Office for National Statistics

Article

Research into the use of scanner data for constructing UK consumer price statistics

Research into using scanner data provided directly from UK retailers to integrate with other data sources in producing UK consumer price statistics.

Contact:
Helen Sands
cpi@ons.gov.uk
+44 (0)1633 456900

Release date:
6 April 2021

Next release:
September 2021

Table of contents

1 . Overview of using scanner data to construct consumer price statistics

In this article we look at how we can produce consumer price indices using scanner data.

Scanner data refers to data that are collected by retailers at the point of sale. These are aggregated by product and do not identify individual consumer purchases. These data will provide us with significantly more prices each month compared with traditional data sources. They also provide us with more information on the types and numbers of products sold.

From 2023, scanner data will be integrated with web-scraped and traditional data sources in the production of UK consumer price statistics. Scanner data will not completely replace the traditional, manual collection as there remains a need for price collection for retailers and services who do not have an online presence or means of providing data.

We are currently receiving scanner data from eight UK retailers and are in ongoing discussions with more retailers to ensure a significant market coverage at the time we begin to implement these data into the live production of UK consumer price statistics from 2023. Our primary focus to date has been on the grocery sector, particularly food and drink items.

We have also been researching how methods for processing these data will differ to the processing of data collected manually in stores, including how we classify products, what the optimum time coverage of data is, how we identify and track unique and relaunched products through time, how we derive a price while accounting for any size changes in a product line, how the treatment of discounts can impact our price indices and how we deal with refunds within the data. These points will be covered in further detail in Section 3 of this report.

2 . Data acquisition and quality assurance

Why are we looking to acquire scanner data?

Scanner data refers to data that are collected by retailers at the point of sale. These are aggregated by product and do not identify individual consumer purchases.

Scanner data offer increased time and product coverage when compared with traditional survey-type methods of data collection; they provide information on transactions for all products in all time periods. They also provide information on transaction prices rather than advertised prices. This means that we can observe the average price consumers have actually paid for each product (for example, accounting for multibuy offers); despite what price the product was being advertised at. They also offer enhanced regional and product coverage relative to our current samples.

A further benefit of scanner data relative to traditional survey-type collection is that they contain details of how many of each product were sold. This allows us to weight individual products relative to their economic importance. For example, if consumers are spending more on Pink Lady apples than Braeburn apples, Pink Lady apples will have a greater influence on the inflation rate for apples.

We have currently progressed scanner data acquisition for groceries as we can cover a large proportion of the market with a smaller number of retailers, compared with clothing for example, where the market is less concentrated.

Progress in scanner data acquisition

We are currently receiving weekly feeds of scanner data from eight UK retailers. These retailers are predominantly grocery retailers, but between them also cover fuel, homeware, personal care, clothing, electronics, and leisure goods. Our research to date, and implementation from 2023, primarily focuses on scanner data for food, drink and tobacco products; however, we look to expand the use of these data in additional categories in subsequent years.

We continue working closely with major UK retailers on data acquisition, and by the end of 2021, we will have coverage of a significant share of the UK food and drink market with these data. For each retailer there are hundreds of millions of rows of data; a substantial increase in the amount of data that we need to process each month.

Quality assurance of scanner data

As part of the regular process of data assurance, each data delivery undergoes a series of quality checks. Initial checks include ensuring that variables are the specified data types and values are within a predefined range and that data size and shape are as expected. If initial quality checks fail, the issues would be raised with the retailer and redelivery of the data arranged. Once data pass these initial checks, further queries are made of the data, including investigations of new entries, identification and investigation of outliers as well as other curiosity checks, depending on earlier findings.

More generally, we are currently undertaking a quality review of all these new data sources to ensure that when they are used in live production, they meet our quality standards. This work will include collecting the information required to update the quality assurance of administrative data used in consumer price inflation statistics article, including identifying any risks involved with these new data sources and any mitigating action that can be taken.

3 . Methods for processing scanner data

Notes

For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared with official measures of consumer price inflation.

Our current proposed index number method[1] to use with scanner data is the Quality-adjusted Geary Khamis (QU-GK) index method using the fixed-base monthly expanding window. We are still working to optimise this method in our processing systems, therefore for the purposes of this article we have used a GEKS-Törnqvist method with a 13-month movement splice; this was our second highest scoring method in our index number methods framework.

Classification of scanner data for food, drink, and tobacco products

Once the data have been received and we are confident that they are of sufficient quality, one of the first steps in the process is to classify the data to the appropriate category, for example classifying a block of cheddar as "cheddar cheese" or a bag of Pink Lady apples as "apples".

Our previous work on classification methods has put a primary focus on using machine learning methods to create a largely automated approach to classifying web-scraped clothing data into categories (such as "women's jeans", "babies' pyjamas" and "boys' shirts"). More detail on this work can be found in Automated classification of web-scraped clothing data in consumer prices.

A largely automated approach was necessary for clothing because of the sheer volume of data, but a different method may be more suitable for working with grocery scanner data because of its different data properties. Details for how we choose a classification approach for each dataset, and in particular our plans for grocery classification, are outlined in our related article Classification of new data in UK consumer price statistics and are therefore not covered further in this article.

Time coverage

We currently receive data aggregated over a week from most retailers, although some provide weekly feeds of daily data. For those who send data aggregated across a week, some weeks fall within two adjacent months; therefore, we are not able to disaggregate the data to determine which month the transactions within a week fall into.

Prices for consumer price indices should be attributable to a single month so the difference arising between months can be seen. The HICP Practical Guide for Processing Scanner data (2017) states: "In principle, as many days as possible should be included, but none should be included that refer to other months. It is important to ensure that the time interval is defined in the same way throughout the year."

However, some international literature suggests that using data from an incomplete time period to construct unit values can lead to an upward bias in the index (Fox and Syed, 2016; Diewert, Fox and Haan, 2016).

We plan to firstly assess any bias arising from only using a partial month of data (for example, using only weeks that fall fully within the month). This can be assessed using retailers who have been providing data aggregated over each day instead of each week, allowing us to include a full month of data when producing indices. If we find that there does appear to be a bias caused by only using partial months, we can explore what data usage option causes the least amount of bias. For example, we could attempt to use at least three weeks of data each month, or we could splice the data that overlaps two consecutive months.

It should be noted that even if we were to use just a single weeks' worth of scanner data from each retailer, this will still be an improvement in time coverage relative to our traditional collection, which is typically carried out over one to three days each month.

Identifying unique products in scanner data

Products tend to be "relaunched" over time, for example, when product packaging changes or there are small changes in a product's size. When constructing price indices, we want to continue to track a product when it is relaunched, providing that the relaunched product is comparable with the original product. Relaunches are often associated with price or minor quality changes. Therefore, it is particularly important that we capture them and adjust for any quality changes, to ensure that we capture any impact that they have on inflation.

All our scanner datasets contain both a global trade item number (GTIN) code and a retailer-defined Stock Keeping Unit (SKU) code that can be used to identify products. The GTIN (global) code is unique to each product and consistent across retailers. However, it can change because of small modifications to the product such as packaging or minor changes in ingredients.

Although the SKU code is retailer-defined, it is typically broader than the GTIN code and therefore a better identifier of a product for the purposes of constructing price indices. The SKU code will typically remain unchanged where there are minor changes in packaging, size or ingredients that may not be immediately noticeable to the consumer, but these changes will usually be accompanied by a new GTIN. GS1 - the provider of industry-standard product identifiers (barcodes) - employ a management standard for GTIN setting, to ensure consistency across retailers and countries.

