Office for National Statistics

Article

Research indices using web scraped price data: August 2017 update

A summary of the ongoing research into using web scraped price data in the production and development of consumer price statistics.

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

Release date:
1 September 2017

Next release:
To be announced

Correction

4 May 2018

An error occurred in the chained Jevons indices due to an error in the way that the indices were calculated. We have corrected this error. You can see all previous versions of this data on the previous versions page. We apologise for any inconvenience.

19 December 2017

An error has been found in Paper 1: Research indices using web scraped price data: clothing data, Paper 2: Research indices using web scraped price data: August 2017 update, Paper 3: 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 . Authors

Authors: Himanshi Bhardwaj, Tanya Flower, Philip Lee and Matthew Mayhew.

2 . Introduction

The past decade has witnessed considerable growth within the information technology sector, as a result of which the amount of information and data that is available to us is growing rapidly and at a speed which is faster than ever. The data available for analysis has not only increased in volume, but has also expanded in the forms in which it is obtainable. The rapid development in the information technology sector and secure payments have led to new purchasing channels, which can offer opportunities to extend the current collection of data for consumer price statistics from the current traditional price collection methods.

One such channel is the online purchase of goods and services through a retailer's website. Online expenditure has increased over the past decade. The latest figures from Office for National Statistics (ONS) show that the average weekly spending online in Great Britain was £1.1 billion in May 2017, an increase of 14.4% compared with May 2016 (Retail Sales in Great Britain: May 2017). The share of retail spending accounted for by online purchases has also increased; the amount spent online accounted for 15.9% as a proportion of all retail spending, excluding automotive fuel, compared with 14.3% in May 2016. This suggests an increasing trend in the amount being spent online. The availability of online data gives National Statistics Institutes (NSIs) across the world an opportunity to review the current methodology for capturing the change in online prices in our measures of consumer price inflation.

Recognising the importance of these new datasets, ONS set up a Big Data project in January 2014 to investigate the benefits and challenges of using Big Data in official statistics. One of the pilots within the project was to collect web scraped data for use within consumer price statistics (Naylor et al., 2014). In October 2015, we received additional funding for this project from Eurostat. This report summarises the work carried out using the funding from this grant, including the studies that were carried out, an assessment of the results achieved and any methodological problems identified, and finally recommends a number of areas for future work for the project.

In particular, Section 3 provides a background to the use of web scraped data in the production of consumer prices statistics, including a discussion of the current method of price collection at ONS and known limitations to the use of web scraped data. Section 4 presents a summary of a recent methodology review that ONS conducted into suitable price index methodology for use with web scraped data. Sections 5, 6 and 7 focus on three different strands of ONS web scraping research: grocery items, items from the central collection and clothing items respectively. Section 8 concludes, and Section 9 discusses some future work propositions, as well as the challenges that remain and possible solutions.

Throughout this report, references will be made to other published work by ONS.

ONS published articles on web scraped data

ONS Big Data Project Quarter 2 Report: The report summarises the Big Data project and provides an overview of the prices web scraping pilot and initial research aims (August 2014).

Initial report on experiences with scanner data in ONS: This report focuses on a sample of scanner data obtained by ONS but also presents some analysis on product churn and index selection which was useful to frame our initial research into web scraped data. Attempts are still ongoing to procure scanner data but this is the only substantial research done by ONS on the subject so far (December 2014).

Trial consumer price indices using web scraped data: The first research article presented experimental price indices from June 2014 to April 2015. The indices included chained Jevons (referred to as the Chained Bilateral Jevons index in this paper) and unit price (referred to as the Fixed Based Jevons index in this paper) indices at different frequencies (June 2015).

Research indices using web scraped price data: The second article extended the time series for these indices to June 2015 and introduced a new index compiled using the Gini, Eltetö and Köves, and Szulc (GEKS) – Jevons formula (GEKSJ) (September 2015).

Web Scraped Data: Extreme price changes: This summary article provided an overview of extreme price changes found in the web scraped data from June 2014 to July 2015. It was used to inform work on cleaning that was done in 2016 (November 2015).

Research indices using web scraped price data: May 2016 update : This article summarises progress made in applying supervised and unsupervised machine learning techniques to the web scraped data. Analysis into using imputation as a means of reducing the impact of missing prices in the data is also presented. Finally, the experimental price indices presented in the previous articles are updated to February 2016 (May 2016).

Imputing Web Scraped Prices: This article looks at the issue of imputation. Imputation is a method of dealing with missing prices, but there are many different techniques to choose from. Following a simulation study it was found that carrying forward the previous price is the best method with respect to minimising imputation bias (May 2016).

Using machine learning techniques to clean web scraped price data via cluster analysis : This article describes an approach to reduce the number of misclassifications in web scraped data by using a clustering algorithm designed to identify them. Preliminary results are presented on the impact to the distributions within each of the grocery categories for which this price information was scraped (May 2016).

Research indices using web scraped price data: clustering large datasets into price indices (CLIP) : This article puts forward an alternative experimental approach to aggregating large datasets into price indices: clustering large datasets into price indices (CLIP). We then apply the CLIP to grocery data that we have web scraped from online retailers between June 2014 and July 2016. The experimental price indices presented in our previous web scraping articles are also updated to July 2016 (November 2016).

Research indices using web scraped price data: clothing data : This article summarises analysis into using web scraped clothing data to produce experimental price indices, including the application of the new CLIP methodology (July 2017).

Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices : This paper explores the nature of product turnover in the web scraped clothing data, utilising a proportional hazards regression model to build a survivor function for clothing items. The paper also constructs price indices using the chained Jevons, Intersection-GEKS (IntGEKS) and Fixed Effects with a Window Splice (FEWS) methods (July 2017).

A comparison of index number methodology used on UK web scraped price data : This article looks at the different approaches to calculating price indices from web scraped data and evaluates them against the test, economic and statistical approaches. It makes a number of recommendations about the use of index methodology in a production environment (August 2017).

3 . Background

3.1 Summary of the current collection methods

The Consumer Prices Index including owner occupiers’ housing costs (CPIH) is produced monthly by Office for National Statistics (ONS). CPIH measures the change in prices for a fixed basket comprising approximately 730 representative goods and services. Of these items, 538 are collected physically by price collectors from various stores across the country, while the prices for the remaining items are collected centrally by ONS price collectors through websites, emails, catalogues and over the telephone.

Methods of collection

There are two basic price collection methods: local and central.

Local collection is used for most items; prices are obtained from outlets in about 150 locations around the country. Around 105,000 quotations are obtained by this method per month. Normally, collectors must visit the outlet, but prices for some items may be collected by telephone. Outlets are categorised as independents (less than 10 outlets in chain) or multiples (chains of 10 or more outlets). For some of the large chains, such as supermarkets and burger chains, which do not have national pricing policies, only one outlet is chosen to represent all outlets in each region. These are known as regional shops. There are some items that are only collected in these regional shops.

Central collection is used for items where all the prices can be collected centrally by ONS with no field work. These prices can be further sub-divided into two categories, depending on their subsequent use: central shops and central items.

Central shops

The prices are combined with prices obtained locally. Central shop prices are obtained from major chains of shops with national pricing policies. Branches of these chains are excluded from the local collection.

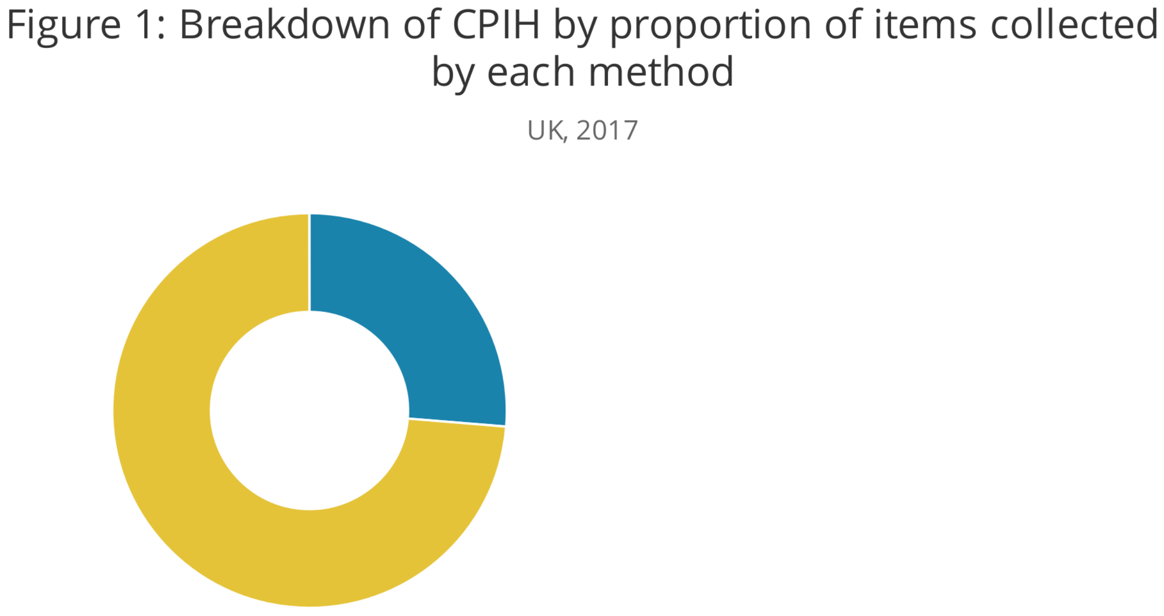
Central items

The prices are used on their own to construct centrally calculated indices. There are about 192 items for which the prices are collected centrally and the index calculation is carried out separately from the main method of index production. Selecting this type of collection and calculation is usually dependent on one or more of the following considerations: sources of data, data presentation, and frequency of price changes, national pricing policies and the possibility of future fundamental changes to pricing methods. Where feasible, price data are collected over the internet. If this is not possible, price data are collected from one central source (such as trade associations or government departments) whenever possible, although market forces do require contact with regional or competing companies in many cases. Data may be requested in writing, by telephone or by email.

The following charts show how the collection is distributed between central and local collections and by mode of collection.

Figure 1: Breakdown of CPIH by proportion of items collected by each method

UK, 2017



Source: Office for National Statistics

Nearly three-quarters of items are collected in the local collection (Figure 1). We can explore this further by looking at this breakdown by expenditure weights and the number of price quotes obtained each month.

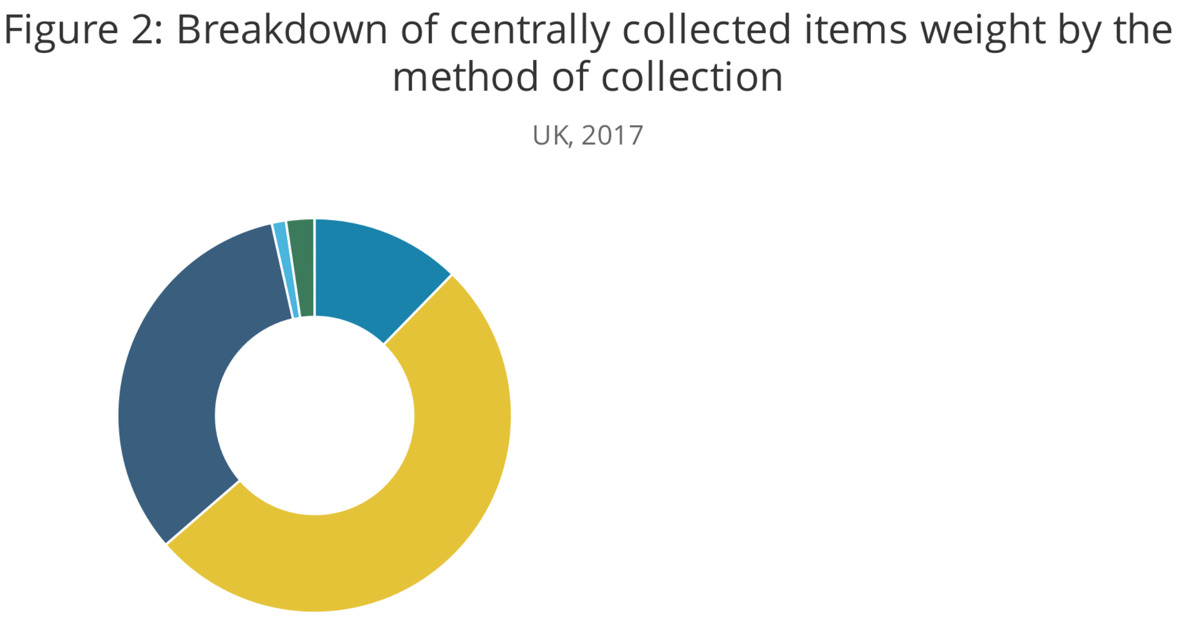
When the expenditure weights of the items are taken into account, around 40% of the basket is collected via the local collection. This is because there are a number of high expenditure categories that are collected centrally. In particular, owner occupiers' housing costs (OOH) comprise 17.4% of the total CPIH basket in 2017, and are sourced from administrative data that are supplied to us by email. However, even if we deduct OOH, the local collection then accounts for 47% of the basket.

In terms of the total number of quotes, the local collection accounts for 15% of the total number of price quotes used to calculate CPIH. This includes the contribution of administrative datasets in the central collection for items such as OOH, which have a large number of quotes[1] .

Figure 2 shows that within the central collection, the majority of items are collected via email. Again, this is partly because this accounts for the OOH component, which is primarily sourced via email. The next largest group is the internet collections followed by phone. Although there is unlikely to be much scope to address the phone and email collections, the remaining 36% of the central collections can feasibly be collected via alternative data sources such as web scraping.

Figure 2: Breakdown of centrally collected items weight by the method of collection

UK, 2017



Source: Office For National Statistics

Notes:

Centrally collected items account for 60.3% of the total CPIH basket of goods and services (that is, they have a weight of 603 out of a total weight of 1000)

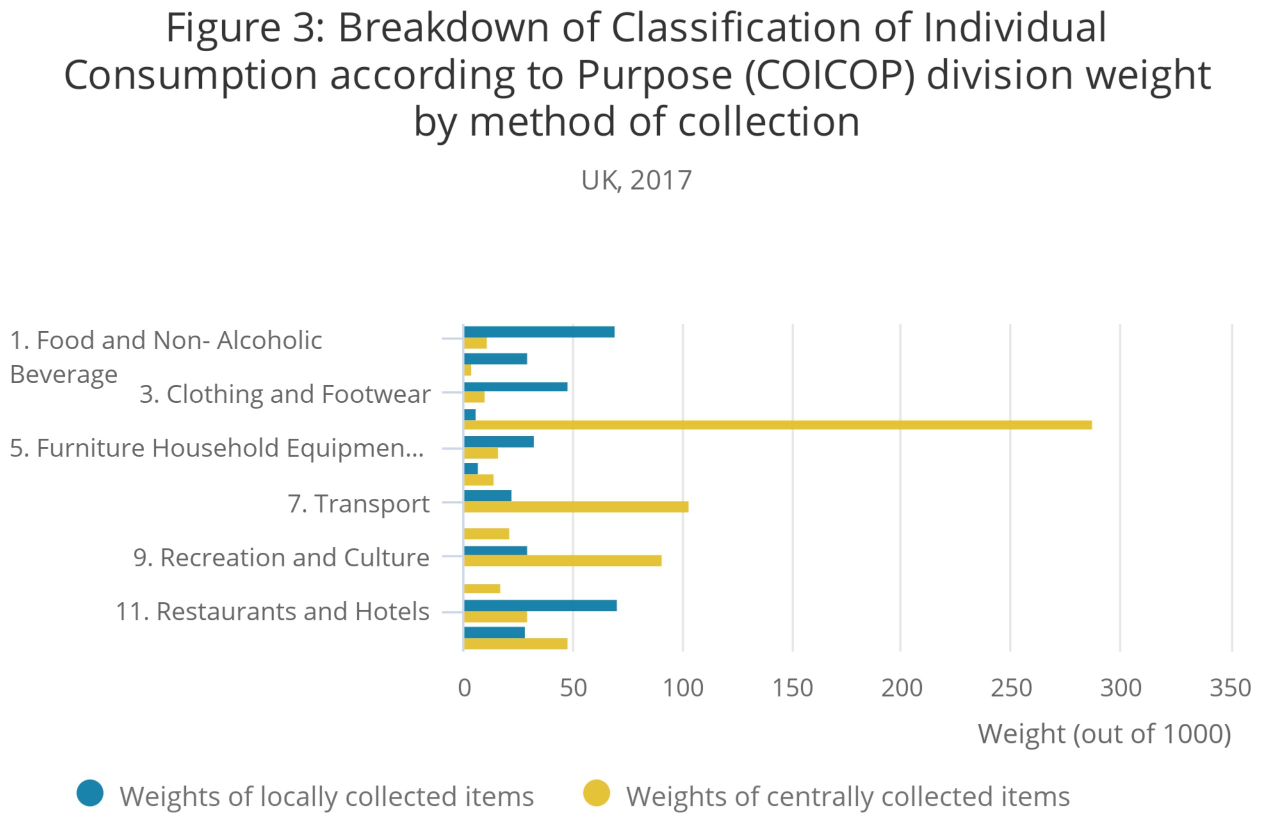
The local collection can also be broken down by type of shop. Around three-quarters of the local collection (in terms of number of price quotes) come from multiple shops, including regional shops. However, this also implies that around one-quarter of local price quotes come from independent shops. These will be more difficult to replace using alternative data sources such as scanner data and web scraped data.

