data

The data used were provided by WGSN (World's Global Style Network), a global trend authority specialising in fashion. They collect daily prices and other information from a number of fashion retailers' websites. When this was supplied in 2015, WGSN obtained data for 37 clothing product categories and 38 retailers in the UK, including a mixture of high street and online only retailers. This has subsequently increased. The web scraped data includes price, product and retailer information. Women's clothing data were provided for the period September 2013 to October 2015 and the men's clothing data were provided for the period August 2014 to October 2015.

The dataset is therefore very large and the product categories do not necessarily match the item descriptions used for the CPIH collection. Therefore, the analysis was restricted to the following nine clothing items, which map relatively closely to items used in the CPIH classification structure. These nine items are listed in Table 1.

Table 1: Clothing items used for analysis

ONS item ID	Item
510106	Men’s Jeans
510124	Men’s Shorts
510131	Men’s Casual Shirt
510413	Men’s Socks
510402	Men’s Pants
510250	Women’s Coat
510254	Women’s Sportswear Shorts
510255	Women’s Swimwear
510415	Women’s Tights

Source: Office for National Statistics, World’s Global Style Network

As discussed in Section 2, clothing is expected to have a high level of product churn due to the nature of the fashion industry. This was found to be the case for the nine products analysed from the WGSN data. The proportion of products in the sample over the whole period (that is, being present in the first and last month of the sample) ranged from 5.12% for men’s jeans to 0.07% for women’s coats. This shows that there is a high level of product churn in the clothing sector with the lifespan of most products being only one season. Further analysis into product churn is given in the article Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices (Payne, 2017).

4 . Constructing price indices

These web scraped clothing data were then used to produce research price indices. Various methods can be used to produce price indices (detailed in Annex A). In the case of web scraped data, the fact that there is no expenditure information limits the methods that can be used. The WGSN data were only considered on a monthly basis due to the size of the dataset and difficulties with processing. To calculate the monthly price for each product, a geometric average of prices is used for consistency with the Jevons formula. The article Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices (Payne, 2017) considered three methods of compiling price indices:

- Monthly chained Jevons Index
- IntGEKS
- FEWS

Further description of these methods is given in Annex A. In this article we also include the following methods to produce price indices from the clothing data:

- CLIP
- RYGEKS
- GEKS

The CLIP method is considered of particular interest as it is based on clustering similar items together and may reduce the problem of product churn associated with web scraped clothing data. It works by grouping together individual products that have similar characteristics using a clustering algorithm. Instead of more traditional methods which track individual products over time, the CLIP tracks the average price of these clusters instead. This reduces the problem of product churn as it means that when products either go in and out of stock or are rebranded, or new products enter the market, they are assigned to the clusters already found in the dataset.

The clusters are the same in each period so they can be compared: even if the exact products the clusters contain are different, they will have the same characteristics as the original products. The geometric mean of the prices of the products in each cluster is taken as the average price for that cluster. Further information on the CLIP is given in the article Research indices using web scraped price data: clustering large datasets into price indices (CLIP) (Metcalfe et al., 2016), which applies the CLIP to our web scraped grocery data. For the clothing data discussed in this article, there are extra characteristics available that help to form the clusters. One such characteristic is the style of the clothing, for example, whether or not a pair of jeans is a skinny jean or a boot cut jean is taken into account when the clusters are formed.

5 . Analysis

In Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices (Payne, 2017), nine clothing items have been investigated (Table 1) and prices indices for these individual items were created. These were the products which best mapped to the Office for National Statistics (ONS) item IDs used in the Consumer Prices Index including owner occupiers' housing costs (CPIH). In this article, we aggregate these items to produce three aggregate indices (Table 2). Charts for each of the items and aggregate indices are presented in the "Data" section of this release.

Table 2: Clothing items included in each aggregation

Aggregates	Items included in Aggregate
Men's clothing	Men's Jeans, Men's Shorts, Men's Casual Shirt, Men's Socks, Men's Pants
Women's clothing	Women's Coats, Women's Sportswear Shorts, Women's Swimwear, Women's Tights
All clothing	Men's Jeans, Men's Shorts, Men's Shirts, Men's Socks, Men's Pants, Women's Coats, Women's Shorts, Women's Swimwear, Women's Tights

Source: Office for National Statistics, World's Global Style Network

Published CPIH expenditure weights are used to aggregate up the item level indices to the higher level aggregates. This method is applied consistently to all index methods described. The CPIH weights are applied to the unchained indices for those indices that undergo a chaining process. These indices are then re-chained to get the indices published in the attached "Data" section of this release.

These experimental indices are early analysis to help us develop our statistical methodology for alternative sources of prices data and we would therefore caution against their use in economic modelling and analysis.

GEKS indices

In this article, we consider three versions of the GEKS index: GEKS, RYGEKS and IntGEKS.

A GEKS index (originally proposed by Gini, Eltetö, Köves and Szulc) is one possible solution to high rates of product churn that can be implemented without introducing significant amounts of chain drift. The GEKS method essentially takes the geometric mean of all bilateral indices connecting all of the periods between the base period and the current period. GEKS indices are free from chaining issues. However, a drawback with using the GEKS index for temporal indices is that whenever a new time point is added the entire index will be revised. This is a significant problem as official price indices are very rarely revised backwards once published.

