Office for National Statistics

Outlier detection for grocery scanner data in consumer price statistics

We are applying data cleaning techniques to grocery scanner data to remove out-of-scope observations and errors when calculating our consumer price indices. Our strategy also identifies dump prices and their effects on the grocery indices.

Contact:
Mario Spina
cpi@ons.gov.uk
+44 1633 456900

Release date:
1 December 2023

Next release:
To be announced

Table of contents

1 . Main points

- We plan to transform consumer price statistics using grocery scanner data from 2025.
- As the new data are too large to manually assure the quality of individual quotes, we intend to apply data cleaning techniques to remove out-of-scope observations, erroneous prices, and end-of-lifecycle products.
- In this article we describe the refinements we are looking to make to our data cleaning strategy when processing grocery scanner data.
- Based on the analysis we present, we propose using a combination of methods for the grocery scanner data, that together appropriately identify and remove extreme monthly price variations and end-of-lifecycle products.

2 . Background to using data cleaning on grocery scanner data

Bigger, alternative data sources, and the methods to use these data sources, are being introduced into our consumer price statistics. As part of our wider transformation plan outlined in our Programme of transformation across UK consumer price statistics, we are looking to introduce grocery scanner data into consumer price statistics from 2025.

In our previous publication on data cleaning, Outlier detection for rail fares and second-hand cars dynamic price data, we presented a variety of different outlier detection strategies, recommending the removal of observations where prices have either increased or decreased more than threefold since the previous month. We describe this method as "price relative fences". We introduced alternative data for rail fares into consumer price statistics in 2023 using this method, and will also use this method for second-hand cars in 2024.

In this article we outline our approach to data cleaning with grocery scanner data. We discuss why we are looking to introduce a targeted filter to remove end-of-lifecycle "dump price" products, alongside the price relative fence method that we have already explored and implemented.

Data cleaning determines the observations within the data that will be used to construct our indices. The main aim is to remove out-of-scope observations and errors that would likely have an undesirable effect on the overall quality of our indices. The overall data cleaning is a two-step approach:

- junk filtering uses variables within the dataset to determine observations that are not in scope and therefore should be removed before index production; the filters applied are pre-defined and are specific to the product category
- outlier detection is used to identify and remove products showing extreme prices or price changes, or extreme numbers of products sold or changes in numbers of products sold, that could potentially indicate either a data error or a product at the end of its lifecycle

For grocery scanner data, we plan to use junk filtering to remove the following types of observations:

- products sold by weight, such as loose carrots, where we do not have the weight of products sold
- those with erroneous or incomplete data for the size or unit of measurement, for example, where different variables have conflicting information
- those which are not genuine sales, for example, products being processed for use in store
- those we cannot link to a UK region
- those without a suitable product identifier

Our price relative fences outlier detection strategy we previously published focused on identifying extreme monthly changes in price, by defining lower and upper fences, flagging price changes outside these fences, then choosing whether to use these flagged observations for calculating the price index. The starting point for the current analysis therefore consists of using the chosen lower and upper fences of a third and three, respectively, meaning if a price increases or decreases more than threefold in the space of a month, these observations will be flagged and removed.

In this article we explore two additional strategies of relative-based outlier detection:

- price-quantity relative outlier detection, which focuses on simultaneous extreme changes in both monthly prices and quantities (number of products sold), represented as $r^{p}_{t-1,t}$ and $r^{q}_{t-1,t}$. $r^{p}_{t-1,t}$ and $r^{q}_{t-1,t}$ are referred to as RP and RQ in Table 1, Table 3 and Table 4.
- the combination of price and price-quantity relative outlier detection methods

These methods are summarised in Table 1.

Table 1: The fence-based methods to be explored in our research

Method	Abbreviation	Keep row if…
Price relative fences	p-dump	RP in [Lp, Up] (E1)
Price-quantity relative fences	pq-dump	RP in [Lp, Up] OR RQ in [Lq, Uq] (E2)
Price and price-quantity relative fences	combined	(E1) AND (E2)

Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Notes

1. Lp Lq (Up Uq) are the lower (upper) fences for price and quantity relatives respectively.

The goal of the price-quantity fences is to attempt to identify and remove prices at the very end of their lifecycle that might affect the index calculation.

Note that alternatives to fence-based approaches were considered, and rejected, within our previous outlier detection article. More about this can be read within Outlier detection for rail fares and second-hand cars dynamic price data.

3 . Why we propose removing dump prices in grocery scanner data

Dump prices occur at the end of a product's lifecycle when the retailer sells the remaining stock of a product at a very low clearance price. Dump prices can be identified by a large price drop alongside a large drop in the number of products (quantities) sold.

While typically we would expect to see an increase in quantity sold following a large decrease in price as consumers switch towards this product, in cases where the product is at the end of its lifecycle, we often observe a steep decrease in quantities sold. This would happen because stocks are not replenished and therefore only a limited amount of stock of the product are still available to consumers. An example are various Christmas chocolate box products, which often have high volumes of sales in December but then exhibit dump price behaviour in January. Another example includes mince pies.

