Office for National Statistics

Compendium

New data sources in consumer price statistics: July 2019

Explaining our research and plans to use scanner data and web-scraped data as alternative data sources in our measures of consumer price inflation.

Contact:
Tanya Flower
tanya.flower@ons.gov.uk
+44 (0)1633 45 5171

Release date:
18 July 2019

Next release:
October 2019

Table of contents

1 . Main points

- Our aim is to incorporate scanner and web-scraped data into our measures of consumer price inflation from early 2023.
- Scanner data have some unique advantages when used to produce price statistics.
- Experimental analysis on web-scraped data has yielded promising results for their future inclusion in consumer price inflation measures.
- When used in combination, scanner data, web-scraped data and data from existing sources will give us an unprecedented level of insight into how prices change across the economy.

2 . Overview

In this article, we will:

- briefly explain what our alternative data sources are
- provide our first status update for scanner data and explain why these data are so valuable
- expand the experimental web-scraped analysis by increasing our time series and providing further commentary
- provide greater detail on our plans for these alternative data sources

3 . Introduction to alternative data sources

We are researching the suitability of two new alternative data sources to inform our measures of consumer price inflation. We plan to incorporate these data sources in the first quarter (Jan to Mar) of 2023. These data sources are:

- scanner data: prices automatically generated from point-of-sale customer transactions
- web-scraped data: prices automatically scraped from websites

We will look to use these data sources in conjunction with the existing (mostly manually collected) data sources. These include local collections (prices obtained from shop shelves), administrative data and some central collections that cannot be replaced by web scraping (such as telephone calls used to collect prices for local services such as hairdressing).

We are aiming to integrate these alternative data sources as part of a phased approach. In the first quarter of 2023, we plan to integrate alternative data sources for spending categories where early data acquisition is possible. These categories account for around 100 of the approximately 700 items covered by the current basket of goods and services. The categories we plan to focus on in 2023 are:

- technological goods (laptops, desktops, tablets and smartphones)
- chart-collected items (CDs, DVDs, Blu-rays and books)
- package holidays
- clothing
- rail fares
- used cars
- groceries (which will likely cover the largest retailers in this sector)

From 2024 onwards, we would then look to expand upon this list, incorporating alternative data sources for additional areas of the basket of goods and services.

In May 2019, we used these alternative data sources and published some early experimental indices and analysis on web-scraped data with a five-month time series.

4 . Scanner data: an update

The UK's largest retailers generate vast quantities of data on consumer spending at the point of sale, through scanner machines in-store or through online sales ("scanner data" will be used to refer to both data sources). These data can provide powerful information on consumer spending patterns and so are a highly valuable source of information for producing consumer price statistics. We are in ongoing discussions with several of the largest UK retailers to make use of the wealth of information that these data sources offer.

Following a period of research, provided the data are suitable, we are planning to produce our first set of experimental results using the scanner data in December 2019.

Crucially, no customer information will be collected in these data feeds. This ensures customers will not be identifiable in these data sources. Instead, these data will solely provide information on the prices and total sales of products.

Scanner data are considered an ideal data source for measuring consumer price statistics because of a unique set of advantages over existing collection methods and web-scraping alternatives.

Scanner data provide greater market coverage

It is estimated that scanner data will provide hundreds of millions of rows of consumer spending data per year. Just a single large retailer can provide an annual dataset with hundreds of millions of rows of data. Such a scale of data has never been used in UK consumer price statistics before. This is achievable as scanner data are not as limited by ongoing data collection costs.

The sample size improvements brought about by scanner data carries numerous potential benefits, such as:

- improved precision for low-level indices
- the potential to use new methods that require more data than is available using existing data sources
- the potential to produce regional analysis by using store location information
- improved market coverage, allowing us to expand the definitions of an item (for example, potentially expanding the definition of a banana to include plantains)
- long-term, we may even be able to expand on the types of items covered by our basket (for example, including home gym equipment)

Scanner data allow us to calculate an exact product average price

Scanner data provide us with the expenditure (that is, total spending) and number of sales for each individual product sold by the retailer over a period (generally a month). Dividing total expenditure by total sales gives an exact monthly average price.