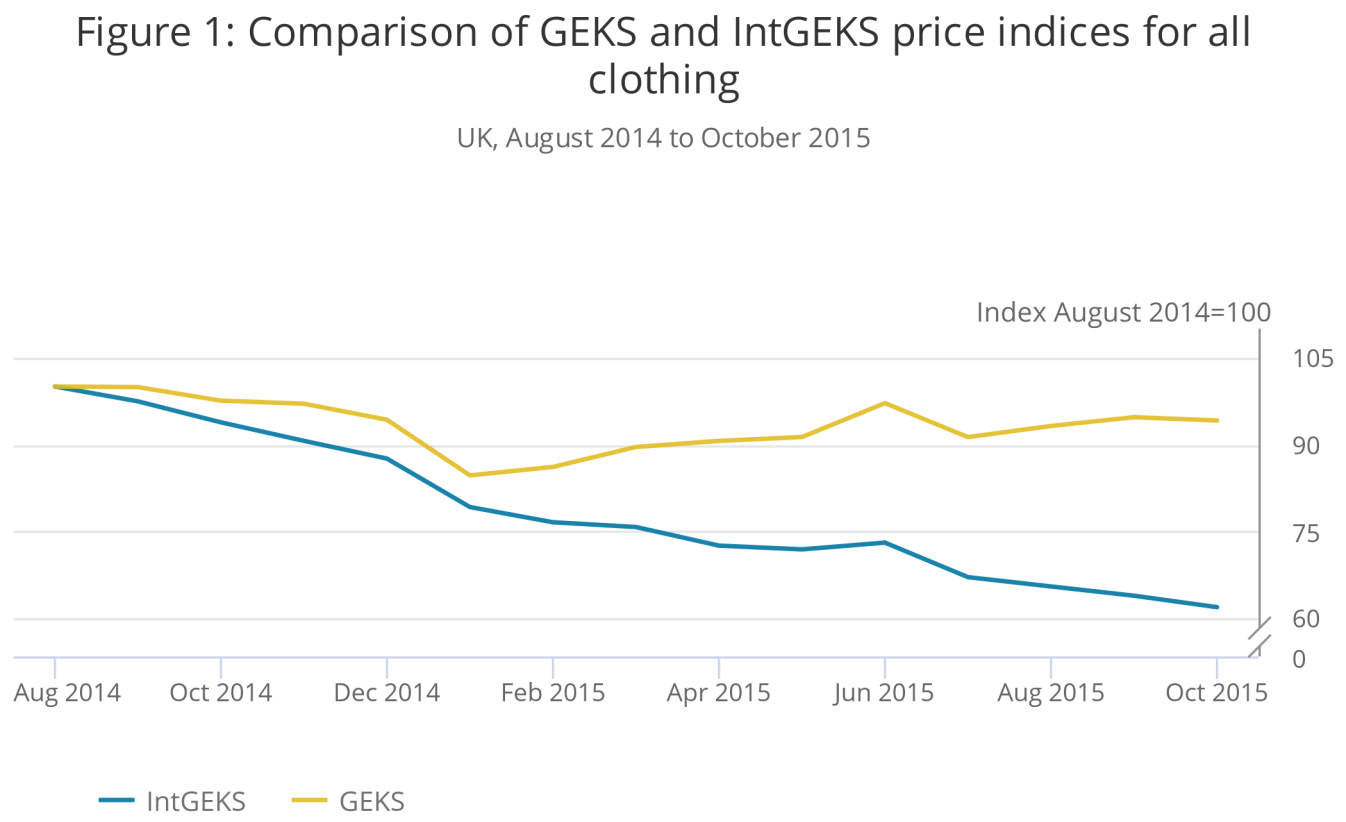
The RYGEKS is a version of the GEKS that was developed to remove this problem as it is based on a rolling year period that allows longer series to be calculated without the need to revise. For a product to be included in the GEKS or the RYGEKS it has to appear in either the base period and the intermediate period or the intermediate period and the current period. For example, if the price change between Monday and Wednesday is calculated then a product must appear in either Monday and Tuesday or in Tuesday and Wednesday.

The Intersection-GEKS (IntGEKS) index, developed by Krsinich and Lamboray (2015) is another version of RYGEKS that uses a matched set of products for bilateral comparison. These products must appear in all three periods: in the example given previously, the products must be in Monday, Tuesday and Wednesday to be included in the calculations. All the GEKS methods in this article use Jevons indices as an input into the GEKS procedure. Further information on these methods is given in Annex A.

Figure 1 and Figure 2 present the GEKS and IntGEKS for all clothing and men's clothing respectively. Figure 3 presents the GEKS, RYGEKS and IntGEKS for women's clothing. RYGEKS is not calculated for all clothing and men's clothing because the time period is too short to allow for a useful comparison to be made.

Figure 1: Comparison of GEKS and IntGEKS price indices for all clothing

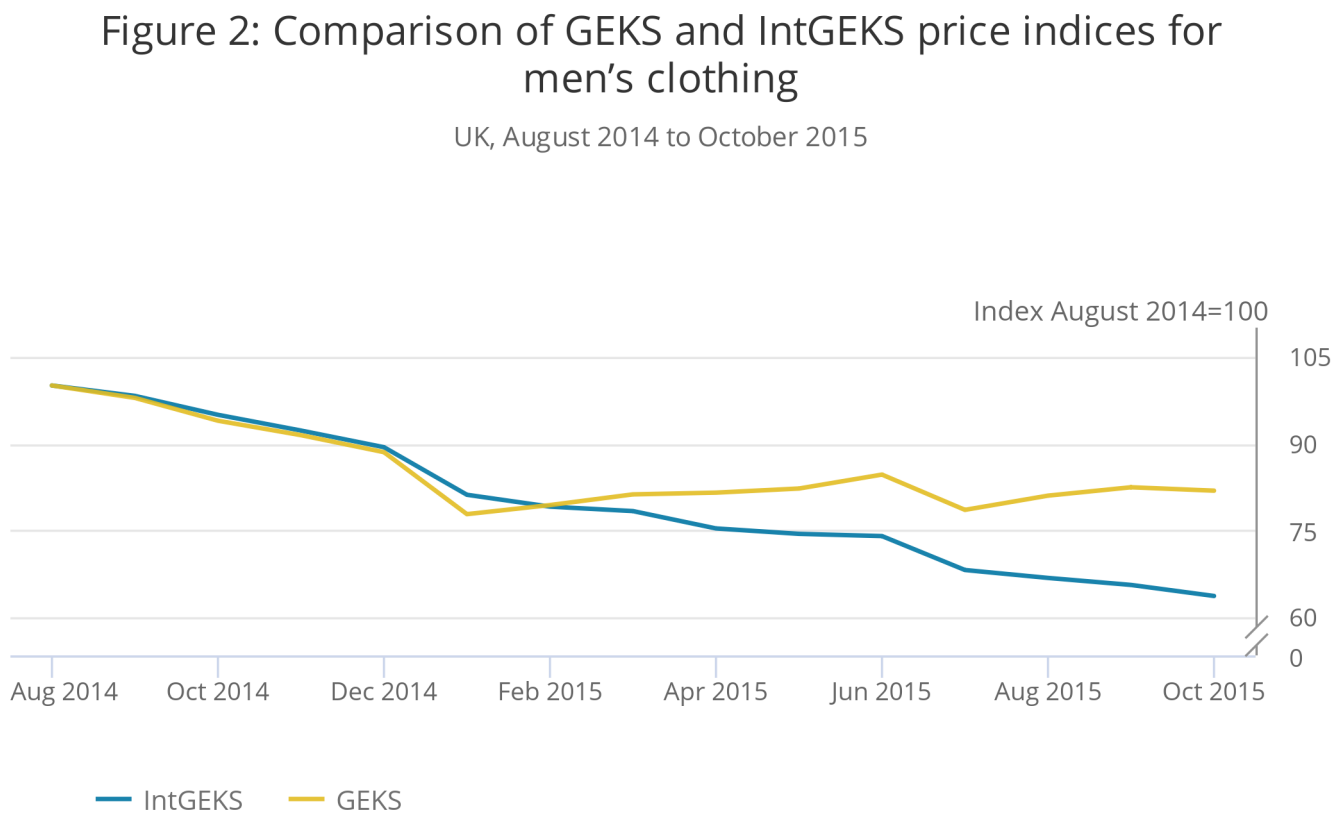
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 2: Comparison of GEKS and IntGEKS price indices for men's clothing

