Research indices using web scraped data

Robert Breton, Gareth Clews, Liz Metcalfe, Natasha Milliken, Christopher Payne, Joe Winton and Ainslie Woods

1. Introduction

The Consumer Prices Index (CPI) is produced monthly by the Office for National Statistics (ONS). The index measures the change in price of a fixed basket of approximately 700 goods and services. Prices for around 520 of these items are collected by price collectors from stores across the country. The remaining prices are collected centrally through websites, catalogues and by phone. This is often referred to as traditional price collection in index number literature.

As part of the drive towards innovation and exploiting existing sources of data we are currently investigating alternative sources of data for consumer price statistics. This was also a key recommendation in Paul Johnson's UK Consumer price statistics: A review (2015). In January 2014 we set up a Big Data Project to investigate the benefits and the challenges of using such data, and associated technologies within official statistics. The prices pilot is one of four practical pilots set up to provide us with first-hand experience in handling big data. The Big Data Project will shortly be releasing a paper on this work. The pilot uses web scraping techniques to collect prices. Web scrapers are software tools for extracting data from web pages. The growth of online retailing over recent years means that price information for many goods and services can now be found on-line.

On 8 June 2015 we published research into using web scraped data to form price indices. This paper presents the next phase of the research. The indices presented in Trial consumer price indices using web scraped data (Breton R, et. al. 2015) were based in June 2014 and ran to April 2015. Here, we extend the time series to June 2015. The previous research presented chained daily and unit price indices at daily, weekly, fortnightly and monthly frequencies. It also included a fixed base index which followed CPI methodology as closely as possible. In this paper we introduce a web scraped index compiled using the GEKS – Jevons formula, which should be more appropriate for high frequency data. Finally, in this stage of research, the unit price and web scraped CPI indices are rebased in January 2015.

In section 2 of this paper we discuss the data collection and cleaning processes. Section 3 discusses the limitations of the data and, in section 4, we present an analysis of each of the different indices. In section 5 we consider the production of a real time price index using a sample of the web scraped data, and we present conclusions and ideas for future work in sections 6 and 7 respectively. Methodologies are presented in Appendix 1, and charts for each of the web scraped items are presented in Appendix 2.

2. Data collection

2.1 Background

Alternative sources of data have the potential to greatly improve the quality of consumer price indices. One such data source is point of sale scanner data. Scanner data are datasets collected by retailers as products are scanned through the till. Scanner data offers many benefits, such as real-time expenditure weights for all transactions. Currently expenditure weights are only available at a lag of two years, and are weighted up to population totals. However, retailers have been largely unwilling to provide scanner data. Our experiences with scanner data to date are reported in Initial report on experiences with scanner data in ONS (Bird D, et. al. 2014).

Web scraping has some similar benefits to scanner data. It could provide an opportunity for us to automate some aspects of price collection. Web scraping also has potential for use in other areas, such as the collection of attribute information for the quality adjustment of technological items, also known as *hedonics*. The hedonic adjustment process is described in more detail in the CPI technical manual, Chapter 8.2 (ONS, 2014).

Price and attribute collection procedures place a heavy burden on official resources. Web scraping has the potential to offer savings in these areas. However, these savings need to be considered alongside potentially high maintenance costs (see section 2. Data collection). Web scraping also provides an opportunity to improve quality by increasing the number of price quotes feeding into the index, and to produce indices on a more frequent basis. Perhaps most importantly, it gives us the opportunity to explore big datasets and develop methodologies that are appropriate for the volume of data. These experiences will be invaluable should other sources of big data (e.g. point of sale scanner data) be introduced into consumer price statistics.

However, unlike scanner data, web scraping does not provide any expenditure information. This means that we are unable to weight indices below the Item level. Moreover, in current local price collection, expert price collectors select products that are representative of what consumers typically purchase. Web scrapers, however, will select all products indiscriminately of expenditure.

Web scraping is also limited to retailers who have an online presence. This immediately rules out several supermarket chains, such as Aldi or Lidl, who have a reasonably large share of the market (approximately 10%), but do not sell online. In this pilot, prices have been collected from three major online supermarkets: Tesco, Sainsbury and Waitrose (with approximately 50% share of the market). As such, these data are not comparable with traditionally collected prices. Further reasons why the data are not comparable are given in section 3.

2.2 Data collection

ONS's Big Data pilot for prices has developed prototype web scrapers for three online supermarket chains: Tesco, Sainsbury and Waitrose. These scrapers were programmed in Python using the scrapy module. Every day at 5.00 am the web scrapers automatically collect prices for 35 items in the CPI basket. The web scraper uses the websites own classification structure to identify suitable products. The number of products collected within each item category varies depending on the number of products stocked by each supermarket. The web scrapers collect approximately 6,500 price quotes per day (approximately 200,000 a month), which is a larger collection of prices than is gathered under the traditional approach.

For example, in current CPI collection, around 140 (70cl) prices for bottles of whisky are collected on a specific day (index day) each month. The traditional collection covers a broad range of shop types and locations. In comparison, the web scrapers collect over 6000 whisky price quotes per month across the three supermarket websites. This is equivalent to hundreds of whisky prices per day; covering a broad range of types, such as single malts or blends. By contrast, and due to the nature of the product, only seven prices for bananas are collected per day across the three supermarkets. Three key variables are collected from each of the supermarkets' websites: the product name, price and offer (discount) information. This paper does not consider the impact of discounts. The data are being used to investigate the effects of discounting in other project work.

Prices have been collected from all three supermarkets almost every day. Missing days were mainly caused by retailers making structural changes to their websites. As a result, the web scraper code needs to be altered. Missing data were also caused by internet outages or other IT system failures. To minimise the risk of this creating breaks in the time series, web scrapers are run from two ONS locations (Newport and Titchfield). It has also been used to investigate potential price differences caused by web scraping in different parts of the country, although, to date, no such differences have been discovered.

Despite the success of the web scrapers in collecting large volumes of data, there are still outstanding issues. These issues are considered below:

- The terms and conditions for Asda imply that web scraping may not be an acceptable use of the website. Further, it is thought that they use blocking technologies to prevent scraping. Therefore, prices are not scraped from this supermarket’s website.
- The Waitrose website has an infinite scrolling system, wherein a maximum of 24 items are listed on a page. If there are more than 24 items, these are loaded sequentially via user activation of the scroll bar. The current web scraper is not able to replicate this user action. This limits the data collection to the first 24 items listed under each category.
- Initially there was a problem web scraping Sainsbury involving cookies. These issues were eventually resolved.

These issues highlight the technical difficulties in web scraping as a method of data collection.

2.3 Maintenance

The web scrapers require daily monitoring. Each day an automated email report is generated. The report has counts for each item category and each supermarket. Other errors are identified through descriptive statistics, or simply by manually inspecting the data.

Since the start of this pilot more intuitive web scraping tools have been developed, such as import.io, which have a simpler interface and lower maintenance. These tools are currently being investigated.

2.4 Data Cleaning and Manipulation

Although the collection of price data through web scraping has been fairly successful, the processing, cleaning and manipulation of the data (also known as data wrangling) have presented additional challenges. Each website stores information in different ways using a variety of formats, descriptions and product classifications. The purpose of the data wrangling step is to process

these diverse raw data in a form which is suitable for analysis. Data wrangling is regarded by data scientists as taking up a significant proportion of the project time when dealing with noisy data. A recent article in the New York Times (Lohr, 2014) terms this 'Janitor work', which can take up to 50%-80% of the project time. The experiences of the Big Data Project confirm this.

Web scraped data collection is not as rigorous as manual collection. It collects high volumes of data in a timely fashion, and relatively cheaply; however, it is difficult to control the accuracy of what is collected since the web scraper will collect all the items under a supermarket's definition. For example, it will select all items under the supermarket's definition of bread. A human collector can distinguish between a white sliced loaf and a tea cake, but the automated wrangling program cannot do this without being programmed to do so. We describe this as the labelling challenge.

Furthermore, key features of interest are buried within the raw data. For example an item description could be 'Cola drink four pack of 335ml cans'. In this case the key numerical information is contained within the description. The wrangling program searches the description string to identify key words and numbers, and the nature of the number (it could be a quantity, size or volume). It is these features along with the supermarket's own definition which are used to classify or label the grocery items.

2.5 Classification

Classifying prices effectively is a key challenge highlighted by this pilot. We do not have a pre-defined map which links the CPI item label to the item's description. Table 1 below demonstrates our current method of classification, based on key word searches and exclusions. The table also demonstrates the limitations of this approach.