The ability to create this exact average figure across the month is unique to scanner data. In existing collections, most prices are collected once per month on a specific day. Price changes within a month will not be accounted for. We can improve our estimation of the product average price by collecting prices weekly and taking an average. This is done in current collections for more volatile areas of the basket (such as motor fuels) and is the standard for indices produced using web-scraped data.

Even so, this does not account for more regular price changes. Nor does it account for instances where spending is spread unevenly over the month and a weighted average is required (for example, in November, more spending may be expected on Black Friday compared with the rest of the month).

In the past, we have only had access to prices and not expenditure or number of sales. We can use these new figures to understand and account for how consumers are reacting to price changes.

This allows us to produce more accurate representative prices for our products, allowing us to better capture changes in market dynamics such as price and sales movements.

Scanner data can provide comprehensive information on the number of each item purchased

When calculating expenditure shares required to produce consumer price statistics, we currently rely on survey data to estimate the relative importance of one item category (for example, laptops) against another (for example, desktops). Scanner data give us another data source to make these types of comparison.

More importantly, scanner data provide a data source to weight at product level (for example, laptop A compared with laptop B). This is a first for price statistics; no other data source can provide this level of detail. This will allow us to see how consumers react in their spending after price changes.

For item categories where we have scanner data from more than one retailer, we can also accurately compare how much consumers are spending at each retailer for each item group. We can supplement this with market research to account for retailers where we do not have scanner data.

It is important to note that we will not publish individual retailer expenditure shares or any other data of a commercially sensitive nature. These will instead be used to produce more precise aggregate figures that will then be published at this higher level.

The improvements in weighting will allow us to better understand where consumers are spending their money: at which retailers, on which types of items and on which products themselves.

Scanner data may provide us with a back series

Retailers may be able to provide a substantial amount of scanner datasets for years and months in the past. We can then conduct long-term analysis from the point at which we receive the data. We will be able to retrospectively compare indices produced using scanner data against historic consumer price statistics measures when producing our planned impact assessments in 2021.

This is not possible with web-scraped data, as we are only able to obtain data from the point at which we have created our scrapers. This means having to wait for a long time series to build up before we can conduct long-term analysis.

Scanner data cannot be our only source of data

Despite their many advantages, scanner data cannot be the only data source for our consumer price inflation figures. The main challenge comes from data availability. Many small and medium-sized enterprises (SMEs) would find it challenging to produce and provide regular data feeds on consumer spending patterns.

Even if they were able to, from most recent estimates there are 5.7 million SMEs (PDF, 818KB) and it would be unfeasible for us to hold discussions with enough retailers to fully represent the SME market. SMEs make up 52% of private sector turnover, so it is important to be able to represent these enterprises in our price statistics.

To provide the best possible coverage of retailers, we will need to use multiple different data sources. For example, for groceries, we may use scanner data first and foremost, use web-scraped data where scanner data are not available, and supplement with local collections to account for SMEs such as independent grocery outlets.

5 . Web-scraped data: experimental analysis

Like scanner data, web-scraped data have the potential to provide substantially more information than manually collected data. However, unlike scanner data, web-scraped data are unable to provide us with exact average prices or expenditure shares.

The main advantage that web-scraped data provides is the ability to quickly obtain data from many different retailers, provided we meet the website's terms and conditions. When used in conjunction with scanner data, the two data sources can cover a substantial portion of consumer spending.

In May 2019, we published experimental indices over a five month time series for four technological goods. Using the same methods, we can now extend this time series. It is important to note that these are experimental results and are subject to change given ongoing research into the suitability of methods applied to the data, some of which will be touched on in the analysis. For this reason, we have also not compared them with the existing published indices for these items.