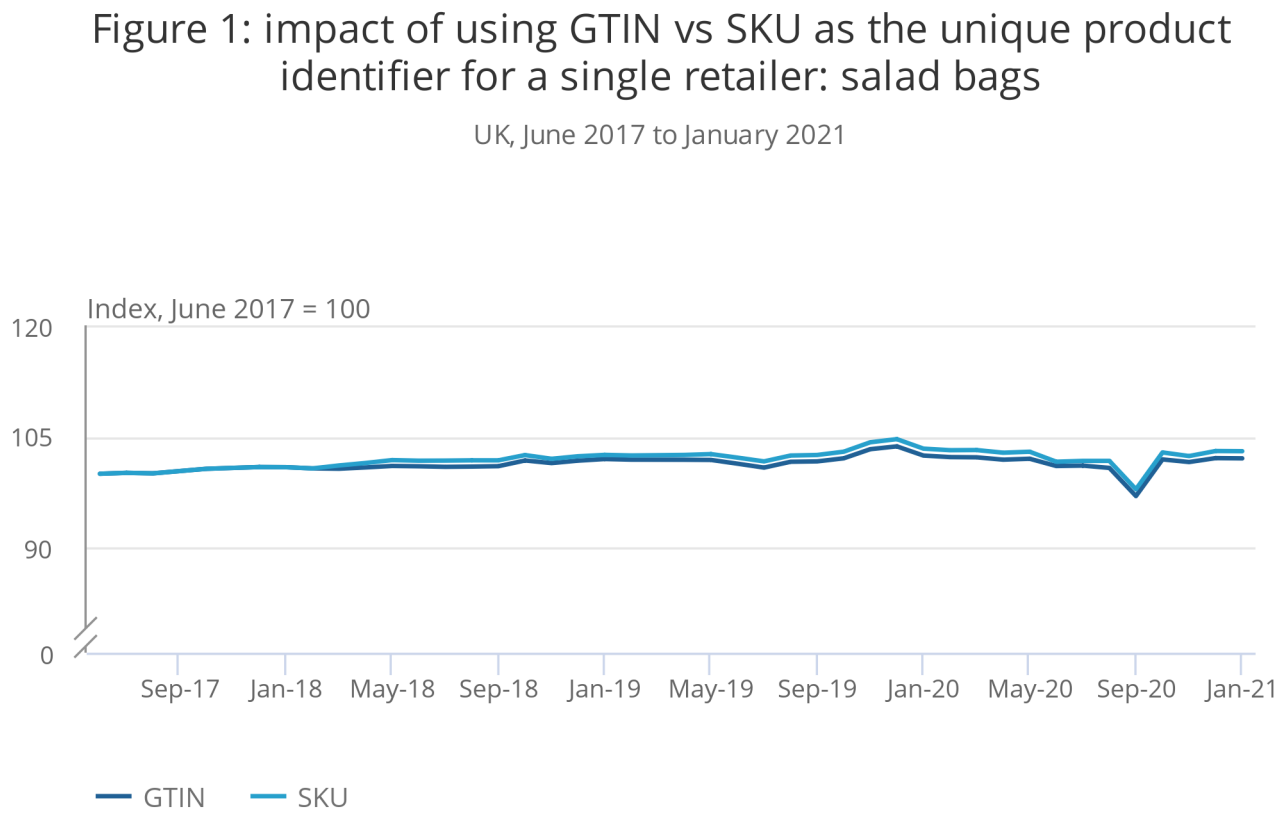
However, there is no guarantee that all relaunches are captured by SKUs. As SKUs are retailer-defined, while one retailer may capture relaunches within their SKU, another may choose to align their SKUs with GTINs at the barcode level.

Other national statistical institutes (NSIs) attempt to match old and new relaunched products using techniques like text mining on product descriptions, matching on product characteristics and detecting trends in expenditure. Some research (for example Belgium) has shown that a downward bias can be introduced when relaunches are not appropriately accounted for.

To begin testing whether this bias is present in our own data, we have produced price indices for a single retailer using the more narrowly defined GTIN code as the unique product identifier, compared with using the retailer-defined SKU code (that we know to be capturing at least some relaunches). These comparisons often do show differences in index values at the elementary aggregate level, however, there is no apparent bias in any direction. Figures 1 to 4 show some case studies with price indices using GTINs as the unique identifier compared with using SKU as the unique identifier.

Figure 1: impact of using GTIN vs SKU as the unique product identifier for a single retailer: salad bags

UK, June 2017 to January 2021



Source: Office for National Statistics – Single UK grocery retailer

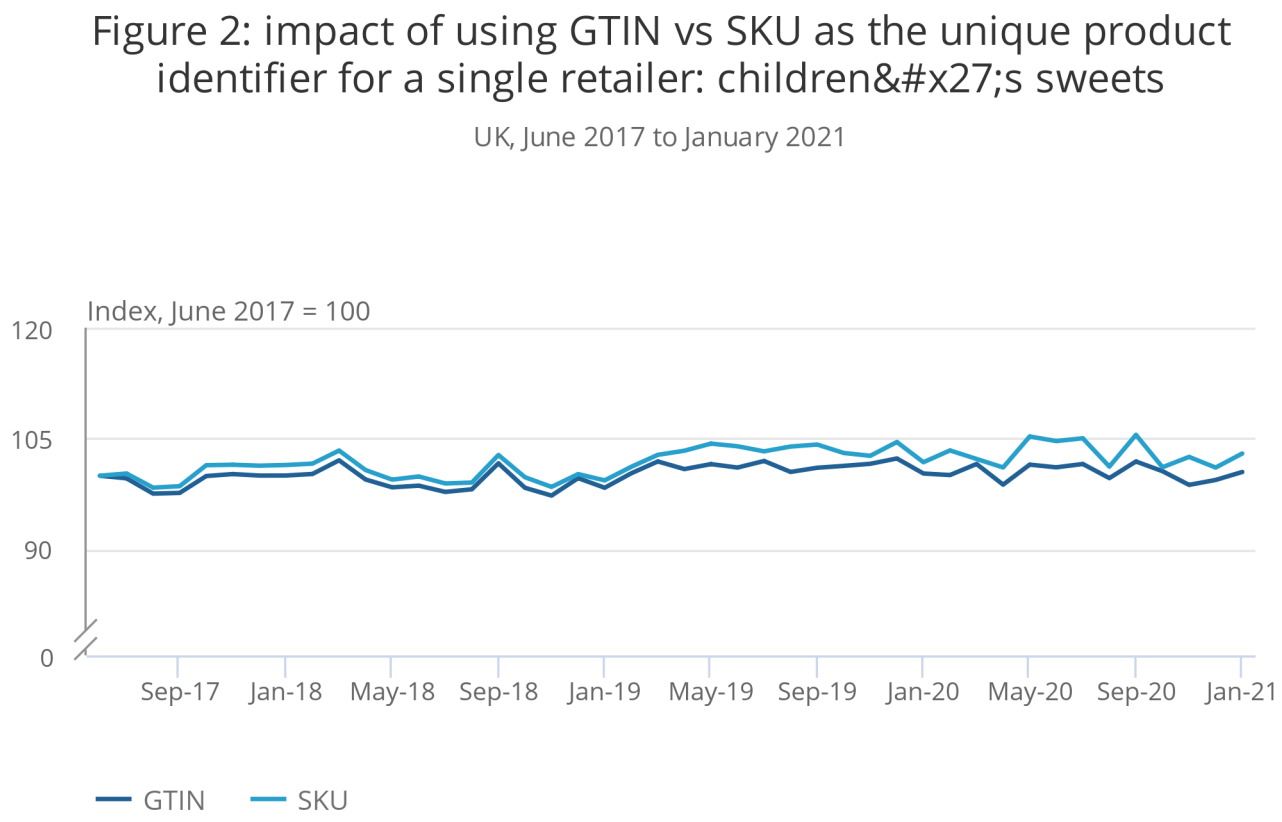
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 2: impact of using GTIN vs SKU as the unique product identifier for a single retailer: children's sweets