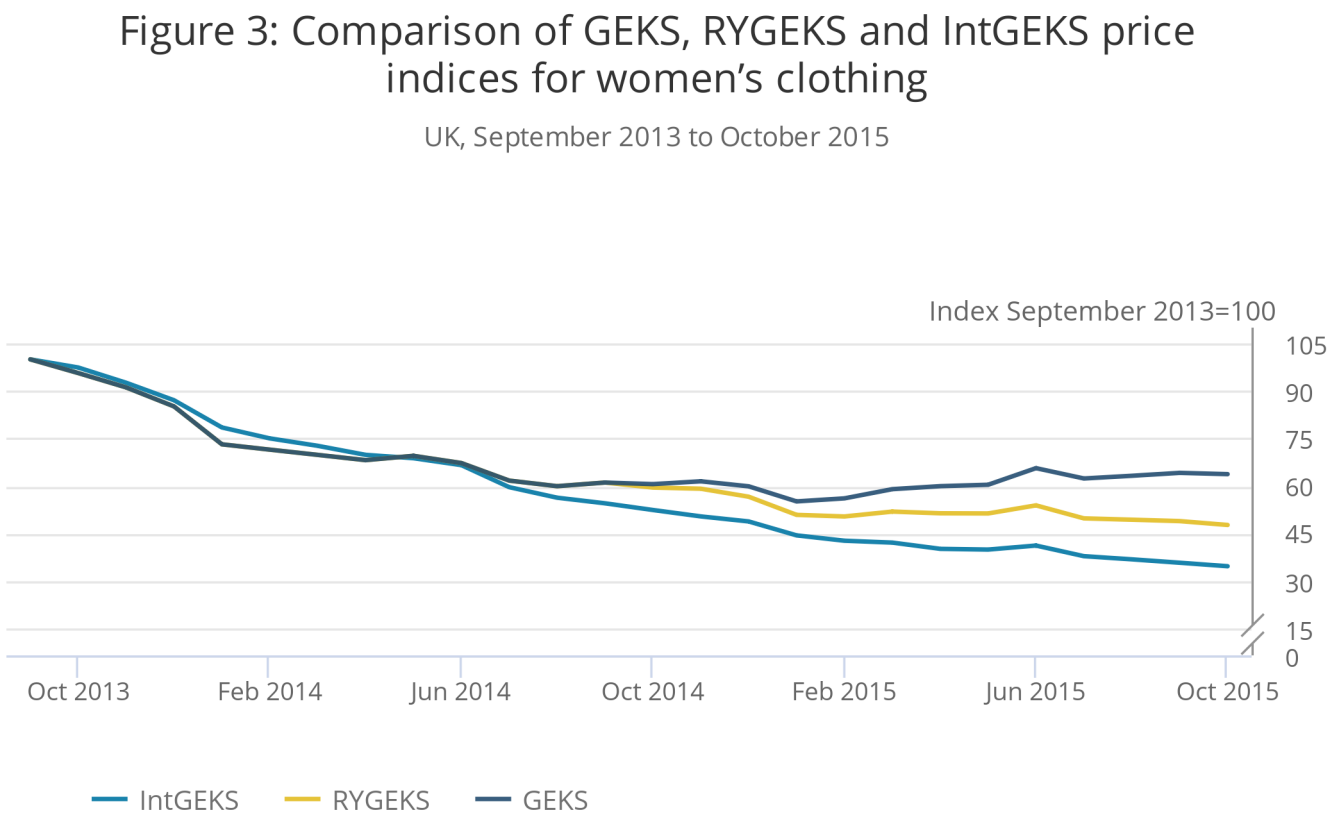
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 3: Comparison of GEKS, RYGEKS and IntGEKS price indices for women's clothing

UK, September 2013 to October 2015



Source: Office for National Statistics, World's Global Style Network

Notes:

1. Data for women's clothing covers the period September 2013 to October 2015, which means that care should be taken when comparing the graphs for women's clothing with men's clothing (August 2014 to October 2015) and all clothing (August 2014 to October 2015), as they cover different time periods.

At the all clothing aggregate level, both versions of the GEKS indices decrease over the period. This is also the case for the men's and women's clothing aggregates (although please note the different time periods covered).

IntGEKS decreases further and at a faster rate than the other GEKS indices. At the item level, some of the IntGEKS series decreased by around 75% since September 2013, such as women's coats. This level of decrease is clearly implausible even given the nature of the fashion industry.

The RYGEKS and GEKS series have also experienced very large decreases for women's clothing. This may be due to the longer run of data available for women's clothing: the men's clothing indices may decline in a similar way if we had a longer run of data. This might also be because a matched index such as IntGEKS means that the longer staying items in the dataset have more influence on the index, and these longer staying items have the possibility of having larger price decreases in order for stock to be shifted. By comparison, the GEKS and RYGEKS incorporate new products in the index and therefore these products may decrease the influence of the longer staying products on the resulting series.

Chained Jevons, FEWS and CLIP

Three other methods were considered besides the different versions of GEKS to construct price indices from the WGSN clothing data. These were the chained Jevons (referred to as the “daily chained” in previous ONS research articles), FEWS and CLIP methods.

The chained Jevons is a simple method that applies a monthly chained index to the web scraped data.