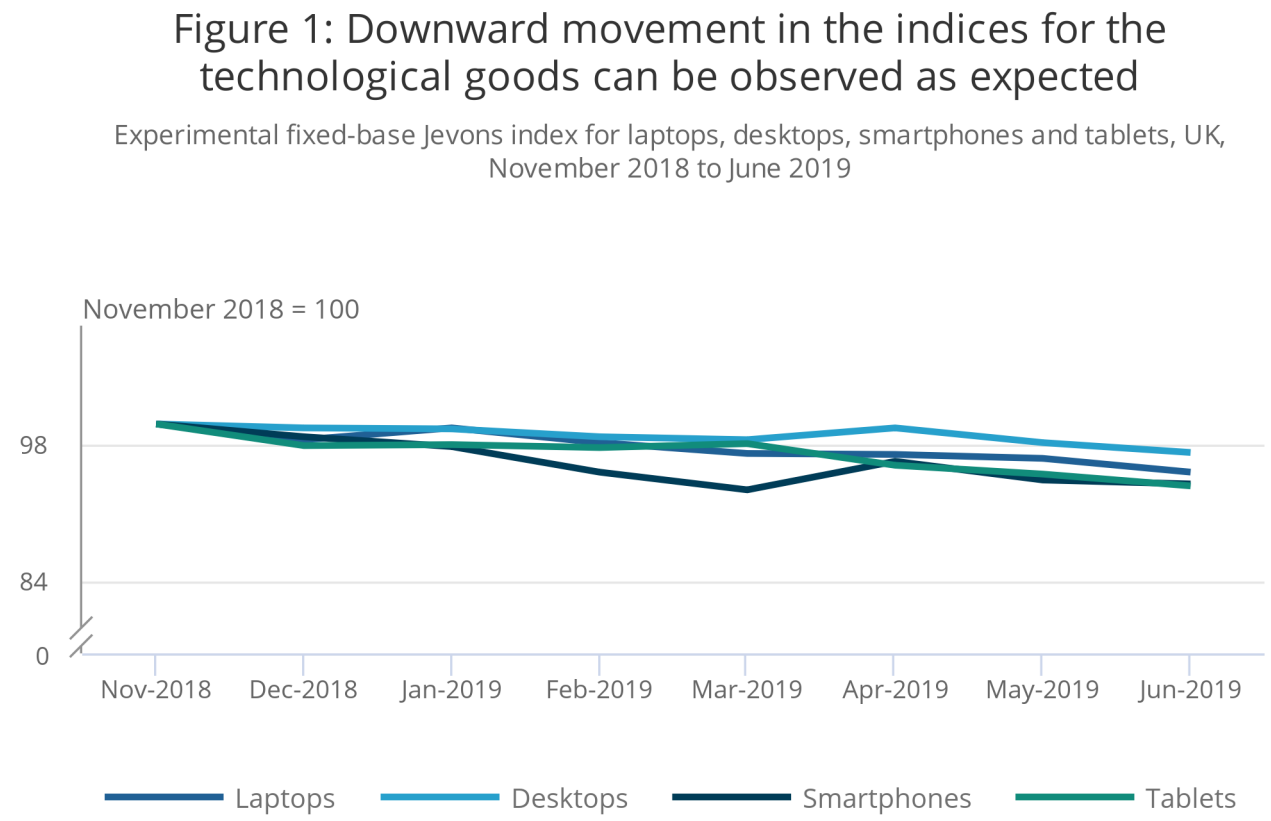
In our previous publication, we noticed some fluctuations in the indices for our technological goods (laptops, desktops, smartphones and tablets). The most noticeable phenomenon was a drop in the laptops index in February 2019. After further investigation, it appears that the performance of the indices was being reduced by a single retailer where small sample sizes and inconsistency in product coverage was causing volatility in the overall index.

We can now present updated indices for these technological goods with this retailer removed for laptops, desktops and tablets. We have retained this retailer for smartphones because of a desire to ensure sufficient retailer coverage. Notice, however, that the smartphones index appears more volatile than the other three as a result.

Figure 1 gives the experimental aggregate indices for our four technological items: laptops, desktops, smartphones and tablets.

Figure 1: Downward movement in the indices for the technological goods can be observed as expected

Experimental fixed-base Jevons index for laptops, desktops, smartphones and tablets, UK, November 2018 to June 2019



Source: Office for National Statistics, Prices web-scraped data

The index presented here is a fixed-base Jevons. This is the most commonly used unweighted index method in the Consumer Prices Index including owner occupiers' housing costs (CPIH). For each month in turn, this method compares each product's price in the current month against the same product's price in the base month (November 2018). Since technological goods become more affordable over time, we would expect that the indices slowly fall over time. Our experimental indices follow this behaviour – a promising sign for use in our consumer price statistics.