UK, June 2017 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

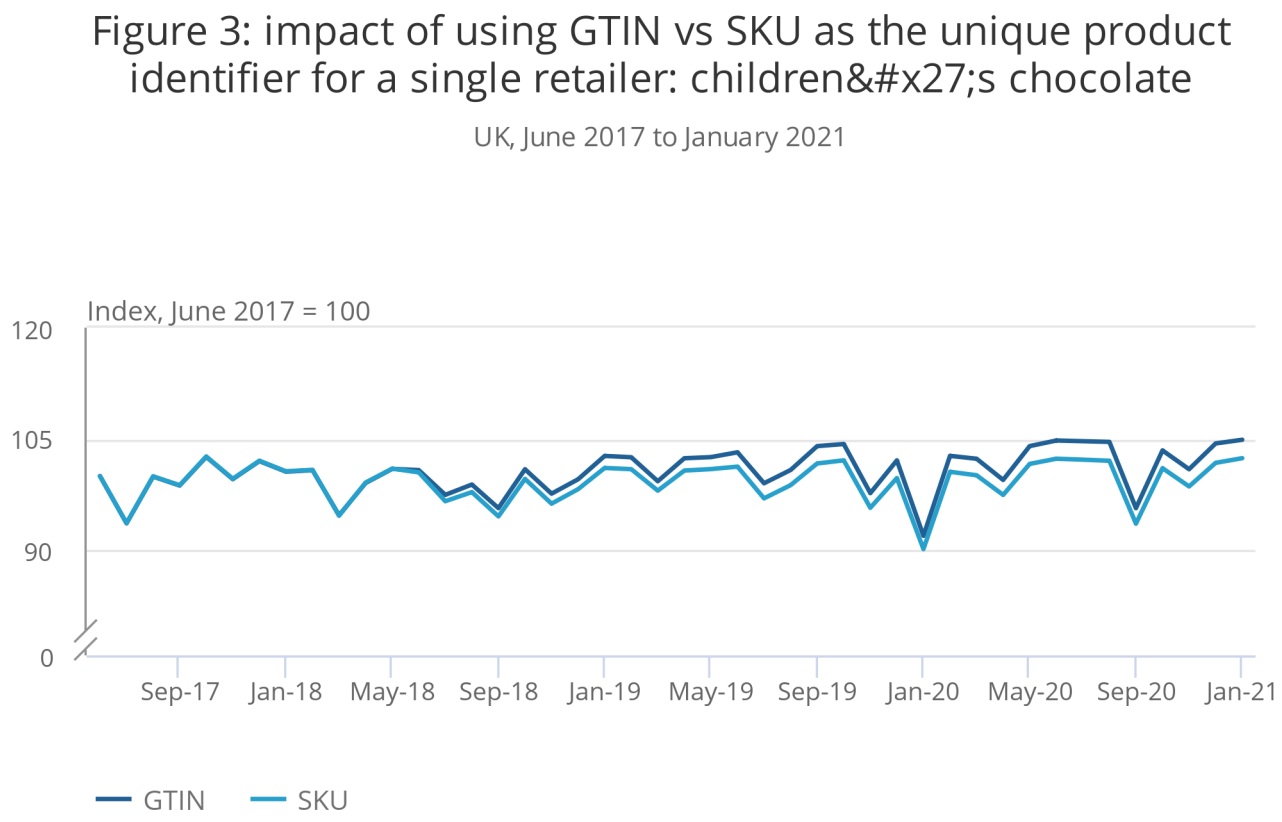
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 3: impact of using GTIN vs SKU as the unique product identifier for a single retailer: children's chocolate

UK, June 2017 to January 2021



Source: Office for National Statistics – Single UK grocery retailer

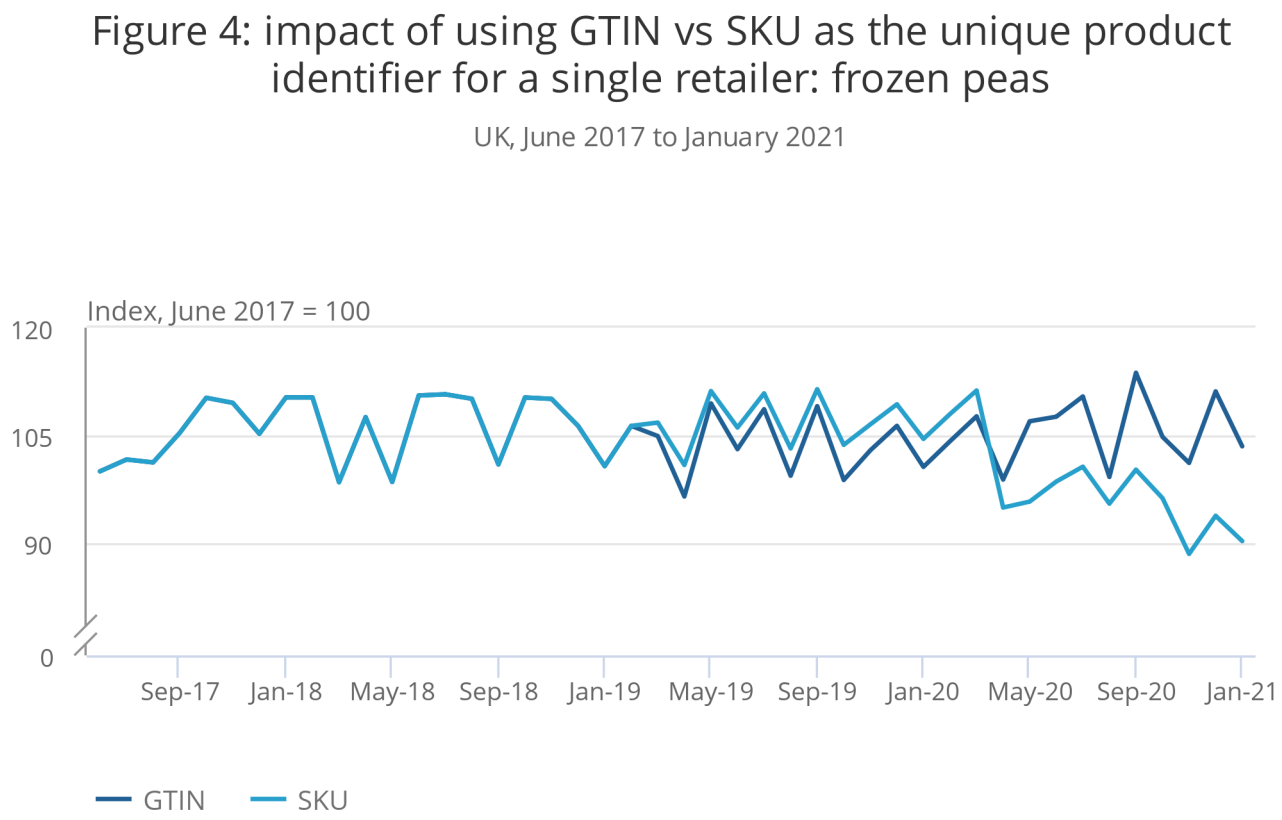
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 4: impact of using GTIN vs SKU as the unique product identifier for a single retailer: frozen peas

UK, June 2017 to January 2021



Source: Office for National Statistics – Single UK grocery retailer

Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

From manual scrutiny of the data and initial investigations into product relaunches we believe it is likely that not all product relaunches are being appropriately captured within the same SKU for this retailer and our remaining retailers. Therefore, we are currently investigating record linking methods for identifying and linking product relaunches. These methods involve assessing each new SKU in the dataset against existing SKUs. If a match is identified within a given timeframe[2], the new SKU is automatically linked to the existing SKU.

For each product with a new SKU that cannot be linked historically, existing products are searched for within the product hierarchy (category) based on similarities in product name and price. Matches are scored quantitively and products with the highest match rates are returned for validation.

We are investigating whether products with a high match rate with an existing product can be automatically validated or whether further manual validation is required. We will likely need to manually validate all products with a moderate match rate to decide whether they are relaunches or not. If new products have a low match rate with all existing products, they can be treated as a new product. We are still investigating the suitability of different thresholds for validating product matches. A stylised example of products with high, moderate and low match rates is provided in Table 1.

Table 1: Examples of SKU-based matching using record linkage

Product with new SKU	Best match from all existing products	Match % on name and price	Match likelihood
BrandX Leek and Potato Mini Pies 4 Pack	BrandX Leek and Potato Small Pies 4 Pack	92.70%	Match
BrandY Mixed Fruit Energy Drink 250ml	BrandY Energy Drink 250ml	87.50%	Likely match
BrandZ Mixed Berry Mints 38G	BrandA Sugar-Free Strawberry Sweets 120G	26.20%	No match

Source: Office for National Statistics

Accounting for inconsistent units of measurement

The size or weight of a product is used in the price calculations (See section 3: Deriving a price from scanner data) to ensure that any changes in size or weight are accounted for in our inflation measures. However, it is possible for a product with a single SKU code to have multiple units of measurement (UoM)[3] over time. In total, between June 2017 and June 2019, SKUs with multiple UoM (across the whole period) accounted for 3.6% of expenditure in one retailers' data.