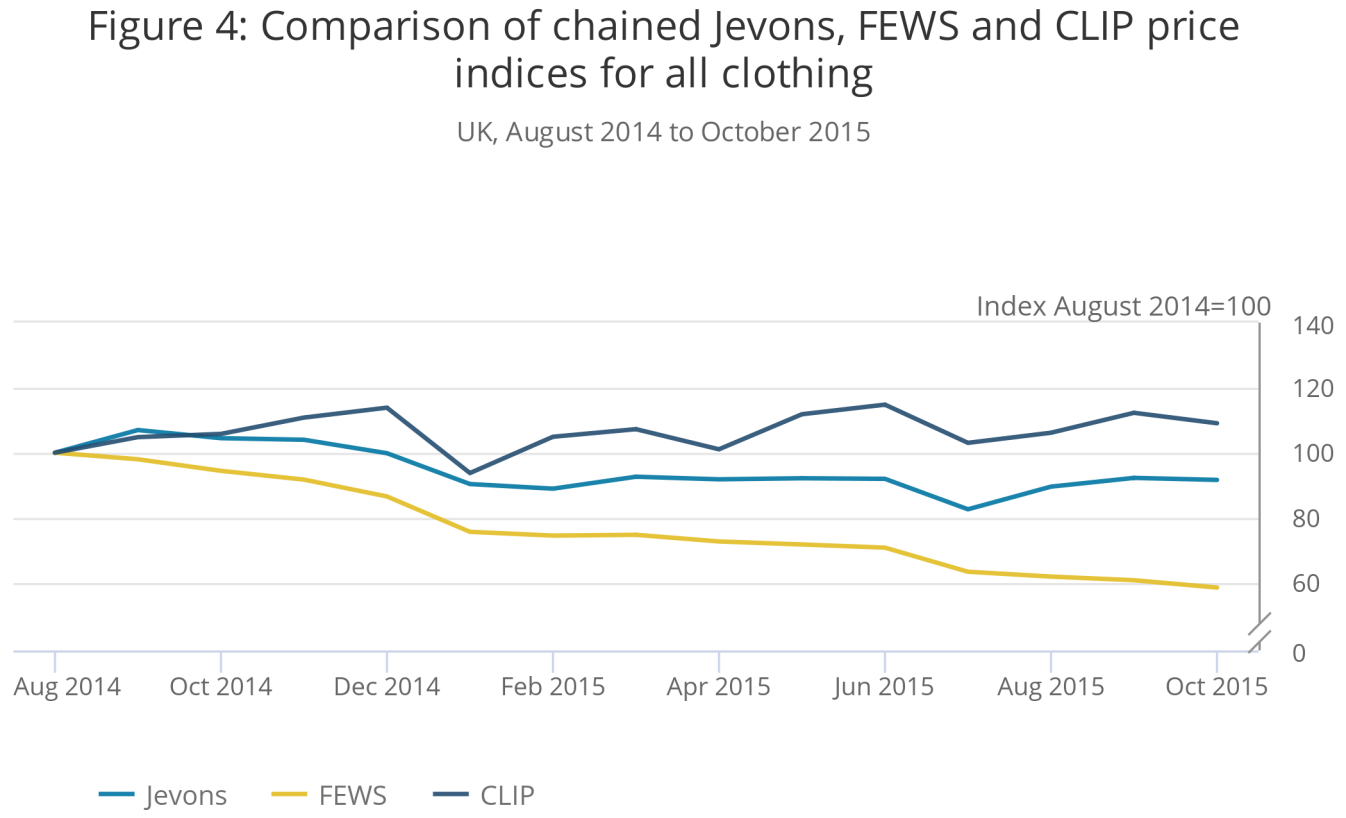
The FEWS method (Fixed Effect with a Window Splice) was developed by Statistics New Zealand (Krsinich, 2014) to account for quality change in big data sources. It works by decomposing log price into a time-based effect and a product-based effect, assuming that the change in the product-based effect accounts for quality change.

CLIP (Clustering Large Datasets into Price Indices) was developed by ONS. It clusters products into similar groups, based on the theory that consumers want to purchase different types of products rather than specific individual products. Further information on the CLIP is given in this article: Research indices using web scraped price data: clustering large datasets into price indices (CLIP). CLIP was designed to help with the problem of high product churn that exists with alternative data sources such as web scraped data and particularly with web scraped clothing data. As such, it may give more credible estimates than the other methods discussed in this article. Further information on all these methods is given in Annex A.

Figure 4, Figure 5 and Figure 6 present the chained Jevons, FEWS and CLIP for all clothing, men’s clothing and women’s clothing respectively.

Figure 4: Comparison of chained Jevons, FEWS and CLIP price indices for all clothing

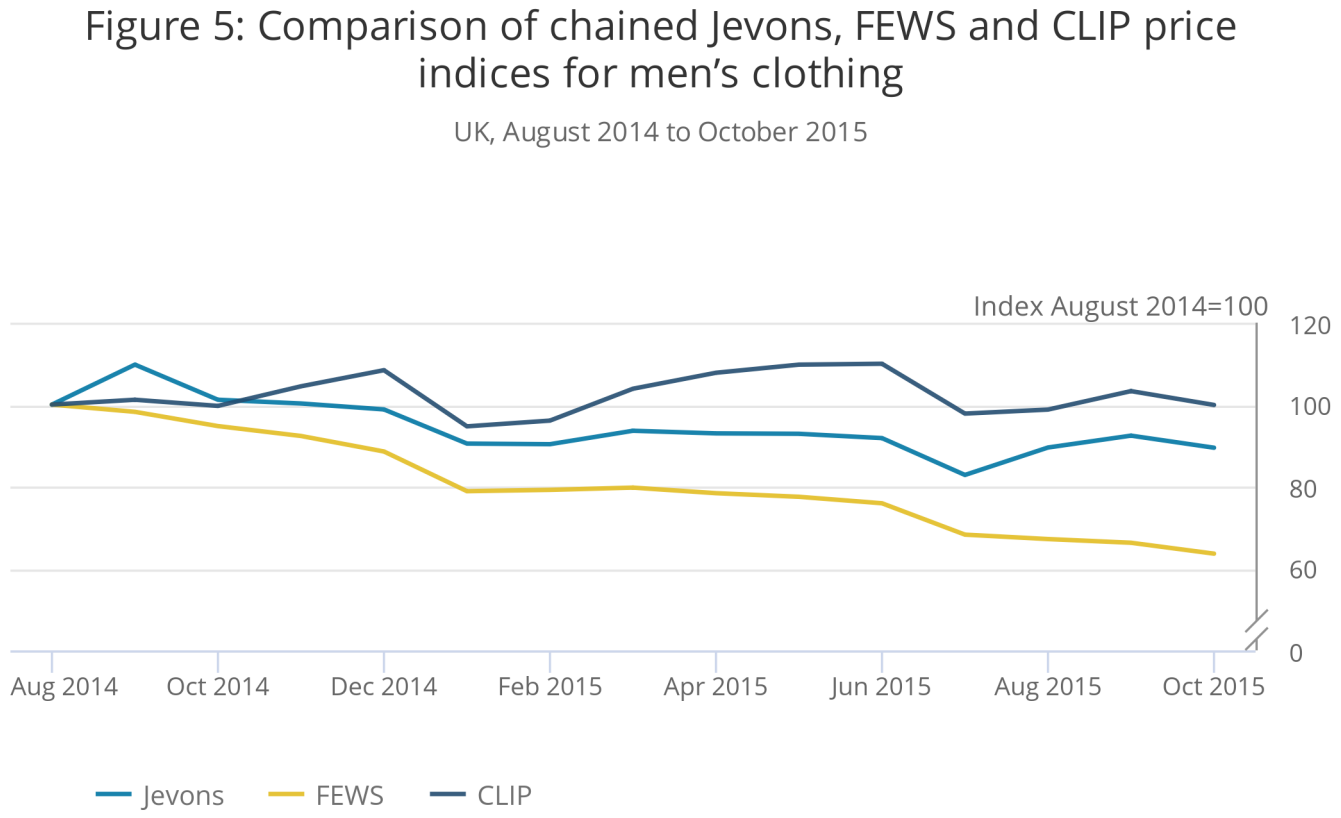
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 5: Comparison of chained Jevons, FEWS and CLIP price indices for men's clothing