Product churn as a potential source of volatility

There are months where the indices rise in value, most notably for laptops in January 2019 and desktops in April 2019. We are currently investigating what may be causing these increases. However, a potential explanation could be because of product churn.

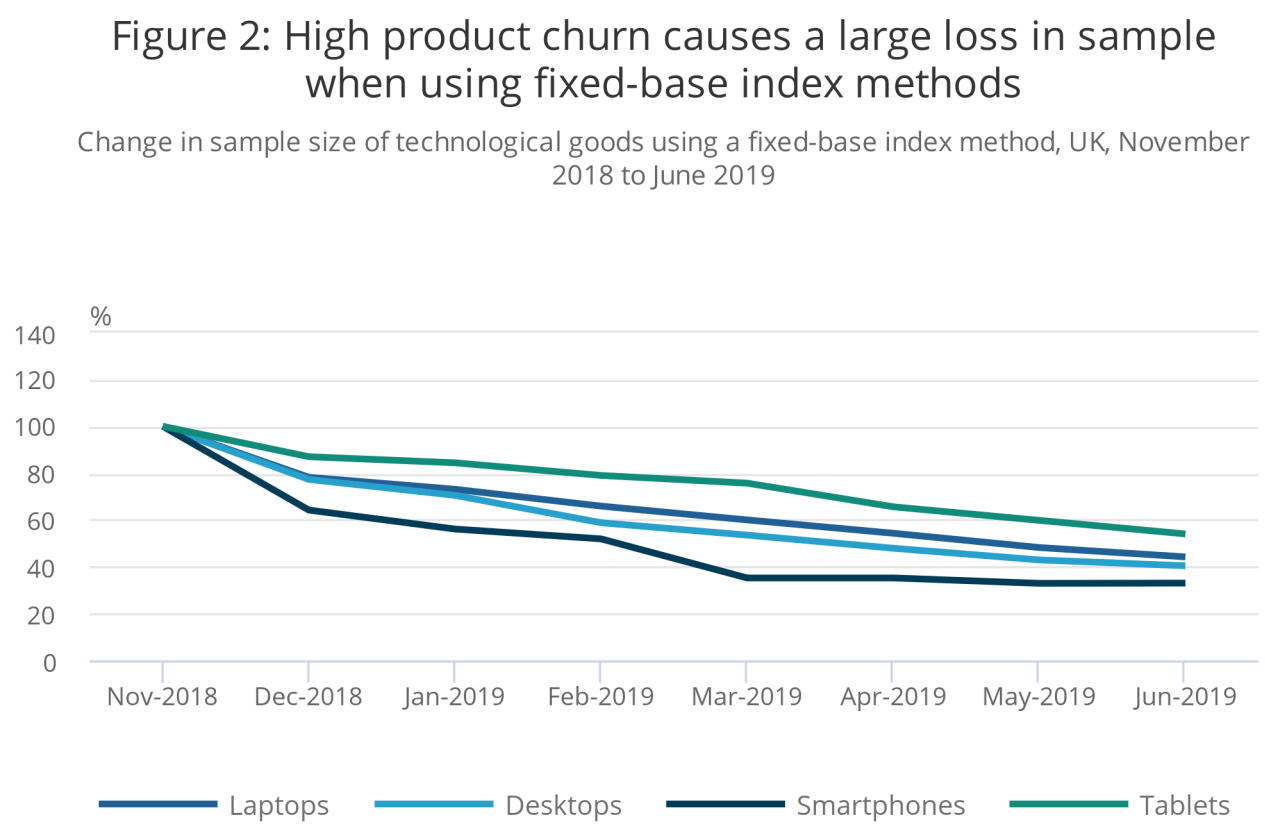
Product churn is the rate at which products enter and leave the market. Technological goods such as laptops have high levels of product churn since the market is constantly refreshing its products. By comparison, gaming consoles have low levels of product churn, as the most popular gaming consoles stay on the market for several years.

In the traditional methodology, discontinued products are manually replaced with a similar product and (if needs be) quality adjusted, allowing a stable sample to be followed over time. However, this manual approach is not feasible given the larger size of the web-scraped data. We are currently investigating methods of automatic comparable replacement but at this stage, if a product is no longer available, it is then dropped from the sample rather than a replacement being found.