Table 1: The Labelling Challenge

CPI Label	Item description	Search term	Match	Correct
Apples, dessert, per kg	PINK LADY APPLES 4S	'APPLE*'	Yes	Yes
Apples, dessert, per kg	APPLE, KIWI & STRAWBERRY 160G	'APPLE*'	Yes	No

In the example above, the web scraper has collected all items under the supermarket's definition of apples (the precision of this classification varies by supermarket). The next step is to narrow the data selection to dessert apples (the CPI label). This is done via simple search and exclusion terms. In this example a simple search for the term 'APPLE*' is applied. This leads to a fruit multipack being incorrectly identified as a desert apple.

One approach to the identification of this kind of error might be to search for price outliers. Misclassified prices which are not price outliers, however, are more problematic to identify. For example, one of the supermarkets includes a bottle of Rum in its Whisky category. Identification of the error is a difficult task since the price of the bottle of rum is not noticeably different from other prices in that category. This can be resolved by creating search or exclusion terms to filter out the incorrect item. However, this can only be done for manually identified misclassifications. The unknown misclassifications present a far greater challenge.

Data wrangling within ONS is still in its infancy. Better methods are needed to classify the data. Machine learning techniques are being investigated to improve accuracy and efficiency. The techniques also have the potential to simplify programming and offer a better system that is easier to test, validate and maintain. Classification challenges do not just occur in web scraped data. For example, scanner data and consumer panel data have the same issues. Hence, knowledge gained in machine learning techniques could be beneficial for other forms of price data (and, potentially, other topic areas within the ONS).

The scraping and cleaning operations described above will be covered in more detail in the Big Data Project's forthcoming release about the pilot for web scraping price data.

3 Data limitations

Web scraping price data clearly offers a number of benefits. It has the potential to reduce collection costs, and increase the frequency and number of products that are represented in the basket. It also has the potential to deepen our understanding of price behaviour through additional and higher frequency indices.

However, the data are different (which does not imply worse) to traditionally collected price data. This means that there are a number of limitations and caveats that should be applied to the data. These prevent conclusions being drawn on supermarket's online price behaviour or national inflation rates:

- Typically, we see very high levels of product *churn* (i.e. 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. In the web scraped CPI (section 4.3), for example, bananas has an average sample size of 3 and, in the unit price index, this is even lower at 2. For the weekly strawberries unit price index (section 4.1) there were no matching products over the period, so an index was not created.
- The volume of the data makes traditional cleaning methods unworkable. Work is currently underway to develop appropriate cleaning and classification procedures. These will be covered in the Big Data Project's upcoming release, and in future updates of this research. In the data used here, however, no additional cleaning (other than the process described in section 2) has been applied (with the exception of the web scraped CPI, section 4.3). This means that there are likely to be a number of misclassifications and erroneous prices.
- Prices have only been scraped from three stores with an online presence. By contrast, in the CPI, prices are collected across a wide range of stores from across the country, accounting for any regional variations in price.
- 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[1], and low expenditure products would not normally be considered.

[1] Price collectors use the shelf space devoted to a product, product knowledge, and engage with the retailer to identify popular items that are regularly in stock.

Moreover, prices are collected daily rather than on an index day each month, as in traditional CPI collection. Whilst 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.

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[2] cannot easily be found for unmatched products, which limits the representativeness and sample sizes of some indices.

- There are a number of missing days in the data. This is caused by internet outages or, more commonly, changes in supermarket's classifications that have made it necessary for the web scrapers to be rebuilt. In the worst case, prices for the last two weeks of June 2015 have not been collected for one supermarket. Where necessary, prices have been imputed by carrying the previous period's prices forward (except in the GEKS index; see section 4.2).

4. Analysis

The web scraped data are used to construct price indices referenced to June 2014 = 100. Data before this date are not considered due to early development issues with the web scraper, which failed to collect prices on several days. Indices are calculated from the standardised price per unit (for example, the price per kilogram or price per litre). Products are not weighted by retailer in the compilation. Published CPI weights are used to produce higher level aggregates of the 35 items.

4.1 Chained daily and unit price indices

There are many ways that web scraped price data could be compiled to form a price index. The high frequency of the data creates additional challenges in this respect. Traditionally, for the CPI, prices are collected once a month on index day (or the day before, or the day after index day). Here we have collected price quotes on a daily basis.

Two basic methods for constructing price indices are presented in Figure 1. One is a chained daily index, where bilateral daily indices are chained together to form a continuous index. The second is a unit price index, where an average price is calculated for the period (weekly, fortnightly or monthly), and used to construct a direct fixed basket index. As a result of using a fixed base, only products which have a unit price in all periods can be used. This significantly reduces the number of prices used in the calculation of the index. It also means that, in the future, the index will be subject to revisions as more periods of data become available. The construction of these indices is covered in detail in Appendix 1.1 and 1.2.

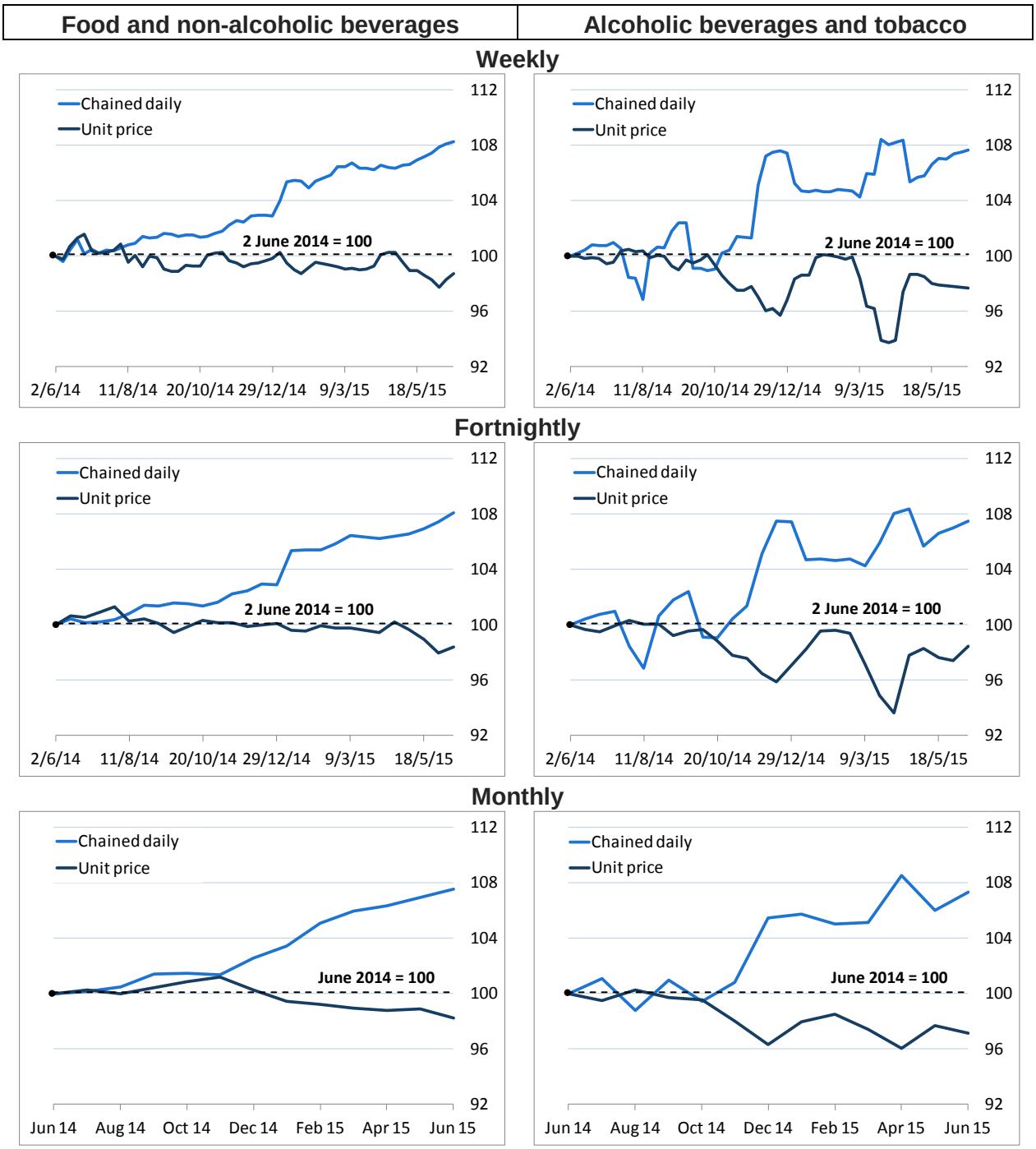
The results are striking. At all frequencies, and for both the Food and non-alcoholic beverages, and the Alcoholic beverages and tobacco COICOP divisions, the chained daily index suggests that, on average, prices are increasing, whereas the unit price index suggests that, on average, prices are decreasing. This is particularly true after November 2014, where the indices diverge greatly. By the end of June 2015 the chained daily index suggests food and non-alcoholic beverage prices have increased by 7.9%. By contrast the unit price index suggests a 1.8% (monthly unit prices) to 1.3%

[2] A product similar enough to the original that it can be considered to be the same, i.e. it will have the same base price.