Where possible, we have standardised the UoM to ensure that they are consistent over time. For example, UoM expressed in the raw data in kilograms or grams are standardised to grams, and UoM expressed in litres, centilitres and millilitres are standardised to millilitres. After standardising units in this way, SKUs with multiple UoM between June 2017 and June 2019 still accounted for 2.8% of expenditure.

These SKUs remain because it is not always possible to convert all UoM used for one SKU into a common UoM. For example, if a product is sometimes measured in grams but other times referred to simply as a "pack", the units cannot be inter-converted. For this reason, a combination of the product SKU code and the standardised UoM are now used to form a unique identifier when matching products for price index calculations. This means that products with the same SKU but a different UoM are treated as unique products, to avoid comparisons being made between products with non-consistent UoMs.

Deriving a price from scanner data

The scanner data do not contain explicit shelf prices for the products, instead they give the total expenditure on each unique product each week (or day if the data are daily), and the number of units sold. From this we can simply derive the average transaction price (p) by dividing the total expenditure (V) for each product (i) at time (t) by the quantity (q) of that product sold.

$$p_i^t = \left(V_i^t\right) / \left(q_i^t\right)$$

Using this equation, if £100 (V) apples (i) were sold in a week (t), and 200 (q) apples had been bought in that week, we would derive an average transaction price for that week of 50 pence per apple (p).

To calculate price indices that are automatically adjusted for changes to product size or weight, the following transformations are made to calculate the prices and quantities used in our index number methods:

1. the total expenditure is divided by the quantity sold and product size or weight to give a price per unit of measurement (for example, price per gram of a chocolate bar, rather than price per chocolate bar)

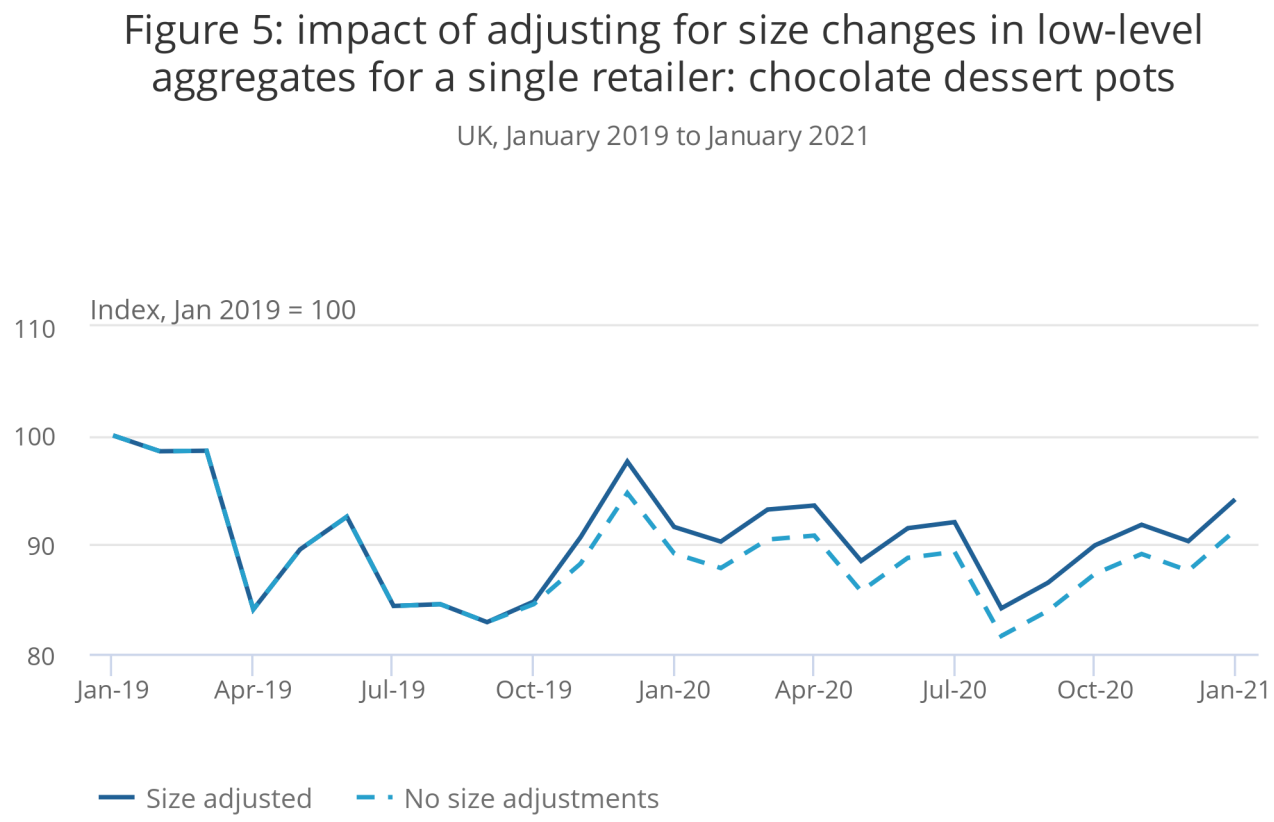
2. quantities sold are multiplied by the product size or weight to give the total size or weight purchased (for example, grams sold of a chocolate bar, rather than number of chocolate bars sold)

These transformations give the same results as the size-adjustment methods used for our current official consumer price indices (see Section 9.3b: Quantity adjustment), where the base prices are adjusted by the ratio of the product size or weight in the base and current period. By using these transformations instead, we can use multilateral methods more efficiently without having to adjust previous prices to account for changes in size.

We have investigated the impact of making size adjustments on low-level aggregates constructed using scanner data from a single retailer. While for many categories there is little to no impact from the use of size adjusted calculations, for some categories we see heightened inflation when we account for the changing size of products. In Figures 5 and 6, the size-adjusted aggregates diverge upwards from the unadjusted aggregates, implying size decreases have contributed to inflation in these categories over this period. However, in Figures 7 and 8 there is negligible difference between the index values, suggesting there have been no size changes for these products during this period.

Figure 5: impact of adjusting for size changes in low-level aggregates for a single retailer: chocolate dessert pots