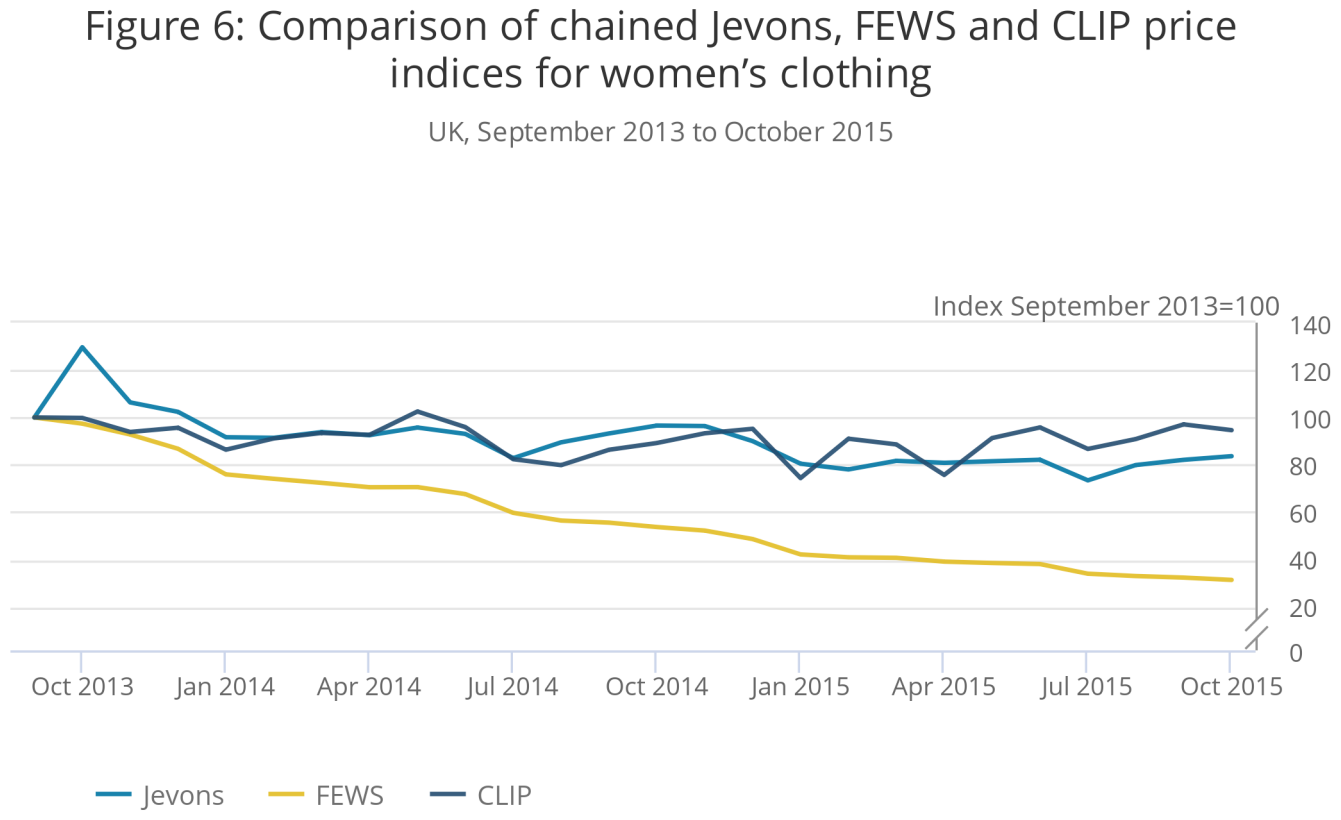
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 6: Comparison of chained Jevons, FEWS and CLIP price indices for women's clothing

UK, September 2013 to October 2015



Source: Office for National Statistics, World's Global Style Network

Notes:

1. Data for women's clothing covers the period September 2013 to October 2015, which means that care should be taken when comparing the graphs for women's clothing with men's clothing (August 2014 to October 2015) and all clothing (August 2014 to October 2015), as they cover different time periods.

For all clothing, the FEWS indices descend in a similar way to the IntGEKS indices and are therefore not plausible over a longer time series. The chained Jevons and the CLIP both show more sensible price movements with some seasonality in the indices.

In general, the chained Jevons indices decrease over the period but show a more plausible decline than the FEWS index. At the item level, some clothing items have seen large increases over the period investigated. From September 2013 to October 2015, women's coats have increased by 25% and women's shorts have increased by 37% as measured by the chained Jevons. By contrast, the CLIP all clothing index increases over the period from August 2014 to October 2015 by 9%. This increase is driven largely by women's coats, women's tights and men's shirts, which have increased significantly over the shorter period for which all clothing data has been available. However, the CLIP indices are also more volatile.