(weekly unit prices) decrease. Similarly, for alcoholic beverages and tobacco, the chained daily indices increased by 7.6% by the end of June, whereas unit price indices decreased by 1.5% (with fortnightly unit prices) to 2.9% (with monthly unit prices).

Decreasing food prices are consistent with what has been observed in the CPI. The CPI entered a period of approximately 0% inflation in February 2015. Downward contributions came from falling food prices among other things. There are, of course, a number of reasons why we may not expect these indices to behave in the same way as the CPI, as discussed in section 4.3.

Figure 1: Comparison of chained daily index with unit price index



The data underlying the chained daily and unit price indices are essentially the same; however, the chained daily indices use more prices in each period. It is possible that the differences in the indices are attributable to underlying price behaviour. Prices for products which come in and out of stock frequently will be better captured in the chained daily indices. This is because products only need to be matched over two days, rather than over the whole period in the case of the fixed base unit price indices. However, this does not necessarily mean that the chained daily indices are a better measure of inflation. More work needs to be done to fully understand the divergent behaviour of these indices and the effect different data cleaning techniques will have on the results.

Each of the methods has its drawbacks. The chained daily indices allow us to make use of as much data as possible by matching only on consecutive days. This method, however, uses sub-annual chaining on a frequent basis. In the series presented above there are 393 separate chain links. One solution to this is to use the Gini, Eltetö and Köves, and Szulc (GEKS) method for chaining. This is explored in more detail in section 4.2.

This is not an issue for a fixed base index. Only products which have a unit price in every period are included in the index calculation (a unit price is calculated if there is at least one price in the period). This reduces the sample size dramatically. This effect increases as the frequency increases. This is unsurprising as a higher frequency means that products need to match in more periods. Table 2 shows the sample sizes that were attained in ONS's published unit price indices. For this article, a refreshed sample has been introduced in January 2015 to increase the sample size in line with current CPI practice. This is, essentially, rebasing the sample and, therefore, one chain link is necessary to form a continuous time series. This increases the sample sizes as shown in Table 2, in the *refreshed sample* column (note, the average of the June to January, and January to June samples is given).

Table 2: Sample sizes after matching unit prices with a June fixed base

Frequency	Previously published			Refreshed sample		
	Product sample	Matched sample	Decrease	Product sample	Mean matched sample	Decrease
Weekly	9,216	1,997	78.3%	10,900	2,277	79.1%
Fortnightly	7,811	2,100	73.1%	10,899	2,542	76.7%
Monthly	9,379	3,108	66.9%	10,801	4,272	60.4%

Note also that some limited manual cleaning was applied to the unit price index dataset. There were some cases where the classification of a product changed during the time series. In this instance, the product was investigated and assigned to the correct category.

4.2 GEKS index

A GEKS index (originally proposed by Gini, Eltetö, Köves and Szulc) is one possible solution to the issues with high frequency data discussed in section 4.1. The GEKS method essentially takes the geometric mean of all bilateral indices between the base period and the current period. GEKS indices are approximately free from chaining issues, such as chain drift[3], but make use of as much available data as possible (see Appendix A1.3 for more detail).

[3] Chain drift occurs when prices return to a previous level, but the chained index does not.

As discussed in section 3, there are dates for which data are not available. For GEKS calculations these dates are not imputed and no indices are calculated for this date. If prices for a missing date are carried forward then they will have additional influence on future indices because their values are replicated in the geometric mean. Imputation should be dealt with carefully when calculating indices from high frequency, high turnover data, in order to avoid bias. The decision was made to leave gaps at this time; future developments will include approaches to impute the values without introducing bias.

Misclassifications in the data will affect a GEKS index at all time points. This is because a GEKS index uses all historic information to calculate the current index. If, at a later date, classification issues are fixed, a rolling window GEKS (such as the Rolling Year GEKS, or 'RYGEKS' index, proposed by Diewert WE, Fox KJ, and Ivancic L, 2009) would remove erroneous classifications once the window moves past these errors. In future publications of this research, indices will be revised as classifications improve, so this should not cause ongoing problems. We cannot produce a RYGEKS index at this stage due to the short length of the price time series.

Figure 2: Comparison of daily GEKS index with chained daily index for food and non-alcoholic beverages