This means that for a fixed-base index in this analysis, a product is only included in the sample for calculating the index for a given month if its price can be observed in both the base month (November 2018) and the month in question. As Figure 2 shows, this leads to a sample size that continuously falls when using a fixed-base method because of product churn.

Figure 2: High product churn causes a large loss in sample when using fixed-base index methods

Change in sample size of technological goods using a fixed-base index method, UK, November 2018 to June 2019



Source: Office for National Statistics, Prices web-scraped data

The high levels of product churn we have observed is what we would expect in a market where continuous improvements cause products to become technically obsolete quickly. To make full use of all the data we are collecting, we may look to use an automatic approach to finding comparable replacements, or indeed find alternatives to fixed-base index methods that allow for new products to be introduced over time rather than just when the base period updates.

We will now look at two of the new methods that may be used to process web-scraped data, and the effects that they have on our indices.

Classification

When we scrape a web retailer's laptops section, we may pick up products that are commonly sold alongside laptops. For example, this may include laptop bags, mice and keyboards. We may also pick up laptops that we do not currently include in our index, such as refurbished laptops. We use classification to sort the products we do and do not want to include in our index.

Figure 3 shows the indices produced based on whether we apply our current method of classification. In our previous publication, classification had a relatively small impact on the indices produced. In the months since, this gap has widened. This suggests that without applying classification, bias may be introduced into our indices.

Figure 3: Classification can affect the indices produced

Fixed-base Jevons index for laptops, desktops, smartphones and tablets, with and without classification, UK, November 2018 to June 2019

Data download

Source: Office for National Statistics, Prices web-scraped data

Our current method of classification is rules-based. We specify a set of keywords and if they are included in the product name, the product is removed from the sample. In the case of laptops, we search for words such as “bag” and “mouse”.

To generate the list of keywords, we take a sample of our data and manually label what each product is. Products that are not to be included in the laptops index are given reasons for exclusion. We then use a program to automatically identify the most common words used to exclude products and these can then be used as our keyword filters.

We can test the quality of this classification method. Price collectors have labelled a sample of the web-scraped data. We can compare what the price collector determines a product to be against what our classifier predicts the product to be. This produces a “confusion matrix”. Table 1 provides the confusion matrix for our first month for laptops and shows the classification method to perform relatively well.

Table 1: The confusion matrix shows our rules-based classifier performs well when there is a lot of labelled data to draw from

	Actual: laptop	**Actual: non-laptop**
Predicted: laptop	3,608	119
Predicted: non-laptop	26	1,493

Source: Office for National Statistics, Prices web-scraped data

Using Table 1, we can calculate the percentage of the time that we accurately predict laptops. This is calculated using the following formula:

$$\frac{3608}{26 + 3608} \approx 99.3\%$$

We can also calculate the percentage of the time that we accurately predict non-laptops. This is calculated as follows:

$$\frac{1493}{119 + 1493} \approx 92.6\%$$

We can take the average of these two figures to give us a metric called balanced accuracy (BA). This can be used as a measure of the quality of our classifier:

$$\mathrm{BA}_{\mathrm{Laptops,\,Month\ 1}} \approx \frac{99.3 + 92.6}{2} \approx \ 96.0$$

We can test the performance of our classifier over time. We can label a sample of our data in a second month and repeat the previous steps as follows:

$$BA_{\mathrm{Laptops,\ Month\ 2}} \approx 96.9$$

Based on these results, it appears that classification is relatively stable over time for laptops. It seems that the keywords identified in the earlier month were generalisable towards the later month.

We can now produce similar results for desktops as follows:

$$\mathrm{BA}_{\mathrm{Desktops,\,Month\ 1}} \approx 92.3$$
$$\mathrm{BA}_{\mathrm{Desktops,\,Month\ 2}} \approx 73.6$$

While the performance for month one for desktops was relatively high (albeit lower than the scores for laptops), this dropped substantially in month two. Upon further research, we found that the classifier predicted desktops well (96.5%) but did poorly at predicting non-desktops (50.7%). This suggests that the keywords generated in the first month were not generalisable enough to perform well in the second month.

We believe that the main reason for this is because we used a much larger sample to generate our keywords for laptops in month one than for desktops. This may have resulted in a more comprehensive list of keywords that were more generalisable over the months.