Additionally, some of the locally collected items are also collected in central shops, and so this might be an over-representation of the local collection. We can separate out items that are collected both locally and centrally, and recalculate the analysis. Out of the 538 items that are collected locally, 322 of these items are supplemented through central collections to some extent. As a result, this reduces the total proportion of the CPIH basket collected locally from 40% to 35%. At least one-fifth of quotes in the local collection come from regional shops.

Figure 3 illustrates the weight of each Classification of Individual Consumption according to Purpose (COICOP) division collected via local and central collection (also accounting for the local items which are supplemented through the central collection). The local collection currently dominates for the collection of food and non-alcoholic beverages, alcoholic beverages and tobacco, and clothing and footwear. For housing, water, electricity, gas and other fuels, transport, and recreation and culture most quotes come from the central collection. Communication and education are both wholly collected by the central collection.

Figure 3: Breakdown of Classification of Individual Consumption according to Purpose (COICOP) division weight by method of collection

UK, 2017



Source: Office For National Statistics

Notes:

After accounting for those locally collected items that are supplemented by the central collection, the total percentage of the basket that is centrally collected is 65.3% (that is, it has a weight of 653 out of a total weight of 1000).

In summary, while there are elements of the current collection basket that will always have to be collected locally by price collectors, there is considerable scope to investigate the use of alternative data sources for both the central collection and elements of the local collection. The current use of administrative data for items such as OOH show the benefit that can be gained in terms of the number of price quotes that can be collected from these alternative data sources. This has a potential impact in terms of quality improvements and efficiency savings.

3.2 Exploring the use of web scraping in consumer price statistics

As part of the drive towards innovation and transformation, we are investigating new alternative data sources to supplement the collection of data for consumer price statistics. As discussed in the previous section, there is considerable scope for this to improve the efficiency and quality of collection methods. This was an important recommendation in Paul Johnson’s UK Consumer price statistics: A review (2015), and was also highlighted in the Independent review of UK economic statistics (2016), led by Professor Sir Charles Bean. The “Bean Review” has also led to the establishment of a new Data Science Campus, which will expand our capability in this area.

There are two main alternative data sources to explore in relation to consumer price statistics. The first is scanner data, which are datasets collected by retailers as products are purchased. These include expenditure and quantity data, from which a unit price can be derived. Scanner data are already used to varying extents by a number of National Statistics Institutes (NSIs) in the production of their consumer price statistics. This includes the Australian Bureau of Statistics , CBS Statistics Netherlands and Statistics New Zealand to name a few. We are continuing discussions with retailers to provide point of sale scanner data. Our experiences with scanner data are reported in Initial report on experiences with scanner data in ONS .

Due to lack of progress with scanner data, we have continued our investigations into using web scraped data. Web scrapers can be defined as software tools that are used for extracting data from web pages. The recent changes in consumer shopping trends, and the growing prominence of online retailing, suggest that the associated price information for a number of goods and services can be found online and obtained through the use of web scrapers.

In January 2014, we set up the ONS Big Data Project, to investigate the benefits and the challenges of using large scale, unstructured data and associated technologies in the production and development of official statistics. The prices pilot was one of four practical pilots that were set up to provide us with first-hand experience in regards to the management and processing of Big Data. The prices pilot uses web scrapers to collect prices (Breton et al., 2016).

Our work builds on existing literature from other NSIs (for example, see the previous papers from Statistics New Zealand and Statistics Netherlands) and academic projects such as The Billion Prices Project (Cavallo and Rigobon, 2016). The cleaning, processing, storage and price index methodologies that are developed using web scraped data will be useful in the event of procuring scanner data in the future.

There are also a number of benefits that web scraping can provide over and above scanner data. For example, if the data are collected centrally, it can be processed in a timelier manner than waiting for data supplied by third parties. Web scraping could also provide an opportunity to collect prices for some goods and services automatically, instead of using manually collected data. Other benefits include reduced collection costs, increased coverage (that is, more basket items and/or products, which may not be covered using just scanner data), and increased frequency of price collection.

Although web scraping can provide a number of benefits, it should also be acknowledged that there are some general challenges to overcome when working with web scraped data.

The terms and conditions for certain websites imply that web scraping may not be an acceptable use of the website. Further, it is thought that some websites use blocking technologies to prevent scraping. Therefore, prices cannot be scraped from all online retailers. This has made it difficult to completely replicate our existing online collection and has limited the research that we can undertake in certain areas.

All prices are scraped regardless of expenditure. This means that we collect the prices of all products that are available, but we do not know which products have been bought and in what quantities. This makes it necessary to treat all products equally. In traditional price collection, price collectors select items that are representative of what consumers buy, and low expenditure products would not normally be considered as this could introduce a bias into the index.

Prices can be collected daily rather than on an index day each month, as in the traditional CPIH collection. While the large volume of data offers many benefits in terms of increasing the number of price quotes, and decreasing the effects of outliers, this limits the extent to which comparisons may be drawn with published CPIH indices.

Products are matched across periods using product names; however, these can change over time. In traditional price collection, a price collector would easily be able to identify if a product is the same, following a change of description. Current matching methods are unable to identify description changes. Again, the volume of the data means that comparable replacements cannot easily be found for unmatched products.

Typically, we see very high levels of product churn (that is, products coming in and out of stock) in high volume price data. This means that, for some items, sample sizes are very small. This problem is particularly acute where the methodology requires items to be matched over the length of the time series, for example, in the Fixed Base Jevons.

The volume of the data makes traditional cleaning methods unworkable. While research has progressed significantly on developing appropriate cleaning and classification procedures, work still remains on testing the robustness of these methods when applied to new datasets.

It is within these constraints that we will evaluate the work carried out so far on using web scraped data within the production and development of consumer price statistics, and suggest potential solutions where possible to overcome some of these limitations.

Notes for Section 3: Background

1. For example, the Valuation Office Agency (VOA) data used to calculate the OOH rate for England are based on around 500,000 rental prices.

4 . Methodology

4.1 Introduction

The treatment of alternative data sources such as web scraped data provides a new set of challenges for National Statistics Institutes (NSIs), including around the issue of index compilation.

Currently, the Consumer Prices Index including owner occupiers’ housing costs (CPIH) uses a traditional fixed basket approach to calculating the index (see the Consumer Prices Index Technical Manual for more information). At the elementary aggregate level a mixture of un-weighted Jevons and Dutot indices are used that compare price change from the given base period to the current period on a product-by-product basis. At the higher levels, a Lowe-type index is calculated from these elementary aggregates, using expenditure weights for the item level and above. If a product changes (for example, the package size decreases) and this change is not captured appropriately, there is a possibility that the inflation rate would change purely because of the change in product and not the change in price. This is unacceptable for an index that should be as much as possible measured at constant quality and to capture genuine price change only.

For web scraped data, traditional methods for compiling consumer price inflation may not be appropriate. The main reason for this is product churn, which is the process of products leaving and entering the sample. This may be due to a product temporarily or permanently going out of stock, new products coming onto the market, or rebranding. The traditional fixed basket approach that relies on being able to follow the same product over time becomes less appropriate when there are high levels of product churn in the data. As discussed in previous work by Office for National Statistics, web scraped data does see more product churn than in the traditional collection, largely as a result of increased sample size.

An increasing number of NSIs have been investigating the use of alternative data sources in the production of consumer price statistics, and several methods have been explored to calculate price indices from these data (see for example, Kalisch, 2017; Chessa et al., 2017). These methods can be divided into bilateral indices (which compare prices from two periods of data) and multilateral indices (which compare multiple periods at the same time).

Well-known price index methods include the following:

Bilateral indices:

- Fixed Based Jevons (this was called the “unit price” index in our previous releases)
- Chained Bilateral Jevons (this was called the “daily chained” index in our previous releases)
- Unit Value Index

Multilateral indices[1]:

- The Gini, Eltetö and Köves, and Szulc (GEKS) family of indices; all the GEKS methods in this paper use Jevons indices as an input into the GEKS procedure (GEKSJ)[2]
- The Fixed Effects index with a Window Splice (FEWS)[3]

For a description of these methods please see Annex A.

4.2 New methodology

One of the main contributions that we have made with regards to methodological developments is the creation of a new index: clustering large datasets into price indices (CLIP) (Metcalfe et al., 2016). This is an innovative experimental index designed to address some of the challenges involved with building an index from web scraped data. The CLIP approach assumes that because a consumer may purchase the same type of product, the price index should therefore reflect changes in the price of the whole set of relatively homogenous products, instead of just following one individual product over time.

The CLIP uses the same classification structure that is used to calculate CPIH. It is applied at the item level. For each item, the available products are clustered together using an unsupervised machine-learning algorithm into similar groups using the information that has been scraped from the website (price, product name, shop, discount marker). The CLIP is calculated by measuring the price change over time between these clusters, implying that a primary difference between CLIP and other index methodologies is that price relatives that construct the index are calculated at the cluster level, rather than the individual item level. To maintain a fixed basket, the clusters are formed and set for the base month (January, to maintain consistency with the UK CPIH), and then the same clusters are formed for each time period over the year (in this case, monthly). The product make-up within each cluster can vary over time, as products move in and out of the market. This therefore allows for high product churn in the data.

Price relatives (the ratio of a price at a given time to the price for the same product at another time) are calculated from the geometric mean price of the cluster in the base period. For each comparison month, these cluster price relatives are then aggregated together using the number of products within each cluster (fixed in the base month) as the weight.

By using the same classification structure as the more traditional approaches to calculating inflation, implies that the same weighting information can be used by the CLIP as that used in the CPIH. This thereby minimises the complexities of obtaining this information, and also maintains consistency. Once the item level indices are calculated, this weighting information is then used to aggregate up the indices to the higher COICOP levels following the same process as the more traditional approach. It can therefore be classified as a “bilateral” index.

The CLIP will be compared against the other indices as part of this review. For a further description of this method please see Annex A.

4.3 Approaches to assessing price index formulae

In index number theory there are a number of approaches to aid a statistician in making an appropriate formulae choice:

1. the Axiomatic approach (see ILO consumer price index manual chapter 16) – the index is tested against some desirable properties

2. the Economic approach (see ILO consumer price index manual chapter 18) – the index is ranked against whether or not it approximates or is exact for a Cost of Living Index

3. the Statistical (Stochastic) approach (see ILO consumer price index manual chapter 16) – each price change is an observation of population value of inflation and the index is the point estimate of inflation

The indices were assessed against these approaches to identify which performs best. In this final report we summarise results of the analysis. Please see the full publication, A comparison of index number methodology used on UK web scraped price data, for more information.

Axiomatic approach

The axiomatic approach assesses each index against a set of desirable properties (axioms). An index will either pass or fail each axiom. If an index passes all of the axioms, then it is deemed to perform well. However, it is important to note that different price statisticians may have different ideas about what axioms are important, and alternative sets of axioms can lead to the selection of different “best” index number formula. There is no universal agreement on what is the best set of reasonable axioms and therefore, the axiomatic approach can lead to more than one best index number formula.

Table 1 shows whether the indices under consideration passed or failed each of the axioms considered. For more information about each of the axioms, please see “A comparison of index number methodology used on UK web scraped price data”, Mayhew (2017).

Table 1: Summary of index performance under the axiomatic approach

Axiom	Fixed Based Jevons	Chained Bilateral Jevons	Unit Value	GEKSJ	RYGEKSJ	FEWS	CLIP
Positivity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Continuity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Identity	Yes	Yes	No	Yes	Yes	No	No
Proportionality in Current Prices	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Proportionality in Base Prices	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Commodity Reversal	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Invariance to change in Units	Yes	Yes	No	Yes	Yes	No	No
Time Reversal	Yes	Yes	Yes	Yes	No	No	No
Circularity	Yes	Yes	Yes	Yes	No	No	No
Monotonicity in Current Prices	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Monotonicity in Base Prices	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Price Bounce	Yes	No	No	No	No	No	No

Source: Office for National Statistics

In summary, only the Fixed Based Jevons index passes all of the axioms. The GEKSJ and the Chained Bilateral Jevons index pass all axioms other than price bounce . Therefore, under the axiomatic approach, the GEKSJ index and the Chained Bilateral Jevons index also perform well.

To account for the high level of product churn that is associated with web scraped data, the GEKSJ index and the Chained Bilateral Jevons index both include an element of chaining. This could lead to a problem with chain drift, where a chained index does not equal the direct index.

Economic approach

The economic approach to index number theory assumes that the observed price and quantity data are generated as solutions to various economic optimisation problems. The quantities are assumed to be functions of the prices and not independent variables. In the context of consumer prices, the economic approach usually requires the chosen index formulae to be some kind of Cost of Living Index (COLI).

Each of the indices was assessed under the economic approach. It was found that the indices are all exact under a certain set of preferences, and therefore no clear conclusions could be drawn. However, one notable result from this analysis was to investigate the possibility of a geometrically weighted CLIP.

It is also the case that CPIH measures the change in prices of goods and services purchased to be consumed. This is different to a COLI because a COLI assumes that consumers change their spending patterns towards goods and services whose prices are increasing relatively slowly. Therefore, even if an index approximates or is exact for a COLI, it might not be appropriate in the context of the CPIH. We therefore do not provide more information about the economic approach in this summary, but for more information please see A comparison of index number methodology used on UK web scraped price data.

Statistical (Stochastic) approach

The statistical approach to index number theory treats each price relative as an estimate of a common price change. Hence, the expected value of the common price change can be derived by the appropriate averaging of a random sample of price relatives drawn from a defined universe.

There are four types of model that can be used to estimate price change [4] :

- price change for each item is equal to some common mean plus some error (or some function of price change): the Carli and Fixed Based Jevons indices fall under this model type
- current price for each item is some common multiple of the previous price plus some error (or some function of price): the Dutot, Unit Value and CLIP indices fall under this model type
- a combination of the first two models: the GEKSJ index falls under this model type
- decomposing prices to determine price changes: the FEWS index falls under this model type

Given these models, the statistical approach becomes a model fitting exercise that selects the index that “best fits” [5] the data collected. This has been done for a selection of data that cover all the different strands of ONS research into web scraping so far, notably; grocery items, clothing items, and items from the central collection.

Table 2 summarises the results over the different categories. Each model was tested in two periods with different rates of inflation. It was found that the FEWS index over-fits [6] for the grocery and clothing items and it was therefore removed from the comparisons.

Table 2. Summary of index performance under the statistical approach

Group	Best Fit
Groceries	Fixed Based Jevons
Clothing	Unit Value
Central collection – chart	Fixed Based Jevons
Central collection – technological	Fixed Effects

Source: Office for National Statistics

From the statistical approach, the Unit Value index was identified as the preferred index to use on clothing items. Stratification improves the precision of the estimate if the sampling units are homogeneous within a stratum and heterogeneous between strata. As the CLIP is a stratified Unit Value index, it follows that it should be a better approach to use for this category.

Assessing the indices against the statistical approach provides evidence to suggest that no single index is suitable to cover the full range of items in the CPIH basket of goods and services that could be sourced from web scraped data.

4.4 Summary and recommendations

In summary, each approach identifies different indices that perform best. Based on this analysis, along with work conducted by other NSIs (see for example, Kalisch, 2017; Chessa et al., 2017), the following recommendations are made from a methodological viewpoint.

A GEKS index should be used for any web scraped data covering the grocery section of the CPIH basket, due to its desirable properties. The GEKS index also limits chain drift compared to other chained indices such as the Chained Bilateral Jevons, but doesn't eliminate it completely (see A comparison of index number methodology used on UK web scraped price data for more information). This is a desirable property given the high levels of product churn found in web scraped data.

A Unit Value index should be used for any web scraped data covering the clothing section of the CPIH basket. The CLIP may be useful approach for this. However, further work is required to compare a geometric CLIP and an arithmetic CLIP, as well as other weighting schemes.

For items that are centrally collected, there are two considerations, as detailed in this section.

For chart collections, the statistical approach along with the results from the axiomatic approach show that a Fixed Based Jevons index or a Chained Bilateral Jevons index should be the preferred methodology. Given the way that the chart-based items are collected (the chart positions are matched over time, rather than individual products), the Fixed Based Jevons index is preferred. The FEWS index did not perform well under the statistical approach for the chart based items collection.

For technological goods, the FEWS index is recommended due to the tendency for new items in the category to have quality improvements. This is in line with international best practice, as detailed in The FEWS index: Fixed effects with a window splice (Krsinich, 2014), and Price indexes from online data using the fixed-effects window-splice (FEWS) index (Krsinich, 2015).

4.5 Future work

Although we have identified these indices as performing best against these methodological approaches, there may be other practical issues that require alternative index methodology to be used in future. For example, the use of multilateral indices may not be compliant with current HICP regulations. This will need further consideration when looking at the impact of using web scraped data in the Consumer Prices Index (CPI), and therefore the CPIH. Eurostat is currently in the process of drafting some practical guidance for processing supermarket scanner data, which includes a section of which index methodology may be appropriate when using this data. Although not exactly comparable to web scraped data, the guidance can provide some useful direction on European compliance issues. Internal system development is another practical issue that should be taken into account.