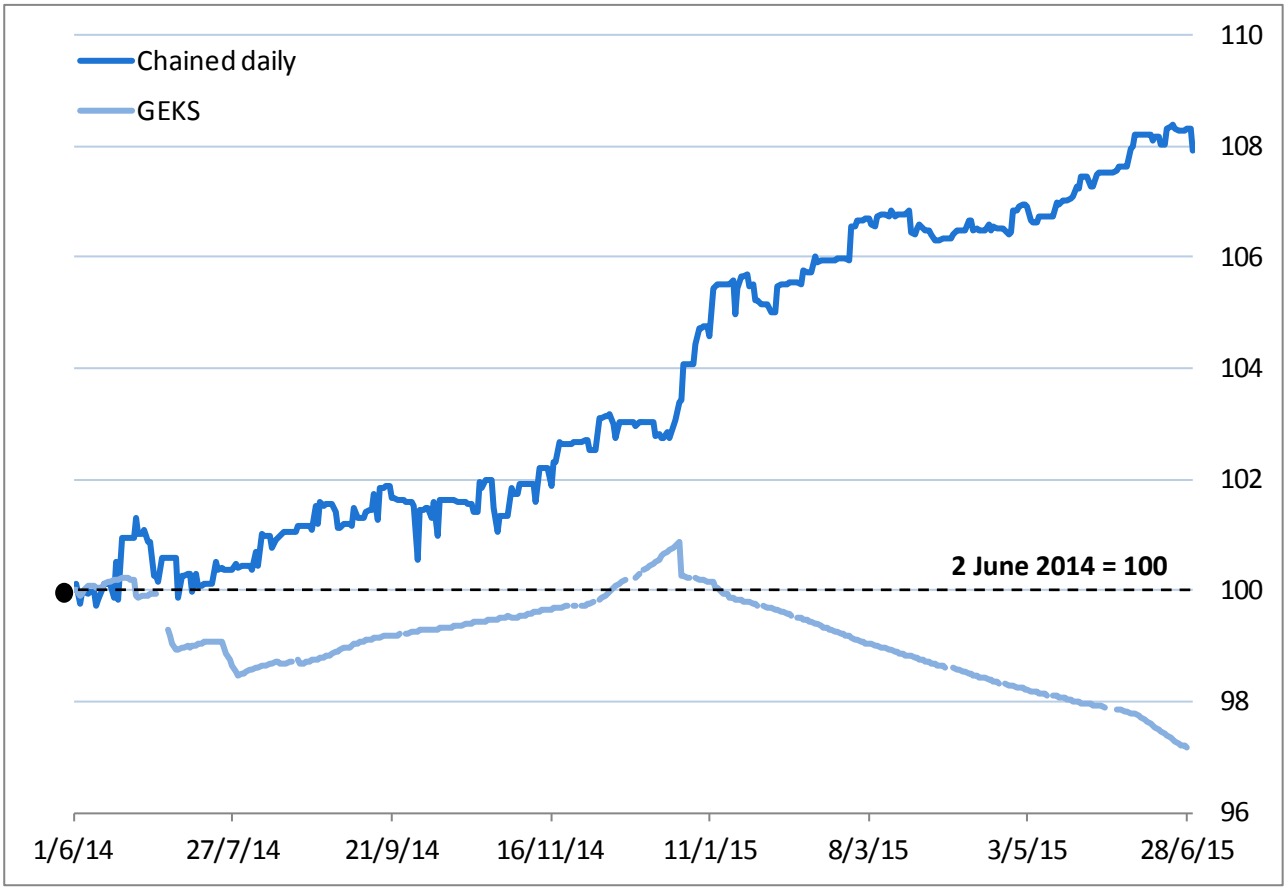
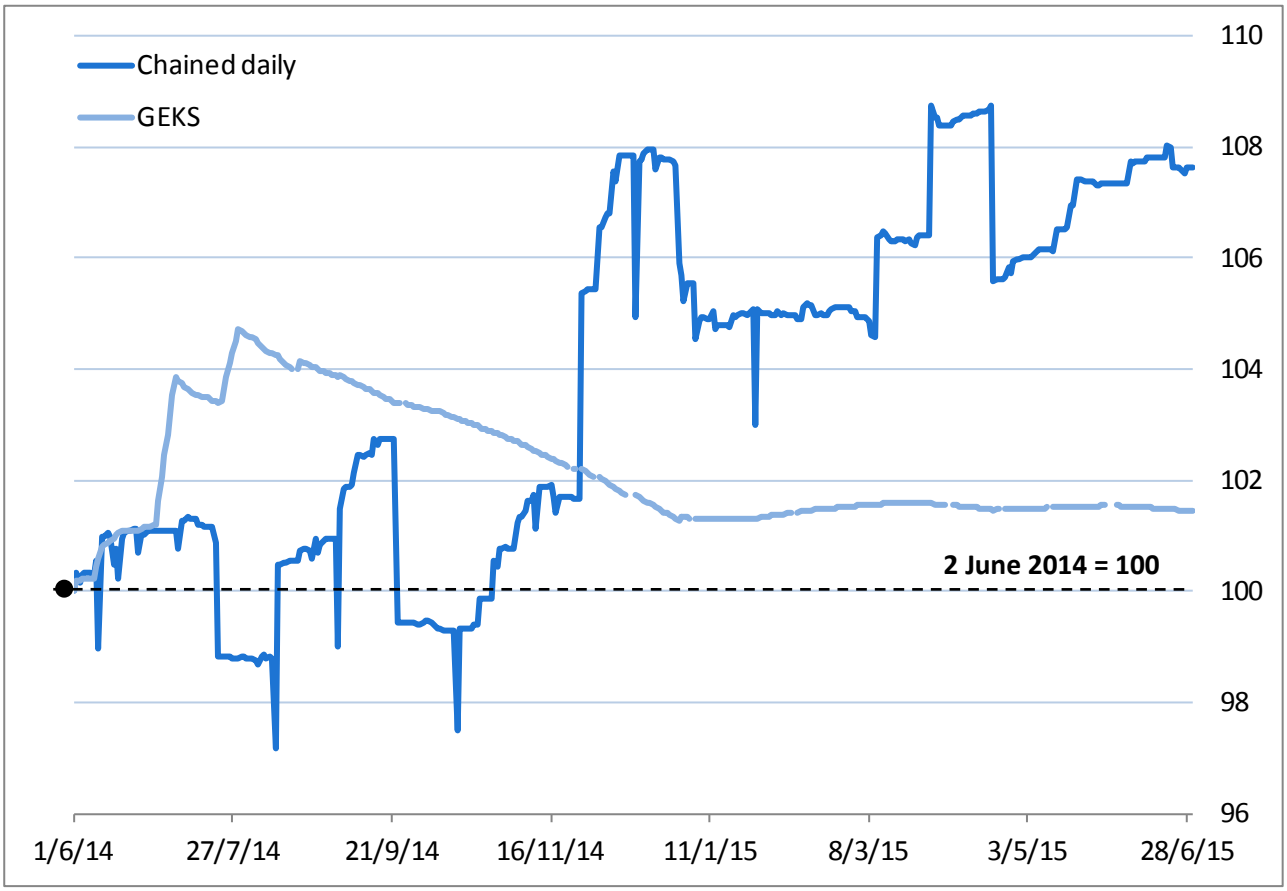


Figure 2 shows the daily GEKS index for food and alcoholic beverages plotted against the chained daily index for the same COICOP division. The GEKS index is much smoother than the chained daily index, which shows many price movements. This is unsurprising as GEKS is effectively averaging over all previous time points, which is similar to using moving averages to smooth time

series data. The GEKS index sits at a lower level than the chained daily and, from January 2015, prices are decreasing on average. This is more in keeping with what we have seen in the unit price index at monthly and higher frequencies.

Figure 3 shows the same indices for the alcoholic beverages COICOP division. Again we see a generally smoother index from the GEKS methodology. For alcoholic beverages we see the index rise over the chained daily index until December 2014, at which point it levels off. This is, perhaps, less in keeping with the unit price indices behaviour.

Figure 3: Comparison of daily GEKS index with chained daily index for alcoholic beverages



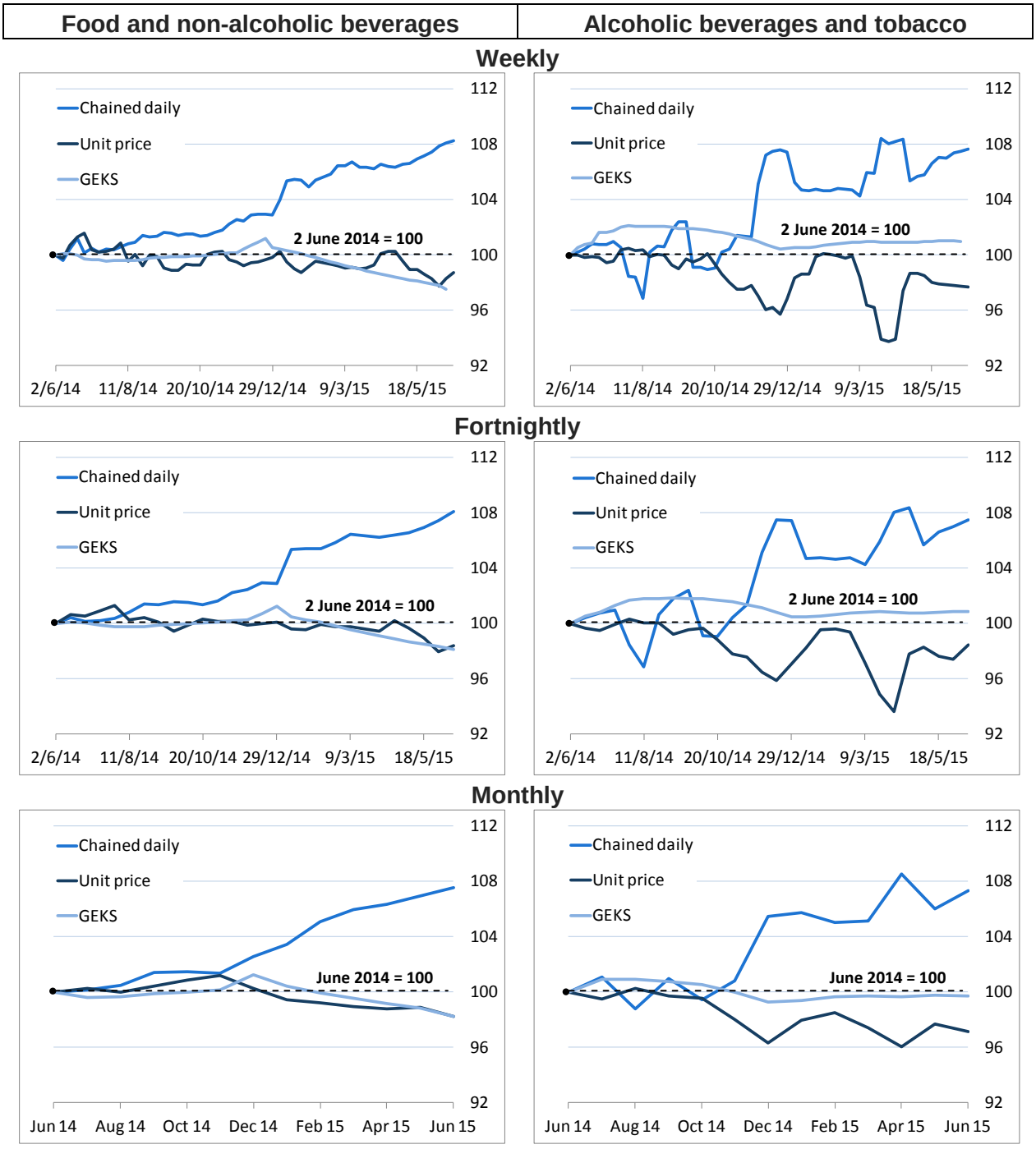
Unit price indices are also compared with the GEKS and chained daily at weekly, fortnightly and monthly frequencies (Figure 4). For food and non-alcoholic beverages we see very similar behaviour to the unit price index. For alcoholic beverages, the GEKS index is generally higher than the unit price index, in keeping with the daily indices; however, the general direction of the index appears to be similar.

