Office for National Statistics

Article

Using alternative data sources in consumer price indices: May 2019

Progress being made on systems and methods to test the feasibility of using alternative data sources in the production of consumer price indices.

Contact:
Tanya Flower
tanya.flower@ons.gov.uk

Release date:
9 May 2019

Next release:
To be announced

Table of contents

1 . Introduction

Alternative data sources such as web scraped and point of sale scanner price datasets are becoming more commonly available, providing large sources of price data from which measures of consumer price inflation can be calculated. The Office for National Statistics (ONS) has ambitious plans to implement these new sources of data in the production of our aggregate measures of consumer price statistics by January 2023.

The implementation plan is based on a staged approach to incorporating alternative data sources for different areas of the Classification of Individual Consumption According to Purpose (COICOP) basket. This allows us to better isolate the impact of changing data sources and methods for particular categories and limits the impact for users. Changes to the first tranche of COICOP categories will be incorporated in January 2023 and then changes to further categories will be incorporated on a rolling annual basis beyond this point (for example, depending on data sources received, new categories will then be incorporated in January 2024, 2025, and so on).

We may also consider a staged approach to introducing these new data and index methods. In the first phase, we could draw a sample from these alternative data sources and apply a fixed base method as we currently do (essentially replicating the current methodology but with a new underlying data source). At the next stage of implementation, we could then move to more sophisticated methods (for example, multilateral methods), which will make better use of the greater coverage of these alternative data sources.

In the first phase of the project, we will focus on categories where we either have, or are confident in receiving, data for shortly. These include:

- technological goods (for example, laptops)
- chart-collected items (CDs, DVDs, Blu-rays and books)
- package holidays
- clothing
- rail fares
- used cars
- groceries (depending on which retailers we are able to secure scanner data from)

This article provides an overview of the IT systems and processing pipeline that have been constructed to test the feasibility of using these data in the production of consumer price indices. This includes details on the individual modules required to process the data (for example, “classification”).

We then showcase some experimental price indices produced using a short time series of web scraped prices data for technological goods (laptops, desktops, smartphones and tablets). The work will also include what the impact is of changing some of the underlying assumptions and methods behind different stages of the pipeline on the final headline index.

Finally, we discuss the next steps planned for this project, including an overview of the research streams we will need to complete that will drive further development of the processing pipeline.

2 . Data sources

There are three data streams that may be used to provide coverage of items in the future basket for consumer prices:

- web scraped data (prices collected automatically from online websites)
- scanner data (point of sale transaction data from retailers)
- data as currently collected (including local collections, admin data and some central collections that can't be replaced by web scraping, for example, phone calls to local services)

The status of the two possible alternative data sources (web scraped and scanner data) are as follows.

Web scraped data

We are now receiving an ongoing supply of web scraped prices data from mySupermarket. These data are for around 25 categories in the basket covering areas such as clothing, technological goods and package holidays (Table 1).

These are generally collected on a weekly basis, but airfares and package holidays are scraped daily. There are no back series with these data, so we will need to build up a sufficient time series before a final impact assessment can be completed. We are also exploring a web scraped data source for used cars.

Scanner data

We are continuing to engage with retailers on receiving some transaction data, targeting some of the largest retailers that we currently collect prices from. Depending on retailers, this will cover areas of the basket such as groceries and clothing.

We may also be able to receive a historical back series with scanner data, which means we will be able to do more meaningful impact analysis without having to wait for a sufficient time series to build up. By summer 2019, we also aim to have access to a database of rail fares transactions.

Table 1: List of categories covered by mySupermarket

One-item datasets	Multi-item datasets
Blu-rays	Airfares
CDs	Books*
Desktops	Carpets*
DVDs	Clothing*
Laminate*	Computer accessories*
Laptops	Jewellery*
Luggage*	Package holidays
Printers	Personal articles*
Routers	Sporting goods*
Rugs*	Sports food supplements*
Smart phones	Stationary*
Tablets	Shoes*

Source: Office for National Statistics

Notes

1. One-item datasets (such as Blu-rays) correspond with one item in the ONS classification structure, whereas multi-item datasets require further breakdown to match the classification structure. Back to table

2. In total there are about 150 representatitive items that are being collected by mySupermarket. Back to table

3. * indicates categories where a large proportion of price quotes are collected locally in the existing collection. Back to table

3 . Systems

The pipeline to produce consumer price indices using web scraped data was built on our in-house distributed system – the Data Access Platform (DAP), which is based on technology provided by the software company Cloudera. The platform is based on a very powerful cluster of computers and provides the users with many software tools to store and analyse data.

The resources of the DAP cluster have been expanding, reaching 57 nodes in total at the moment of writing. Each node is a very powerful computer with processing power delivered by 40 cores, memory size of 1 terabyte (TB) and storage capacity 28TB. All these nodes work collaboratively to process the data and deliver the final output of our pipeline. Data are distributed and stored across the cluster using the Apache Hadoop framework and Apache Spark is used to distribute the computational tasks across the cluster.

The Cloudera Data Science Workbench (CDSW) is the environment we are using inside DAP to build the prices pipeline. Within CDSW, PySpark, which is the Python interface to Spark, was chosen as the language to develop and test our code. Our data are currently stored as individual csv files on the Hadoop Data File System although with the arrival of big data, they will be migrated onto a Hive database on the platform.

Even though the web scraped data currently used are not that big to necessitate a distributed system, the strategic choice to use big data technology was made to be ready to process the big scanner data we expect to receive in the near future.

These web scraped and scanner datasets will be stored in a virtual environment called the Data Access Platform (DAP), and users at the Office for National Statistics (ONS) connect to this remote (and isolated from the internet) environment through a virtual machine, providing extra data security. Permissions to the various data sources as well as individual projects are managed very strictly in order to protect any sensitive data. This includes requiring suitable security clearances to gain access to the platform.

4 . Pipeline and methods

Once the source datasets are ingested onto the Data Access Platform (DAP), they then need to be processed and aggregated to a format that can be used by the final production platform. In practice, this means that we need a pipeline that takes the raw input data, processes it, and outputs item-level [1] indices, which are required as inputs into this final platform.

Figure 1 includes all the modules of the processing pipeline. There are three main elements.