The GEKS index has been shown to be sensitive to products that have periods with atypical prices (for example, products on clearance) (Chessa et al., 2017). The cleaning approach that we currently adopt for our grocery web scraped data (Section 4) should exclude these clearance products from the cleaned dataset that are then used for index compilation, subject to further refinement and testing. These products will also be queried by the Tukey Algorithm that is currently used to validate the traditional collection, as it being explored with regards to data from the central collection. The exclusion of products at clearance prices is consistent with current practices adopted in the CPIH.

Another drawback of the standard GEKS index is that whenever a new time point is added, the entire index will be modified. Although this can be resolved by merging the latest movements in the GEKS index to the existing time series, this may lead to a loss of “characteristicity” (De Haan and van der Grient, 2009). Characteristicity is the extent to which an index is based in relevant data. Consequently, the more time points that are added, the less and less characteristic the GEKS index will become. To resolve this, researchers have developed methods for extending the index series without the need for revisions, which will also minimise the loss of characteristicity (Kalisch, 2017).This includes methods such as Rolling Year Gini, Eltetö and Köves, and Szulc (RYGEKS), which have been explored as part of this review.

Another approach to overcome these issues has been developed at ONS, which we call the “in period” method. The in period method requires that the GEKSJ is calculated using only the data that are available up to that period, and when new data are added from the next period the index is not recalculated. For example, if we have five periods worth of data, the standard GEKSJ methodology for calculating the index for period four would use all five periods of data, whereas the in period GEKSJ would only use the first four periods. This is the definition we have used in Annex A to define GEKS: our intermediate comparison periods are between the base and current periods, whereas in the standard GEKS, the intermediate periods are all periods on the dataset, including future periods. However, further work is required to test other extension methods in the context of our web scraped data.

Further research is also required on the impact that the lack of expenditure weights for web scraped data will have on the index. Expenditure information can be used to remove products that may not have been purchased by many consumers (and therefore should not be included within a representative basket of goods and services). Early research by Statistics Netherlands (Chessa and Griffioen, 2016) that compares web scraped and scanner data for the same retailer and products shows that using equally-weighted web scraped data could introduce some bias into the index, although this can be mitigated by using expenditure proxies such as the number of days that a product is available on the website.

The frequency and coverage of the data should also be investigated further. For example, web scraped data can be collected on a daily basis for a near-census of products from a retailers’ website. This is in contrast to the traditional collection, which focuses on a representative sample of products for a given item, which is selected to best cover consumer expenditure for a given class of Classification of Individual Consumption according to Purpose (COICOP). Prices for these products are collected at most once a month. Questions remain for the indices identified as part of this review, for example, how do they perform when they use monthly, weekly or daily prices; should a sample of the web scraped data be taken instead; and would this better suit the use of bilateral methods?

Finally, this research has shown that different methods are suitable for different areas of the CPIH basket. This implies that if we wanted to extend our coverage of web scraped data to other areas of the basket, further research is necessary to understand which index methodology is suitable for this particular area. As a next step, this will focus on understanding the requirement for package holidays (discussed in Section 6: Central Collections Project).

Notes for Section 4: Methodology

1. For this work, the Geary-Khamis index wasn't considered due to the lack of an acceptable quantity proxy available at this time.

2. The IntGEKSJ was not included in the comparisons as it uses the same structure as GEKSJ but just includes a different set of products. The IJRYGEKS (a Jevons version of the ITRYGEKS) was also not included due to the lack of characteristics to perform the imputations required.

3. The Fixed Effects index is also known as the Time-Product Dummy index.

4. In the presence of product churn, the Chained Bilateral Jevons and the GEKSJ do not pass the price bounce axiom.

5. The Chained Bilateral Jevons and the RYGEKSJ weren't assessed against the statistical approach due difficulties specifying an appropriate model, because of incorporating the chaining that is involved.

6. Best fit in terms of minimising the information lost when using this model. The Akaike Information Criterion was used to determine best fit.

5 . Grocery Prices Scraping Project

5.1 Introduction

The first strand of our research into web scraping data focuses on the online grocery sector. Since June 2014, our Big Data pilot has been running bespoke web scrapers to collect prices for three of the largest supermarket chains: Tesco, Sainsbury's and Waitrose, which together make up approximately 50% of the grocery market share in the UK (Kantar World Panel, 2016)[1] .

Supermarket grocery prices were identified as an initial area for investigation for a number of reasons. Food and beverages are an important component of the Consumer Prices Index including owner occupiers' housing costs (CPIH) basket in terms of both expenditure weights and number of items. They are also relatively easy items to collect: only a small number of retailers are required to cover a significant proportion of the market.

Before price indices can be compiled using the raw web scraped data there are a number of data processing steps that need to be performed. Issues such as data storage, classification, cleaning and imputation need to be considered. This section summarises our research into these topics, as well as presenting research indices from Section 4 that have been produced using the grocery prices data.

5.2 Data collection

The web scrapers used for the grocery prices project have been developed using the Scrapy package in Python. The project identified 33 grocery items that have been collected from each of the three supermarket websites on a daily basis since June 2014. The number of products extracted within each item category varies depending on the product churn (number of products coming in and out of stock) of each supermarket. On average, around 7,000 price quotes per day are extracted by the web scrapers (approximately 200,000 a month), which far exceeds the volume of prices collected via the traditional manual methods of price collection (approximately 6,800 a month for the specified 33 items).

Public cloud infrastructure and data storage

For the majority of the project, the web scraping software has been deployed on an Office for National Statistics (ONS) internal private cloud. The environment is designed for research and development; in particular, it enables experimentation with novel software. However, the infrastructure is not fully supported by the office: there are no uptime guarantees and researchers have to fix problems as they occur. This has been particularly challenging for this project where the aim has been to collect price quotes on a daily basis.

Towards the end of the project, work was undertaken to port the scraping code and data storage over to the public cloud computing services provided by Amazon (AWS). This entailed some development work around updating the way the system runs. The resulting system is more robust than the version deployed internally. The following is a list of benefits gained from the AWS infrastructure:

- reproducibility: it is easier for other teams to run their own version of the system without the upfront costs associated with in-house servers
- AWS provides a file storage solution S3; (Simple Storage Service).This has fine-grained access control, and strong guarantees for data durability
- hourly billing for computer resources allows for greater control over costs
- doesn’t require in-depth technical knowledge of infrastructure to set up and maintain the system
- AWS data centres are far less likely to experience problems than our unsupported environment
- stable, industry standard, platform for future development

There is still work to be done if we want to take the system further, but having the system deployed in AWS gives us options for improving system reliability and scaling the project up. Some of the main areas of development would be to automate backups of the data and improve system resilience and monitoring. There is also the potential to scale the project across a cluster of machines, for example, by using a queue to distribute the work. The open source project scrapy-cluster does this with the distributed streaming platform Apache Kafka.

Extension of scrapers

To give a richer dataset for analysis, work has been undertaken to extend the scrapers from the initial 33 items to perform a full site collection of prices that covers the whole CPIH grocery basket. Another benefit of this approach is to reduce the maintenance burden. In the current set-up the scrapers are vulnerable to changes in the supermarket product hierarchy; they can stop collecting prices for an item altogether when a change happens. This means that urgent maintenance has to happen whenever such a change is detected. This kind of reactive maintenance would be less frequent when the default is to collect every price and process it afterwards. However, it should be noted that the implementation isn’t a generic crawler; it still makes certain assumptions about the website structure.

The full site scrape is running and stable for one of the supermarkets. The main challenge to extending this further comes from an earlier technical decision in the project. To deal with problems like infinite scroll (that is, products on the page keep appearing as the user scrolls down in the browser) and cookies, the decision was taken to use Selenium in two of the scrapers. Selenium is a browser automation tool; it plugs into a browser and mimics the actions of a user. It is commonly used for testing and scraping websites. This was an expedient solution for the 33 item scrapers. However, the technical problems it solved come at the cost of a slower run time. This makes full site scrapes for those two supermarkets far slower than hoped for. There are a number of approaches that we could take to solve this but this has been left for future research (for example, we could augment the Scrapy code to perform similar fixes for infinite scroll).

The move to full site collection has necessitated research into a general classifier for the products (see “Progress towards a general classifier” in Section 5.3).

5.3 Classification

While we have made substantial progress with collecting and analysing web scraped data over the course of this project, challenges still remain around classification of the data.The CPIH aggregation structure is based on the European classification system for household consumption expenditure known as E-COICOP (Classification of Individual Consumption according to Purpose).COICOP is an internationally used hierarchical classification system comprising of divisions, groups and classes. The E-COICOP version introduces another layer to the classification structure underneath the class level. More detail is then provided by ONS specific-item and product-level components.

For the purpose of producing and developing measures of official consumer price statistics, products need to be classified in accordance to their item category.

When scraping data from websites, each retailer has their own product classification structure that in general does not follow the COICOP classification structure, and also tends to get updated frequently. For example, the item “apples (dessert), per kilogram” could be found by navigating to “groceries” then “fresh fruit” then “apples and pears”. However, there would be a number of products within this category that would not be consistent with the item definition. For example, it may contain fruit multipacks, or seasonal and promotional products that the retailer wants to promote (for example, Easter eggs). Therefore, it is not possible to simply use the retailer’s classification to construct price indices.

Instead, we can use the retailer’s classification structure as a starting point in our classification method, which would need to classify products into the correct COICOP category, and also remove products that do not fit the item description. Although exclusion terms could be used to solve this issue, inconsistencies may still remain, in particular when dealing with complex datasets (Breton et al., 2016). For example, a simple method would be to include products from the retailer’s “apples and pears” category, which contains the word “apple” in the product description. However, this suggests that inconsistent products such as “toffee apples” could also be included. This could be solved by creating a list of filter words for each class, for example, “toffee” and “pear”; however, this is an expensive solution in terms of resource.

Therefore, a more efficient solution to this problem is required, necessitating the use of machine learning algorithms. A number of options were explored before choosing to use a Support Vector Machine (SVM) model, as it performed best in terms of classification performance as measured by the F1 score. The F1 score represents the performance of a classifier as a single number: the harmonic mean of precision (proportion of predicted positive which are true positive) and recall (proportion of true positive which were predicted positive).

The quality and size of the training data are also important. Training data have been used in this instance to teach a series of binary classifiers to predict the correct products for the CPIH item classifications. For each item the products are scraped from close matching areas of the supermarket product hierarchy. To form these training datasets, expert price collectors were given a large number of examples for each of the 33 CPIH items currently scraped. The collectors assigned a label of “1” for a consistent classification and “0” for an inconsistent classification (Table 3).

Table 3: Generating training data for supervised machine learning; apples (dessert), per kilogram

CPIH item label	Product description	Manual classification (0 for inconsistent, 1 for consistent)
Apples, dessert, per kilogram	PINK LADY APPLES 4S	1
Apples, dessert, per kilogram	APPLE, KIWI & STRAWBERRY 160G	0
Apples, dessert, per kilogram	TESCO PACK 4 APPLES	1

Source: Office for National Statistics

Once the assignment was complete, 80% of the data are used to train the SVM. The remaining 20% of the data are used to test accuracy and generate the F1 score. This is required because the algorithm needs to be tested on unseen data that have been correctly classified.

As a result of the training data, the SVM learns to classify products based on certain words and word combinations in the product name. For example, for apples, it learns terms that have a positive relationship (for example, “apple”) and negative relationship (for example, “kiwi”, “strawberry”). Common words that add no predictive value are excluded from the algorithm.

The process detailed creates a classifier for each of the 33 CPIH item indices that we scrape price data for. As new data are scraped from each website, they are fed through a classifier and assigned a flag if the SVM believes they are incorrectly classified. Those with an “incorrect” flag can be identified in further cleaning steps, and potentially be excluded from index creation and analysis. Furthermore, the classifier can be periodically updated by fitting the SVM with new training examples.

The classifiers have an average accuracy of around 85% (F1 score) over the 33 CPIH items. The speed and the low resource required to build such a system has demonstrated the usefulness of applying supervised machine learning to the classification of alternative data sources. However, there are a number of ways in which the classifiers can be improved, such as using more training data and including a wider use of natural language processing, for example, investigating links between common sets of words. These issues are explored in the next section.

Issues with the classification approach

The approach to classification has been quite robust over the course of the project and it is easy to explain but it is not without drawbacks.

A common problem in machine learning applications is that of dataset shift, where the deployment context of the model no longer matches the training data. This can happen for a number of reasons; often the evolving nature of a system causes the difference to grow over time. For example, this happens in our project when we have to change the supermarket category that we’re picking from to adapt to changes to the websites. If a retailer changes the hierarchy for potatoes from “potatoes and leeks” to “potatoes and sweet potatoes”, then there is a high probability that the false positive rate of the classifier will increase, especially as the word “sweet” may not have been present in the training data. In principle, we should retrain the SVM each time we see a noticeable loss in the number of quotes that are collected for each item.

Keeping track of performance over time would be an important part of a production version of this system. For example, the micro-averaged F1 score of the classifier in the second month of operation (in 2014) was 0.93 plus or minus 0.023 (percentile bootstrap 0.95 confidence interval), whereas by March 2017 this had dropped to 0.87 plus or minus 0.033 (note that these intervals don't overlap). While this drop doesn't necessarily imply dataset shift, it is a perfectly plausible explanation. Retraining the algorithm can be expensive if relying on manually generated training data.

There are some other limitations to this classification approach:

- it is a manual process to identify areas of website for each basket item and each retailer; this makes it harder to scale the approach to more basket items and retailers
- we can measure the product hierarchy recall for the classifier, but we can't know how many valid matches for a product have been missed
- the products that don't match the targeted item are flagged; there is nothing in the system to check whether they could be a valid match for another basket item
- the data can only be used for a CPIH type fixed basket approach; other analyses that categorise all consumers spending can't be applied, which reduces the value of the dataset
- there is no way to retrospectively change which products are picked up by the classifier and therefore we don't have the ability to model changes to the CPIH basket going back in time

Progress towards a general classifier

As we move towards a full catalogue web scraper the current classification approach becomes less feasible, as it relies on manual matching of retailer hierarchies to the COICOP structure. There has been some initial work to build a general classifier. This is a harder task than the targeted approach. An initial hurdle is the gathering of training data for any supervised learning process. A number of sources of data are being evaluated (Table 4).

Table 4: Possible sources of training data for general classifier

Method	Notes
Manual classification of scraped data	Expensive to perform but could use semi-supervised methods to get as much value from people's time as possible.
The instruction document used by staff to manually classify survey data	This could work well for generic products, but does rely on human levels of knowledge to apply to all products.
Coded survey data, for example, the Living Costs and Food Survey (LCF)	The way products are represented in survey data is very different to scraped data (for example, extensive use of abbreviations in names).
Raw price collection data, which goes into the CPIH	Not a long-term solution. Suffers from similar data issues as consumer surveys.
Third-party scraped data (either with its own classification or COICOP)	Work needs to be done to guarantee the quality of external data.

Source: Office for National Statistics

It is important to note that some of these data sources are not directly comparable to items in the CPIH basket. For example, the Living Costs and Food Survey (LCF) uses a different classification hierarchy called COIPLUS. While this is also an extension of COICOP, it maintains full coverage of consumer spending at the lower levels instead of using representative items.

As well as investigating data quality and coverage for each source we have been looking into appropriate linking methodology to be able to label samples from the full scraper output.

Model reframing

There has been a recent piece of ONS research on providing automated assistance to manual coding of survey data for products according to COIPLUS. In that context, the approach achieved both a recall and precision of around 98%. We investigated the possibility of applying the model built directly to our data (that is, reframing the model for a new deployment context). By simply taking the same training data and building the model in the same manner as the other project we were able to achieve a classification accuracy of 0.80 plus or minus 0.051 (95% Wilson score interval with continuity correction) on the data gathered by the 33 item scraper. This is a promising start, while not enough to operate alone this could potentially form part of an ensemble of models trained in different contexts. There is also scope to refine the performance by standardising the data further. For example, the survey data that the model was built on has a lot of common abbreviations. Assessing the performance on our dataset is difficult as we don’t have a large training set of examples of general classification to rely on.

The performance detailed previously is achieved by classifying everything; it is possible to trade off coverage for classification accuracy by only classifying when confident in the result. This can be done by calibrating the model; we have experimented with calibration from the original deployment context. The results are quite variable but the model is able to achieve higher accuracy by limiting the number of predictions it makes.

This demonstrates the feasibility of drawing directly on work from other contexts, although as noted this is against a slightly different classification structure.

5.4 Data cleaning

Following the classification step, further checking of the data takes place. This step is required because while the classification algorithm used does classify the data effectively, it is still possible that there are a number of misclassifications contained within the data that could introduce a bias. This procedure also identifies anomalous data.

Extreme price changes

At the beginning of the project, we conducted a piece of analysis to examine volatile prices in the web scraped grocery prices. The analysis highlighted products that saw notable price changes. A number of the extreme high and low price relatives flagged by the analysis could be attributed to one-off errors in the data collection. Therefore, these were removed from the dataset as they did not represent volatile prices. For each product left in the dataset, the underlying prices were investigated and found to be genuine price changes, usually as a result of a product coming on or off sale.

Cluster-based anomaly detection

Given the existence of extreme prices and misclassifications in the web scraped grocery data, a more robust anomaly detection method is required. Therefore, an automated method of error detection has been developed using unsupervised machine-learning techniques. The method and results are summarised in this section.