The only difference between the unit price index and the GEKS index is in the sample. The differences between the unit price indices and GEKS, then, can be seen as the impact of including price movements for items which appear in some, but not all, of the periods. Due to the fixed base, this is not possible in the unit price indices.

At the lower levels behaviour is mixed. Often, the GEKS indices will track closely with the unit price index, albeit much smoother. This is notable in brandy, cheddar, cola, spaghetti, vodka, whisky and

small individual yoghurt indices. In other cases, the GEKS index lies in the middle of both the chained daily and unit price indices (for example, in apple cider, bitter, strawberries, and white slice loaf indices), lower than both indices (breakfast cereal 1 and whole milk indices), or higher than both indices (bananas and yoghurt / fromage frais indices). In one case, the GEKS index tracks the chained daily index more closely (onions). The charts for these indices are available in Appendix 2.

Figure 4: Comparison of GEKS with Chained daily and unit price indices



The variety of different behaviours in these indices makes it hard to draw any firm conclusions. Work to understand the divergence between GEKS, chained daily and unit price indices is ongoing.

4.3 Comparison with CPI

There are many reasons why it is not appropriate to draw direct comparisons between the price indices presented in sections 4.1 and 4.2, and the CPI. These reasons include differences in:

- the data collection techniques,
- where the data is collected from,
- when the data is collected,
- the magnitude of data collected, and
- the different sampling techniques.

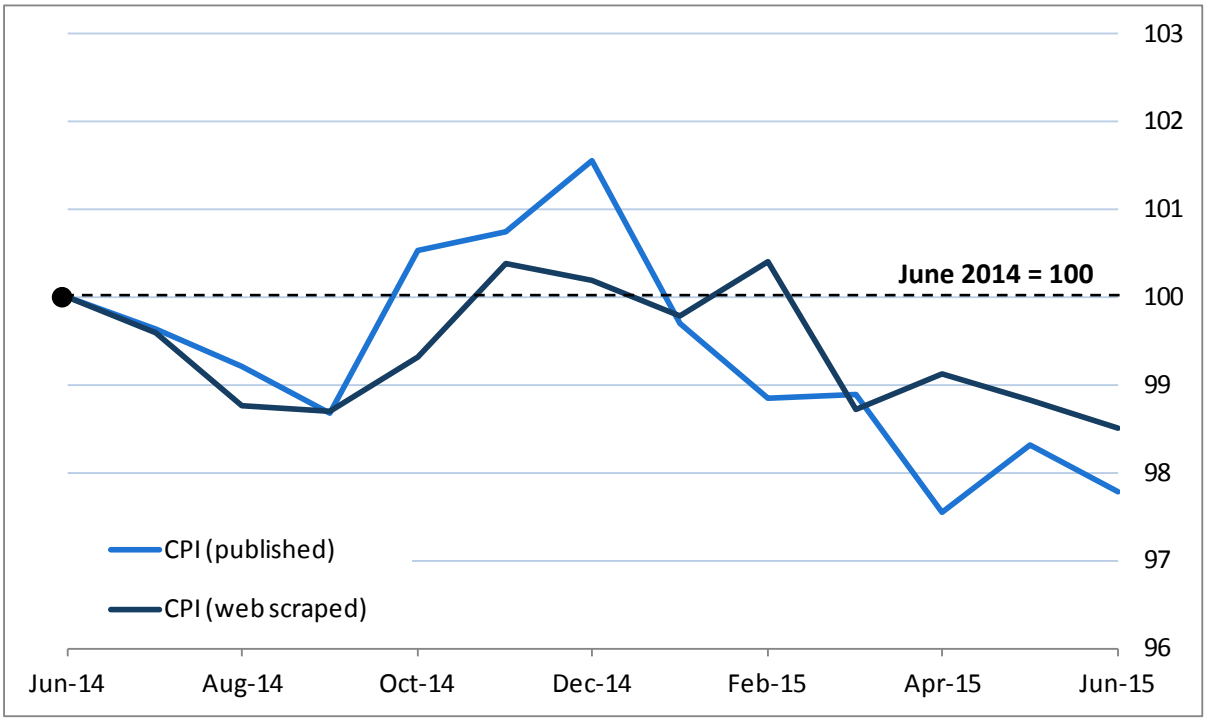
A web scraped CPI has been developed to minimise these differences and, hence, be as comparable as possible. There are, however, still limitations to the extent to which published CPI data can be compared with the web scraped CPI. The web scraped index is constructed from data collected on index day, which are matched over the period. Any unmatched products are removed, although several measures are used to increase the matched sample (see Appendix A1.4 for more details). This includes the use of comparable replacements which are found using big data techniques. The replacements are sourced from both the matched and unmatched data (so, over time, duplicate products may be introduced).

As with traditionally collected prices, a new sample of products is selected in January. In traditional CPI collection the product sample is chosen in the base month and followed through time, with comparable replacements used as substitutes when products become unavailable. This differs from the web scraped CPI, where the product sample is chosen based on the entire time series. This effectively matches products retrospectively over time and, as a result, the web scraped CPI will be subject to revisions when new data are collected. Misclassifications in the data are manually cleaned. This means that the web scraped CPI is based on a different dataset to the indices presented in sections 4.1 and 4.2.

A special aggregate of published CPI item indices is then constructed, using only the items that have been collected in the web scraping pilot. This allows us to compare the web scraped CPI with published CPI data directly. Nevertheless, despite the steps taken, we would expect the web scraped CPI and the published CPI to be different, given that many methodological differences remain.

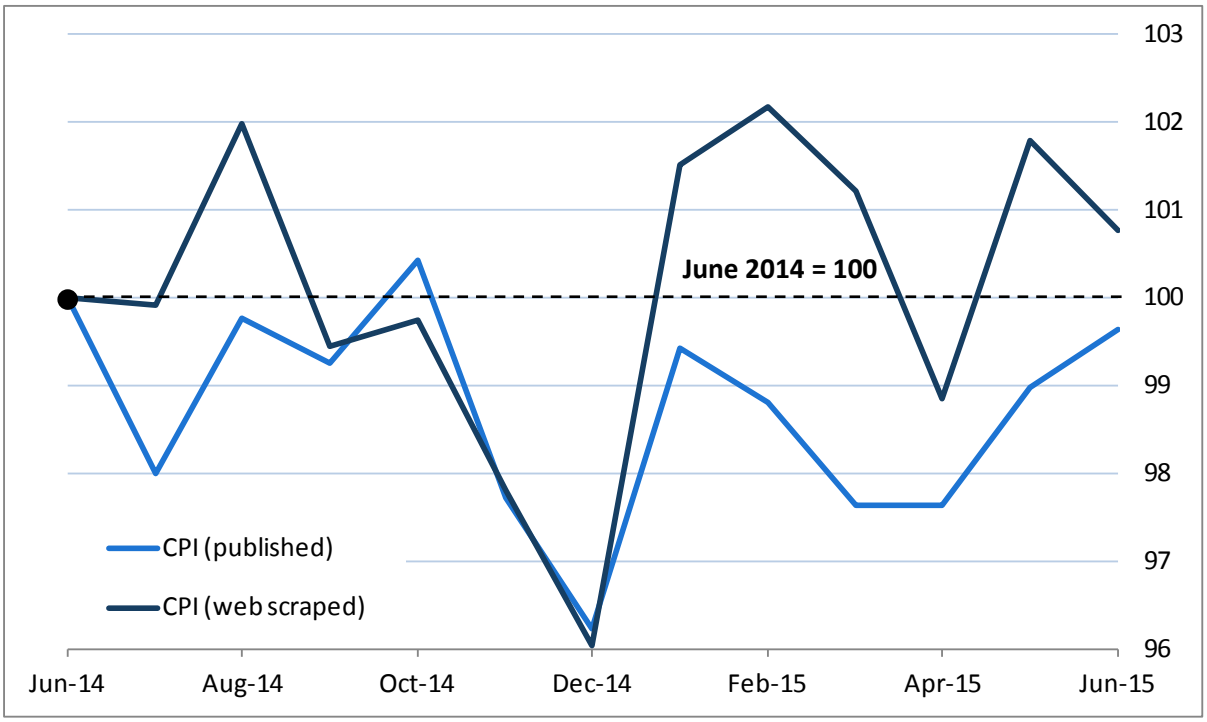
The results for the Food and non-alcoholic beverages Division, shown in Figure 5, suggest similar trends within both the published CPI and the web scraped CPI. At the item level, this is also the case for apples, cheddar cheese, onions, cola flavoured drinks, plain biscuits, and small, individual yoghurts (see Appendix 2). The average difference between the two Food and non-alcoholic drinks indices is small (with a mean of 0.6 and a standard deviation of 0.6). In both indices there is a similar drop in price between June 2014 and September 2014. After this, price movements seem to occur a month earlier in the web scraped CPI than in the published CPI. Between September 2014 and January 2015 the published CPI rises quicker than the web-scraped CPI, making greater step changes. This shows that, though the long-term story is consistent, the month to month story may not be. From January 2015 onwards a lag develops between the two indices. Price changes in the web-scraped index occur a month earlier than in the published CPI index. Shifting the web scraped CPI forward by one month would show very similar price change movements after January 2015.