We are now exploring alternative methods of classification drawing on machine learning methods, which we believe may improve classification performance for more complex cases when rules-based keyword filtering may not be suitable (for example, clothing). We can use metrics such as balanced accuracy to compare the performance of different classification methods to ensure high-quality classification.

Imputation

As previously discussed, product churn can cause a loss in sample size, potentially increasing volatility.

Product churn can be:

- temporary – for example, where a retailer runs out of stock for a product
- permanent – for example, where a product is discontinued

We would not want to include discontinued products in our indices as these products can no longer be said to represent consumer spending in the market. However, there may be a case for imputing missing prices when a product has been temporarily taken off the market.

This leads to the option of imputing data (where a missing product price is filled). While there are several different methods available, we are currently exploring a fill forward method. This means a missing price is filled with the latest available price, up to a maximum of a specified number of months. An example of this is given in Table 2.

Table 2: How different imputation methods react to a product price (£30) disappearing in February before returning to the market in April at a higher price of £35

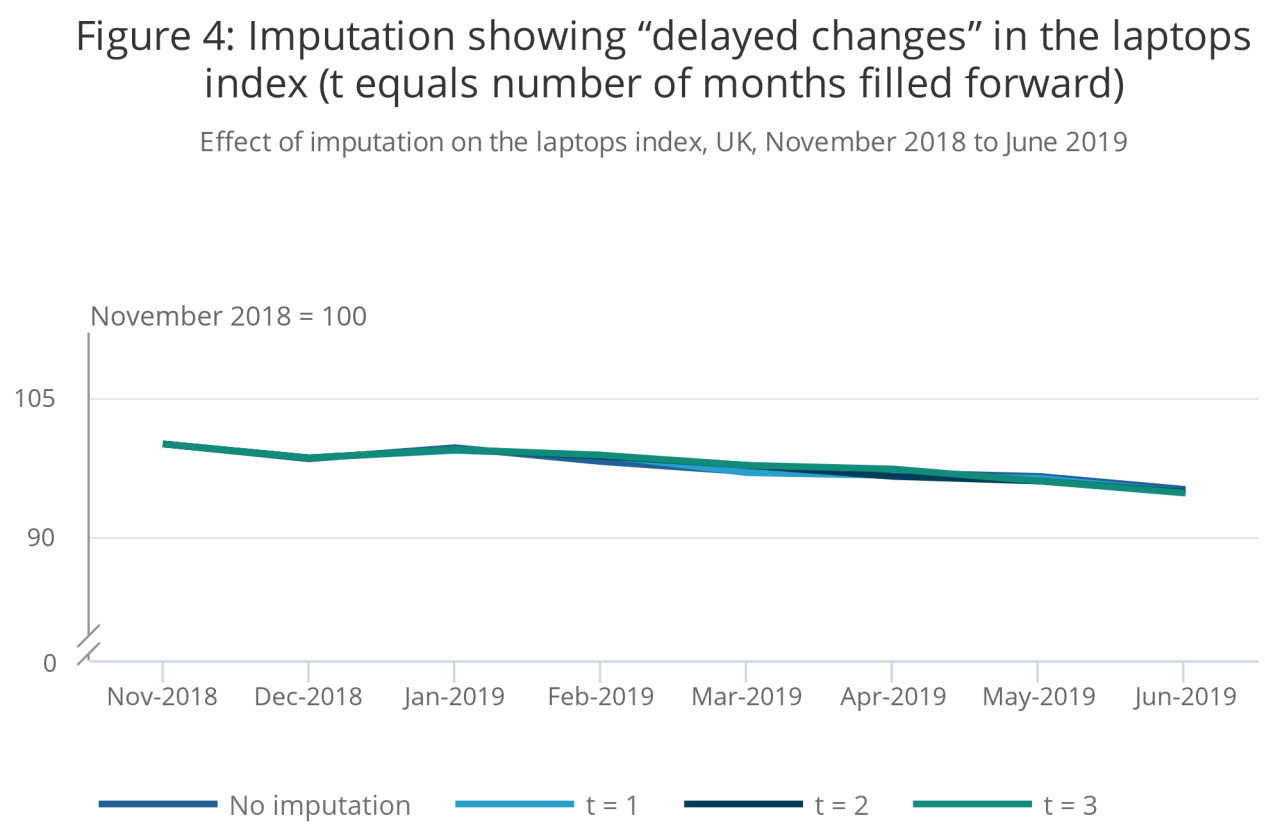
Imputation method	January	February	March	April
No imputation	30	Missing	Missing	35
One month fill forward	30	30	Missing	35
Two month fill forward	30	30	30	35