For the method to be applied, the assumption is made that the anomalous prices follow a different distribution to correctly classified data (that is, the prices fall in the tails of the distribution). Therefore, if these different distributions can be identified it will then be possible to label the data as potential misclassifications or anomalous prices. The identification of the different distributions can be done by using an unsupervised machine-learning technique called cluster analysis. The clusters are sensitive to heterogeneity in the underlying data, which means that data from different distributions will be assigned to different clusters. While there are many different clustering algorithms to choose from, it was decided to use Density-based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN is not as sensitive to the choice of distance measures, initial "means" or the shape of the clusters as other popular clustering algorithms are.

The products were clustered according to their price, with similarly-priced products being placed in the same clusters. This is appropriate here because we are not measuring price growth of these clusters, rather that misclassified products and erroneous prices are identified correctly. For more information on how the DBSCAN algorithm works and an example of its application to our data, please see Research indices using web scraped data: May 2016 update.

To clean the data, two sets of clusters are calculated: a set of "training" clusters, which were produced using DBSCAN on the first month of collection, and a set of "full" clusters, which were calculated for the rest of the period. The first month was chosen to train the clusters because during the development of the web scrapers the first period of data were closely checked for errors. Additionally, for the full clusters to pick up extreme prices in the dataset, the training clusters would initially need to contain anomalous prices.

The limitation to the approach to calculate "full" clusters is that given more data, the clusters are likely to change in size and shape. Therefore, it is entirely possible that clusters may merge, or a new cluster may form between existing clusters. This could happen due to a new product being introduced on the market at prices between the old products. To take account of this effect, the "full" clusters are matched to the "training" clusters so products retain their initial cluster assignment across time. From this, any clusters that did not exist in their entirety in the training month are labelled as errors. The problem with this approach is that new products may be classified as errors because they do not fit into previously-defined clusters. Future works shall look into optimising the parameters of the clustering algorithm to resolve this issue, and to refresh the "training" cluster annually.

After the algorithm has been applied, the labels given to the product during the clustering step are compared to those given in the classification step. This comparison creates four different categories: agreement in correct classification, agreement in correct misclassification, new product, and disagreement in classification. The last two then require further manual checking to see whether the product is actually new, or to understand why the machine-learning algorithms disagree with each other. For future work this manual checking will be automated, and reduce as the parameters in the model are optimised. Over time, the combined validation process removes on average around one-quarter of the collected prices from the dataset.

5.5 Imputation

The last stage of the data processing stage is imputation. There are a number of reasons why there might be missing prices in the web scraped dataset: a product is no longer available, a product is temporarily unavailable (out of stock), or there was an error in the collection ("scraper break"). Where a product is out of stock or there has been a scraper break, imputation can be used to deal with temporarily missing prices.

A number of imputation methods have been tested on the web scraped dataset (Mayhew, 2016) and as a result the following rules have been applied to missing prices:

1. if a product is unavailable then a price will be carried forward for a maximum of 3 days
2. if a scraper break is identified then a price will be carried forward for a maximum of 7 days
3. otherwise, prices should not be imputed and if all product prices are missing from certain days, this will be shown as a dotted line in the series

Carrying forward prices performed best in the simulation study conducted in Mayhew (2016). This makes sense as there are relatively few price changes day to day. In fact, it can be seen that across the web scraped dataset, prices only change on average every 120 days.

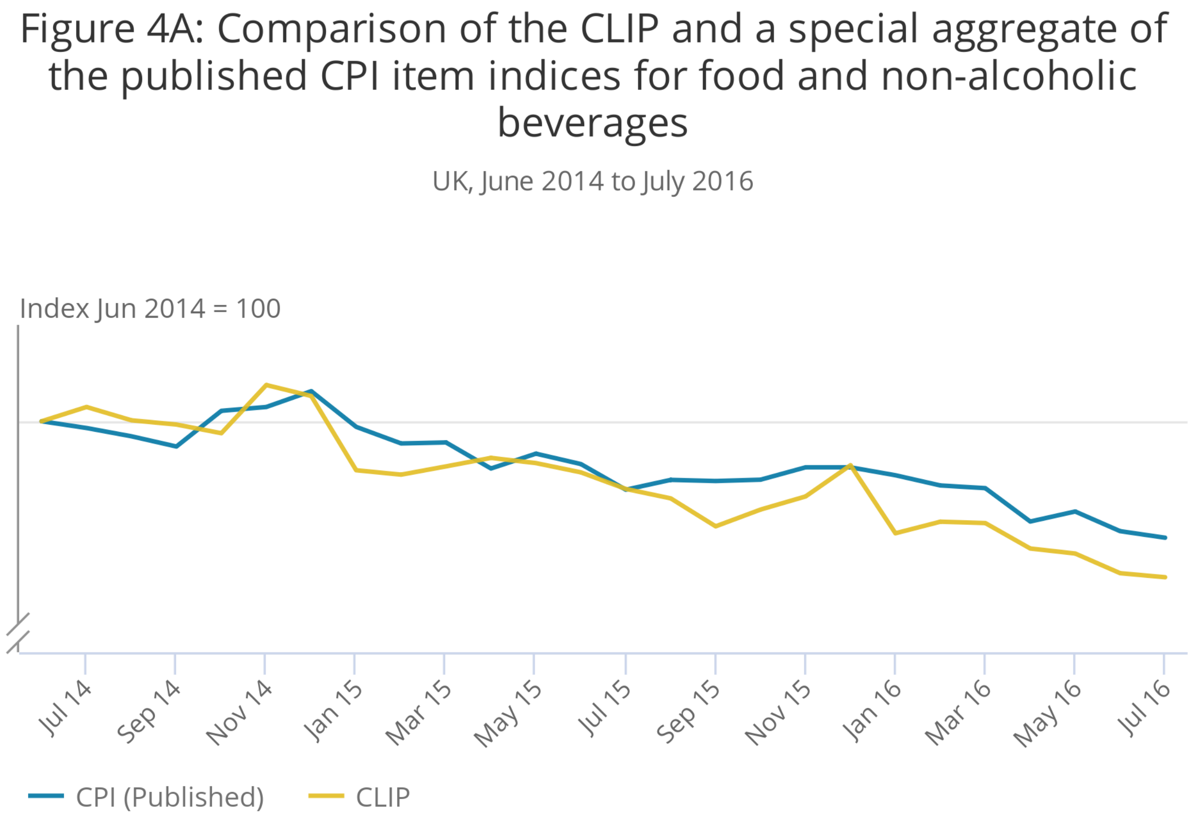
Imputing missing prices in the case of a scraper break is necessary as we are compensating for not being able to collect data due to a technical issue. The majority of scraper breaks do not last longer than 7 days, so imputation of prices in these instances helps us reduce breaks in the series. The decision to impute prices in the case of out of stock items is less clear cut. In measuring inflation, we aim to record the price of items that are available to consumers on a particular day and imputing for out of stock items may seem contrary to this. However, while the online supermarkets scraped by ONS operate national pricing policies, the stock availability is local. The decision to impute for missing prices aims to address this, and produce a price index representative of UK prices.

5.6 Indices

The web scraped data are then used to construct price indices. Figures 4a and 4b present the clustering large datasets into price indices (CLIP) index for the higher-level aggregates of food and non-alcoholic beverages, and alcoholic beverages respectively (reproduced from the November 2016 paper [Research indices using web scraped price data: clustering large datasets into price indices (CLIP)]()). Although these indices are not produced on a comparable basis with the published Consumer Prices Index (CPI) for a number of reasons, including data source and methodology, it is still a useful exercise to examine the trends shown in the different indices. Therefore, the CLIP indices are produced here alongside special aggregates of the published CPI item indices, which only include items that have been collected in the web scraping pilot.

Figure 4A: Comparison of the CLIP and a special aggregate of the published CPI item indices for food and non-alcoholic beverages

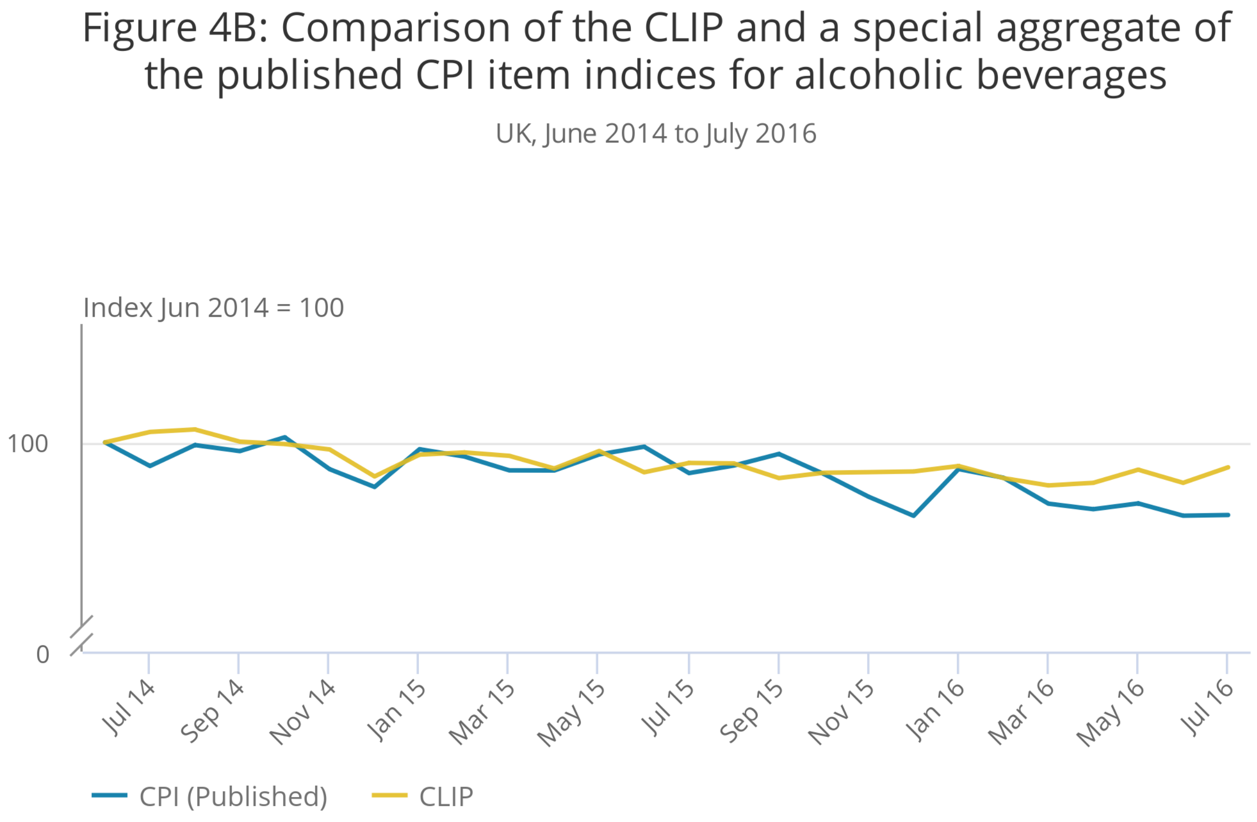
UK, June 2014 to July 2016



Source: Office For National Statistics

Figure 4B: Comparison of the CLIP and a special aggregate of the published CPI item indices for alcoholic beverages

UK, June 2014 to July 2016



Source: Office for National Statistics

Similar downward trends are shown for both the published CPI and the CLIP price indices. CPI has seen largely negative contributions from grocery prices over the period since February 2015. While we may not expect the CLIP to behave in the same way as the CPI due to its different methodology and source data, supermarkets have been engaged in a price war since the beginning of 2015 and have therefore reduced prices accordingly to attract consumers. The CLIP also provides evidence for this behaviour.

The similarity in trend is particularly true for food and non-alcoholic drinks, for which the CPI and the CLIP have very similar dynamics over time especially for the period June 2014 to July 2015. For alcoholic drinks, the CLIP index is smoother than the published index. This may be due to the larger amount of data being used within the calculation.

Our analysis has also compared grocery price indices constructed using the CLIP to other methods that could be used to compile a price index, notably the Fixed Based Jevons index and the GEKSJ index. The CLIP tends to follow the GEKSJ more closely than the Fixed Based Jevons. This may be due to the fact that the fixed-base methodology used in the Fixed Based Jevons means that it is not possible to include products with high product churn within the index. The deviation therefore demonstrates the impact of including price movements for items that appear in some, but not all, of the periods, which the CLIP and GEKSJ can better account for.

Results from the application of other methods to the web scraped groceries data can be found in Research indices using web scraped data: May 2016 update and Research indices using web scraped price data: clustering large datasets into price indices (CLIP). These studies have been used to feed through into the methodology review (Section 4) and inform what indices should be used for the other projects (Sections 6 and 7).

5.7 Future work and recommendations

The recommendations and plans for future work below largely relate to the development of the scrapers and data processing steps for the grocery web scrapers. In terms of index creation, please refer to the future work section of Section 4.

Collecting richer data from websites

In our current implementation we collect the product name, the crumb trail for the product hierarchy, whether the product is on offer, price and size information. These data are available across all supermarkets and are relatively standard between them. However, there are more data that we are currently not collecting, such as product descriptions, ingredients, nutrition information, product reviews and product images. An avenue for future research would be to combine a richer scraped dataset with an extensive training set to improve classification. It should be noted that in a production system we would probably need to screenshot the pages we are scraping (for audit purposes), so the burden of collecting and storing these data may well be a necessary part of a production system.

Investigating different approaches to automatic classification

There are many different ways to classify the web scraped data into the Classification of Individual Consumption according to Purpose (COICOP) aggregation structure. A number of suggestions for further work are detailed in this section, and these will be explored in conjunction with work carried out by other National Statistics Institutes (NSIs) and Eurostat (these projects are currently in the context of scanner data but the same principles will apply). At the moment our other projects do not require automated classification but in future these may require more technical methods if we expand the collection in these areas.

Classifier calibration

The current system produces a binary decision on whether a product is correctly classified. At the moment we don’t output any measure of how certain that classification is. This has been a perfectly valid approach for this project, but does mean that any changes to the decision threshold would require a wholesale reclassification of the dataset. An alternative approach would be to use some of the training data to calibrate the model and express certainty in the classification as a probability. This is an avenue that is currently being explored by other NSIs and Eurostat in the context of scanner data but the same principles will apply. Well calibrated probability measures of the estimated chance of a classification being correct also help build trust in a system.

Unsupervised clustering of products

There are multiple points in the project where we use clustering to process data, for example, when detecting anomalous price data, or in the CLIP experimental price index (where we track prices of clusters of products rather than the products themselves). A perhaps natural extension of this is to allow clustering to drive our classification of products. For example, we could cluster products, effectively forming a custom classification, and then match clusters to COICOP categories (or basket items). To do this we would need to pick a clustering approach that reliably grouped similar products together, we could then label the whole cluster (possibly using weights to allow for fuzzy mappings).

Structured learning – exploiting structure in COICOP

There is a school of machine-learning approaches that exploit structure in the data that they are learning. In our case we could potentially exploit the hierarchical nature of COICOP to improve classifications. Rather than training a general classifier as a multi-class classifier, where the algorithm picks the best label from the list of basket items, we could instead train a multi-label classifier. This would pick the set of labels from all levels of COICOP that best describe the item. A structured learner can learn the correlations between labels that are present in the full COICOP hierarchy. An added advantage of this would be that even if the algorithm can't get the exact basket item, it might be able to at least guess the correct class. This could form part of a system working in concert with manual classifiers.

Notes for Section 5: Grocery Prices Scraping Project

1. Extending the scrapers to cover other supermarkets was not possible due to terms and conditions on the relevant websites that prohibit web scraping.

6 . Central Collections Project

6.1 Introduction

The second strand of our research on web scraped data aims to assess the feasibility of using point and click web scraping tools to facilitate the central collection of prices that currently takes place manually within the Prices division at Office for National Statistics (ONS). This includes looking at the impact of using larger datasets, which can be collected more frequently than the traditional monthly collection. This may lead to greater coverage of items that are currently collected centrally, but also adds a number of challenges in terms of data processing and storage, as well the calculation of indices.

Out of the 730 items currently collected within the Consumer Prices Index including owner occupiers' housing costs (CPIH) basket, prices for 192 of these items are collected centrally, accounting for almost 26% of the total CPIH basket (Figure 1). Prices for these centrally-collected items are obtained manually by the ONS price collectors through websites, phone calls, emails, CD's and brochures. The process of obtaining prices via these methods is resource intensive and therefore the introduction of alternative data sources such as web scraping has the potential to greatly improve the quality and efficiency of price collection. Web scraping can also increase the number and frequency of price quotes that can be collected, which may reduce volatility in the existing index that is caused by only being able to collect prices once a month.

The Central Collections Project is initially scheduled to run from January 2016 to the end of 2017. This section summarises work completed so far on the project, finishing with some next steps that we would like to concentrate on in the next stage of the research.

6.2 Data collection

Since January 2016, we have been collecting web scraped data for a number of items in the CPIH basket (Table 5) as a proof of concept.

