Date trimming for consumer prices alternative data sources

Exploration of the theory behind date trimming and its implementation in the calculation of consumer price statistics in the UK, using new data sources for grocery products.

Contact:
Laura Christen
cpi@ons.gov.uk
+44 1633 456900

Release date:
26 July 2023

Next release:
To be announced

Table of contents

1 . Main points

- Alternative data sources, and methods to use these data sources, are being introduced into consumer price statistics, as detailed in our Transformation of consumer price statistics article series; for example, new data for rail fares were first introduced in March 2023.
- In this article, we explore how date trimming, which is where indices are compiled from data using specific dates within a month, may be used with grocery scanner data, as first discussed in Research into the use of scanner data for constructing UK consumer price statistics.
- We present some preliminary analyses, which show that most products have the same, or very similar price, when averaged over three or four weeks.
- We plan to follow up this article with further analyses looking at the impact on our indices as well as a final decision on trimming.

2 . Overview

In traditional consumer price statistics, we typically measure inflation through point-in-time price collection, where prices are collected for most products once a month. An advantage of alternative data sources is to use information beyond a single day to give a better representation of the average transaction price paid by the consumer throughout the month. However, we must decide on which days to use each month when calculating representative prices used in our monthly price indices.

From a methodological perspective, ideally, we would calculate representative prices using every day of the month. However, from a practical perspective this may not be possible or preferred for the following reasons.

Some grocery retailer datasets are provided daily-aggregated, whereas some retailer datasets are provided weekly-aggregated. A limitation of weekly-aggregated datasets is that some weeks can overlap two consecutive months. Since we do not have daily information, it is difficult to separate these weeks and so including an entire month of data in every month is not possible. Note that Eurostat guidance (PDF, 942KB) advises that data pertaining to other months should not be included in the measurement month.

Within a monthly production round, it may be beneficial to only use an earlier portion of the month so that index compilation can begin earlier, giving more time for quality assurance.

Some retailers may only be able to provide data on a lag.

One method of overcoming these practical barriers could be date trimming. Date trimming involves filtering datasets down to a timeframe where transactions within that timeframe are in scope for measuring inflation, and where transactions outside of that timeframe are dropped from index calculations. For example, we could decide to use the first three full weeks of data for each month.

3 . Different scenarios for date trimming

Before introducing date trimming, we will research the impact of different scenarios on our indices. There are three scenarios we will consider:

- Scenario 1 is to use all days within the month to calculate the indices, for example, for July 2023, this would mean using all 31 days; it is worth noting that we will only be able to use this approach for retailers that provide daily-aggregated data

- Scenario 2 is to use all weeks falling fully into each month; this would mean using three or four weeks for each month dependent on how the days fall in each month, for example, for July 2023, this would mean using four full weeks covering 2 to 29 July

- Scenario 3 is to use a consistent, fixed timeframe every month; this would entail using the first three full weeks in each month to calculate indices and disregarding the rest of the data, for example, for July 2023, this would mean using three full weeks covering 2 to 22 July

Each of the scenarios has various advantages and disadvantages. We describe these with reference to the five quality dimensions set out in the Quality Assurance Framework of the European Statistical System (PDF, 916KB).

Accuracy

All three scenarios result in different amounts of data being used, which can affect accuracy. Scenario 1 uses all data available, resulting in no data loss. Scenario 2 involves a data loss of approximately 19% since we only use full weeks in each month. Scenario 3 results in the most data loss, as we lose approximately 31% of data over the year by using a fixed three-week time period.

Disregarding any amount of data may have the potential to introduce some bias to our indices. Note that this would be a concern if prices earlier in the month are systematically higher or lower than prices later in the month. This could potentially occur in months where large seasonal events such as Easter occur, or where weather events affect the weight of different products in the market. Further analysis is needed to ascertain whether indices could become biased because of date trimming.

Furthermore, the scenarios differ in the amount of time that they allow for quality assurance. We receive data on the same day of the week every week, meaning that Scenario 1 gives us reduced, and variable, time to scrutinise our indices before publication as we may have to wait multiple days between the end of the month and receiving the data. Scenario 2 also, in some cases, gives us reduced time to scrutinise, depending on where the last day of the month falls, whereas Scenario 3 provides us with the most time in between receiving the data and publication. We can mitigate this risk substantially by producing and scrutinising interim indices as we accumulate data throughout the month, giving additional time to quality assure the data.

Relevance

There is potential for Scenarios 2 and 3 to be slightly less relevant or interpretable to the end user. This is because our consumer price indices are interpreted as monthly indices, but Scenarios 2 and 3 will provide indices constructed on a smaller timeframe. However, as our preliminary analyses show, most products have a very similar price regardless of whether three or four weeks are used and so our current understanding is that indices are likely to be very similar between these scenarios. We also note that all scenarios provide greater time coverage than our traditional data sources.

Timeliness

None of the three scenarios will affect the punctuality or timeliness of the published indices.