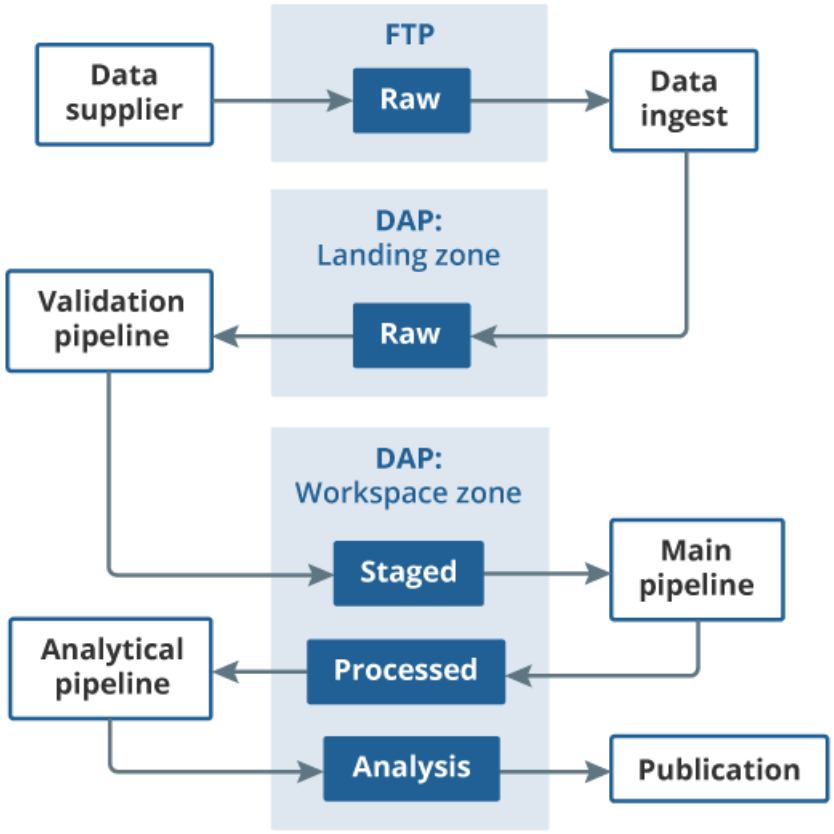
The first is the validation pipeline, which checks the data supplied by mySupermarket and flags a warning if there are any inconsistencies with previous deliveries (for example, if the total number of price quotes has dropped by more than 10%).

Once the data passes the validation checks (which are run every time new data are ingested), the data moves to the staging area. The main pipeline is then run on these data. The main pipeline includes modules such as classification and outlier detection. The output from the main pipeline is item-level indices.

The last element, the analytical pipeline, is where any additional outputs are generated (for example, the price distribution charts that are useful in understanding why some of the indices are behaving in the way that they are).

It should be noted that the final version of the pipeline will be dependent on some of the decisions we need to make in the next phase of the work. It is envisaged, however, that the underlying framework used to build this pipeline (that is, the different modules) is applicable to all items and data sources that may be used to construct consumer price statistics in future, although specific items and data sources may require some additional fine-tuning.

Figure 1: Pipeline for processing alternative data sources



Source: Office for National Statistics

Notes:

1. The File Transfer Protocol (FTP) is a standard network protocol used for the transfer of computer files between a client and server on a computer network. It is the method of transfer we currently use to receive web scraped data. Other data sources may require a different method of transfer in future, but the basic principal will be the same.

Data validation pipeline

Before the main pipeline can be run, the data must be checked to make sure the delivery is as expected from the supplier. Without it, the main pipeline may crash, or we may get spurious results.

This validation pipeline should run as soon as possible after data ingestion. It includes checks on:

- if the list of datasets and attributes are as expected
- if there are any duplicate values
- the total count of products and levels of product churn each month
- the number of null values for each column
- a quick check on price distribution to make sure there is nothing out of the ordinary happening in the data

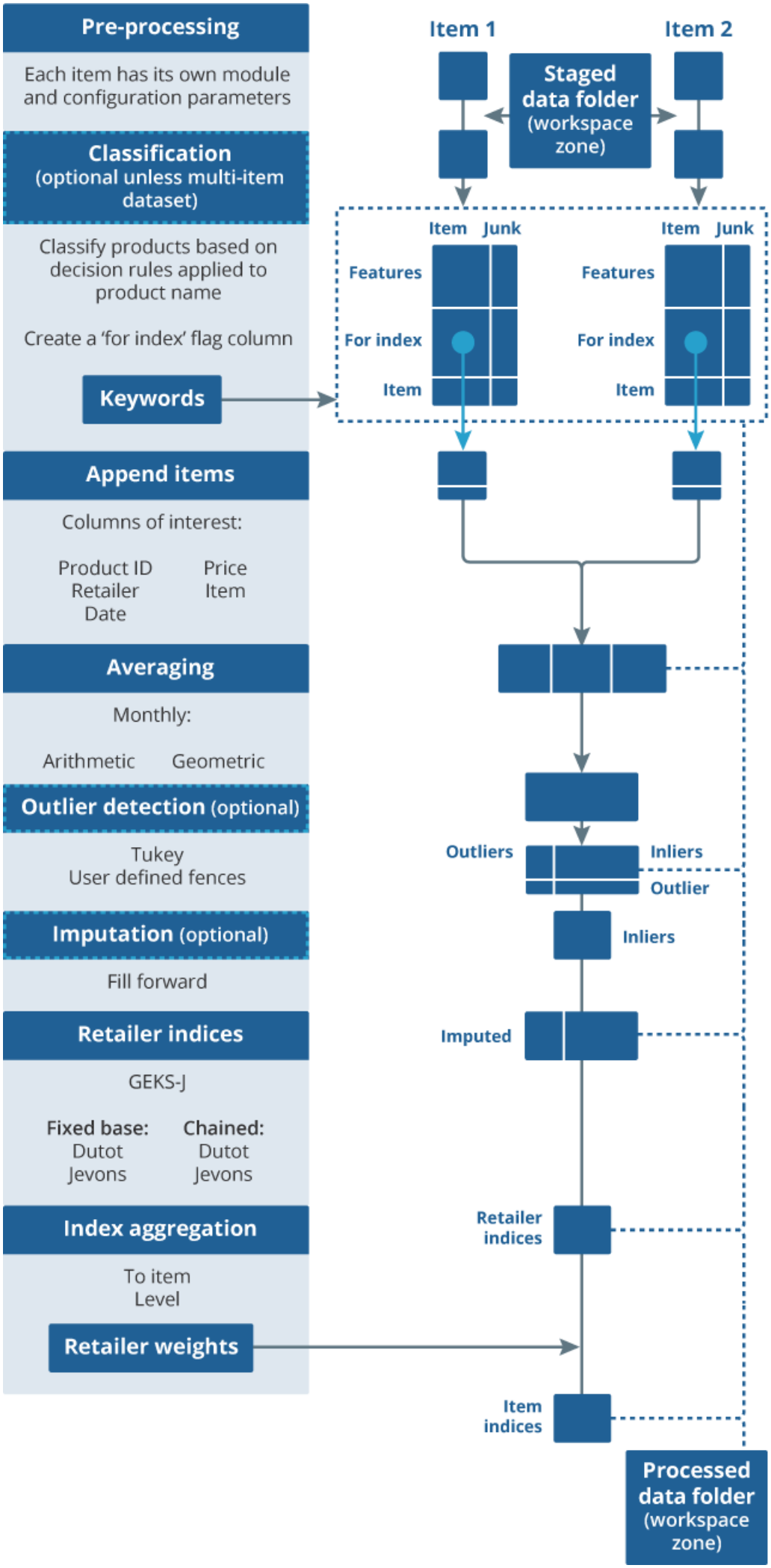
This information is flagged up to analysts via a text document, which is produced for each new months of data.