Table 5: Items that are currently collected by the central collections pilot

Item ID	Item name	Date collected from
630232	Blu-Ray Disc, purchased over the internet	Jan-16
630224	DVDs, purchased over the internet	Jan-16
630223	CDs, purchased over the internet	Jan-16
630122	PC Peripherals (printers and routers)	Jan-16
640304 to 640309	Package Holidays	Nov-15
630129	Laptops (for quality adjustments)	Feb-17

Source: Office for National Statistics

These items were chosen for a number of reasons. Chart collections (Blu-Rays, DVDs and CDs) and package holidays were identified as being areas of the basket where web scraping could improve both the coverage and frequency of data collected.The current collection method for chart collections captures the top 10 chart positions each month from two online retailers, therefore twenty price quotes each month. Whereas, web scraping gave collectors the chance to increase the sample to cover the top 100.

For package holidays, the ability to both increase the size of the data collection and the frequency provided big incentives to include this in the central collections pilot. In December 2014, Eurostat released their recommendations for the treatment of airfares and package holidays, which highlighted several possible areas for improvement, including last minute deals, booking fees, treatment of seasonal items and booking method. The current aim for the package holiday scrapers is to increase the number of type of bookings collected, including for countries such as; “Caribbean”, “France”, “Greece”, “Spain”, “Portugal”, “Turkey”, “USA” and “Italy”. This collection has initially been limited to once a month, but in future we would like to begin looking at increasing the number of times per month the scrapers are run to improve our understanding of the fluctuations in price seen across the period. This work stream is closely linked to the project run by ONS on package holiday compliance more generally.

Laptops were chosen as an example of an item where hedonic regression is currently used for quality adjustment. Hedonic regression is a technique that uses a set of ordinary least squares regressions to relate the price of an item to its measurable characteristics. For these regressions to provide statistically useful results, a large number of product observations (including price and suitable attribute information) are required, which creates a large resource burden on price collectors. Initially, the aim was to use web scraped data both to calculate these regression models and also to replicate the regular collection of prices.

These data were largely collected from one popular retailer’s website. Unfortunately, the terms and conditions of this website changed shortly after the project was initiated. This meant that we weren’t able to test the feasibility of creating regression models from the web scraped data for a long enough time period to provide any meaningful comparisons, although initial results showed that using web scraping instead of manually collecting the products made a considerable efficiency saving for the price collectors. We have been able to continue the collection of prices but at the moment this only represents a small proportion of the total price quotes that are collected for laptops, due to the lack of website coverage.

PC peripherals (printers and routers) were chosen as a useful pilot study for a number of reasons. The item has fairly stable prices that are collected from a number of different websites manually, therefore the intention was to explore if the current collection methods can be replicated with less resource requirements. Secondly, the specifications for printers seem to change fairly regularly, which implies that products in the sample tend to change more often and require replacement with either comparable or non-comparable products, the latter requiring significant resource to ensure that any quality change is captured appropriately. Therefore, web scraping larger datasets containing additional product attributes could potentially alleviate the difficulty of making replacements for the price collectors, and at the same time increase the accuracy of these replacements.

Data are currently collected using Import.io, a point and click web scraping tool available using a paid online subscription. Prices are collected on a monthly basis on index day (usually the second Tuesday of each month). The reason Import.io was chosen as a platform to extract price data for this project was due to its easy-to-use visual interface and the minimal programming skills required to use the tool. It has also been used by a number of other national statistical institutions such as Statistics Norway, as detailed in Keeping up with the modern consumer – online data in price statistics (Hov Nyborg K, et al., 2016).

The project so far has focused on the set up and maintenance of these scrapers, as well investigating some of the issues around website access and terms and conditions. Data validation and processing issues have also been explored. For the purpose of this summary article, the analysis will be focused on the data collected for the following items: “Blu-Ray Disc”, “DVDs”, “CDs”, “PC peripherals” and “laptops”.

6.3 Data cleaning

Because of the different set up of the scrapers, there is less need for this project to consider advanced classification methods such as those presented in Section 5. In terms of cleaning and validation, for some of the centrally collected items (in particular, the chart based collections), there is not enough data to fully utilise the clustering technique used in the anomaly based error detection method in Section 5. It may be more appropriate for items such as PC Peripherals and Laptops, where a larger number of products are collected. For this report, we have focused on understanding the distribution of the underlying datasets to determine what methods of cleaning can be applied to the data. Application of the various cleaning techniques to the centrally collected items is left for future work.

Fundamentally, the correct choice of data cleaning techniques is dependent on the properties of the underlying distribution modelling any given dataset. If the underlying distribution can be assumed to be parametric (that is, data can be modelled as being drawn from a distribution – for example, normal – based on a fixed set of parameters), parametric statistical methods and tests can be used to clean and validate the dataset. Since parametric models rely on a fixed set of parameters they make strong assumptions about the distribution. However, if the underlying distribution can’t be assumed to be parametric, then non-parametric statistical methods and tests should be used instead. These assume far less about the underlying distribution.

Distribution of the underlying data

The first step to clean and validate the data is therefore to understand the underlying distribution of the data. Distribution plots for both web scraped and manually-collected data have been produced for each item in this pilot collection. Figure 5a is a histogram plot for web scraped and manually-collected prices of “Blu-ray Discs” over the period January 2016 to May 2017. Figure 5b is a continuous Kernel density distribution of the same data. The graphs for the other items can be found in Annex B.

Figure 5A: Histogram for web scraped and manually collected prices of ‘Blu-ray Discs’

UK, January 2016 to May 2017

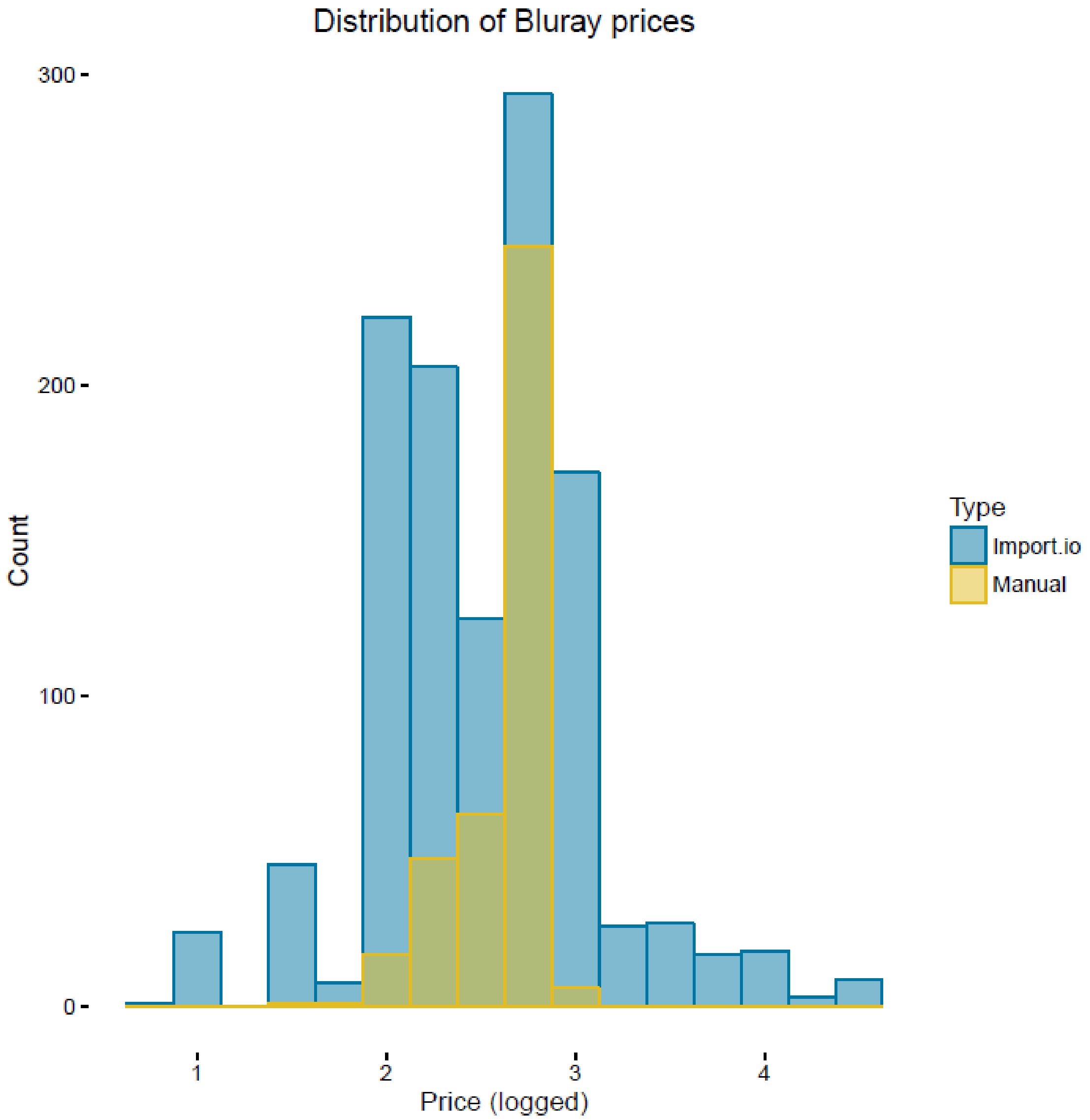
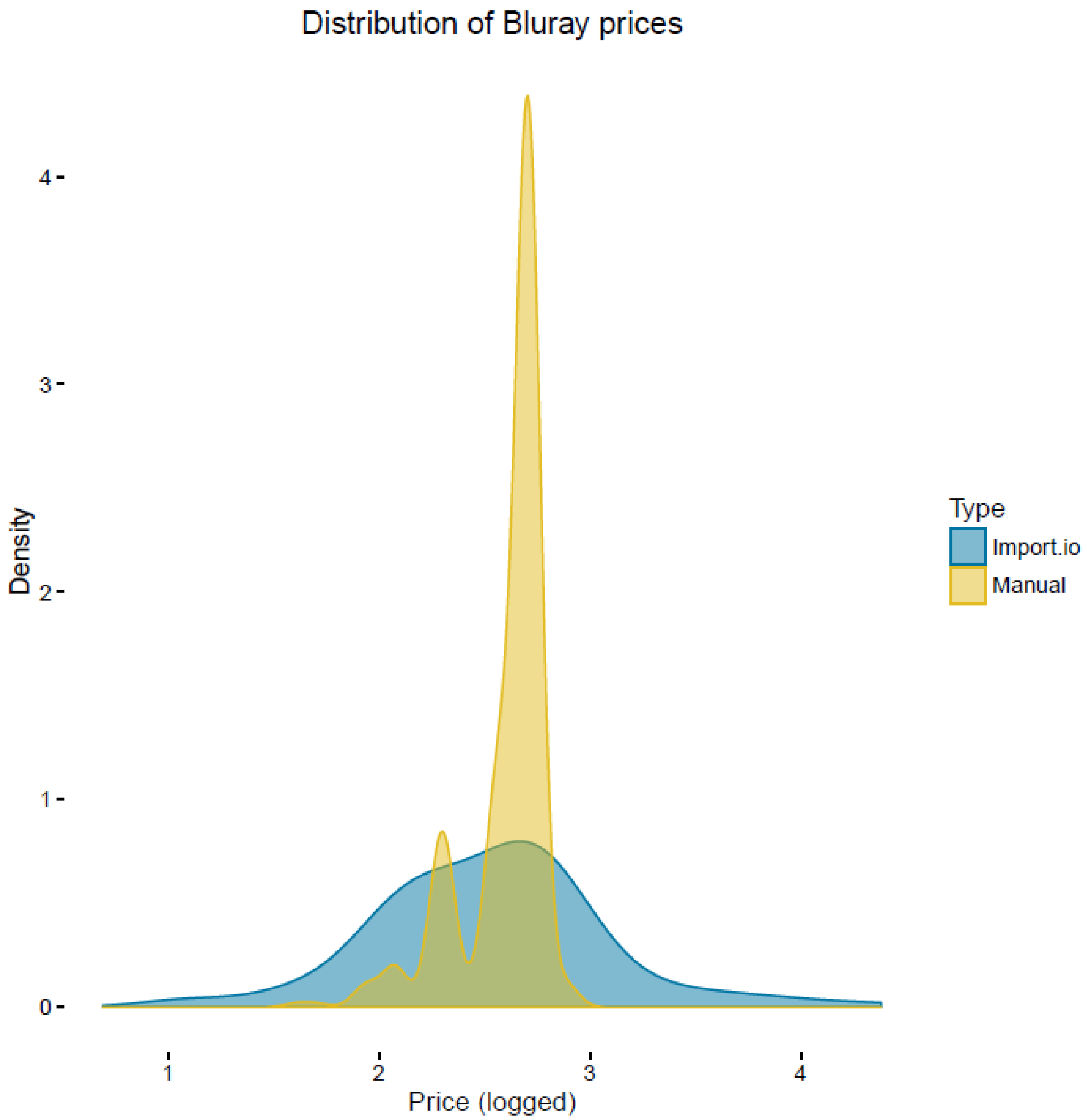


Figure 5B: Kernel density distribution for web scraped and manually collected prices of ‘Blu-ray Discs’

UK, January 2016 to May 2017



From the distribution charts, it is apparent that both datasets follow a multimodal distribution. These distributions are drawn from logged prices. While log-transforming the data did not result in a normally distributed dataset, it did to an extent reduce the asymmetries and positive skew that existed in the distributions initially (particularly for the raw web scraped data).

However, it should also be noted that these distributions are drawn from un-cleaned data. This means that there may still be anomalous prices that remain in the data, resulting in longer tails to the distributions. This is particularly evident in the case of web scraped data for “CDs” and “DVDs” (Annex B). The occurrence of anomalous prices is relatively low for the manual data compared with the web scraped data, primarily because these prices are manually obtained from websites and are scrutinised by expert price collectors at the time of collection. This results in the manually-collected price distributions having substantially smaller right-hand tails compared to the web scraped data.

It is also clear from the distributions that the web scraped data have a wider range and are less sparse when compared with the manually-collected data. This is particularly the case for the media chart collections (Blu-Rays, CDs and DVDs). This may be due to the differences in the sample sizes for the different data collection methods: the manual collections are based on a sample size of 20 each month, whereas the web scraped data are based on a sample size of 100. Additionally, the narrow peaks for the manually-collected media items could also be explained by the type of prices that the dataset contains. For instance, the pricing pattern for some of these datasets, in particular "DVDs", is quite irregular. This pattern, combined with a smaller sample size, results in narrow peaks in the distributions for the manually-collected media items.

Outlier detection using the Tukey algorithm

Having identified the datasets as suitable for a non-parametric approach, the next task is to identify a suitable non-parametric statistical test to clean and validate the data. For example, the Tukey algorithm is commonly used in these situations (for example, it is currently used in the traditional CPIH price collection to identify outliers for further scrutiny by price collectors - see Chapter 6 in the Consumer Prices Index Technical Manual for more information).

The algorithm operates as follows:

1. the ratio of current price to previous valid price (the price relative) is calculated for each price (in the case of items tested by price level rather than price change, this stage is omitted)

2. for each item, the set of all such ratios is sorted into ascending order and ratios of 1 (unchanged prices) are excluded (in the case of items tested by price level rather than price change, the prices themselves are sorted)

3. the top and bottom 5% of the list are flagged for further investigation and removed

4. the trimmed mean is the mean of the residual observations

5. the upper and lower "midmeans" are the means of all observations above or below the trimmed mean

6. the upper (or lower) Tukey limit is the trimmed mean plus (or minus) 2.5 times the difference between the trimmed mean and the upper (or lower) midmean

7. price relatives, or price levels, outside the Tukey limits are flagged for further investigation

The Tukey algorithm has been used in the traditional collection since the 1980s. It produces limits that are; intuitively reasonable, consistent from month to month, robust in the presence of outliers (in other words, adding in one or two rogue observations does not affect the limits set by the algorithm very much), and robust as data volume changes (that is, limits calculated from a subset of the data do not vary much from those calculated on the full dataset).

Due to the difference between the web scraped and manually-collected datasets, these parameters will have to be adjusted and tested in accordance to the needs of the different data. At present, the algorithm generally operates on price relatives, which will not be feasible to obtain for the web scraped data. This is because the data contains high product churn, which results in the web scraped data consisting of missing values over various months. The algorithm will therefore need to operate on price levels rather than price changes.

In addition, the top and bottom 5% of the observations are currently flagged for further investigation and removed. Because of the additional range of prices captured by the web scraped data, this may introduce a risk of missing out on genuine prices.

Lastly, the algorithm currently uses “midmeans”. With web scraped data, there is a possibility that “midmeans” could be influenced by more extreme price changes at the higher end. Therefore, other measures could be used instead, such as taking the median.

In summary, we intend to test and explore various different parameters that are used within the Tukey algorithm. We will also look to test other cleaning techniques as part of this work, to find a method that best suits the needs of the different central items collected using web scraping.

6.4 Data access and legal and ethical issues

This project has opened up a number of challenges with regards to the collection and processing of web scraped data. Due to the increase in the number of websites required to collect prices compared to the three websites that were chosen for the grocery prices project, further consideration was required to ensure that web scraping is carried out consistently, ethically and legally. We are currently in the process of drafting web scraping guidance. Adopting this guidance and making it publically available will ensure that we have a consistent and transparent approach to web scraping.

There are some websites that are currently used in the manual collection that do not allow web scrapers to be used to download data. As already discussed, this has made it difficult to completely replicate our existing central collection and has limited the research that we can undertake in certain areas, for example, hedonics. In future, it may be possible to contact the retailers and ask for data to be shared directly, for example, using Application Programming Interface (API) access. However, until this is resolved it is a significant disadvantage to note in the feasibility of using web scraped data to replace the central collection.