Source: Office for National Statistics, data for illustrative purposes

In the results discussed so far, we have not used imputation. We can now explore what happens if we do use this method of imputation. If t equals the number of months from which we can fill forward, then Figure 4 shows what happens as t varies from 0 (that is, no imputation) to 3.

Figure 4: Imputation showing “delayed changes” in the laptops index (t equals number of months filled forward)

Effect of imputation on the laptops index, UK, November 2018 to June 2019



Source: Office for National Statistics, Prices web-scraped data

Imputation generally does not have a large impact on the indices produced, as seen in Figure 4. However, it can cause interesting effects. Where rises and falls in the index are because of product churn, imputation can delay the change in the index.

For example, consider the laptops index shown in Figure 4. Without imputation, a fall in the index is observed between January 2019 and February 2019. It appears that this drop is because of product churn as imputation reduces this drop substantially, suggesting the dropped products accounted for the rest of the movement in the index.

However, when imputation is used, this drop is not prevented entirely, but rather delayed. For example, when t equals 1, the drop is displaced to March; when t equals 2, the drop is displaced to April; and when t equals 3, the drop is displaced to May. The reason for this is that these are the months from which the imputation methods stop being able to pull forward values from January that went missing in February.

If imputation primarily imputed temporarily missing data, then such drops in the index would have been prevented rather than delayed. This suggests that imputation is mostly imputing discontinued rather than out-of-stock products. This means that imputation may not be preferable for the technological goods.

Which index method should be used?

Up until now, all indices presented in this article have used a fixed-base Jevons index. This is the most commonly used unweighted index method in the Consumer Prices Index including owner occupiers' housing costs (CPIH). However, there are many different index methods that could be applied to alternative data sources, such as chained versions of the fixed based indices and new multilateral methods that use prices from more than two periods (for example, GEKS-J). These alternatives are shown in Figure 5.

For more information on these index methods, please see ONS methodology working paper series number 12 – a comparison of index number methodology used on UK web-scraped price data.

Figure 5: Differences are emerging in the values of different index methods

Alternative methods for calculating price indices, UK, November 2018 to June 2019

Data download

Source: Office for National Statistics, Prices web-scraped data

In our previous publication, the limited time series available meant that the choice of index did not seem to have a large impact on the results. In the months since, larger gaps have opened, particularly for the laptop and smartphone indices.

We have previously discussed fixed-base methods (particularly Jevons), where prices are compared in the current month against the base month. These can be contrasted to the high-frequency chaining methods (chained Jevons; chained Dutot), where prices in the current month are compared with the previous month. Chained methods have been considered as a means of getting around the product churn problem. By only requiring prices in two consecutive months, sample sizes are kept high.

Note, however, that high-frequency chained methods show the largest drops in the indices for three of the four technological goods. The price decrease is significant even though new items are being introduced into the sample over time because of the chaining. This is because we are not making any comparable replacements at the moment, so even if a comparable item comes in at a higher price to one that went out of stock, this increase in price will not get captured in the period that the new item is introduced. Instead, it will take until the second period for any price change to be captured for the new item, and in most cases for technological goods, this will be a downward movement. This means that chained indices may not be suitable for this type of item in future.

As part of our research on alternative data sources, we will be producing a framework for assessing the quality of consumer price indices produced using alternative data sources. This work will summarise the properties of a desirable index calculated using these big datasets and provide recommendations on how a final index method (out of the many alternatives available) could be selected for our initial specified categories. The recommendations from this will feed into our final decision on which method/s to choose for implementation.

6 . Plans for alternative data sources

We will be looking to integrate alternative data sources into consumer price statistics in the first quarter (Jan to Mar) of 2023. In this section, we present a high-level roadmap that summarises the main phases of our implementation plan (Figure 6).