Dump prices are considered separately to "clearance sticker products" where some prices are reduced for a continuing product line because of nearing their expiry date. We also aim to exclude these clearance prices from the data as there is a difference in quality compared with the full-price products, but these are excluded at the point the data are compiled and delivered by the retailer, rather than through outlier detection methods.

Dump prices are particularly common in the grocery market since retailers need to make shelf space for new product lines and need to sell perishable end-of-lifecycle products. Retailers can drop prices rapidly to encourage end-of-line sales and avoid completely missing out on a sale. It is recommended by international guidance to remove dump prices from index calculation, as they could have an undesirable influence on the final index.

We have chosen to use a GEKS-Törnqvist index with alternative data sources. Similar to the underlying Törnqvist index which it is based upon, a limitation of this method is that it can be sensitive to dump prices, because it gives an overly high weight to the price decrease of the dump product as a result of the much larger expenditure shares of that product in past period(s).

An extreme example of how the Törnqvist (and by extension, the GEKS-Törnqvist) index is affected by dump prices is given in Table 2. The second product drops substantially in price from £3 to £0.50, but this sale price was not widely available to the average consumer because of clearance behaviour, and only one consumer was able to benefit from it. Note that the method for calculating the Törnqvist index is explained in Section 4 of Introducing multilateral index methods into consumer price statistics. However, the Törnqvist calculation gives product 2 a non-trivial weight since it had a high weight in the base period, January. This results in the index dropping to 0.64, suggesting a large drop in prices, despite very few people experiencing price savings. Product 2 is considered a "dump price" and our methods are targeting removing these products.

Table 2: The Törnqvist index method can be sensitive to dump prices.

Product	Price, Jan	Price, Feb	Quantity, Jan	Quantity, Feb
1	3	3	10000	10000
2	3	0.5	10000	1
			Törnqvist	0.6389

Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Note that the drop in price for product 2, with a price relative of 0.1667, would be removed by the price relative fences method, independently of changes in quantity. Therefore, the approach we developed for second-hand cars and rail fares data, when applied to grocery data, would already provide some protection against the most extreme dump prices. However, we will show how we may look to combine a price fence with a price-quantity fence in Section 4: Analyses of outlier detection, to provide additional protection against dump prices.

4 . Analyses of outlier detection

Analysis overview

This section covers our outlier detection analyses on grocery scanner data. It is broken down into several sub-sections:

- Methods explored: covers the candidate outlier detection methods we are exploring
- Indices analysis: explores the impact that our outlier detection methods have on our indices, both at high-levels of aggregation, and on particularly impacted lower-levels of aggregation
- Consumption segment analysis: explores dump price behaviour through distributional analysis in one of our most highly-impacted consumption segments, "Chocolate assortment"
- Seasonal dump prices analysis: investigates the seasonal distribution of our outlier detection methods, showing how most of the outliers we detect are likely caused by dump price behaviour

Methods explored

Table 3 summarises the fence-based methods explored in this article and defines the exact thresholds being considered. Note that the price-quantity methods only use a lower threshold, resulting in asymmetric use of fences. This choice allows us to target dump prices.

The "combined" methods apply simultaneously the price and price-quantity dump filters, so for example, the method "combined 1" is a combination of "p-dump 3" and "pq-dump 0.01", where transactions are kept when [0.3334 $r^{p}_{t-1,t}$ 3] AND [0.5 $r^{p}_{t-1,t}$ OR 0.01 $r^{q}_{t-1,t}$]. The table also shows the percentage of transactions removed both in terms of expenditure and row counts for the whole grocery category.

Table 3: The fence-based methods explored in our research and the percentage of expenditure and rows flagged.

Fencing method	Abbreviation	Keep row if…	% removed: expenditure	% removed: rows
No outlier detection	benchmark	All rows kept	NA	NA
Price	p-dump 3	0.3334 RP 3	0.00852%	0.01671%
	p-dump 4	0.25 RP 4	0.00295%	0.00729%
Price-quantity	pq-dump 0.01	0.5 RP OR 0.01 RQ	0.00015%	0.00082%
	pq-dump 0.1	0.5 RP OR 0.1 RQ	0.00131%	0.00577%
Price and	combined 1	p-dump 3 AND pq-dump 0.01	0.00860%	0.01708%
price-quantity	combined 2	p-dump 3 AND pq-dump 0.1	0.00949%	0.02038%
	combined 3	p-dump 4 AND pq-dump 0.01	0.00308%	0.00781%
	combined 4	p-dump 4 AND pq-dump 0.1	0.00414%	0.01194%

Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Table 3 shows that none of the outlier strategies investigated remove a large fraction of the total dataset. When looking at the combined methods, "combined 1" and "combined 2" remove a very similar percentage of transactions. This happens because most of the flagged transactions are identified by the p-dump 3, which seems to potentially remove too many rows. Given that large price reductions are more common for groceries than other product categories to which we have applied similar methods, such as rail fares, we would prefer using less stringent criteria for the groceries category. The last two combinations instead seem to have a more significant contribution of the price and quantity filters.