Figure 5: Web scraped CPI and published CPI for food and non-alcoholic beverages



At the lower levels, the differences between the two indices are particularly large for the following items: breakfast cereals (not sugar/chocolate), bananas, dry spaghetti, fresh/chilled orange juice, grapes, new potatoes, wholemeal sliced branded loaf, tea bags, strawberries, and yoghurt/fromage frais.

Figure 6: Comparison of web scraped CPI and published CPI for alcoholic beverages



The mean average difference between the Alcoholic drinks indices is 1.5 (with a standard deviation of 1.3). This suggests that the difference between web scraped and published CPIs is, therefore, larger for the Alcoholic drinks division than for Food and non-alcoholic beverages. From Figure 6, however, we see that the Alcoholic drinks indices have very similar dynamics over time, with the two indices matching very well between September 2014 and December 2014. After this the two indices continue to show similar movements, but with greater differences in the extent to which the price changes. The main contributions to these differences are from rum, bitter, and white wine (see Appendix 2). For apple cider and red wine, the published and web scraped indices price indices are very similar.

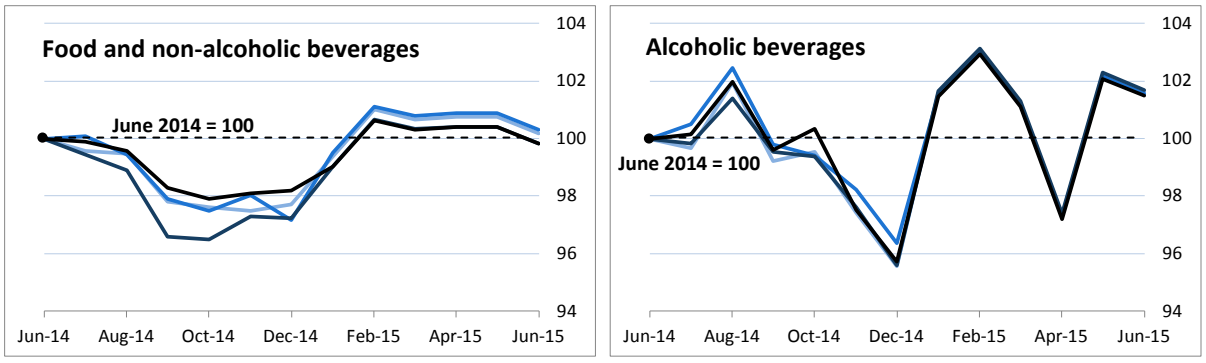
There are some surprisingly strong similarities between the web scraped CPI and the published CPI results, despite the many differences in the collection methods. There are, however, still many problems with the web scraped data and the collection process. A longer time series of results is needed to improve our understanding of the relationship between the web scraped data and the published CPI data.

5. Matching products in real time

5.1 Split sample approach

The web-scraped CPI was developed using as similar method as possible - within the restrictions of the data - to the published CPI. One of the main challenges with this approach is being able to automatically follow the same set of products from one month to the next, across a whole year. For example, in 2014 a single day's collection of plain biscuits produced, on average, 292 prices. It is unlikely that the same set of products will be available throughout the year. In the web-scraped CPI products are matched retrospectively to resolve this problem, and comparable replacements are sourced from both the matched and unmatched data.

Figure 7: Iterations of the 50% sample approach



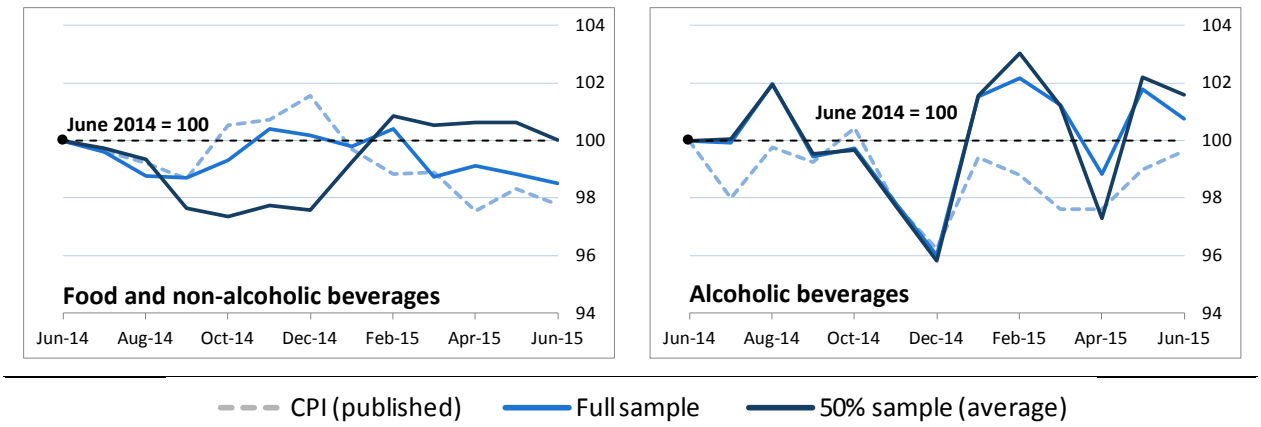
Another approach would be to split the collected prices into two equally sized random samples. One sample would then be used as the price index basket, and the other would be used to draw comparable replacements from. This is similar to the approach used to measure rental prices. Figure 7 shows the results for the split sample approach[4] for four different random samples. It shows that, when taking different samples, similar changes in price are observed over time. The magnitude of change, however, differs depending on the sample. This is particularly noticeable

[4] The collected prices are only split if there are more than four products within a category, e.g. there is only one matched strawberry price collected in 2014 and, hence, it is not possible to split the basket in half.

between September and October 2014. This emphasises the importance of gaining a representative, sample on which to base a price index.

Figure 8 below compares the split sample price index (the average of the four random samples) to the full sample web scraped CPI from section 4.3, and the published CPI. The split sample index for Food and non-alcoholic beverages shows more extreme price changes than the full sample web scraped and published CPIs. Moreover these price changes are often in the opposite direction to either the full sample web scraped CPI or the published CPI. The split sample index for Alcoholic beverages has more similarities with the full sample web scraped and published CPIs although, again, there are more extreme changes in prices.

Figure 8: Comparison between the full sample approach and the 50% sample approach



It may be that these differences stem from the reduced sample size, but without more data it is difficult to draw a firm conclusion on why this is the case. The two approaches vary only in how they follow items through time. The differences in the results, then, clearly show the difficulties that are faced when developing a methodology for use with web scraped data.

6. Conclusions

Since May 2014 we have been scraping prices from the websites of Tesco, Sainsbury and Waitrose. This represents a significant achievement in terms of developing innovative techniques to make use of alternative sources of price data. Over a period of 13 months we have collected around 6,500 price quotes daily for 35 CPI items. This gives us access to unprecedented volumes of price data.