6.5 Future work and recommendations

Once the data have been cleaned and validated, they can be used to construct price indices. At present, different index methodologies are being tested and trialled on both the web scraped and manually-collected price datasets, incorporating the recommendations made in A comparison of index number methodology used on UK web scraped price data and summarised in Section 4.

The project also aims to explore other web scraping tools in the market. The intention is to run the different platforms in parallel to examine the differences in ease of use and output obtained.

We are also looking to extend the list of items that are currently included in this project. We have made significant progress recently in setting up scrapers for air fares and hope to be able to process these data in the next couple of months. We have also identified a number of items that don’t see frequent price changes but still require collectors to manually check the website each month. A work stream to look at the use of the CBS Robot Tool is set up for this summer.

Finally, this collection has initially been limited to once a month, but in future we would like to begin looking at increasing the number of times per month the scrapers are run to improve our understanding of the fluctuations in price seen across the month.

7 . Clothing Data Project

7.1 Introduction

The final strand of our research on web scraped data focuses on web scraped clothing data provided by World's Global Style Network (WGSN), a global trend authority specialising in fashion. Clothing items generally experience much higher rates of product churn (that is, products coming in and out of stock) compared to 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. These difficulties mean that there is a particular interest in investigating generating clothing price indices using alternative data sources.

This section summarises our analysis into using web scraped clothing data. For more information, please see the papers Analysis of product turnover in web scraped clothing data and its impact on methods for compiling price indices and Research indices using web scraped data: clothing data.

7.2 Data analysis

The data used were provided by WGSN. They collect daily prices and other information from a number of fashion retailers' websites. The web scraped data include 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 product categories do not necessarily match the item descriptions used for the Consumer Prices Index including owner occupiers' housing costs (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 6.

Table 6: 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

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 (including by type of retailer) is given in the paper Analysis of product turnover in web scraped clothing data and its impact on methods for compiling price indices.

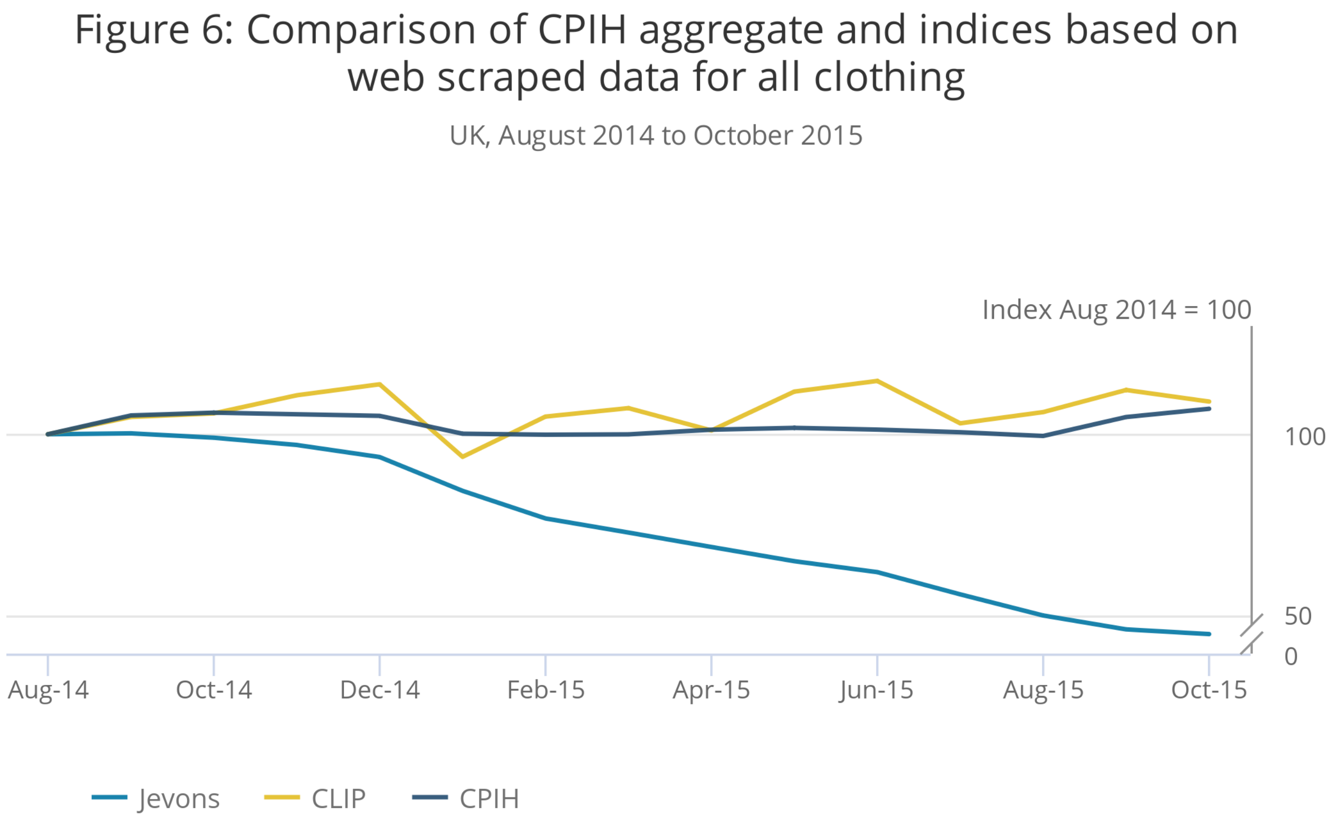
7.3 Indices

These web scraped clothing data were then used to produce research price indices. Following a similar structure to analysis on web scraped data for groceries and items collected centrally, a number of methods were applied, summarised in Research indices using web scraped data: clothing data.

Figure 6 presents the CPIH special aggregate, Chained Bilateral Jevons and clustering large datasets into price indices (CLIP) index for all clothing. The FEWS, RYGEKS, IntGEKS and GEKS series are excluded from this graph as these gave more implausible results over the time series. The CPIH special aggregate presented here is constructed in a similar way to the special aggregate presented for the groceries data, using only the items (and respective weights) that are included in the web scraped indices. There are many reasons why it is not appropriate to draw a direct comparison between the price indices presented in the previous sections, 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.

Figure 6: 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

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 Bilateral Jevons, which decreases over the period investigated.

7.4 Future work and recommendations

This work has contributed to our understanding of the online clothing sector, in particular with regards to how rates of product churn can be affected by different types of products and retailers. We have also taken the opportunity to apply price index methodology suitable for Big Data sources to a new section of the consumer prices basket.

This research has also highlighted the difference in product replacement behaviour of some online-only retailers, who do not have issues such as shelf space to limit stock availability. As expenditure on clothing is seeing one of the largest increases in online spend, further investigation is required to understand if we are currently adequately capturing price behaviour for clothing in our consumer price statistics.

Many issues discussed in Section 4.6 (Future work) are particularly relevant to clothing prices data. Web scraped data contain no expenditure information, which means that all products are counted equally in the production of these indices. Studies by other National Statistics Institutes (NSIs) have shown that this may lead to downward bias in indices that are calculated from web scraped data, compared to 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.

These clothing data can also be used to further test our data processing techniques discussed in Section 5, including the development of a general classifier. Similar principles can be applied across the different sections of the CPIH basket.

8 . Conclusions

Office for National Statistics (ONS) is at the forefront of international research on the use of web scraped data in the production and development of consumer price statistics. Since 2014, we have made significant progress in a number of areas including the development of in-house scraping capability and associated data processing issues such as data storage, classification, cleaning and imputation. We have also made important contributions to the price index methodology literature, including the development of a new index: clustering large datasets into price indices (CLIP).

Our work has focused on three areas of the Consumer Prices Index including owner occupiers' housing costs (CPIH) basket of goods and services: grocery items, items that are currently collected centrally by our price collectors, and clothing items.

Our research into grocery items has enabled us to explore methods of collecting web scraped prices in-house, including the development and maintenance of custom-built scrapers in Python. This has led to wider benefits for ONS in general, in particular an increase in knowledge and experience that has contributed to the success of other Big Data projects such as web scraping job vacancies. We have also made a number of improvements to the way that web scraped data are cleaned and classified, which can be applied to other alternative data sources.

The second strand on centrally-collected items assesses the feasibility of using "off the shelf" point and click web scraping tools to facilitate the collection of these prices. The project so far has given us valuable experience in navigating the use of web scraped data. In particular, the need to scrape a wider range of retailers, compared with the grocery items collection, has required us to develop new guidance around the legal and ethical aspects of web scraping.

The treatment of clothing prices in consumer price statistics is a topic of interest for many National Statistics Institutes (NSIs) and the use of alternative data sources can potentially reduce a number of measurement challenges. It has also allowed us to explore web scraped data provided by a third-party, with external cleaning and processing already applied to the data.

Finally, there is still much discussion internationally over the best way to calculate price indices from these data. Our methodology review has shown that different methods are suitable for different areas of the CPIH basket. Practical issues should also be taken into account before deciding on which index should be used.

This work contributes to a growing body of international 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 across the consumption sector, and in the long-term, improve the way prices are collected for national consumer price statistics. There remains, of course, much work to be done in this area.

9 . Future work and recommendations

This report has highlighted a number of areas for future work including the continued development of in-house scraping capability and associated data processing issues, and extended research and testing of how these price quotes are incorporated into official production systems.

In summary, these are the main areas identified for future research by Office for National Statistics (ONS) into web scraped data:

Technological developments

- Continue to explore developments of our in-house scrapers, including research into more general classifiers, which can also be used in other areas of the Consumer Prices Index including owner occupiers' housing costs (CPIH) basket.
- Investigate different sources of web scraped data (for example, the use of third-party web scraped data may allow us to cover a larger proportion of the basket that can currently be maintained by ONS resources).

Methodological developments

- Further consideration is required to the extent of which web scraped data adequately captures the price trends that are seen in the local collection; it is only when this is better understood that we can explore the use of web scraped data to replace existing parts of the local collection.
- Practical issues such as European compliance and existing system constraints and requirements will be taken into account and used to inform the final decision on which index methodologies are suitable for use with web scraped data.
- The impact that the lack of expenditure weights for web scraped data will have on the index should be tested, with appropriate expenditure proxies.
- Further research is also required on the frequency of data collected; at the moment, our grocery prices data is collected on a daily basis and we would like to extend this to our central collection – how this data is used within the index remains to be tested.

There are also two additional considerations that may affect the use of alternative data sources in the development and production of consumer price statistics in future.

The first is that recent legislation (the Digital Economy Act 2017) will give us faster and easier access to data from external sources such as government departments, other public bodies, charities, and large- and medium-sized businesses. This may open up the possibility of accessing scanner data for certain areas of the CPIH basket. The areas identified in this section for the most part will also be relevant when processing scanner data, but there may be further issues that have not been considered yet in the context of web scraped data that will also need to be prioritised.

The second is that the contract for the local price collection, which is done by a market research firm, is due for extension in the next stage of the project. The next contract will likely be different to the current, to take into account the evolving modes of collection. This means that work to supplement the local collection may need to be prioritised.

To take these considerations into account, we are currently drafting a Data Sources Collection strategy for Prices. This will be used to inform our work plan on alternative data sources over the next 2 years. Research areas will be prioritised and taken forward accordingly.

10 . Acknowledgements

This research is partially funded by a Eurostat grant. The authors want to express their gratitude to Eurostat for the possibility of carrying out the research under the grant assigned. The authors would also like to thank the contributions of Rob Bucknall, Ben Hayes, Chris Payne and Edward Rowland at Office for National Statistics.

11 . References

Bean, Charles. Independent review of UK economic statistics: final report, 2016.

Beeson, Joshua. Web Scraped Data: Extreme Price Changes. 2015.

Bentley, Alan et al. Towards a Big Data CPI for New Zealand. Statistics New Zealand, 2017.

Bird, Derek et al. Initial report on experiences with scanner data in Office for National Statistics, 2014.

Breton, Robert et al. Consumer Price Indices, research indices using web scraped price data: May 2016 update, 2016

Breton, Robert et al. Consumer Price Indices, research indices using web scraped price data, 2015.

Breton, Robert et al. Trial Consumer Price Indices using web scraped data, 2015.

Consumer Price Indices - Technical Manual. Office for National Statistics, 2014.

Cavallo, Alberto et al. Our Research. The Billion Prices Project, 2016.

Chessa, Antonio G et al. A Comparison of Price Index Methods for Scanner Data. CBS Netherland, 2017.

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

De Haan, J. and van der Grient, H. Eliminating chain drift in price indexes based on scanner data. Journal of Econometrics 161 (1), 2009. pages 36 to 46.

Hayes, Ben et al. Research indices using web scraped data: clothing data. Office for National Statistics, 2017.

Johnson, Paul. UK Consumer Price Statistics: A Review. UK Statistics Authority, 2015.

Kalisch, David W. Information paper: An implementation plan to maximise the use of transactions data in the CPI, 2017.

Kantar Worldpanel. Grocery Market Share, 2016.

Krsinich, Frances. FEWS Index: Fixed Effects with a Window Splice, 2014.

Krsinich, Frances. Price Indexes from online data using the Fixed-Effects Window-Splice (FEWS) Index, 2015.

Mayhew, Matt. A comparison of index number methodology used on UK web scraped price data. Office for National Statistics, 2017.

Mayhew, Matthew et al. Using machine learning techniques to clean web scraped price data via cluster analysis, 2016.

Mayhew, Matthew. Imputing Web Scraped Prices. Office for National Statistics, 2016.

Metcalfe, Elizabeth et al. Research indices using web scraped price data: clustering large datasets into price indices (CLIP). Office for National Statistics, 2016.

Murphy, Rhian. Retail sales in Great Britain: May 2017. Office for National Statistics, 2017.

Naylor, Jane, et al. The ONS Big Data Project. Office for National Statistics, 2014.

Nieminen, Kristiina et al. Small scale “Big Data” in the Finnish Pharmaceutical Product Index Compilation. Statistics Finland, 2017.

Nyborg Hov, Kjersti, et al. Keeping up with the modern consumer – online data in price statistics. Statistics Norway, 2016.

Payne, Chris. Analysis of product turnover in web scraped clothing data, and its impact on methods for compiling price indices. Office for National Statistics, 2017.

12 . Annex A: Methodology

1. Fixed Based Jevons index

The Fixed Based Jevons fixes the base period to the first period in the dataset and matches the products common to all periods. It compares the current period price back to the base period. The formula is defined as follows:

$$P_{FBJ}^{0,t} = \prod_{j \in S^*} \left(\frac{p_j^t}{p_j^0} \right)^{\frac{1}{n^*}}$$

where p_j^t is the price of product *j* in period *t*, S^* is the set of all products that appear in every period and n=*S* is the number of products common to all periods.

2. Chained Bilateral Jevons index

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 $S^{i-1,i.}$.

3. Unit Value index

The Unit Value index is normally defined as the ratio of the unit value in the current period to the unit value in the base period. However, there is no quantity data in the web scraped data so a true Unit Value index can't be calculated. Instead, the ratio of geometric means of unmatched sets of products will be used, that is:

$$p_{UV}^{0,t} = \frac{\left(\prod_{j \in S^t} p_j^t\right)^{\frac{1}{n^t}}}{\left(\prod_{j \in S^0} p_j^0\right)^{\frac{1}{n^0}}}$$

where S^0 is the set of products in period *0*, and n^0 is the number of products in S^0, S^t is the set of products in period *t*, and n_t is the number of products in S^t. The geometric mean has been used so that it is consistent with the other indices presented in this article.

4. 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.

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 WE, Fox KJ 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*.

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:

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

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).

Imputation Törnqvist RYGEKS – ITRYGEKS

As new products are introduced on the market and old products disappear, an implicit quality change may occur, for example, this often happens in technological goods. Hence, there is an implicit price movement that isn't captured in the standard RYGEKS method because it doesn’t account for quality change. There is an implicit price change when these goods are introduced, and if the consumption of these goods increased then these implicit movements need to be captured.

De Haan and Krsinich (2012) propose using an imputed Törnqvist as the base of the RYGEKS. An imputed Törnqvist is a hedonically adjusted Törnqvist index, where the prices of new or disappeared products are imputed using a hedonic regression in the base or current period respectively. A hedonic regression assumes that the price of a product is uniquely defined by a set of *K* characteristics. The imputed Törnqvist index is defined as follows:

$$P_{IT}^{0,t} = \prod_{j \in S^{0,t}} \left(\frac{p_j^t}{p_j^0}\right)^{\frac{w_j^0 + w_j^t}{2}} \prod_{j \in S^t_{N(0)}} \left(\frac{p_j^t}{\hat{p}_j^0}\right)^{\frac{w_j^0}{2}} \prod_{j \in S^0_{D(t)}} \left(\frac{\hat{p}_j^t}{p_j^0}\right)^{\frac{w_j^t}{2}}$$

where:

- w_j^0 is the expenditure share of item *j* at time 0
- w_j^t is the expenditure share for item *j* at time *t*
- p_j^t is the estimated price for a missing product at time *t*
- $S^{0,t}$ is the set of products observed in both periods
- ${S_{N(0)}}^t$ is the set of new products at time t but weren't available at time 0
- ${S_{D(t)}}^0$ is the set of products at time 0 that have disappeared from the market at time t

De Haan and Krsinich (2012) suggest three different imputation methods: the linear characteristics model, the weighted time dummy hedonic model and the weighted time-product dummy method. These are discussed further in this section.