Main pipeline

The main pipeline consists of a number of modules required to calculate price indices from the validated data stored in the staging area. Figure 2 is a more detailed diagram of the stages involved in the current main processing pipeline, which we have used to output the item-level indices produced in the results section.

Additional functionality will be added into this main pipeline in the next phase of research (for example, more sophisticated classification methods). Each module is discussed in turn in this section – a more detailed discussion of each module from an earlier stage of research (PDF, 383KB) is also available.

Figure 2: Main pipeline processes item datasets into price indices


Pre-processing
Each item has its own module and configuration parameters
Classification (optional unless multi-item dataset)
Classify products based on decision rules applied to product name
Create a 'for index' flag column
Keywords
Append items
Columns of interest:
Product ID
Retailer
Date
Price
Item
Averaging
Monthly:
Arithmetic
Geometric
Outlier detection (optional)
Tukey
User defined fences
Imputation (optional)
Fill forward
Retailer indices
GEKS-J
Fixed base:
Dutot
Jevons
Chained:
Dutot
Jevons
Index aggregation
To item
Level
Retailer weights
Item 1
Item 2
Staged data folder (workspace zone)
Item
Junk
Features
For index
Item
Outliers
Inliers
Outlier
Imputed
Retailer indices
Item indices
Processed data folder (workspace zone)

Module 1: Pre-processing

Once the monthly data for each category has passed validation, it is then appended to a dataset that contains validated data from previous months. This is then used as an input into the main pipeline.

Pre-processing is then applied to specific columns in the data. Many item and data source combinations are going to have their own functions for pre-processing depending on their own unique characteristics. For example, a function that corrects the format for the attribute column “RAM size” is applicable only for certain items, for example, laptops and desktops. It is likely that these features will be used in future to help with the classification module. They can also be used in any hedonic regressions we apply to the data as part of quality adjustment.

Module 2: Classification

Classification is about ensuring that we have the right products in each individual dataset to produce an index for a specific Office for National Statistics (ONS)-defined item. In the current phase of research, these items are those that are defined in the current consumer basket although there is scope to expand item coverage in future. Currently, we are producing indices for technological goods, specifically: laptops, smartphones, desktops and tablets.

The current method used in the pipeline is a simple rules-based filter. The keywords used for the filter have been generated by looking at a small sample of labelled data and automatically identifying which words are most often associated with a product not being classified as that particular item. For example, our current item definition for laptops excludes refurbished products. The supplied dataset can also sometimes contain other products such as laptop bags and other accessories. Given these conditions, the generated keywords used for filtering for laptops make intuitive sense: refurbished, grade, box, mouse, bag, certified and headset.

While the results of this method look promising for these particular items, it will not be suitable for multi-item datasets such as clothing. An alternative method for classification (PDF, 1.61MB) using label propagation and supervised machine learning may be able to cope better with more complicated datasets.

The output from this module is an individual dataset that only contains the specific products that relate to a particular ONS item, for example, laptops. These individual item datasets are then appended to create a single dataset that is taken forward through the rest of the pipeline. This dataset contains the unique product ID, retailer, collection date, price and item category for each product; any additional product information is dropped at this stage.

Module 3: Averaging

After classification, the data are still at the observation level, that is, each row relates to a specific collection date for that product. Laptops, tablets, smartphones and desktops are all collected at a weekly frequency. Therefore, to produce a monthly price index we will need to combine these weekly prices into a single monthly price. There are two methods that are currently built into the processing pipeline: geometric averaging and arithmetic averaging.

Module 4: Outlier detection

The item datasets may contain errors in the price column, for example, because of collection issues, or prices listed incorrectly on websites. For example, we identified an issue with one of our web scrapers where it was only scraping up to the comma for prices above £1,000 that included a thousand-pound separator. Therefore, a price could have been listed as “£1,200.00”, but the price column only contained “£1”. These errors can be identified by using anomaly detection techniques and removed.

In the current pipeline, we have two methods coded up. This is the Tukey method (currently used in the traditional collection to identify outliers) and we also have an option for the price collector to input user-defined upper and lower fences. For the experimental indices produced in this article, we use these user-defined fences and remove any prices that are above or below these fences, but other methods may be more suitable subject to further testing and analysis.

Some index methods use prices as inputs, whereas others use price relatives. Therefore, there is also further work required on whether we need to do these outlier checks on prices, price relatives, or both. Price relatives are also defined over different time periods for certain index methods. A price relative is essentially a ratio of two prices. This ratio could be the ratio of current price over the base price (fixed-base Jevons), or current price over previous price from the month before (chained Jevons).