Compiling high frequency data of this nature into price indices presents a unique set of challenges to the price statistician, which must be resolved before the data can be put to effective use. The construction of basic chained daily and unit price indices demonstrates this problem. Typically, we see prices increasing in the chained daily indices, but prices falling in the unit price indices. Clearly, the method of compilation is a very important factor in how high frequency price data should be treated! The GEKS index is one proposed solution to high frequency index number problems. Unlike the unit price index, it makes the most use of available data and, unlike the chained daily index, it does not rely on multiple chain links. GEKS indices are much smoother than chained daily

or unit price indices. Whereas, in section 4.1, we typically saw increasing chained daily and decreasing unit price indices, the behaviour of the GEKS index tends to be less predictable.

We also considered how the data could be used to make comparisons with CPI data. A web scraped CPI was developed for food and non-alcoholic beverages, and alcoholic beverages. These followed traditional CPI methodology as closely as possible to minimise differences due to the data collection method. In this index we saw a similar long-term trend to published CPI data, with some differences in price movements. A split sample approach was used to understand how the web scraped CPI might evolve in real time. Under this approach, we saw that price movements appear relatively similar between samples, whereas the magnitude of the movements can vary. When compared to the full sample methodology (and published CPI data) we see some surprising differences in both the direction and magnitude of price change.

This work contributes to a growing body of research into large alternative sources of price data, and its results are useful in developing methods for scanner data, as well as web scraped data. Despite the issues faced in producing price indices, web scraped data have the potential to deepen our understanding of price movements in the groceries sector in the medium term and, in the long term, improve the way prices are collected for national consumer price indices. There remains, of course, much work to be done in this area.

7. Future work

7.1 Machine learning

Better methods are needed to classify the data. Unsupervised and supervised machine learning techniques are being investigated to improve accuracy and efficiency. These techniques also have the potential to simplify programming and offer a better system that is easier to test, validate and maintain.

Supervised techniques require a training dataset. This is created by manually inspecting and classifying a sample of data. This training dataset is used to train classification algorithms. The trained algorithms can then be used to classify unseen data. Examples of supervised learning techniques are logistic regression, neural networks or support vector machines. These techniques can be used to systematically classify prices based on their features. For example, a standard unit feature such as litres could be used to identify an item as a drink. These features can then be used to assign the item to a particular CPI category.

Unsupervised learning techniques do not require the manual creation of a training set. Two key examples of unsupervised learning are k-means clustering and principle components analysis (PCA). They can be used to infer structure from the data. K-means clustering could be used to group prices with similar features. This grouping could be then used to form a pool of substitutes to replace an item if goes out of stock or is discontinued. PCA could be used to identify the most important features. This could then be used alongside supervised classification.

7.2 Improved cleaning

Improvements to matching techniques will also be investigated. Matching on product descriptions comes with many problems, as described in Appendix A1.4. A more appropriate approach for big datasets might be to group product descriptions into subsets that are as homogeneous as

possible. We can then match these subsets over time, rather than individual product descriptions. Again, K-means clustering is one technique that would allow us to do this.

In addition, the web scrapers are being edited to collect retailer's product codes[5]. Using these codes along with the product description should help to improve the quality of standard matching, as well as utilizing more of the available data.

A data cleaning strategy will be developed for dealing with extreme price changes, imputation and classification. All indices will be based on the same dataset, ensuring that the reasons for divergence between different indices are clear.

7.3 Additional collection

The web scrapers have been adapted to collect additional items. We will shortly begin collecting web scraped prices for all CPI food and beverages items. However, it will be some time until enough prices have been collected to form a meaningful index.

7.4 Techniques for compiling high frequency data into price indices

We will continue to explore methods for compiling high frequency indices. This includes further work with the GEKS index, possibly incorporating a rolling year window (RYGEKS), as well as other recent methods that have been developed, such as Fixed Effects Window Splice (FEWS). In addition, we have been considering how web scraped prices might be combined with representative prices chosen by expert price collectors to calculate price indices.

Finally, more research will be conducted to understand why indices compiled using chained daily, unit price, and GEKS methodologies are so different from one another.

7.5 Future publication

We will continue to explore ways to make use of the web scraped data, whether through automating aspects of price collection, or assisting in collection of attribute information for hedonic quality adjustment. We will continue to update price indices on an ad-hoc basis, as and when progress is made in any of the above areas.

8. References

Bird D, et. al. (2014): 'Initial report on experiences with scanner data in ONS', [online], [accessed 14 August 2015], available from: http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices/initial-report-on-experiences-with-scanner-data-in-ons/index.html

Beeson J (2015): 'Consumer prices index and retail prices index: updating weights, 2015', [online], [accessed 24 August 2015], available from: http://www.ons.gov.uk/ons/rel/cpi/cpi-and-rpi-index--updating-weights/2015/index.html

Breton R, et. al. (2015): 'Trial consumer price indices using web scraped data', [online], [accessed 14 August 2015], available at: http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices/experimental-consumer-price-indices-using-web-scraped-data/index.html

[5] Product codes can be changed over time. Old product codes can also be reassigned to new items. For this reason product codes are unsuitable as the sole identifier for matching.

Diewert WE, Fox KJ, Ivancic L (2009): 'Scanner Data, Time Aggregation and the Construction of Price Indexes', *Journal of Econometrics* 161 pp 24-35

Johnson P (2015): 'UK Consumer Price Statistics: A Review', [online], [accessed 14 August 2015], available at: http://www.statisticsauthority.gov.uk/reports---correspondence/current-reviews/range-of-prices-statistics.html

Krsinich F (2014): 'The FEWS index: Fixed effects with a window splice; non-revisable quality-adjusted price indexes with no characteristic information', [online], [accessed 18 August 2015], available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf

Lohr S (2014): 'For big-data scientists, "janitor work" is key hurdle to insights', [online], [accessed 14 August 2015], available at: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

ONS (2014): 'Consumer price indices – technical manual', [online], [accessed 14 August 2015], available at: http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices---technical-manual/2014/index.html

Appendix 1: Methodology

A1.1 Chained daily index

For a time series $t = \{0,1,2,\dots,T\}$, we construct daily bilateral indices. For n matched products in both periods t and $t-1$, priced at P pounds, we have:

$$B_t = 100 \cdot \left(\prod_{i=1}^{n} \frac{P_{i,t}}{P_{i,t-1}} \right)^{\frac{1}{n}},$$

for $t = \{1,2,\dots,T\}$, where $B_0 = 100$. On days where the scrapers have failed to run, prices are carried forward from the previous day. This is equivalent to imputing an index value of 100 (no change) for the missing day, t:

$$B_t = 100 \cdot \left(\prod_{i=1}^{n} \frac{P_{i,t}}{P_{i,t-1}} \right)^{\frac{1}{n}} = 100 \cdot \left(\prod_{i=1}^{n} \frac{P_{i,t-1}}{P_{i,t-1}} \right)^{\frac{1}{n}} = 100.$$

The indices are then chained together on a daily basis. For $t = \{1,2,\dots,T\}$ we have:

$$C_t = \frac{B_t \cdot C_{t-1}}{100},$$

where $C_0 = 100$. For indices of frequency F, where F is a weekly, fortnightly or monthly period of length f, we take the value of the daily index on the first day of the period.

A1.2 Unit price index

The unit price is the total expenditure on a product, i, over the period F, divided by the total quantity, Q, sold. Web scraped data do not provide expenditure information at the item level. Therefore we assume that expenditure shares are equal over time:

$$\bar{P}_{F,i} = \sum_{j=1}^{f} \frac{Q_j P_j}{\sum_{j=1}^{f} Q_j} = \sum_{j=1}^{f} \frac{Q}{fQ} P_j = \frac{1}{f} \sum_{j=1}^{f} P_j.$$

The unit price, therefore, can be thought of as simply an average price. For consistency in aggregation, we use the log of prices:

$$\bar{P}_{F,i} = \frac{1}{f} \sum_{j=1}^{f} \ln(P_j) \Rightarrow \bar{P}_{F,i} = \left(\prod_{j=1}^{f} P_j \right)^{\frac{1}{f}},$$

which is the geometric mean. We can then construct a fixed base index using the Jevons formula. To increase the sample size, the sample is refreshed in a secondary base period. Essentially, this means creating two indices: one index is based in period 0, the other is based in the secondary base period, b, where $t = \{1,2,\dots,(b-1),b,(b+1),\dots,T\}$. The indices are chained together to create a continuous index:

$$U_{F,t} = \begin{cases} 100 \cdot \left(\prod_{i=1}^{n} \frac{\bar{P}_{F,i,t}}{\bar{P}_{F,i,0}} \right)^{\frac{1}{n}} & t \leq b \\ 100 \cdot \left(\prod_{i=1}^{n} \frac{\bar{P}_{F,i,b}}{\bar{P}_{F,i,0}} \right)^{\frac{1}{n}} \cdot \left(\prod_{i=1}^{n} \frac{\bar{P}_{F,i,t}}{\bar{P}_{F,i,b}} \right)^{\frac{1}{n}} & t > b \end{cases}$$

where $U_{F,0} = 100$. For $t \leq b$ the base period for the monthly index, is June 2014. For weekly and fortnightly indices the base period is the week commencing 2 June 2014. For $t > b$ the secondary base period for the monthly index is January 2015. The secondary base period for the weekly and fortnightly indices is the week commencing 29 December 2014.