The linear characteristics model:

This method estimates the characteristic parameters using a separate regression model for each period. The imputed price is calculated as follows:

$$\widehat{p_j^t} = \exp\left(\widehat{a^t} + \sum_{k=1}^{K} \widehat{\beta_k^t} \mathrm{z_{jk}}\right)$$

where ^{t} is the estimate of the intercept, $_k^t$ is the estimate of the effect characteristic *k* has on the price and z_{jk} is the value of characteristic *k* for product *j*.

The weighted time dummy hedonic method:

This method assumes parameter estimates for characteristics don't change over time, and includes a dummy variable, D_j^i, for in which period the product was collected. In this method the imputed price is calculated by:

$$p_i^j = \alpha + \sum_{j=1}^{t} \delta^j D_i^j + \sum_{k=1}^{K} \beta_k z_{ik}$$

The weighted time-product dummy method:

This method can be used when detailed characteristic information is not available, and a dummy variable, D_j, for the product is created. The missing price is then estimated using:

$$\hat{p_j^t} = \exp\left(\hat{\alpha} + \sum_{i=1}^{t} \hat{\delta^i} D_j^i + \sum_{j=1}^{N-1} \hat{\gamma}_j D_j\right)$$

where

$\hat{\gamma}_J$

is the estimate of the product-specific dummy and the N^{th} product is taken as the reference product. This method assumes that the quality of each distinct product is different to the quality of other products to a consumer. It is a reasonable assumption as the number of potential characteristics is large and not all of them are observable.

For each of these methods, a weighted least squares regression is used, with the expenditure shares as the weights.

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 in 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.

5. FEWS

The Fixed Effects Window Splice (FEWS) produces a non-revisable and fully quality-adjusted price index where there is longitudinal price and quantity information at a detailed product specification level (Krsinich, 2016). It is based around the Fixed Effects index, which is defined as follows:

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

where
$\hat{\gamma}^0$
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 FEWS index: fixed effects with a window splice.

6. CLIP

Clustering large datasets into price indices (CLIP) is a recently developed price index from Office for National Statistics (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:

$$P_{CLIP}^{0,t} = \frac{\sum_k \mid C_(k,0) \mid \frac{\prod_{i\in S_h^t}(p_i^t)^{\frac{1}{n^t}}}{\prod_{i\in S_h^0}(p_i^0)^{\frac{1}{n^0}}}}{\sum_k \mid C_(k,0)|}$$

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 Research indices using web scraped price data: clustering large datasets into price indices (CLIP).

Notes for Section 12: Annex A: Methodology

1. Unit Value is usually defined as the value of a product divided by the quantity bought, but since the data aren’t available, an average price is taken for the web scraped data.

13 . Annex B: Distribution charts

1. Compact discs (CDs)

Figure 7: Histogram for web scraped and manually collected prices of ‘Compact Discs’

UK, January 2016 to May 2017

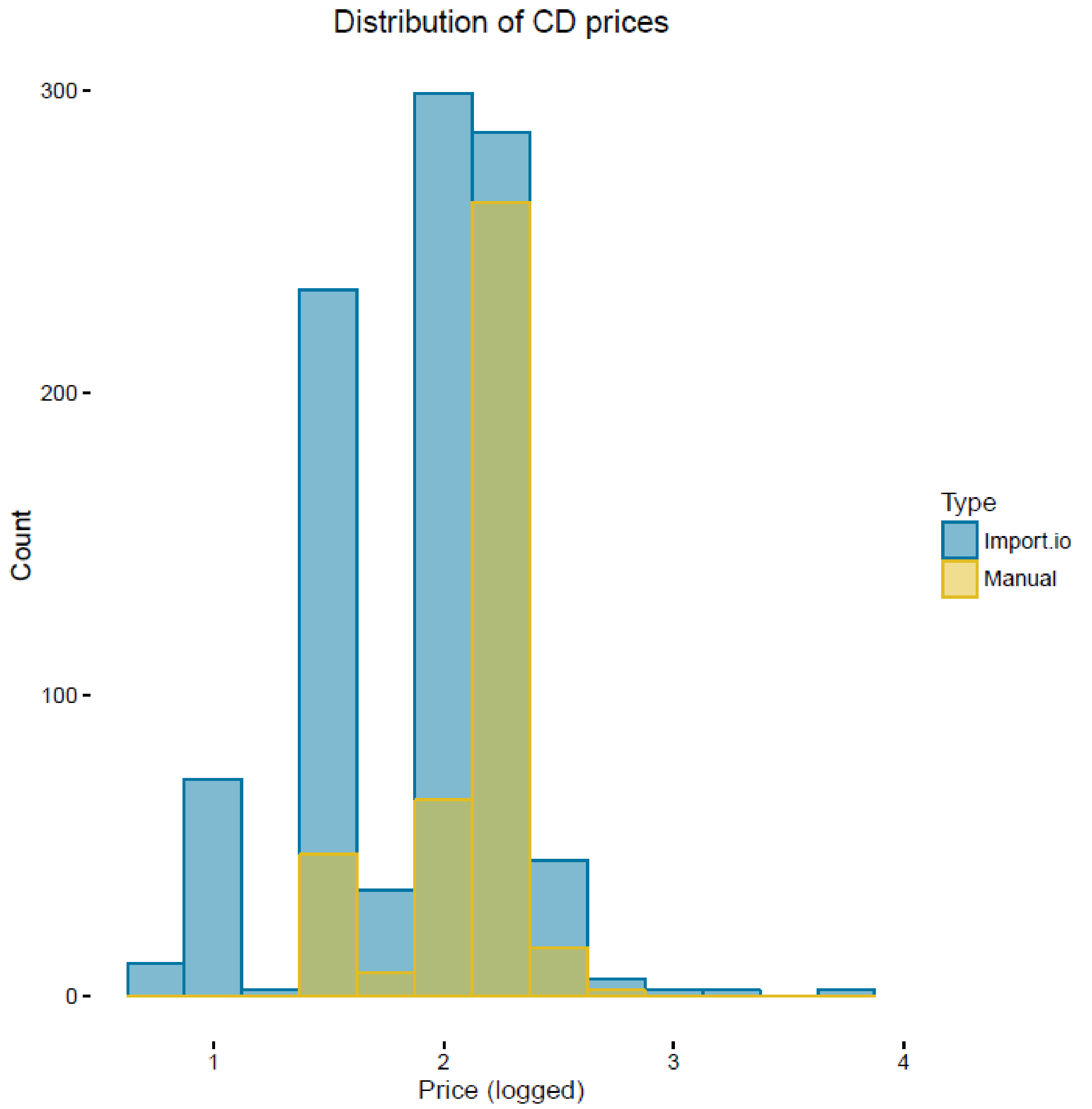
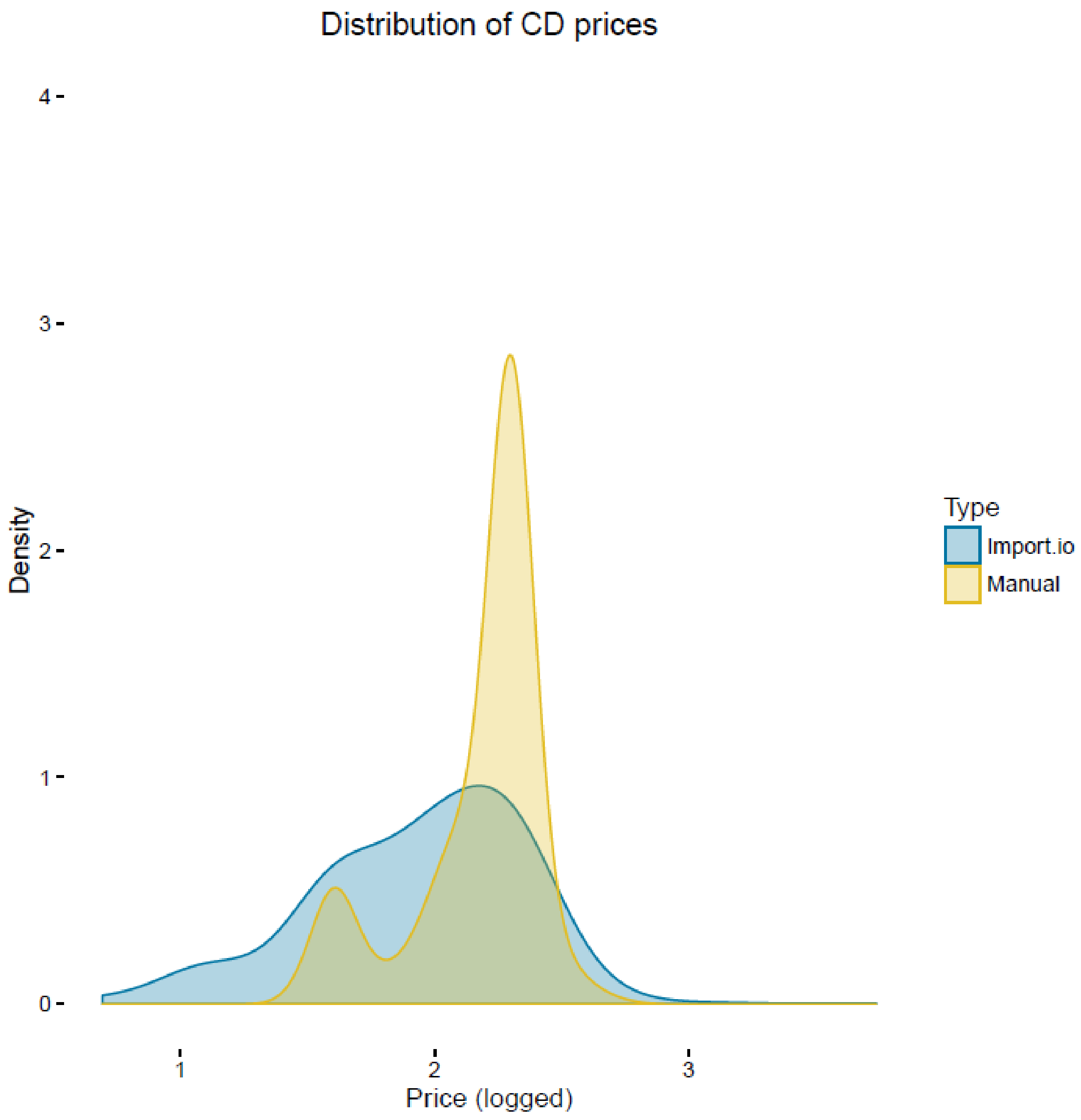


Figure 8: Kernel density distribution for web scraped and manually collected prices of ‘Compact Discs'

UK, January 2016 to May 2017



2. DVDs

Figure 9: Histogram for web scraped and manually collected prices of 'DVDs'

UK, January 2016 to May 2017

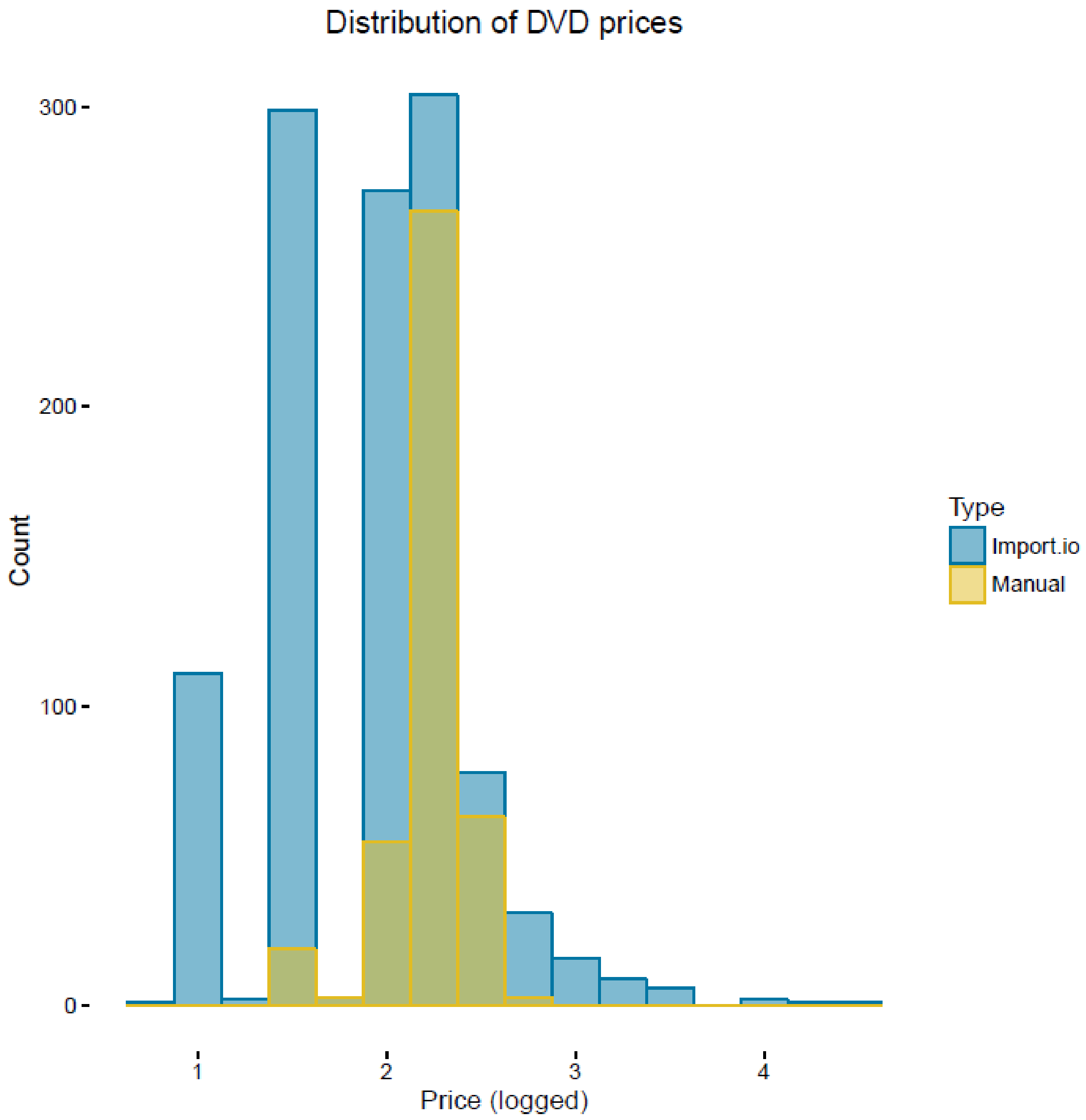
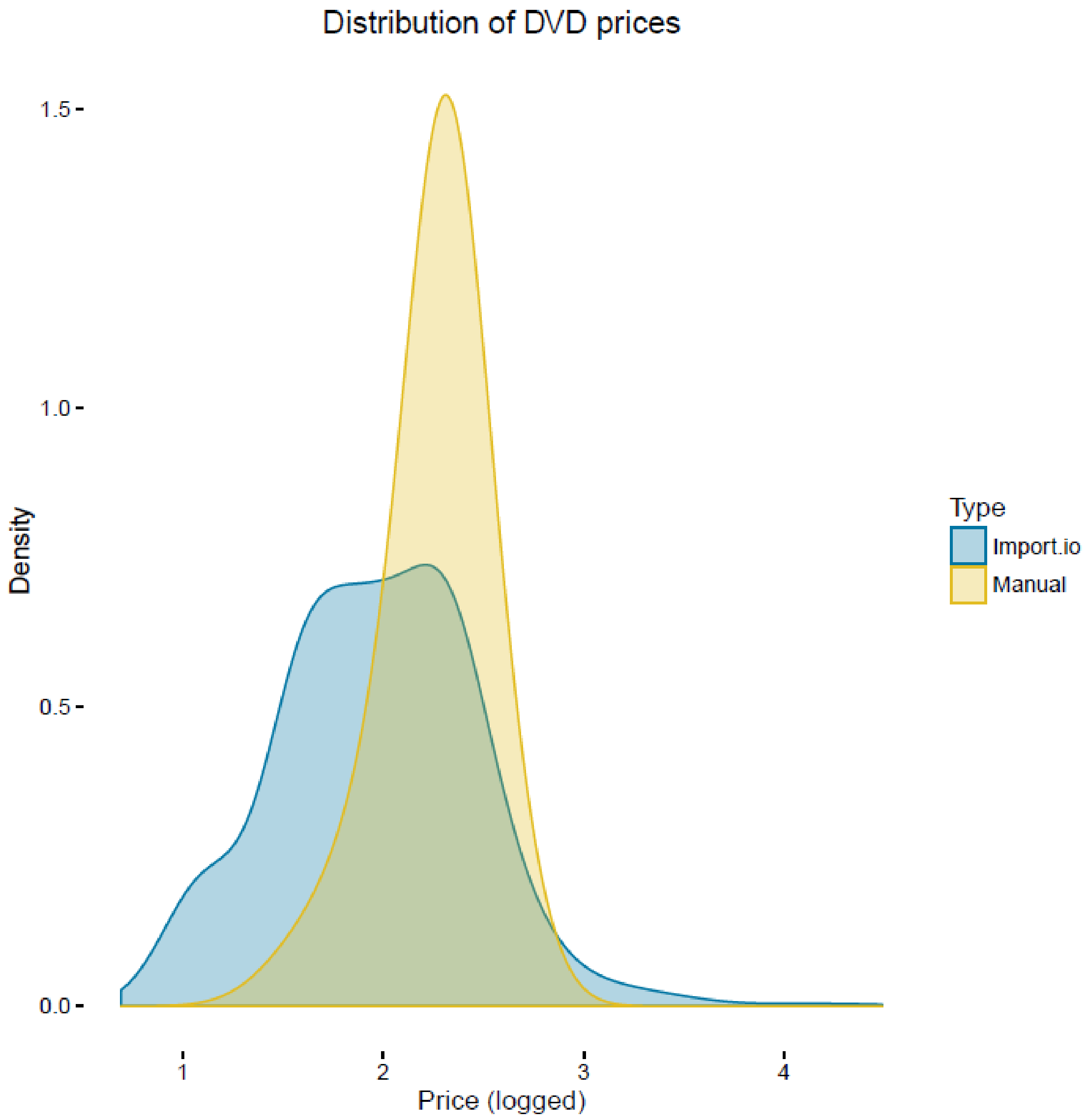