Module 5: Imputation

Missing prices can be a problem particularly for price indices constructed from alternative data sources, which may have higher rate of product churn than in the current collection. Most price indices use price relatives as an input into their series, so if a price is missing from one period because it is out of stock on the website, the imputation module allows the product to remain in the population for a period in case it comes back into stock.

Imputing prices is a commonly accepted way to deal with missing prices, ensuring that a consistent sample size is kept over all of the period of interest, but sometimes a product may go out of stock for a significant period and either get reintroduced or removed from the market altogether. Therefore, it may be unwise to continually impute the prices in either of these situations, as the index may not be representative of actual price movements.

Previous research work found that carrying forward the previous price minimised the relative imputation bias (PDF, 205KB). While this work looked at grocery data, we have used the same method in the current pipeline until alternative methods can be tested.

The choice of whether to impute or not is highly interdependent with the index methodology chosen. Indices that use a fixed basket approach and use the ratio of prices from the current month over the base month will require imputation. However, some indices use price levels, in which case, imputation is not necessarily required. Multilateral indices also may not require imputation as they take multiple periods into account and do not follow a fixed-basket approach. Additionally, some multilateral indices (for example, the ITRYGEKS, FEWS) impute automatically while implicitly or explicitly quality adjusting prices.

Module 6: Index calculations

The index number methods module takes the cleaned and classified data from previous modules in the pipeline. It uses price and expenditure data (where available) to calculate a price index at the elementary aggregate level (unpublished indices at the lowest level of aggregation below item). For these web scraped data, elementary aggregates are defined at the retailer level, so we will produce separate indices for each retailer. For scanner and locally collected data, we can add a regional dimension to these elementary aggregates that can be weighted up to the item level using suitable expenditure weights (see Aggregation section).

Many of the different existing methods that can be used to create price indices from alternative data sources (PDF, 658KB) have been well documented in the international literature over the last few years. One of the main areas of our future research is to set out a framework for how we decide what a suitable index method is and then apply this framework to decide what method we should use for the Classification of Individual Consumption According to Purpose (COICOP) categories we are looking to implement into production in 2023.

As we are still in the early stages of testing the feasibility of using these data sources in production, the current pipeline only produces a subset of possible indices that may be used in future. These are: chained and fixed base versions of Dutot and Jevons, as well as a version of the GEKS-J index.

These indices are currently unweighted, although there is scope to introduce different methods that also require expenditure weights as inputs, which will be highly desirable when we apply the pipeline to scanner data. Functionality will also be built to include different methods of splicing and chaining these index methods as required.

Module 7: Aggregation

Indices are produced at the elementary aggregate level (in this case retailers) and then weights are applied to aggregate these up to an item level. In the current pipeline, we weight all retailers equally, but this is an area for development in the future. For scanner data, we may also be able to introduce regional weights at this stage before an item index is calculated.

Once item-level indices are created for a particular data source (for example, web scraped data for laptops), it can then also be combined with indices from other data sources if available (for example, data from the local collection, or scanner data) using suitable expenditure weights, to produce a final item-level index.

Analysis pipeline

While price indices are our main focus, we will also need additional analytical outputs to help us understand and interpret these results. At every stage of the main pipeline, intermediate outputs are saved in a processing folder that we can then use for analysis purposes.

For example, we can look at the distribution of price for each item and retailer. If the price distribution in the most recent data is substantially different from the price distribution in the data from the previous month, we may expect this to have an impact on the index, which we can then explain given these results.

We can also produce analysis on product churn and summary metrics on the various processing modules. For example, we can look at the proportion of misclassified unique products (that is, products classified as not being the item of interest) out of the total number of products per month per retailer.

Notes for: Pipeline and methods

1. Item-level refers to the lowest level that ONS publishes price indices at and sits under the defined COICOP5 level set by Eurostat. More information about ONS representative items is available.

5 . Results

At this stage of the research, it is important to focus on the feasibility of our approach and test appropriate methods. The web scraped data we are using are also only available from November 2018, so any experimental price indices that can be produced are only for five months.

Nevertheless, this section showcases some experimental price indices that have been produced using the current version of the processing pipeline. It is important to note that these are still at the research stage and the indices may change in future publications due to further developments of the processing pipeline and underlying methods. For this reason, we have also not compared them with the existing published indices for these items.

Baseline scenario

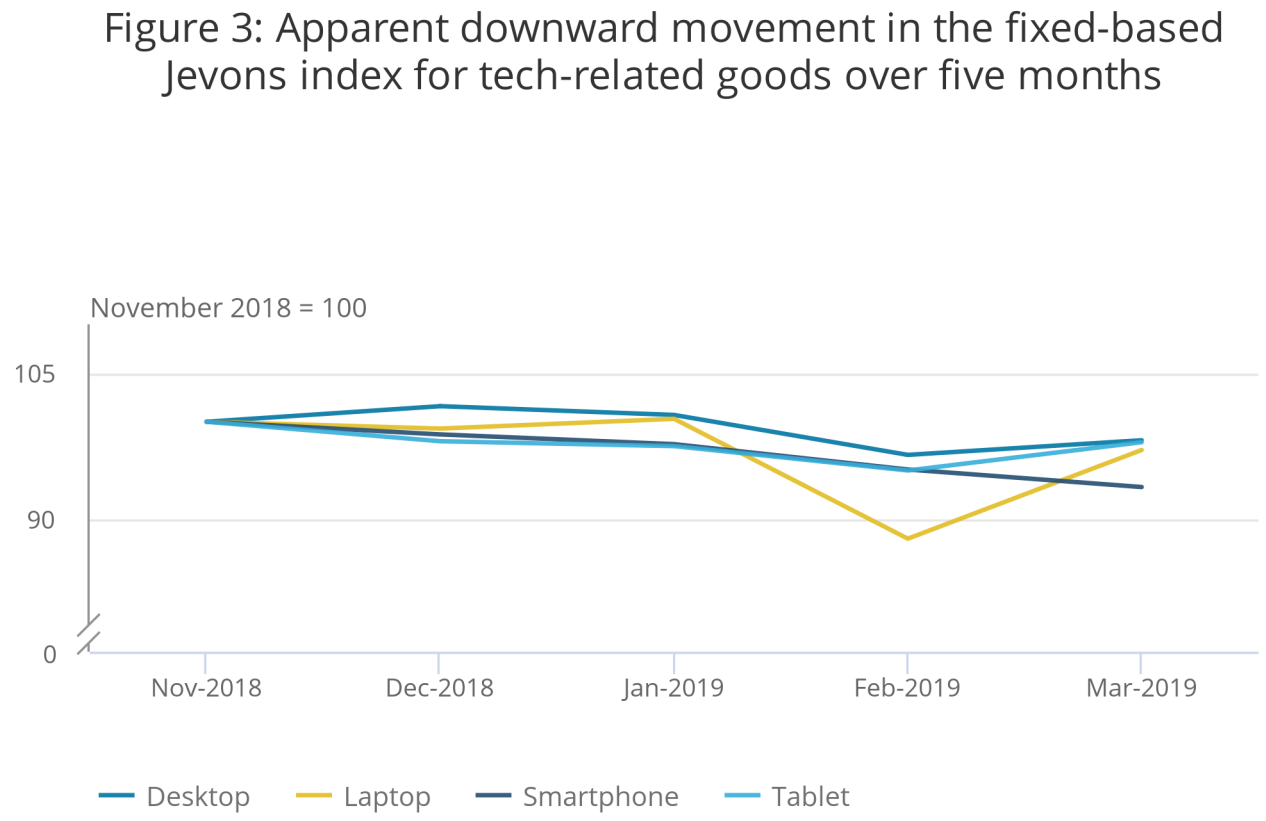
The main pipeline is run using a configuration file. This allows us to change or alter the parameters used by each of the main modules. The results in this section are produced using a baseline scenario:

- averaging: geometric mean
- classification: rules-based classifier (on list of keywords)
- outlier detection: user-defined fences
- imputation: OFF
- indices: all indices should be produced (chained and fixed-based Dutot and Jevons, GEKS-J)

The next section will test the impact of changing some of these assumptions on the aggregate indices.

Figure 3 shows a fixed-base Jevons index for laptops, smartphones, tablets and desktops. We are not applying any imputation in this baseline scenario and a Jevons index does not include any form of quality adjustment. Therefore, it is reasonable to expect some form of downward movement in prices for these technological goods, as prices tend to decrease over time and a fixed base index does not allow for new products entering the market.

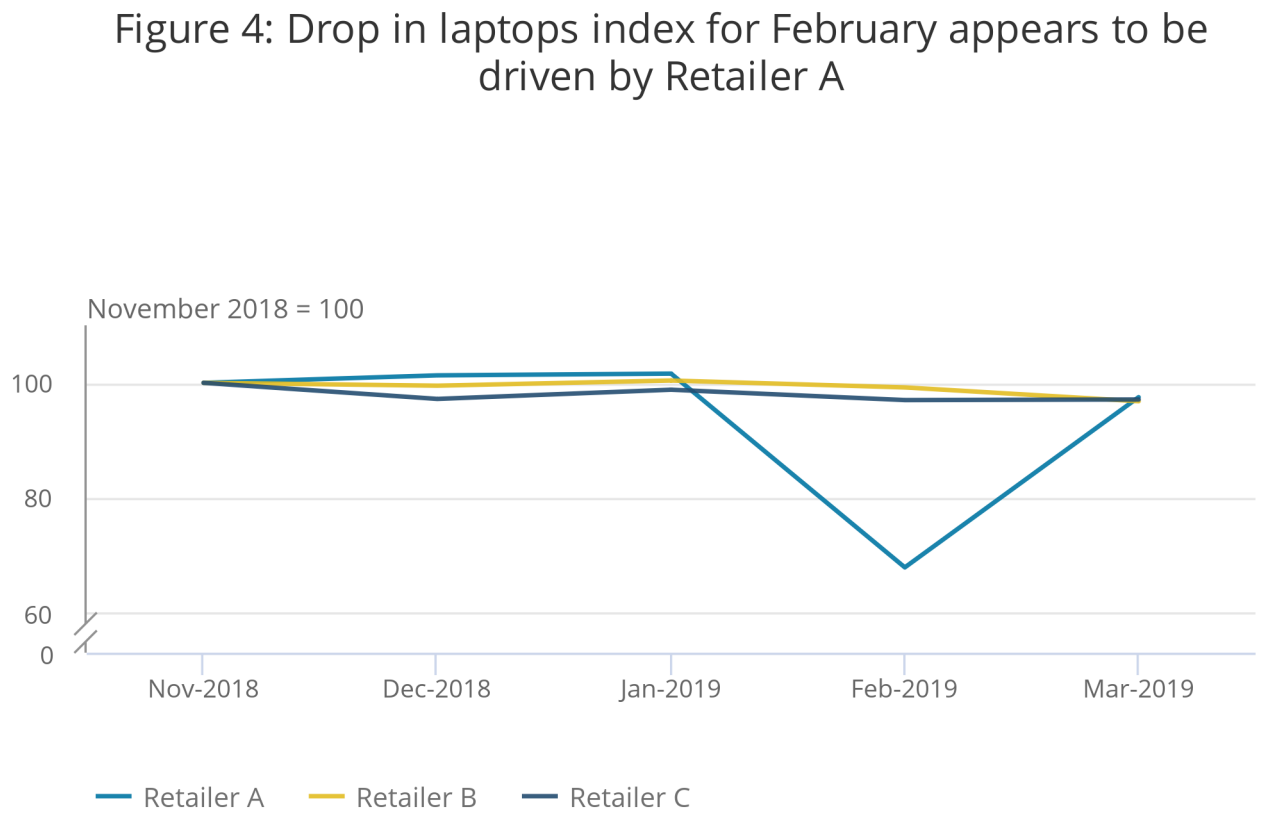
Figure 3: Apparent downward movement in the fixed-based Jevons index for tech-related goods over five months



Source: Office for National Statistics

The laptops index falls in February compared with the other indices. The output from the pipeline allows us to also look at the individual retailer indices to see if they can offer any clues for why this may have occurred. Figure 4 shows that Retailer A saw an unusually large decrease in this particular price index in February.