Clarity

All three scenarios offer similar levels of clarity because the methods are easy to explain, and the format of the indices outputted from each scenario will remain the same.

Coherence and consistency

Scenarios 1 and 3 involve using a consistent amount of data for all retailers in all months. Scenario 2 involves using an inconsistent amount of data within retailers, since each retailer can be represented by a different number of days or weeks each month. This could affect the comparability of our indices across different months and years, because in one year a month could have three full weeks falling into it, and in another year have four full weeks.

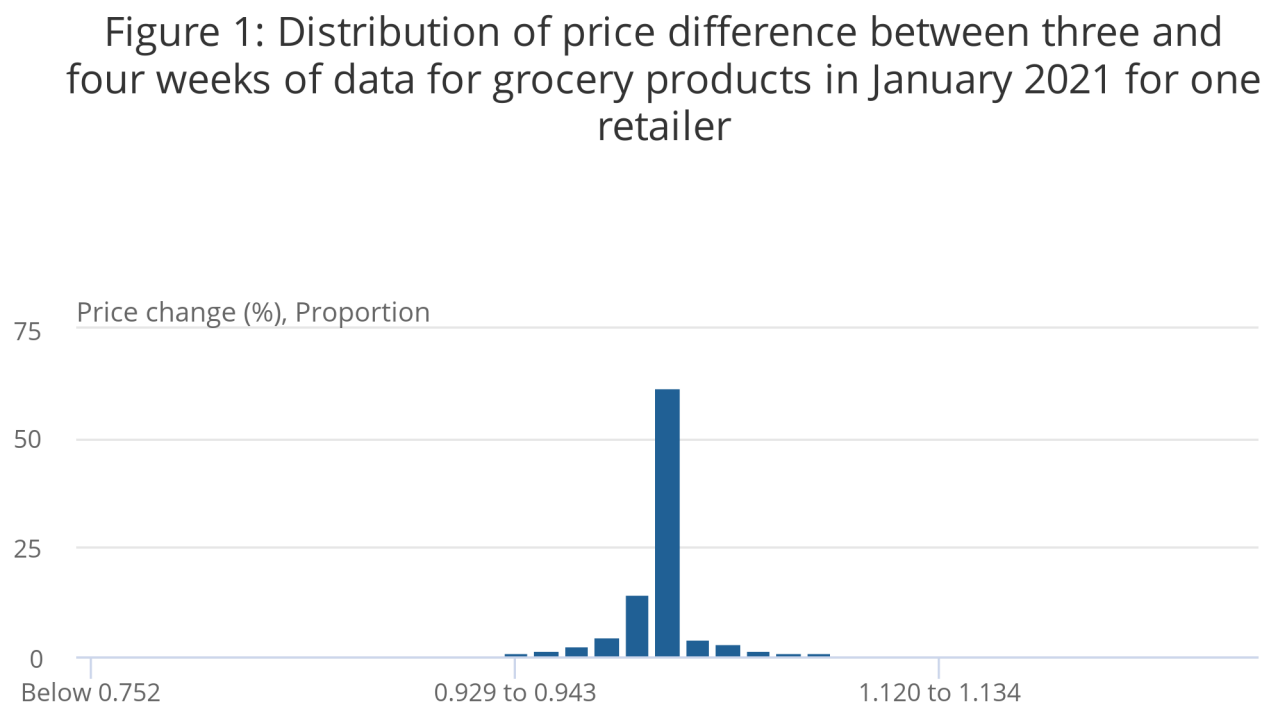
4 . Preliminary analyses

To examine whether representative prices calculated using three weeks of data would be similar to those calculated using four weeks of data, we calculated a ratio between these two prices by dividing the three-weekly price by the four-weekly price. A price difference ratio of 1 would indicate no difference in price, and a price difference ratio of 0.75 would indicate a 25% reduction in average price when calculating over three weeks compared with four.

As indicative analysis, we plotted the distribution of price difference ratios for all grocery products in one retailer in two seasonally different months, January and July 2021. The results are shown in Figures 1 and 2. As can be seen, the three-weekly and four-weekly prices are extremely similar for most products.