Figure 10: Kernel density distribution for web scraped and manually collected prices of ‘DVDs’

UK, January 2016 to May 2017



3. PC peripherals

Figure 11: Histogram for web scraped and manually collected prices of ‘PC Peripherals’

UK, January 2016 to May 2017

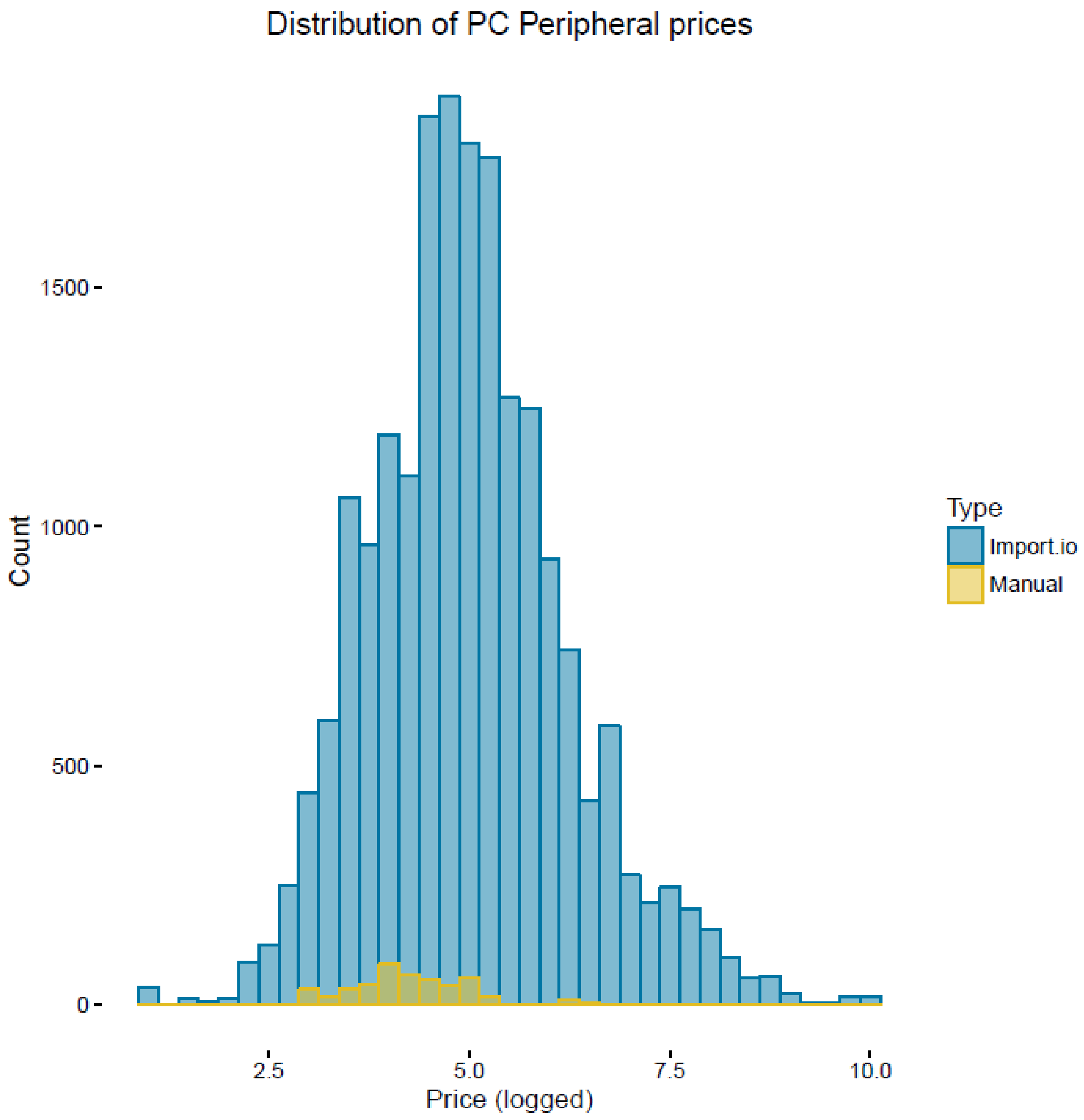
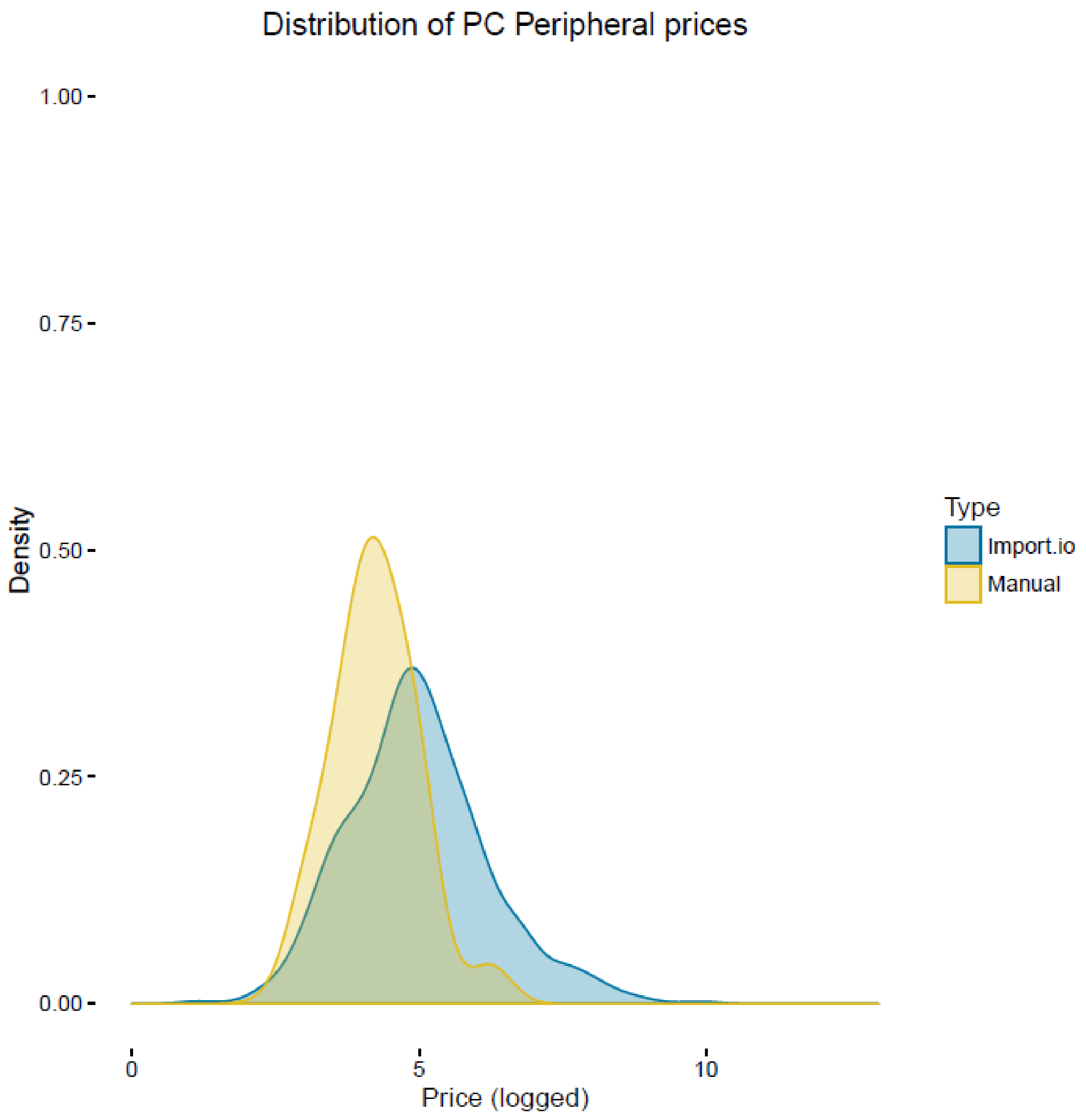


Figure 12: Kernel density distribution for web scraped and manually collected prices of ‘PC Peripherals’

UK, January 2016 to May 2017



4. Laptops

Figure 13: Histogram for web scraped and manually collected prices of 'Laptops’

UK, February 2017 to May 2017

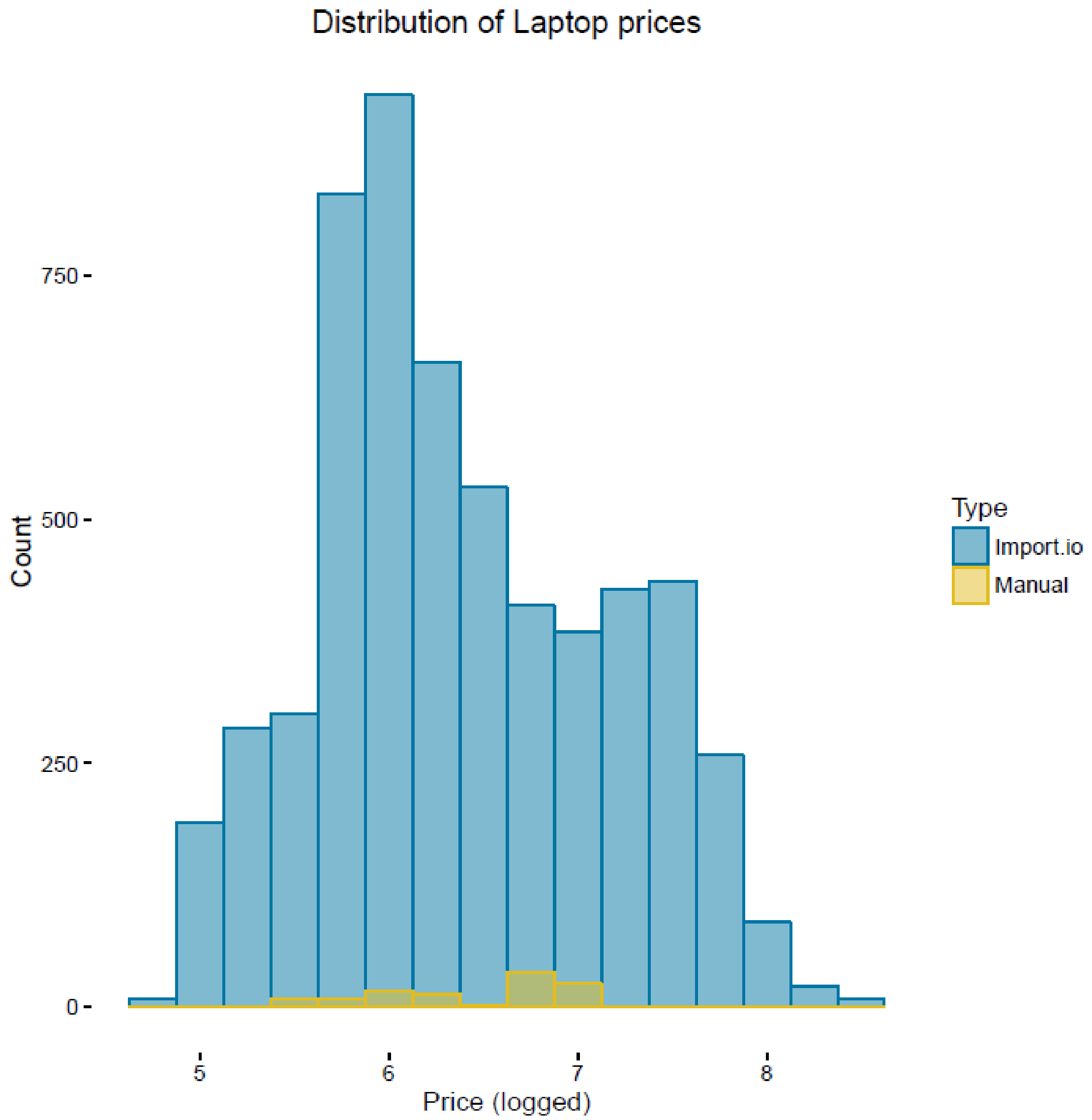
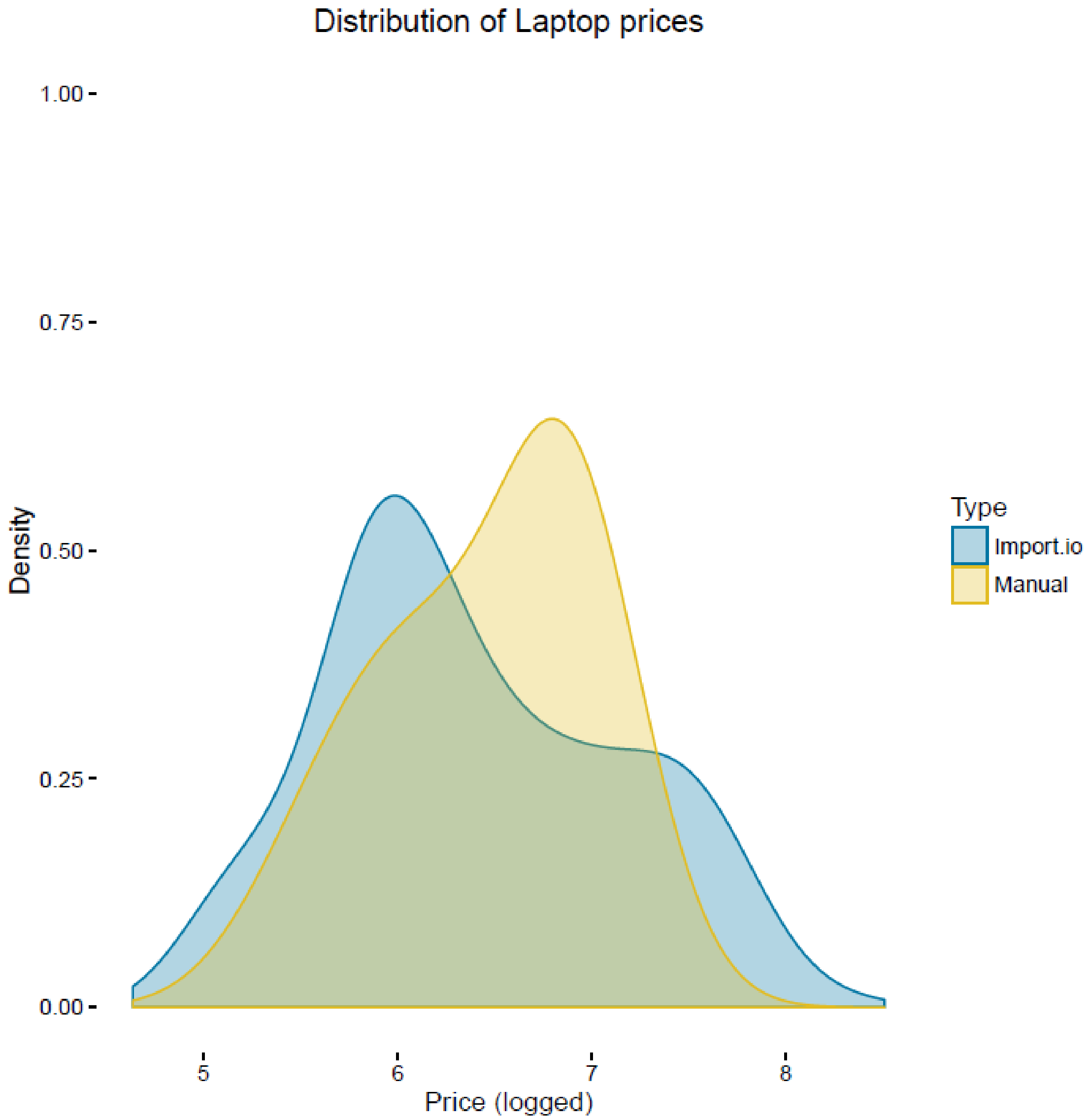


Figure 14: Kernel density distribution for web scraped and manually collected prices of ‘Laptops’

UK, February 2017 to May 2017


Distribution of Laptop prices
Density
Price (logged)
Type
Import.io
Manual