Figure 4: Drop in laptops index for February appears to be driven by Retailer A

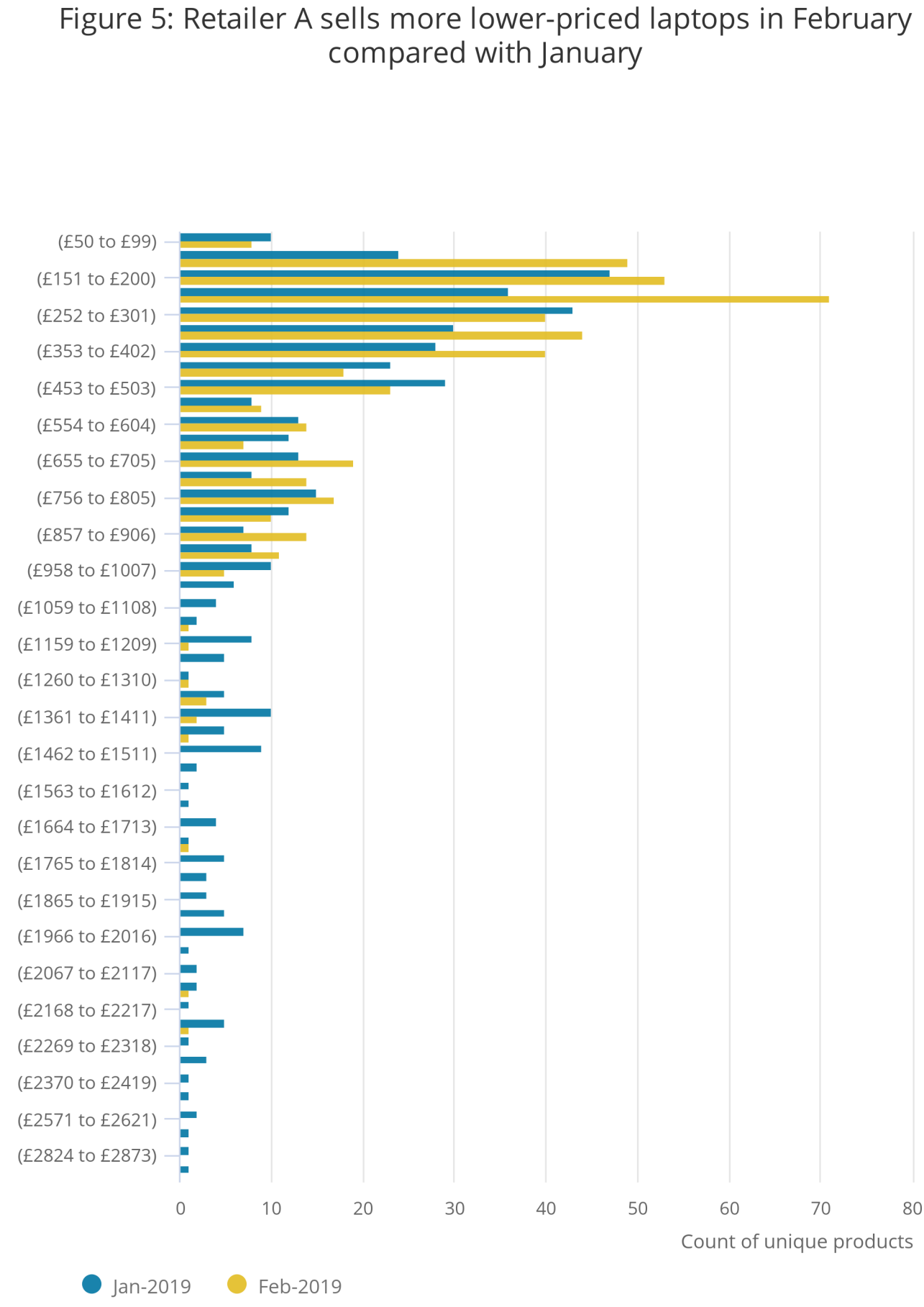


Source: Office for National Statistics

We can use our analytical pipeline to explore this further. For example, Figure 5 compares the price frequency distribution for laptops for January and February 2019 for Retailer A. It shows that February has a higher number of lower-priced laptops, which may explain this fall in prices.

In a future iteration of the analysis pipeline, it is envisaged we will also be able to plot the distribution of price relatives for each month as well as price levels. For a fixed-base Jevons index, it is these price relatives that drive changes in the index, and therefore understanding the distribution of these will help us better understand the movements in the aggregate price indices.

Figure 5: Retailer A sells more lower-priced laptops in February compared with January



Source: Office for National Statistics

Our analytical pipeline can also produce summary metrics on product churn to see if that can offer any additional insight into the indices. Table 2 shows the number of unique laptops for each month, and the number that enter or are dropped from the sample. While there is no noticeable effect on the index values, a large number of products are dropped from the sample in December 2018. This is due to the web scrapers being restricted by one of the retailers.

Future research will consider the impact of using web scraped data for the production of consumer price indices, including any possible mitigation required in case of website structural changes, or changes to the terms and conditions, which then limit our collection.