UK, January 2019 to January 2021



Source: Office for National Statistics -Single UK grocery retailer

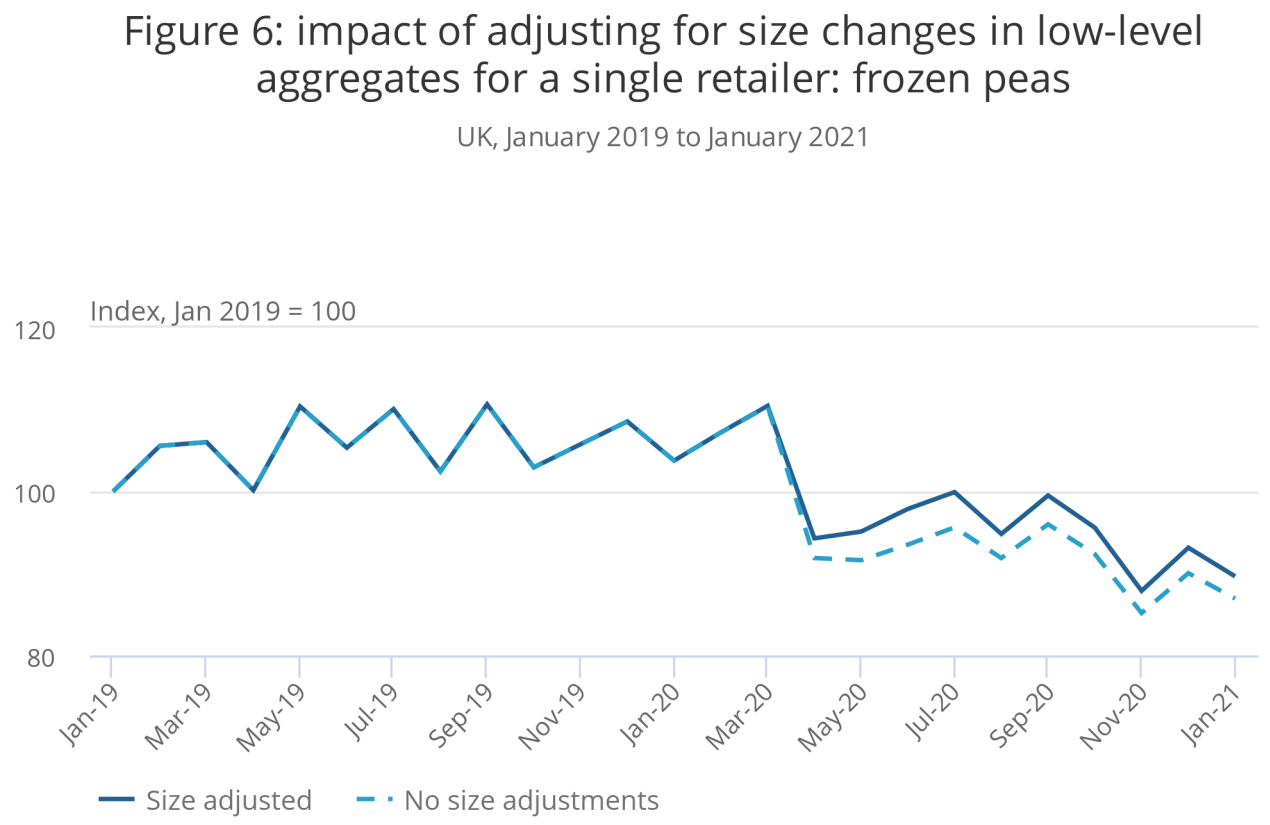
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 6: impact of adjusting for size changes in low-level aggregates for a single retailer: frozen peas

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

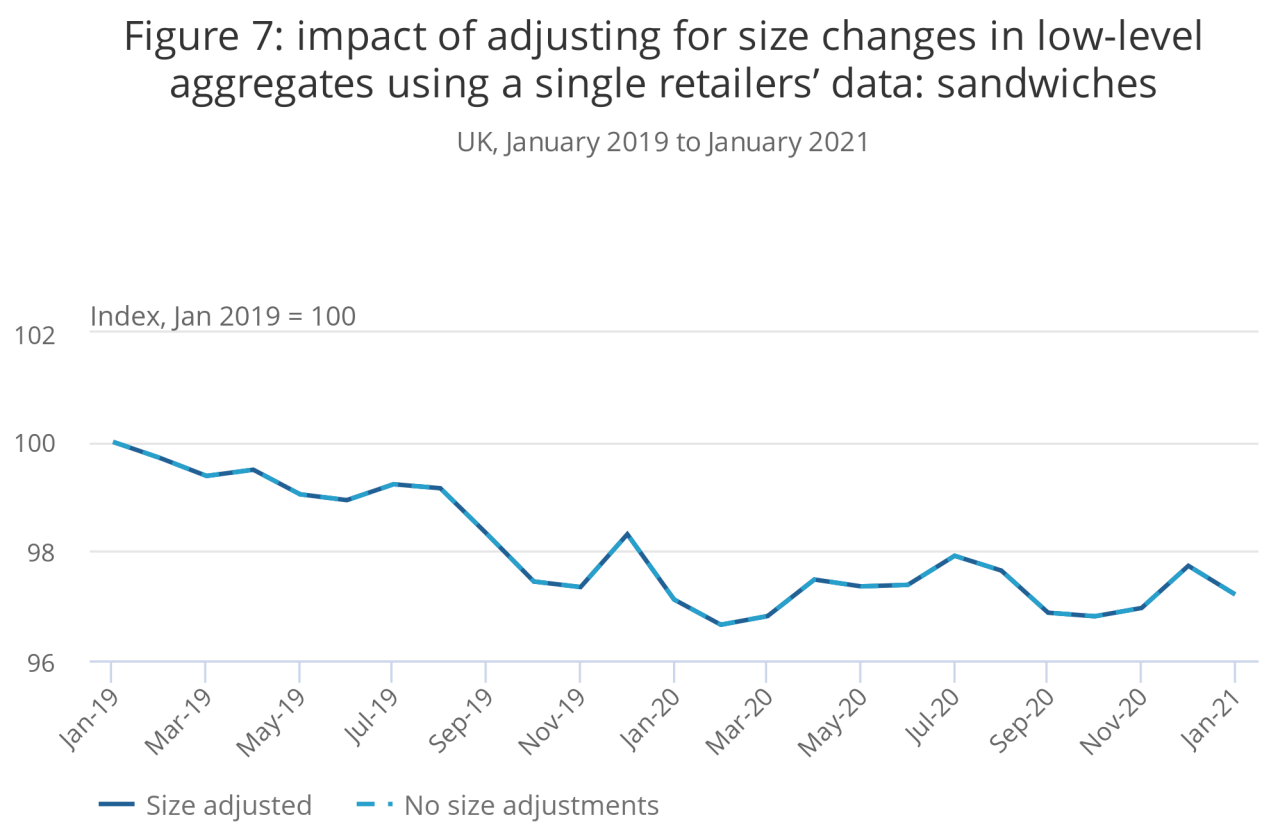
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 7: impact of adjusting for size changes in low-level aggregates using a single retailers' data: sandwiches

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

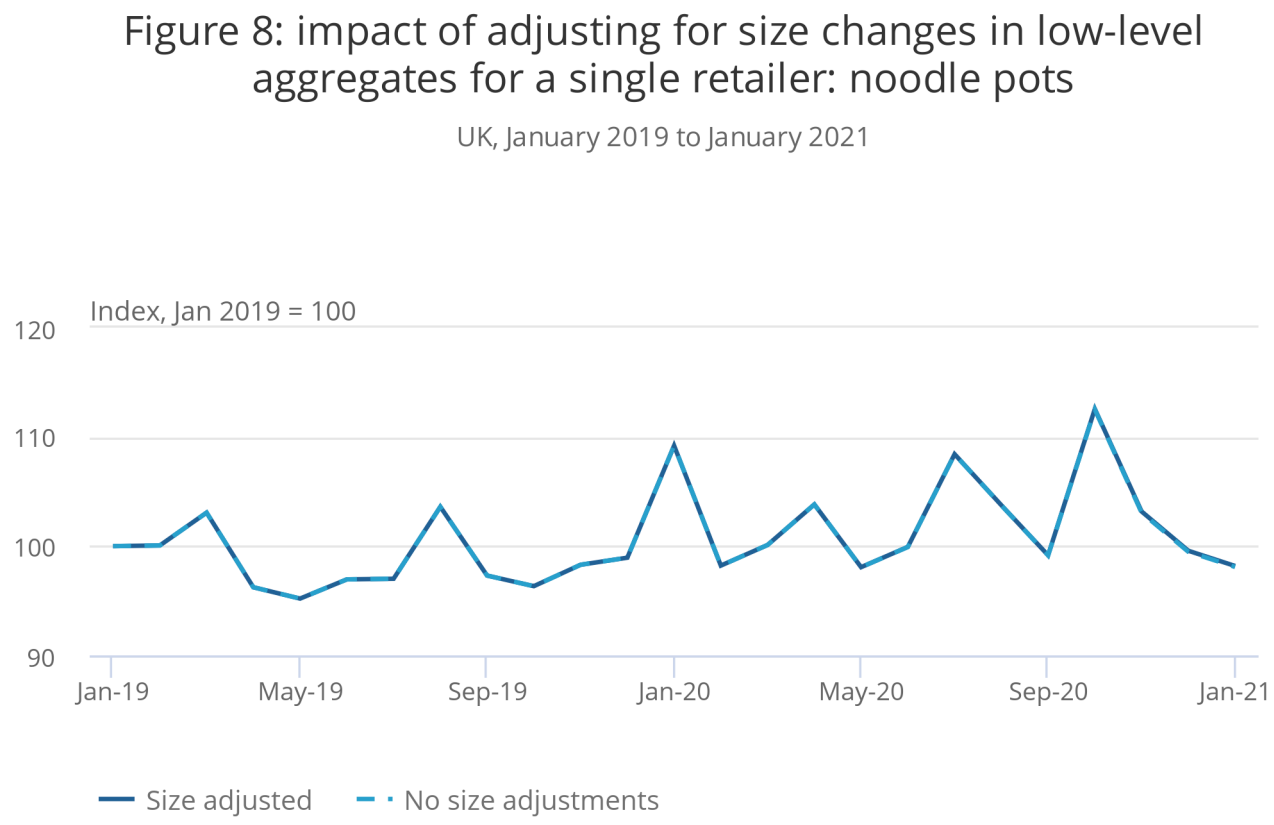
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 8: impact of adjusting for size changes in low-level aggregates for a single retailer: noodle pots