Throughout the process we will be working with our users and the Office for Statistics Regulation, and we plan to hold a formal consultation in 2022 prior to the final implementation phase. We will notify users of progress or changes to this roadmap subject to further exploration of the data and methods.

Figure 6: Our aim is to use alternative data sources in the production of consumer price statistics by Quarter 1 (Jan to Mar) 2023

Timeline of integrating alternative data sources into consumer price statistics

Source: Office for National Statistics

Notes:

1. Our aim is to use alternative data sources in the production of consumer price statistics by Quarter 1 (Jan to Mar) 2023.

2019

In May 2019 we released our first set of experimental results using web-scraped indices, which have now been expanded upon in this Economic review.

We are expecting imminent scanner data feeds (see Scanner data: an update section) and will now focus on ensuring that these data feeds can be used for producing price statistics. Provided they are suitable, we intend to produce our first set of experimental indices from scanner data in December 2019.

To produce indices from these new alternative data sources, a new processing pipeline is being developed in a secure in-house virtual environment (the Data Access Platform), allowing us to process vast quantities of data. 2019 will see prototype builds of the system being produced, allowing the experimental results to be produced. For more information on the Data Access Platform, please see the Systems section of our May 2019 release.

2020 and 2021

Several stages of data processing are required to transform price quotes (and quantities where available) into price indices. Examples include needing to:

- classify products to match the Office for National Statistics (ONS) basket of goods and services
- choose an index method for a product group
- determine how to weight retailers to account for product market share

Ongoing research is exploring how we choose and then parameterise these methods that underpin the construction of price indices (for more information, please see the Further research section in Using alternative data sources in consumer price indices: May 2019).

The first half of 2020 will see the release of several methodological articles on the main areas of research. This puts us in a position to begin making recommendations to our advisory panels on the methods and parameters that we should use for constructing price indices for our initial specified categories. Upon agreement with our advisory panels, we will then apply these methods to publish initial impact assessments for these categories over the period to mid-2021.

Subject to having a sufficient time series in place, these assessments will allow us to understand how using these alternative data sources for these specific categories will affect our headline measures of consumer prices over time.

During and following the publication of the research reports, the statistical methods required for processing prices data will continue to be built into the processing pipeline, ensuring the system contains all the methods needed to produce our indices. Systems development will finish in December 2021, resulting in a final system for producing prices data.

2022

In 2022 we will run a parallel year where existing systems and new systems run concurrently. We will produce quarterly experimental results on the alternative data sources using established business processes, drawing comparisons with existing data sources and methods. We will run a consultation open to all users for users to provide feedback.

2023 and beyond

Our goal is to see alternative data sources used in consumer prices for the first time from Quarter 1 (Jan to Mar) 2023 as part of a phased approach. In 2023, we will focus on integrating alternative data sources for the categories listed in the background section.

2024 and beyond will see alternative data sources used in additional areas of the basket. Longer-term, we may also be able to use alternative data sources to expand the basket to cover new items that are not currently included in the existing consumer basket.

7 . Conclusion

Web-scraped and scanner data will allow us to collect more data than ever before, allowing us to improve the precision of our statistics, especially at more granular levels. This may also allow us to adopt methods and report on levels that were previously not possible because of limitations in sample size.

Scanner data have some useful unique properties that can allow us to address some of the limitations that current data sources have. For example, we can:

- calculate exact product price averages, accounting for change in price throughout the month
- understand how price changes (including price increases and sales) affect the number of sales and expenditure
- improve our understanding of where consumers are spending their money (on what products, on which types of item, and at which retailers)

The experimental web-scraped analysis we have conducted so far shows promising results. Technological goods generally lose value over time and this effect is reflected in a downward movement in the indices that we have produced.

The main challenge observed in our web-scraped data is in instances where product churn is high. When using fixed-base index methods, this can reduce sample size over time and increase volatility. When using chained methods, this can cause chain drift. We are exploring other index methods that we hope will allow us to account for this problem.

Our plan is to incorporate these alternative data sources in consumer price statistics by the first quarter (Jan to Mar) of 2023. Between now and then, we will conduct research, build IT systems, conduct impact assessments and engage with stakeholders. This will allow us to be transparent with any changes we intend to make and how they may affect our consumer price statistics.

8 . Author

Liam Greenhough, Office for National Statistics