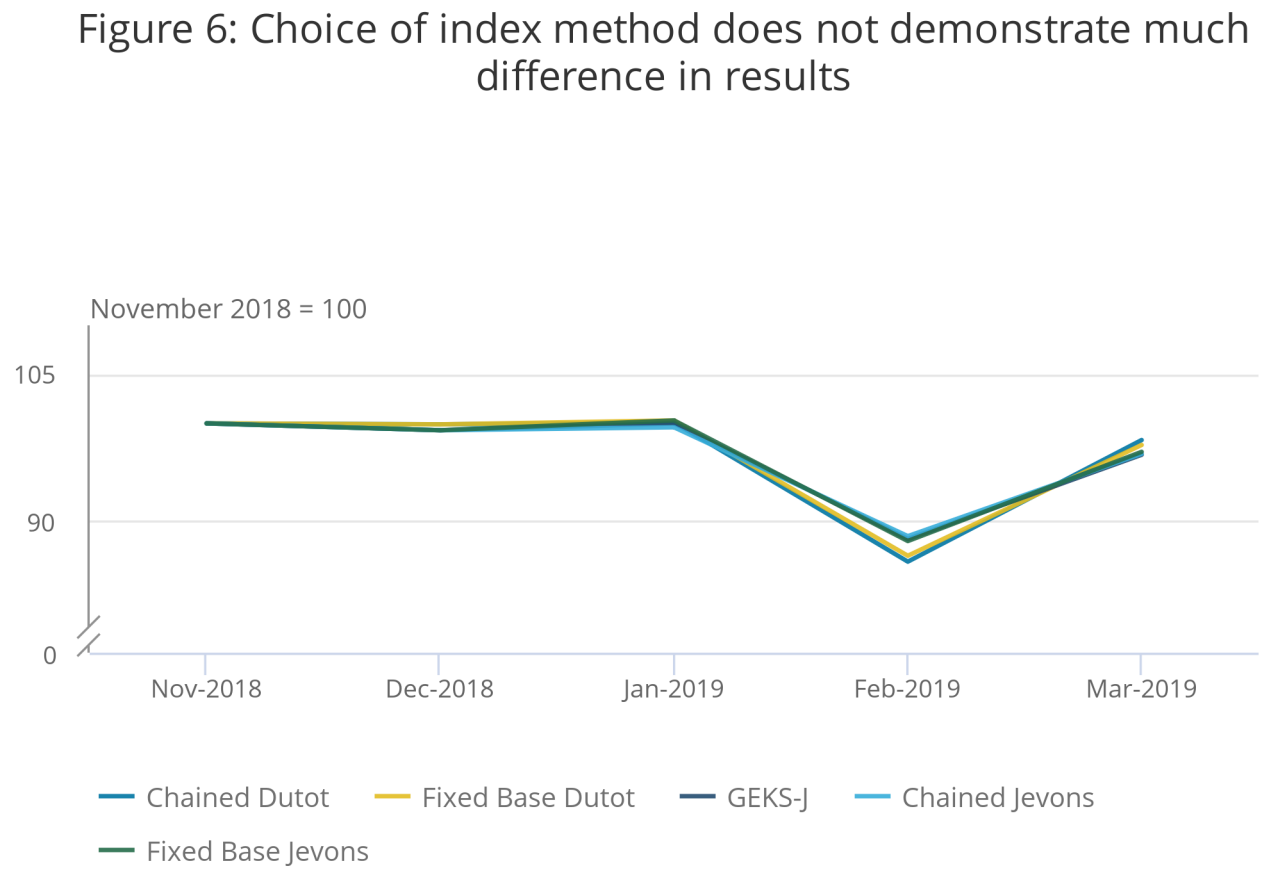
Table 2: Product churn statistics for laptops

	Total unique products	Unique products entering sample	Unique products dropped from sample
Nov-2018	3683		
Dec-2018	1441	281	2523
Jan-2019	1446	528	523
Feb-2019	1489	550	507
Mar-2019	1982	950	457

Source: Office for National Statistics

Finally, our pipeline allows us to produce a number of different price indices. Figure 6 shows the price index for laptops using a fixed base and chained Jevons, a fixed base and chained Dutot, and a version of the GEKS-J index. With the short time period we have available, there is not much difference between these index methods at the moment.

Figure 6: Choice of index method does not demonstrate much difference in results



Source: Office for National Statistics

Alternative scenarios

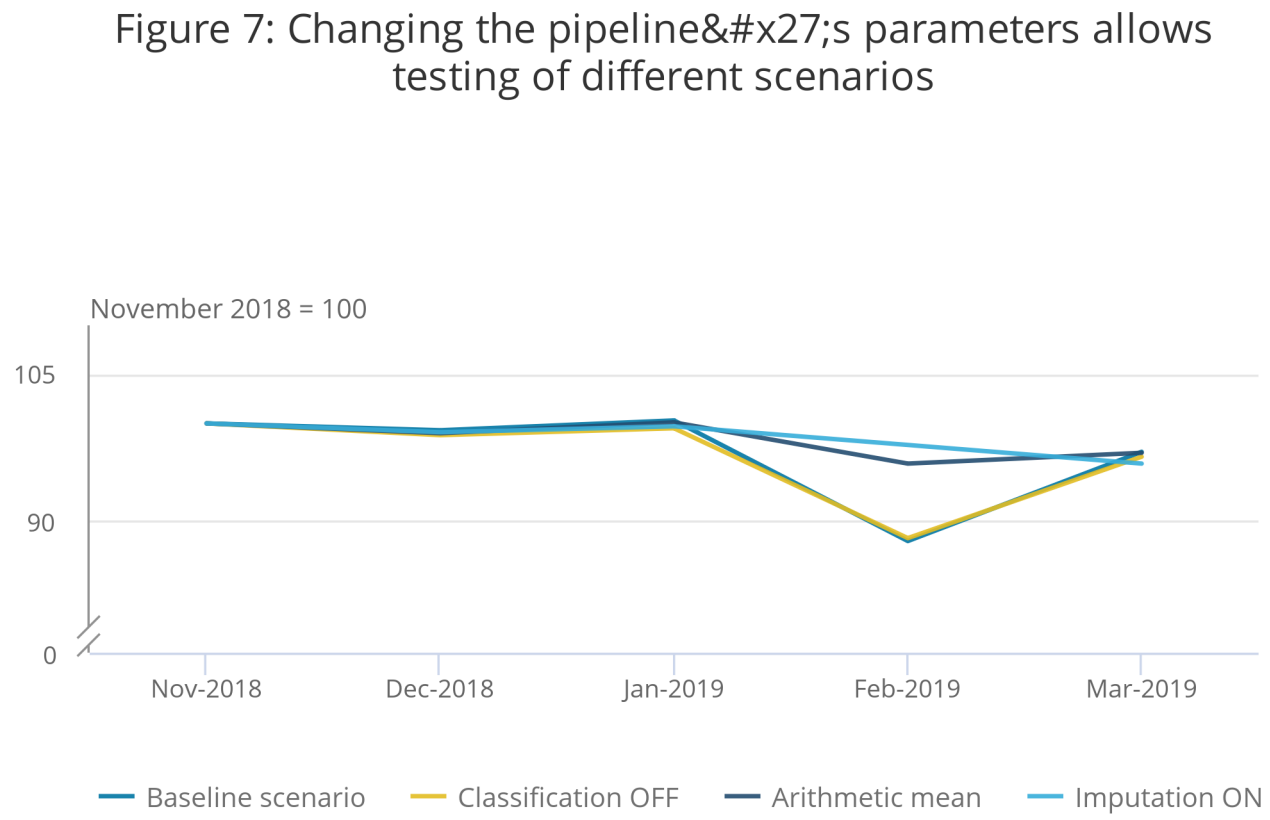
We have also been able to test different scenarios using the main pipeline. All of these are built on the baseline scenario and change one parameter (holding the rest constant), to allow us to identify the difference in a controlled manner.