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Treatment of discounts in scanner data

In our current consumer prices indices, price promotions are accounted for but multibuy promotions are currently not included given that we do not know the rate at which multibuy offers are taken up by consumers (as discussed in paper APCP-T(19)15). We also do not currently account for discriminatory discounts (those which are not available to all customers, such as loyalty card discounts or staff or student discounts) for the same reasons. Prices for reduced to clear ("yellow-sticker" type) products are not collected, as the quality is not deemed comparable with a product being sold at its full or promotional price.

Scanner data contain information on all types of promotional offers, including price reductions, multibuy discounts and discriminatory discounts (although the information regarding discounts that we can extract from scanner data varies by retailer). This means that by using scanner data we can typically better reflect the actual prices paid on average by consumers including any impact from changes in promotional offers over time. How we treat discounts will in part be influenced by the ability of retailers to break down the value of their sales according to different discount types.

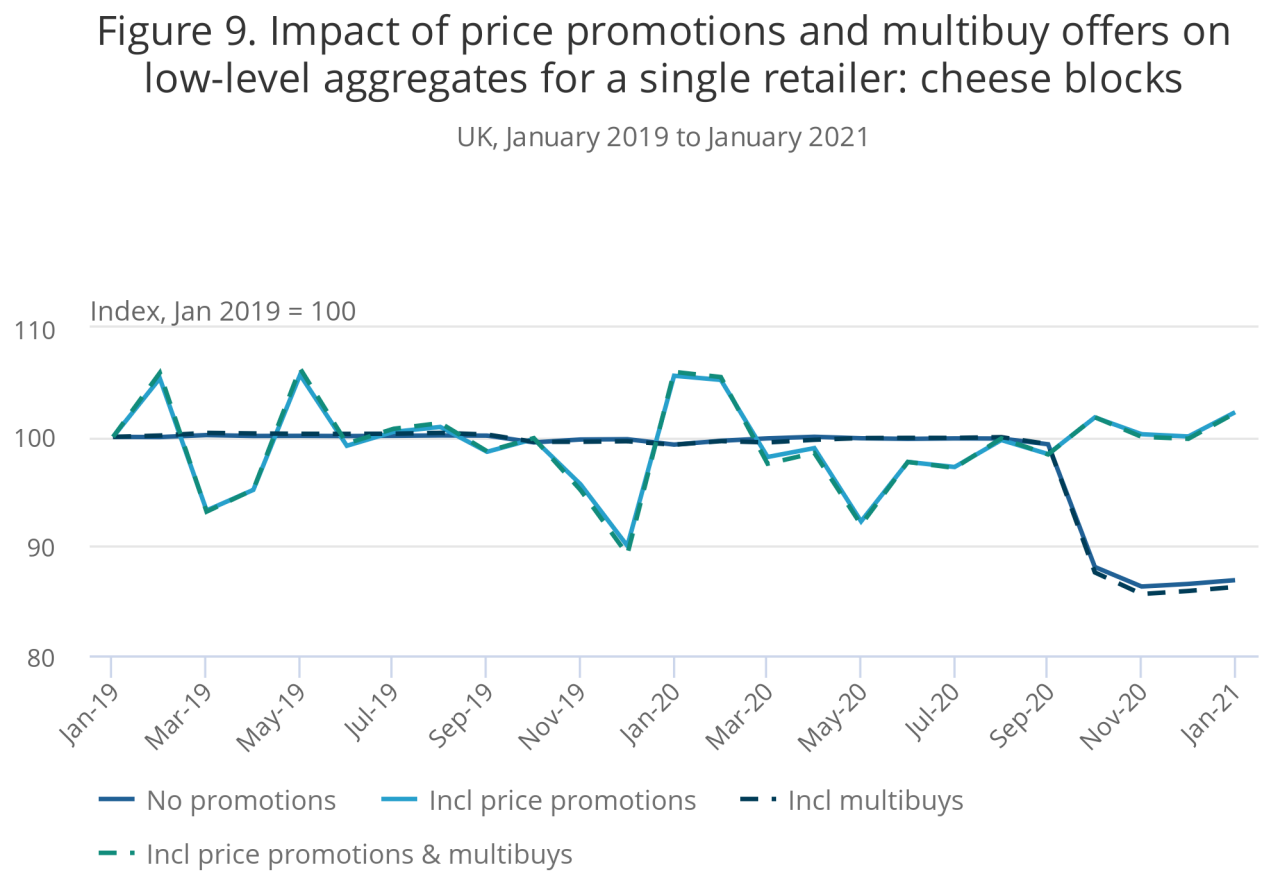
We have investigated the impact of price reduction and multibuy discounts on low-level aggregates using scanner data from a single retailer between January 2019 and January 2021.

In many low-level aggregates, the inclusion of promotional offers introduced volatility relative to when these offers were not accounted for (Figures 9 and 10). In some cases, trends were significantly altered. For example, for some product categories, there was evidence of an increase in prices during the nationwide lockdown because of the COVID-19 pandemic in 2020; however, this effect was removed when the impact of price promotions was excluded (for example, Figure 10: Olive oil). This implies that a reduction in the number (or magnitude) of promotional offers caused inflationary pressure for these product groups over the pandemic period. More generally this implies that, if all price promotions were excluded from price indices, we may not be appropriately capturing the consumer experience of changes in transaction prices.

In most cases observed within this single retailer, price promotions had a larger impact on price indices than multibuy discounts. However, this varied by product category; for example, multibuy discounts had very little impact compared with price promotions in the cheese blocks and olive oil categories (Figures 9 and 10), while multibuy discounts had a larger impact than price promotions for berries and, to a lesser extent, chocolate bars (Figures 11 and 12). This suggests that the prevalence of multibuy offers varies by product category although this will also largely depend on individual retailer pricing strategies.

Figure 9. Impact of price promotions and multibuy offers on low-level aggregates for a single retailer: cheese blocks