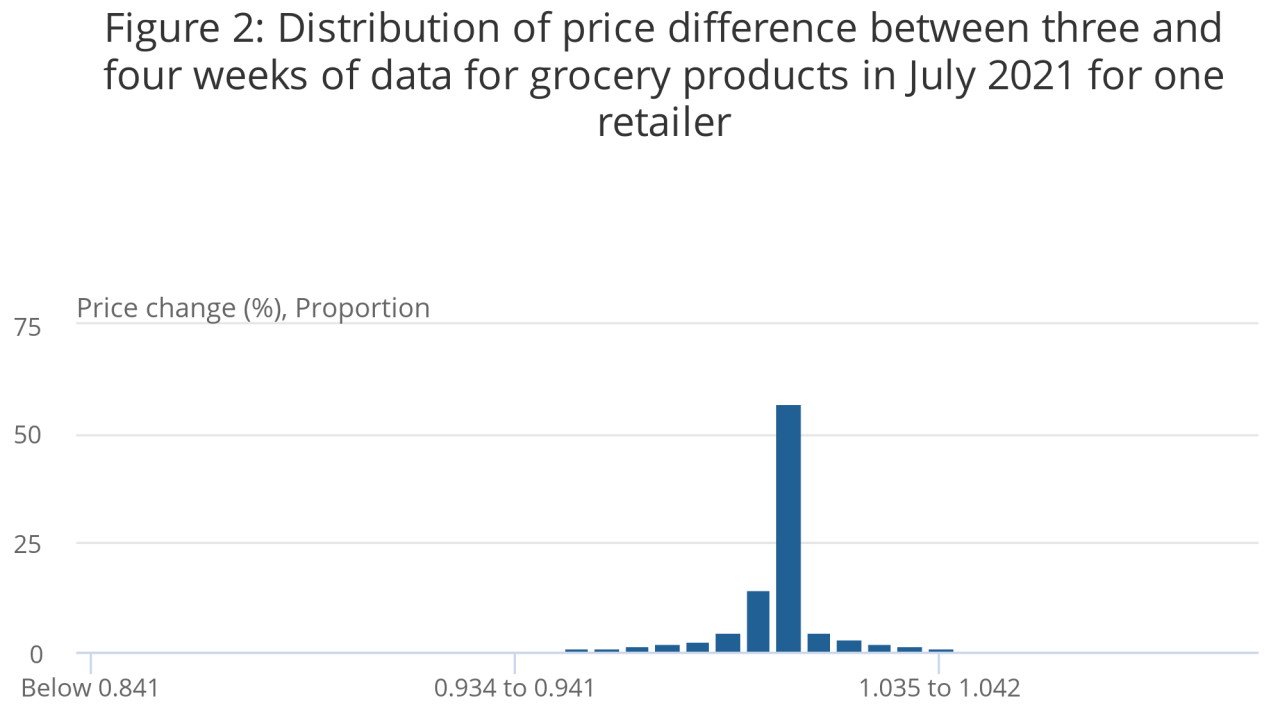
Furthermore, summary statistics of these differences can be seen in Table 1. The 10th and 90th percentiles show that 80% of products do not differ by more than around 2% in price when comparing three- and four-weekly prices.

Figure 1: Distribution of price difference between three and four weeks of data for grocery products in January 2021 for one retailer



Source: Office for National Statistics

Figure 2: Distribution of price difference between three and four weeks of data for grocery products in July 2021 for one retailer



Source: Office for National Statistics

Table 1: Descriptive statistics for the price difference between three-weekly and four-weekly prices for grocery items in 2021

Measure	Jan 2021	July 2021
Mean	1.002	0.9995
S.D.	0.038	0.021
1st percentile	0.909	0.938
10th percentile	0.982	0.985
Median	1.000	1.000
90th percentile	1.021	1.013
99th percentile	1.132	1.059

Source: Office for National Statistics

5 . Future developments

Our preferred option is to maximise the use of data through using Scenario 1 for retailers who provide daily-aggregated data and Scenario 2 for retailers who provide weekly-aggregated data. This is on the basis that doing so does not introduce an unacceptable level of risk because of the reduction in scrutiny time in the monthly production round. Scenarios 2 and 3 will be considered for all retailers if monthly production round timescales are considered a concern.

For future work, we plan to finalise our timescales for the monthly production round and study the impact on indices of using the various scenarios, considering the interaction of this timing with our data processing timelines each month. We then plan to publish an updated article with these analyses and a final decision on whether, and how, to trim the data, ahead of introduction of grocery scanner data in 2025.

6 . Related links

Research and developments in the transformation of UK consumer price statistics: July 2023
Article | Released 26 July 2023
Research to modernise the measurement of consumer price inflation in the UK: sixth in a series of biannual articles to update users.

Using Auto Trader car listings data to transform consumer price statistics, UK
Article | Released 26 July 2023
Car listings data will improve measurement of consumer prices from 2024. This article updates our methods and research indices using these data.

Transformation of consumer price statistics: July 2023
Article | Released 6 July 2023
We are undertaking a programme of transformation across our consumer price statistics, including identifying new data sources, improving methods, and developing systems to improve both the Consumer Prices Index including owner occupiers' housing costs (CPIH) and the Consumer Prices Index (CPI).

Consumer price inflation, UK: June 2023
Bulletin | Released 19 July 2023
Price indices, percentage changes, and weights for the different measures of consumer price inflation.

Research into the use of scanner data for constructing UK consumer price statistics
Article | Released 6 April 2021
Research into using scanner data provided directly from UK retailers to integrate with other data sources in producing UK consumer price statistics.

Consumer Prices Indices Technical Manual, 2019
Methodology | Released 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.

7 . Cite this methodology

Office for National Statistics (ONS), released 26 July 2023, ONS website, methodology, Date trimming for consumer prices alternative data sources