The three scenarios we have tested are:

- switching off the classification module
- using the arithmetic average instead of geometric when calculating the monthly price
- switching the imputation module on and carrying the price forward for up to three months where a product is no longer available

Figure 7 shows the fixed-base Jevons index for laptops under the three alternative scenarios.

Figure 7: Changing the pipeline's parameters allows testing of different scenarios



Source: Office for National Statistics

Turning the classification module OFF does not appear to have a significant impact on laptops. This is despite on average 40% of products each month being classified as not belonging in this item category (one of the retailers sells a lot of refurbished laptops). This suggests that the price movements for misclassified products are similar to laptops in this instance.

Turning imputation ON, however, does appear to have an impact, most notably in February. The product churn analysis in Table 2 can possibly explain this. The large drop off in unique products from November to December means that these products are not included in February in the baseline scenario. However, when we carry prices forward for three months, this means products that were dropped now appear in the sample for December, January and (crucially) February. These products are given unchanged prices, which is why the imputed index is much closer to the 100 line, which indicates no price change (it is not exactly equal to 100 as there are still products that exist throughout the time period without the help of imputation that experience price change). Changing from geometric to arithmetic mean also seems to have an impact in February.

Further research is required to understand what is driving this difference, for example, the price relative analysis mentioned previously would be useful to understand the difference between the two scenarios.

6 . Further research

The previous sections set out the initial phase of the research, alongside some experimental statistics that have been produced using this processing pipeline. This work has provided some early indications that the production of consumer prices using these data sources is feasible given our new IT systems, and therefore we will continue with our implementation plans.

This section provides detail on the main research strands for this work and links them to the areas of the pipeline that they will impact on. It is envisaged that these research projects will provide regular updates via our advisory panels throughout 2019, with final recommendations due towards early and mid-2020.

These project streams include the following.

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 and staged implementation plan could be selected for different Classification of Individual Consumption According to Purpose (COICOP) groups.

The recommendations from this will feed into our final decision on which method/s to choose for implementation. The subsequent index methods will be coded up as part of the module on index methods in the processing pipeline. The first update for this work is due to be presented to the Technical Advisory Panel on Consumer Prices (APCP-T) meeting in May 2019.

Expenditure weights for web scraped data

One of the limitations of web scraped data is that it does not provide information on expenditure. This work will identify if the lack of expenditure weights introduces any bias into any index based on web scraped data and if we can approximate expenditure weights using alternative indicators like page rankings. Any recommendations from this work will feed into the index methods module. More information about this project is available.

Classification

Work on the classification module is split into two projects, the first is to review the existing item hierarchy and definitions. Our current methodology is based on representative items, which are generally quite narrowly defined. This work will evaluate the existing product hierarchy considering the increased coverage of products in alternative data sources.

Classification techniques

The second classification project looks at the new methods required to process these new “big” data sources, including new ways of automatically classifying products to a specific COICOP category. This work will recommend which methods are suitable for particular categories. A possible solution for multi-item datasets (PDF, 1.61MB) is summarised.

Product definition

In the current methodology, an individual product is followed over time and compared back with the base period. An alternative approach would be to follow the average price of a defined group of homogenous products instead. Research has shown this to be a viable alternative for categories such as clothing (PDF, 860KB), which experience high rates of product churn over time.

This will feed into how we define a unique product in the pipeline, for example, the averaging module could produce an average price for a homogenous product group rather than an individual product over the month. The first update for this work is due to be presented to the APCP-T meeting in September 2019.

Expenditure weights for different data sources and retailers

This work will recommend methods and suitable data sources that will allow us to aggregate together data sources from different collection methods for the same category (for example, locally collected data for bread from local bakeries alongside scanner data from a large retailer). This will feed into the aggregation part of the processing pipeline.

The impact of product returns and discounts on alternative data sources

The issue of returns affecting expenditure weights for particular categories (for example, clothing) may impact on how we can use expenditure weights in a final item index.

Processing pipeline

Further research will also be required to refine existing parts of the processing pipeline, for example, looking at the impact of using different outlier detection or imputation methods on the final item indices. This is where the configuration of the main and analysis pipelines will be very helpful to test the impact of these different scenarios.

Future work

There are also two workstreams that could be implemented before January 2023 as they do not rely on the development of the processing pipeline.

The first is a report published by the end of 2019 on whether or not the new data sources can be used as a direct replacement for our existing sample of prices in the short-term. This would also mean using existing systems and replicating the existing method of choosing comparable and non-comparable replacements.

This is not the best solution in the long-term as it does not make full use of the benefits these new data sources can bring in terms of increased coverage and higher frequency collection. However, it may be helpful in a staged approach if we wanted to replicate the existing methodology to ease collection burden in the short-term, before moving onto more complex methodology and systems by January 2023.

The second is exploring whether we can use the data as part of our existing process of hedonic regression in the short-term, while we explore alternative methods of quality adjustment. Hedonics regression has two parts.

The first part is to use “test data”, collected online separately from the monthly collection, to build a regression model to quantify the effect that individual components of a product have on the total price of a product (for example, the effect that RAM, processor speed and so on has on laptops). This collection requires several collectors two to three days to collect additional data each time the model is updated, which is currently up to three times a year.

The second part is to then use this regression model to apply quality-adjustment to the monthly sample when products are replaced due to product churn. Web scraped data could be used instead of the manually collected “test data” to reduce the burden on collectors. It will not directly feed into the monthly sample of products used for index calculation, however.

7 . Authors

Tanya Flower and Eleftherios Karachalias, Office for National Statistics.

8 . Acknowledgements

The authors would like to thank the contributions of:

Emerging Platforms Team; especially Andrew Banks, Lewis Edwards, Ian Thomas, Victor Irles and Nick Drake

Economic Microdata Research; especially Andrew Sutton and Daniel Ollerenshaw

Methodology; especially Alex Rose, Edward Rowland and Hazel Martindale

Prices; especially Liam Greenhough and Chris Payne