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

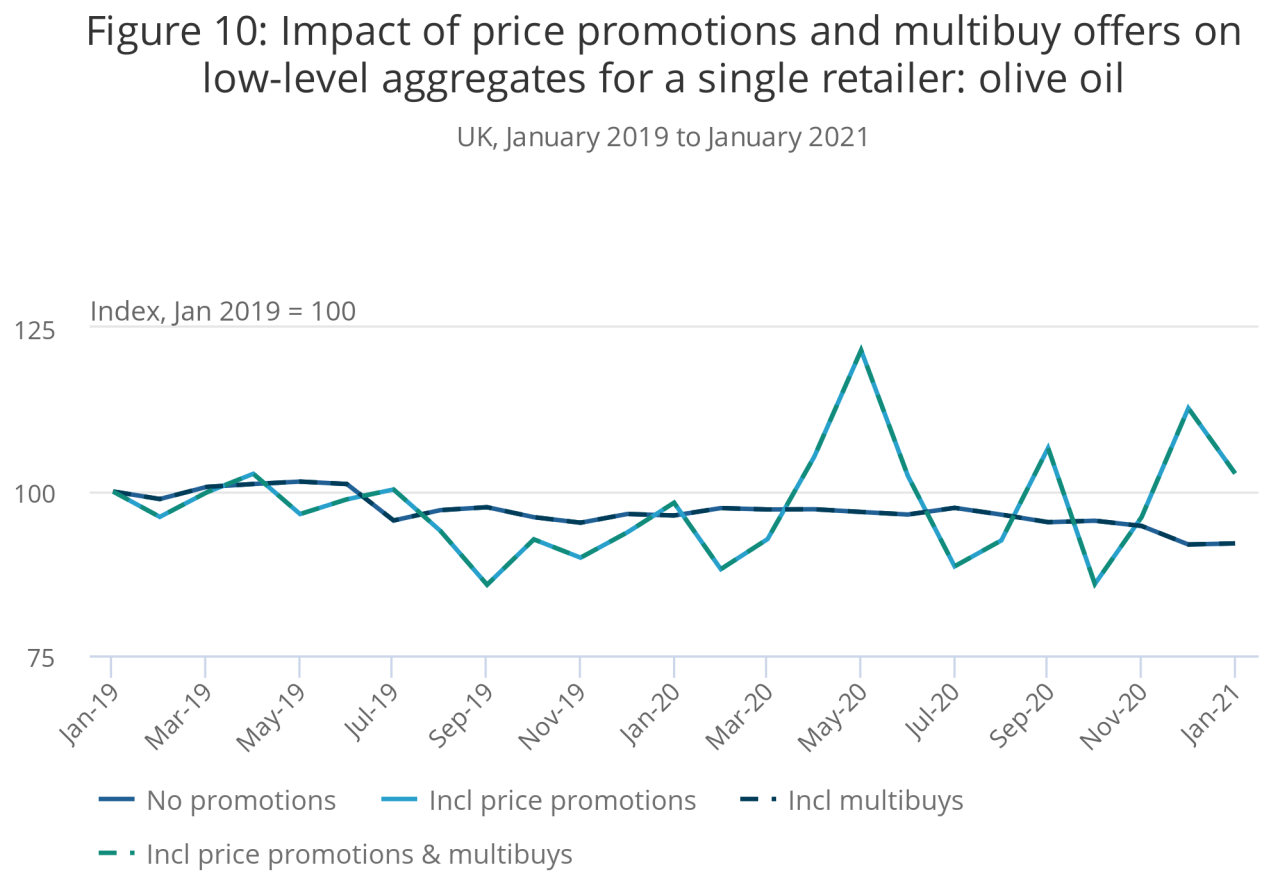
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 10: Impact of price promotions and multibuy offers on low-level aggregates for a single retailer: olive oil

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

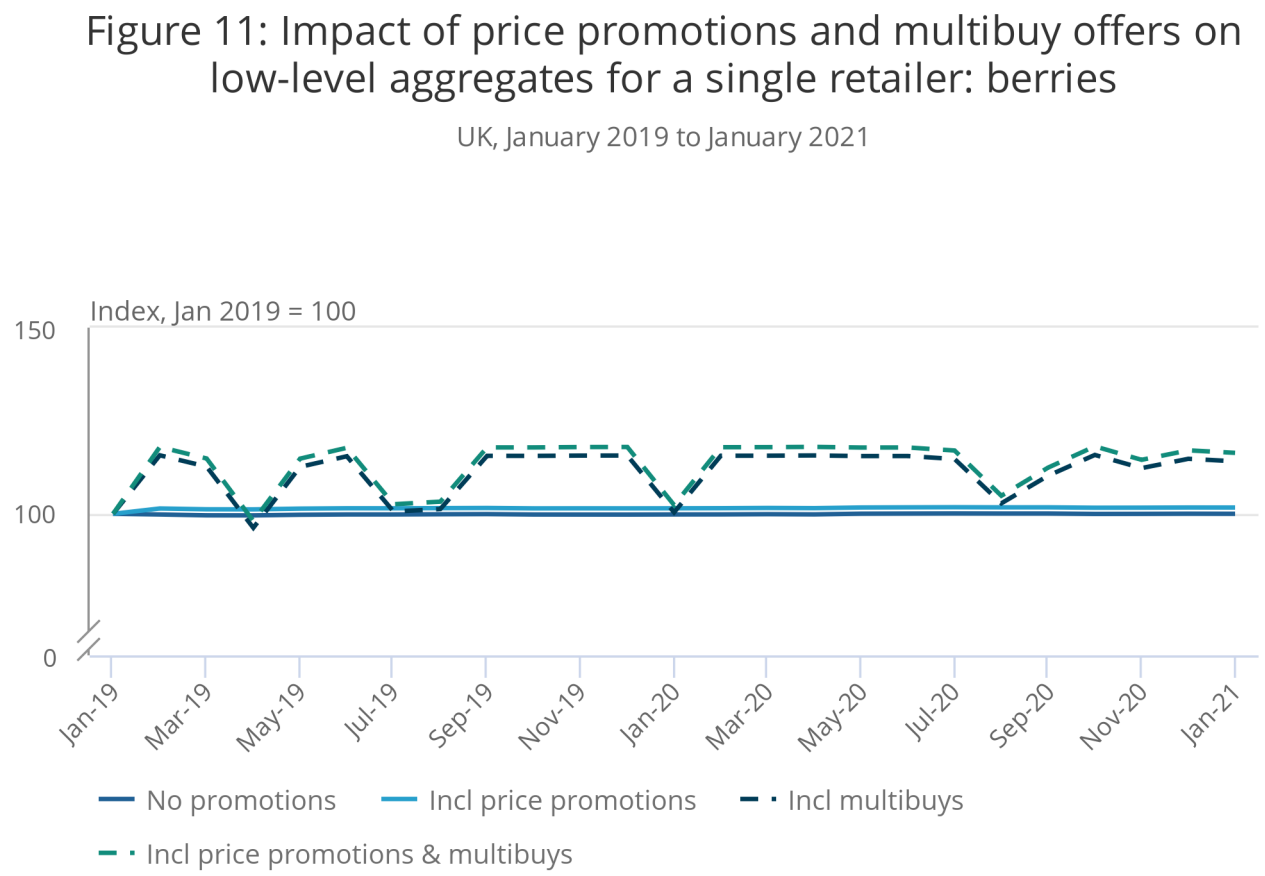
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 11: Impact of price promotions and multibuy offers on low-level aggregates for a single retailer: berries

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

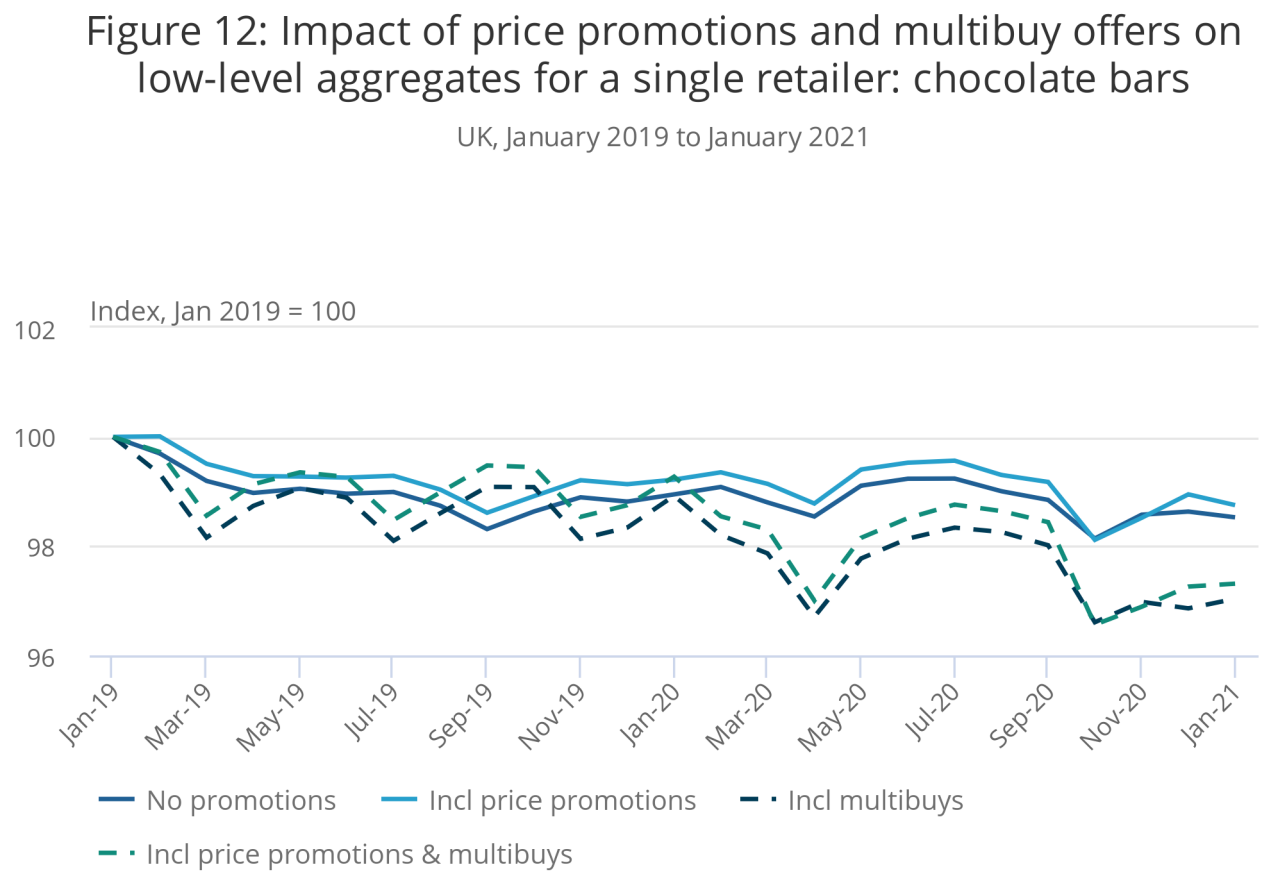
Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