Comparison with CPIH

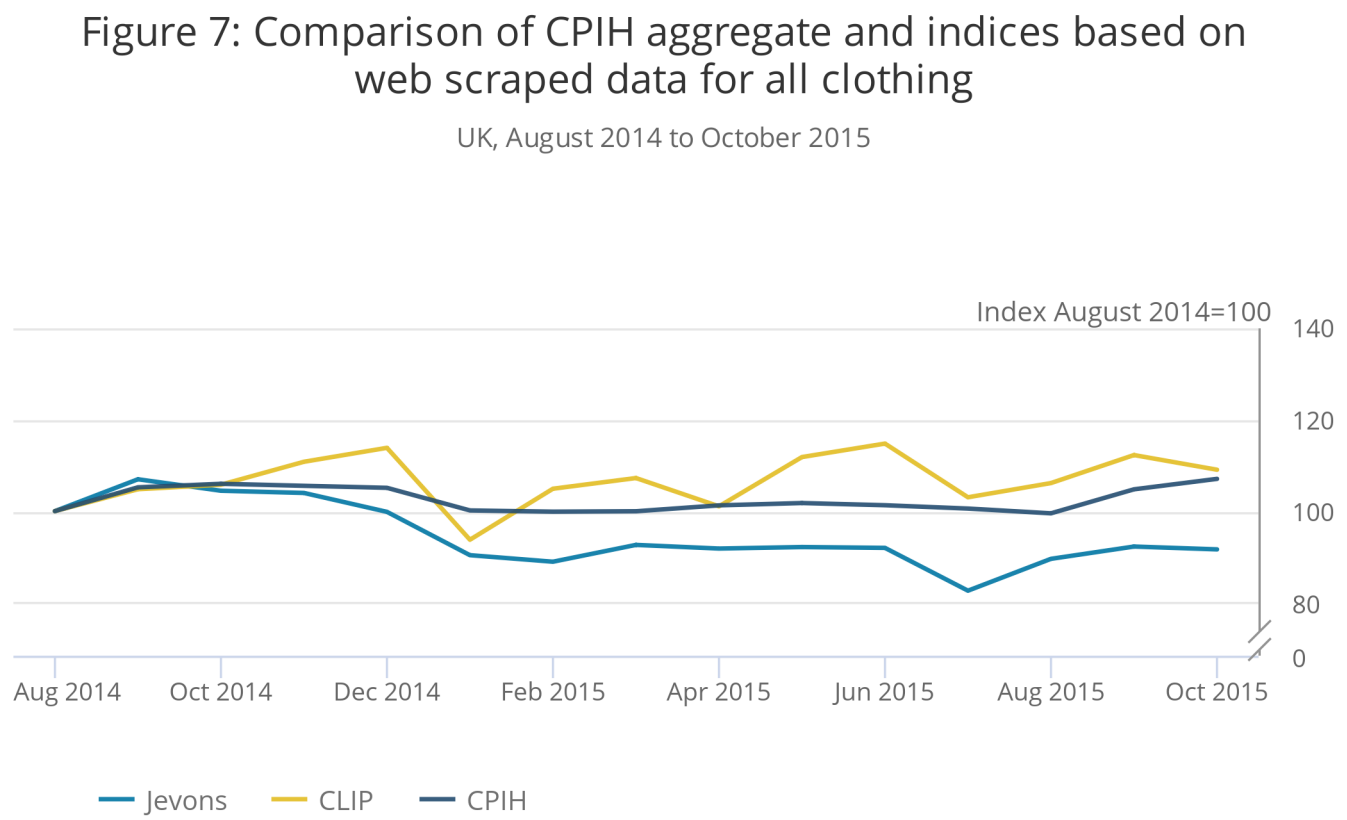
There are many reasons why it is not appropriate to draw a direct comparison between the price indices presented in this section and the published CPIH. These include differences in data sources and methodology used. Further information on these differences is given in Research indices using web scraped data (Breton. R. et al, 2016).

However, it is still a useful exercise to examine the trends shown in the different indices and therefore in the final part of this section we construct special aggregates of published CPIH item indices, using only the items and weights that have been used in this analysis. This allows us to compare the price indices presented earlier in this section with published CPIH data using similar items. Nevertheless, despite the steps taken, we would expect these indices and the published CPIH to be different, given that many methodological differences remain. The FEWS, IntGEKS, RYGEKS and GEKS series are excluded as these gave more implausible results. However, it should be noted that the CPIH data is not necessarily a benchmark due to the difficulties outlined previously for the traditional collection.

Figure 7, Figure 8 and Figure 9 present the CPIH special aggregate, chained Jevons and CLIP for all clothing, men's clothing and women's clothing respectively.

Figure 7: Comparison of CPIH aggregate and indices based on web scraped data for all clothing

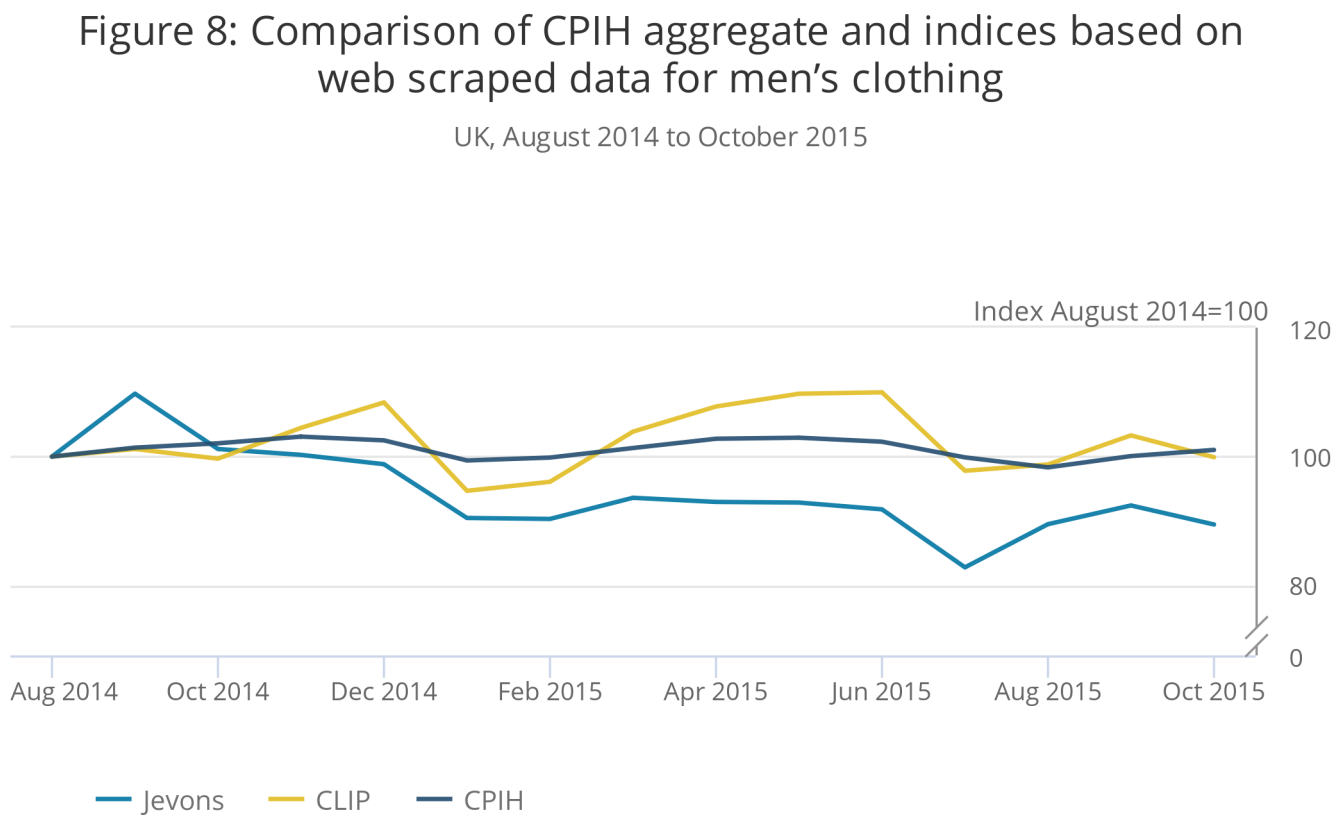
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 8: Comparison of CPIH aggregate and indices based on web scraped data for men's clothing