Note that in weeks commencing 15 June 2015 and 22 June 2015, a web scraper failure led to prices not being collected for Sainsbury. To ensure that the supermarket was represented in the matched sample from January to June 2015, prices from week commencing 8 June 2015 were imputed. This was only necessary for the weekly and fortnightly indices.

A1.3 GEKS

The GEKS index, named after its creators Gini, Eltetö, Köves and Szulc, allows spatial comparisons of properties of defined regions. Comparisons of each of output of each region with all of the others are taken, and then all of these comparisons are averaged to give an overall view of the combined area. In 2009, Diewert, Fox and Ivancic modified this approach to calculate price indices in the time domain. The idea is to combine bilateral price indices between two periods with all of the historic movements in the indices. The GEKS index at time t with base period $0, 0 < t$ is:

$$GEKS_{0,t} = \prod_{j=8}^{t} \left(J_{0,j} J_{j,t}\right)^{\frac{1}{t}}.$$

As noted previously, weighting information is not available for the web scraped price data at the lowest level of collection. In our calculations this necessitates the use of unweighted averaging of changes in price levels. We use the geometric mean of price relatives (or Jevons index number formula) to calculate each of these:

$$J_{0,t} = \prod_{i=1}^{n} \left(\frac{P_{i,t}}{P_{i,0}}\right)^{\frac{1}{n}}.$$

Compiling indices using traditional methods on high frequency data is known to cause large chain drift (Diewert, Fox and Ivancic, 2009). Multilateral indices such as the GEKS are generally free of this, except in very specific cases. In those cases where there may be tangible chain drift and churn of products is high, it is possible to adjust the GEKS methodology to be approximately free of chain drift.

A1.4 CPI (using web scraped data)

Sampling by index day

The CPI collection takes place on the second or third Tuesday of every month. To emulate this, web-scraped data is sub-setted on these dates. Before any data cleaning has taken place the sample sizes are 35,980 in 2014 and 25,040 in 2015.

Classification problems

The web scrapers collect all products under a supermarket's product classification; for example, the scraper will navigate to 'Whisky' on the supermarket's website and then collect all the products under this classification. Unfortunately, the supermarkets frequently put other products under 'Whisky', such as Rum. This is a common issue across the supermarket categories. For these series, the misclassifications were removed using the experimental automated machine learning algorithm discussed in section 7.1. The classification was then checked manually for this sample to ensure accuracy.

Basket update

The CPI basket of goods and services is updated once a year in February, and the associated weights above item level are updated once a year in January. To emulate this process the same approach is taken with the web scraped CPI.

Tracking products through time

In an attempt to emulate the CPI local collection, each specific web scraped product collected from a specific retailer is followed from one month to the next. This restricts the web scraped products used to those that have been collected every month (for the 2014 and 2015 baskets separately). If a product is considered to have missing data, then it is not used in index construction. This enables every product to be followed. It is important to note that this approach effectively retrospectively matches products across time, and it will result in revisions as more time periods are included. This

approach is not employed in the production of the CPI. The matching of products reduced the sample size from 35,980 to 8,640 in 2014, and 25,040 to 8,860 in 2015.

Missing data

In order to increase the number of products that can be followed through time, a variety of techniques are employed to handle missing data. These approximate current CPI processes for handling missing data as closely as possible. This also increases the sample size and reduces the probability of Not Missing At Random (NMAR) products introducing bias into the sample. These techniques are:

- *Product description. P*roduct descriptions collected by the web scraper are used to follow items through time. However, supermarkets make regular changes to their product descriptions, and this does not necessarily imply a change in the products quality, function or appearance. To reduce the number of times this occurs, any unnecessary words within the product description are removed; for example, the shop's name. This marginally reduced the number of unique products in 2014 and 2015; however, the sample size of matched products remained the same.
- *Collect across three days.* The Consumer Price Indices Technical Manual (ONS, 2014. p.30)] states that, "in practice, local collection for the CPI, CPIH, RPI and RPIJ is carried out on the day before and day after Index Day as well as Index Day, as it is not practically possible to collect every price in one day. If it is not possible to collect a product on index day, the CPI collection would collect on the next day" (apart from fresh fruit and vegetables, which are always collected on index day itself). For the web scraped data, when a product is not collected on index day, the product is searched for the day after Index day, and then the day before index day. This increased the final sample size (after all missing data techniques had been applied) from 8,640 to 21,280 in 2014 and 8,860 to 19,310 in 2015.
- *Comparable replacements.* ONS (2014, p.32) also states that, "if a chosen product is temporarily out of stock, no price is recorded and a T code is used. If it is out of stock for three consecutive months, the collector should choose a replacement product which matches the item description". The replacement product chosen by the CPI price collectors are as similar as possible to the original item chosen in that store.

  A similar process using big data techniques is put in place for the web scraped data. When a product is not available on either index day, the day after index day, or the day before index day, a comparable replacement is identified within the same category of product, and from the same supermarket, from these three days. The replacement can be selected from the pool of matched or unmatched products. To choose a comparable replacement, the product description of the missing product is compared to each of the potential replacement items' product description. This comparison is conducting using fuzzy string matching (from the Fuzzywuzzy library in Python). This measures the size of the differences between strings (product descriptions). In cases when a sufficient match is not available from any of the three days around index day, no comparable replacement is made and the product is still classed as missing. This increased the final sample size (after all missing data techniques had been applied) from 21,280 to 21,610 in 2014 and from 19,310 to 19,450 in 2015.
- *Temporarily missing.* As stated above, if a product is missing from the CPI local collection, it is considered temporarily missing for three consecutive months. When this occurs the price from the previous month is used as a temporary price for this product. If there are still missing data

after each of the previously discussed methods have been applied, then this approach is used. For the web scraped data, however, only one month of prices is brought forward. This is to minimise the introduction of fabricated stability, caused by the high churn rate of web scraped products. This increased the final sample size (after all missing data techniques had been applied) from 21,607 to 25,850 in 2014, and from 19,453 to 22,755 in 2015 (a 16.4 % and 14.5% increase respectively).

- *Complete-case.* If, after applying each of the previous methods, a product is still considered to have missing data, then this product is not used. This enables every product to be followed across time and, hence, broadly comparable to CPI methodology.

Compilation

Price indices are compiled at the lower level using the geometric mean of price relatives (the Jevons formula):

$$S_t = \prod_{i=1}^{n} \left(\frac{P_{i,t}}{P_{i,0}} \right)^{\frac{1}{n}}.$$

A1.5 Aggregation

For $I_{F,k,t} \epsilon \{C_{F,k,t}, U_{F,k,t}, GEKS_{F,k,t}, S_{k,t}, CPI_{k,t}\}$, where k is the COICOP Class (or Group) belonging to COICOP Group (or Division) K, higher level aggregates are constructed by using published CPI expenditure weights, w_k:

$$A_{F,K,t} = \sum_{k=1}^{K} \frac{w_k}{\sum_{k=1}^{K} w_k} I_{F,k,t}.$$

Expenditure weights are applied to the *unchained* indices. 2014 weights are used from June 2014 to January 2015, and 2015 weights are used from January 2015 to June 2015. The resulting aggregates are then chained at COICOP Division level. In the weekly and fortnightly series for $I_{F,k,t} \epsilon \{C_{F,k,t}, U_{F,k,t}, GEKS_{F,k,t}\}$, 2015 weights are introduced in week commencing 29 December 2015. At the Division level, a single chain link is applied in January 2015. For $I_{F,k,t} \epsilon \{S_{k,t}, CPI_{k,t}\}$ a double chain link is used in January and February, as in the CPI. The single and double chain links are described in detail in the CPI technical manual (ONS, 2014).