Figure 12: Impact of price promotions and multibuy offers on low-level aggregates for a single retailer: chocolate bars

UK, January 2019 to January 2021



Source: Office for National Statistics - Single UK grocery retailer

Notes:

1. For the purposes of this report, we have received permission to use data from a single retailer to demonstrate the impact of different processing decisions on the resulting indices. The indices for this retailer are experimental and should not be compared to official measures of consumer price inflation.

2. We have used a GEKS-Törnqvist method with a 13-month movement splice for this analysis.

In some instances, the index that includes price promotions shows greater inflationary pressure than the index that excludes promotions. This is because a reduction in the number or magnitude of promotional offers between periods can lead to an increase in the transaction price.

While price promotions may introduce volatility into price indices, they may also offer valuable insight into where inflation may be higher or lower for consumers because of the increase or reduction in the number or magnitude of available offers. Given that we are trying to measure changes in consumer prices, and these can be influenced by discounting behaviour, we are proposing to include price promotions, multibuy offers and discriminatory discounts when producing price statistics using scanner data. However, we will continue to remove reduced to clear “yellow sticker” type products, where possible, as the quality of these products is not comparable with products selling at full or promotional prices.

In the short term, the inclusion of all discount types will only be for the retailers whom we have scanner data for. But longer term we may be able to apply modelled take up rates of multibuy discounts for certain product categories in the scanner data to traditional data sources, to ensure consistency in our treatment of multibuy discounts across the different collection methods. However, we would not be able to apply the take-up rates of discriminatory discounts to the locally collected data as discriminatory discounts in one store are not directly generalisable to another.

Treatment of refunds in scanner data

Refunds relate to products that have been returned to the retailer by the consumer. The product is not consumed, and within our price indices it would be optimal to treat the product as if it had never been bought.

As we aggregate data over a month, provided that the product is bought and subsequently refunded in the same month (where the data are included in construction of the price indices; see section on time coverage), then the refunded product will not be included in the aggregate sales and quantity figures. However, if the product is bought in one month, and refunded in a subsequent month (or in a period where we have not been able to use the data, because of the week falling in two consecutive months for example) this could be problematic for the construction of consumer price indices.

In some retailers' data, refunds and sales for each product are displayed in separate rows, allowing us to easily identify when products have been refunded. In other retailers' data this is not the case, and sales and refunds are aggregated onto the same row. We have asked retailers to provide disaggregated sales and refunds where possible, but some retailers have been unable to separate out this information in their data.

We have investigated the prevalence of refunds for retailers who provided disaggregated sales and refunds information for different product categories. For food and drink categories (where we are primarily focussing our use of scanner data) the prevalence of refunds is typically extremely low, with categories predominantly showing less than 1% refunds as a percentage of total sales value and volume. However, for other categories, such as clothing and entertainment, we see refunds of up to 30% of the total sales value (although refunds as a percentage of sales volumes are typically lower at up to 15%, suggesting it is higher value products that are generally being refunded).

We have also produced indices for food and drink items using data for a single retailer who disaggregate their refunds into different rows. We compared these indices to those produced had the refunds been aggregated together with sales. This comparison showed that there were only differences in the indices to 14 decimal places, suggesting that the impact of refunds on indices for food and drink items for this retailer is negligible. We assume these patterns are generalisable to other grocery retailers given the dynamics of the grocery market and prevalence of refunds.

Based on this research, we do not think refunds will significantly hinder our ability to produce scanner data indices for food and drink categories, however, we will need to further understand the influence of refunds for other categories (where refunds are more prevalent) when we decide to increase our use of scanner data in these areas in future.

Notes for Section 3: Methods for processing scanner data:

1. Note that this framework and its findings are still under review.

2. The timeframe is limited so that if retailers recycle their SKU codes after a time of being unused, then new products with recycled SKUs are not automatically linked to entirely different products. This timeframe will be retailer-specific dependant on how often, if at all, they recycle their SKU codes.

3. Examples of the units of measurement (UoM) in the data include grams, kilograms, litres and packs.

4. Similar findings were presented in a report investigating the impact of coronavirus on the CPI by the Institute for Fiscal Studies in April 2020.

4 . Future developments

As discussed throughout this report there are several workstreams that need further consideration, including:

- ongoing data acquisition and quality assurance
- further development of classification techniques for scanner data as discussed in Classification of new data in UK consumer price statistics
- investigations into any bias caused from using a subset of data from within a month rather than using a full month (and mitigation of bias, if found) when calculating consumer price indices
- further development of record linkage methods to link product relaunches, including research into suitable thresholds for automatic and manual linking
- consideration of the impact of refunds beyond the scope of price indices for food and drink items

In the coming months we will test the impact of these different methods, as well as the inclusion of different discount types, across indices produced for several different retailers. Once we have classified data for several retailers to a consistent hierarchy, we will also be able to assess the impact of the inclusion of these data on our headline consumer price statistics. Our next progress update in autumn 2021 will likely include these impact analyses.

More broadly, we are considering how we identify and validate outliers in scanner data, as well as whether any imputation methods are needed for missing products, items and/or retailers. We are also continuing research into different index number methods for web-scraped and scanner data in consumer price statistics and are working with some external researchers on this topic. We will update users on these workstreams in due course. Our full development plan for alternative data sources can be found in our Consumer Prices Development Plan.

5 . Related links

Research and developments in the transformation of UK consumer price statistics: April 2021
Article | 6 April 2021
The second in a series of biannual articles to update users on our research to modernise the measurement of consumer price inflation in the UK.

Classification of new data in UK consumer price statistics
Article | 6 April 2021
Classification of new data in UK consumer price statistics A broad outline of the different methods used for classification of alternative data sources, including a more detailed set of methods that will be applied in the classification of grocery scanner data.

Research and developments in the transformation of UK consumer price statistics
Article | Released 1 September 2020
The first in a series of biannual articles to update users on our research to modernise the measurement of consumer price inflation in the UK.

Introducing alternative data sources into consumer price statistics
Articles
Our plans to include alternative data sources into the production of consumer price statistics.

Consumer price inflation, UK statistical bulletins
Bulletin | Monthly
Our plans to include alternative data sources into the production of consumer price statistics.

Consumer Prices Indices Technical Manual, 2019
Methodology | Published 18 September 2019
This technical manual is a reference tool for anyone wanting to understand how measures of consumer price inflation and associated indices are compiled.