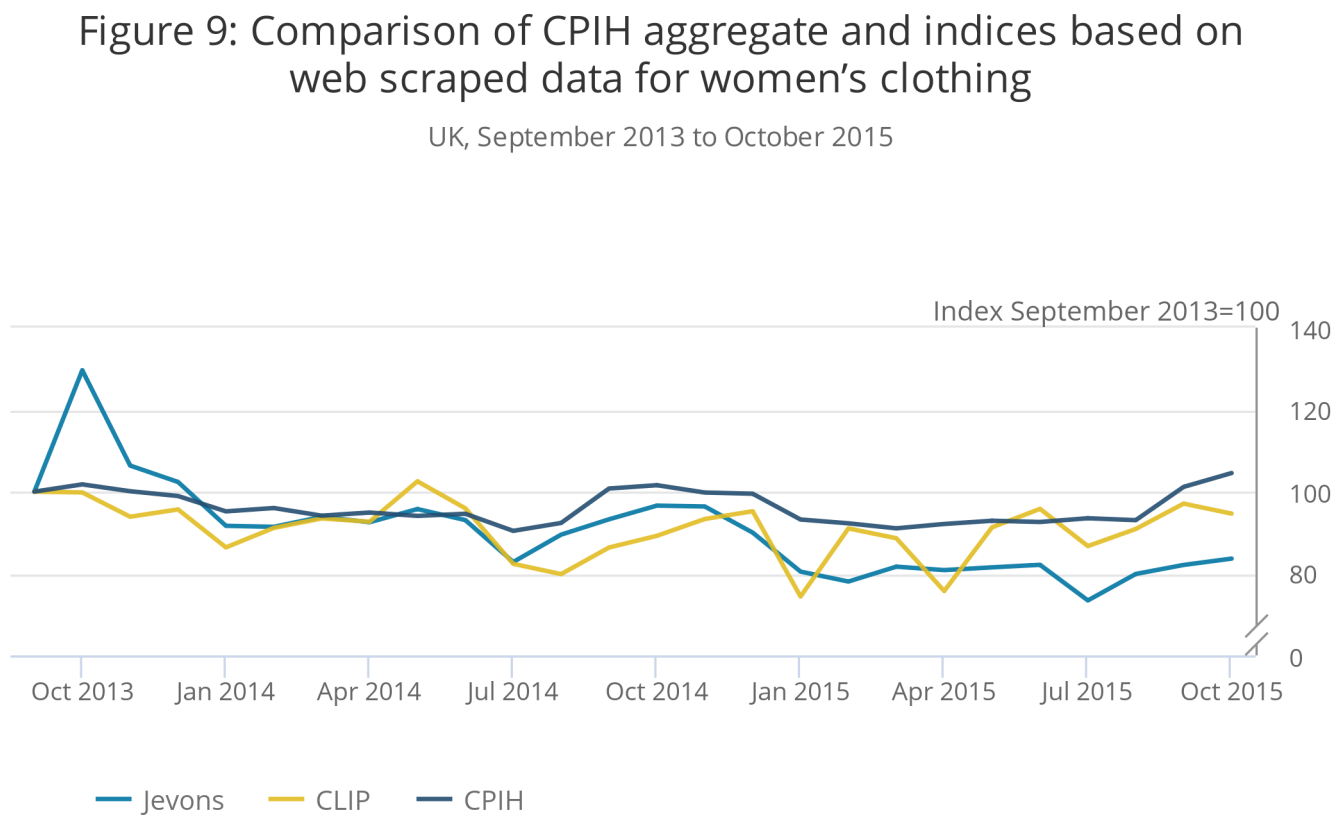
UK, August 2014 to October 2015



Source: Office for National Statistics, World's Global Style Network

Figure 9: Comparison of CPIH aggregate and indices based on web scraped data for women's clothing

UK, September 2013 to October 2015



Source: Office for National Statistics, World's Global Style Network

The CPIH all clothing aggregate index remains relatively stable over the period, with a slight upward trend. This is also the case for the men's and women's clothing aggregates. The CLIP matches the CPIH trend, but also appears to be more volatile. However, this may be a better reflection of the seasonality displayed in the fashion industry. This is distinct from the chained Jevons, which decreases over the period investigated.

6 . Conclusion

Previous ONS research carried out on alternative data sources for consumer prices has focused on web scraped grocery data. Due to the growth of online retailing, it is important to extend this analysis to other retail sectors. The web scraped clothing data obtained from WGSN gives us the opportunity to apply price index methodology suitable for big data sources to a new section of the consumer prices basket. Web scraping offers the potential to improve greatly the quality and efficiency of consumer price indices. In particular for clothing, web scraped data may also offer the opportunity to overcome problems associated with the traditional collection of clothing prices, such as high seasonality and product churn.

However, there remain a number of limitations to using this data to construct price indices, including problems with processing and cleaning large datasets. Questions remain around whether all web scraped data should be used (that may reduce the representativeness of the products included within the analysis) or whether a sample should be taken. These must be resolved before the data can be put to effective use within consumer prices.

Six different methods for constructing research price indices from the web scraped clothing data were investigated: CLIP, chained Jevons, GEKS, IntGEKS, RYGEKS and FEWS. These were then compared against a special aggregate of the published CPIH index. These are early analyses using experimental techniques to help us develop our statistical methodology and are not comparable with the headline estimate of inflation. We would strongly caution against their use in economic modelling and analysis.

In all cases, the IntGEKS and FEWS methods decline rapidly over the period and therefore do not give plausible price indices from these data. While the GEKS and RYGEKS indices better reflect the seasonal movement in clothing prices, the decline is still implausibly large. This is particularly true for women's clothing, for which we have a longer run of data. By contrast, the chained Jevons index is actually rather successful in identifying the seasonal peaks and troughs of these particular items and the general level of the indices seems reasonable. The magnitude of the index movements, however, seems greatly exaggerated.

A new approach trialled by ONS called the CLIP was also applied to the web scraped clothing data. This approach clusters products into similar groups, based on the theory that consumers want to purchase different types of products rather than specific individual products. The CLIP aggregate for all clothing increases over the period investigated, matching the CPIH trend, but it also appears to be more volatile. This may be a better reflection of the seasonality displayed in the fashion industry.

This work contributes to a growing body of research into large alternative sources of price data and its results are useful in developing methods for scanner data, as well as web scraped data. Despite the issues faced in producing price indices, web scraped data have the potential to deepen our understanding of price movements in the clothing sector in the medium-term and, in the long-term, improve the way prices are collected for national consumer price indices. This particular piece of research also contributes to work we are doing to improve the collection of clothing price data, outlined in the Consumer prices development plan.

7 . References

Breton R, et. al. (2015): Research indices using web scraped data

Breton R, et. al. (2016): Research Indices using web scraped data: May 2016 update

Chessa, A.G., and Griffioen, R. (2016): "Comparing Scanner Data and Web Scraped Data for Consumer Price Indices". Report, Statistics Netherlands.

Krsinich F (2014): The FEWS index: Fixed effects with a window splice; non-revisable quality-adjusted price indexes with no characteristic information

Metcalfe E, et . al. (2016): Research Indices using web scraped data: clustering large datasets into price indices (CLIP)

Office for National Statistics (2011): "Increased Impact of the Formula Effect in 2010"

Office for National Statistics (2014): Consumer price indices – technical manual

Payne C (2017): Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices

8. Annex A: Outline of the methodologies used to produce the prices indices

Chained bilateral Jevons indices

The chained bilateral index involves constructing bilateral Jevons indices between period *t* and *t-1* and then chaining them together. The formula is defined as follows:

$$P_{CJ}^{0,t} = \prod_{i=1}^{t} P_J^{i-1,i} = \prod_{i=1}^{t} \left(\prod_{j \in s^{i-1,i}} \frac{p_j^i}{p_j^{i-1}} \right)^{\frac{1}{n^{i-1,i}}}$$

where $P_J^{i-1,i}$ is the Jevons index between the current period and the previous period, p_j^i is the price of product *j* at time *i*, $S^{i-1,i}$ is the set of products observed in both period *i* and *i-1*, and $n^{i-1,i}$ is the number of products in $S^{i-1,i}$.

GEKS

All the GEKS indices in this article use Jevons indices as an input into the GEKS procedure.

The GEKS Family of Indices is a set of indices that is based on a formula devised by Gini, Eltetö, Köves and Szulc:

a. The GEKS-J Index:

The GEKS-J index is a multilateral index, as it is calculated using all routes between two time periods. It was originally developed for purchasing power parities but adapted for the time domain in Diewert, W.E., Fox K,J., and Ivancic, L. (2009) The GEKS-J price index for period t with period 0 as the base period is the geometric mean of the chained Jevons price indices between period 0 and period t with every intermediate point (i = 1,...,t-1) as a link period. The formula is defined as follows:

$$P_{GEKSJ}^{0,t} = \prod_{i=0}^{t} \left(P_J^{0,i} P_J^{i,t} \right)^{\frac{1}{t+1}}$$

A product is included in the index if it is in the period *i* and either period 0 or period *t*.

b. RYGEKS-J:

RYGEKS-J or Rolling Year GEKS-J extends the GEKS-J to allow for a moving base period and allows for a longer series to be calculated without the need to revise the back series constantly. The formula is defined as follows:

$$P_{RYGEKS-J}^{0,t} = \begin{cases} \prod_{i=0}^{t} \left(P_J^{0,i} P_J^{i,t} \right)^{\frac{1}{t+1}} & t < d \\ \prod_{i=0}^{d-1} \left(P_J^{0,i} P_J^{i,d-1} \right)^{\frac{1}{d}} \prod_{k=d}^{t} \left(\prod_{i=k-d+1}^{k} \left(P_J^{k-1,i} P_J^{i,k} \right)^{\frac{1}{d}} \right) \end{cases}$$

where *d* is the window length, for a monthly series d=13. A formal definition of RYGEKS is in De Haan and van der Grient (2009).

c. The Intersection-GEKS-J or IntGEKS-J:

The IntGEKS was devised by Krsinich and Lamboray (2015), to deal with an apparent flattening of RYGEKS under longer window lengths, though this was found to be an error in applying the weights. It removes the asymmetry in the match sets between periods 0 and *i* and between periods *i* and *t*, by including products the matched sets only if they appear in all three periods, the set $S^{0,i,t}$. The formula is defined as follows:

$$P_{IntGEKSJ}^{0,t} = \prod_{i=0}^{t} \left(P_{J,j \in s^{0,i,t}}^{0,i} P_{J,j \in s^{0,i,t}}^{i,t} \right)^{\frac{1}{t+1}}$$

If there is no product churn (products coming in and out of stock) then the IntGEKS-J reduces to the standard GEKS-J. The IntGEKS-J has more chance of “failing” than a standard GEKS-J as the products need to appear in more periods.

CLIP

Clustering Large datasets Into Price indices is a recently developed price index from ONS. The CLIP groups products into clusters and tracks those clusters over time. In the base period the products are clustered according to their characteristics, for example, if the product was on offer, as it assumes consumers would buy within a certain set of products on offer. Clusters are formed using the same rules over time, but the products that form the cluster can change over time, allowing for product churn. The geometric mean of the clusters in two periods are compared, creating a unit value index for each cluster, which are then aggregated using the size of the cluster in the base period. Mathematically, the formula is defined as follows:

$$P_{CLIP}^{0,t} = \frac{\sum_k |C_{k,0}| \frac{\left(\prod_{j \in k,t} p_j^t\right)^{\frac{1}{|C_{k,t}|}}}{\left(\prod_{j \in k,0} p_j^t\right)^{\frac{1}{|C_{k,0}|}}}}{\sum_k |C_{k,0}|}$$

where $C_{k,0}$ is cluster k in period 0, $C_{k,t}$ is cluster k in period t, and $|C_{k,0}|$ is the size of a cluster. For a full description, please read Metcalfe et al. (2016).

FEWS

The Fixed Effects Window Splice produces a non-revisable and fully quality-adjusted price index where there is longitudinal price and quantity information at a detailed product specification level. It is based around the Fixed Effects Index, which is defined as follows:

$$P_{FE}^{0,t} = \frac{\prod_{j \in s^t} \left(p_j^t\right)^{\frac{1}{n^t}}}{\prod_{j \in s^0} \left(p_j^0\right)^{\frac{1}{n^0}}} \exp\left(\overline{\widehat{\gamma}^0} - \overline{\widehat{\gamma}^t}\right)$$

where ^{0} is the average of the estimated fixed effects regression coefficient at time 0. Using a fixed effects regression overcomes some of the disadvantages of using the time dummy ITRYGEKS, whilst being equivalent to it. Like the RYGEKS, after the initial estimation window, the new series is spliced onto the current series for subsequent periods; this is called a window splice. The window splice essentially uses the price movement over the duration of the estimation window, rather than the price movement in the latest period. This approach has the advantage of incorporating implicit price movements of new products at a lag. There is a trade-off, then, between the quality of the index in the current period and in the long-term. Over the long-term, the FEWS method will remove any systematic bias due to not adjusting for the implicit price movements of new and disappearing items. A full description of the method can be found in Krsinich (2014).