Note that we describe "no outlier detection" as a benchmark, not because it is a preferred index, but because it represents a baseline comparison to evaluate how the other outlier detection strategies affect the indices.

Indices analysis

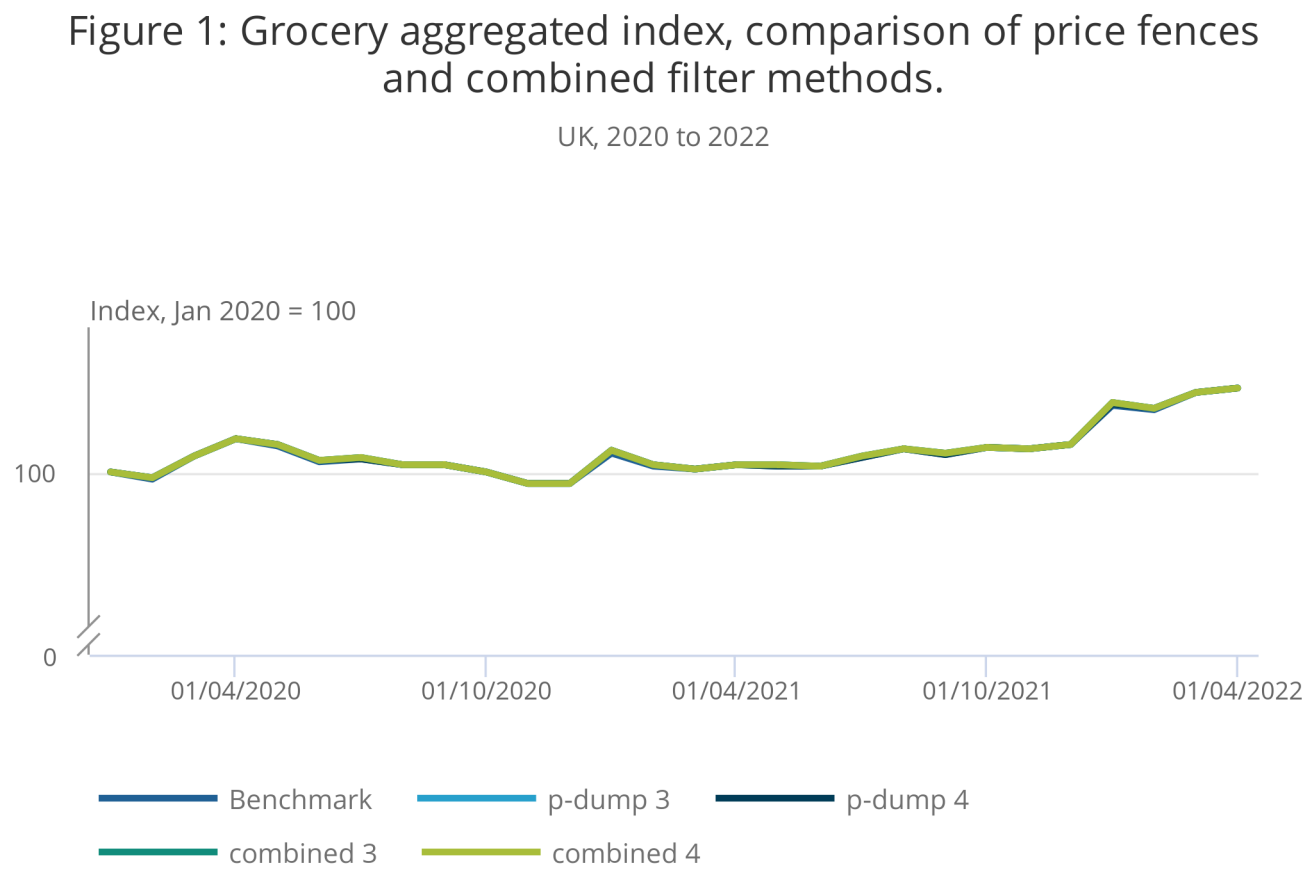
In Figure 1 and Figure 2 we show, respectively, the indices trends for different outlier detection methods in the time span considered for the overall grocery index and the difference between the outlier detection methods and no-outlier benchmark.

The two figures show the impact of the two price fences and the combination of the price-quantity dump filters with a price filter, with price fences of [0.25, 4]. These methods are called "combined 3" and "combined 4" according to the definition in Table 3. The figures show that the overall impact of the outlier detection methods explored is to marginally increase the index in some months, and have a negligible impact in others, compared with the case when no outlier detection methods are applied. This is true for all methods shown in Figure 1 and Figure 2.

It is likely that most of the transactions removed have a price relative below 1, which might suggest those transactions are because of dump prices. This explains the increase of the index observed. In Figure 2, it can be observed that the largest, albeit still mild, differences between indices happen in January and May, with negligible differences in other months. This may be caused by retailers clearing their remaining Christmas or Easter stocks, causing seasonal dump prices. We also observe that p-dump 3 has a similar impact on the index compared with combined 4, which shows the largest difference with the benchmark index.

Figure 1: Grocery aggregated index, comparison of price fences and combined filter methods.

UK, 2020 to 2022

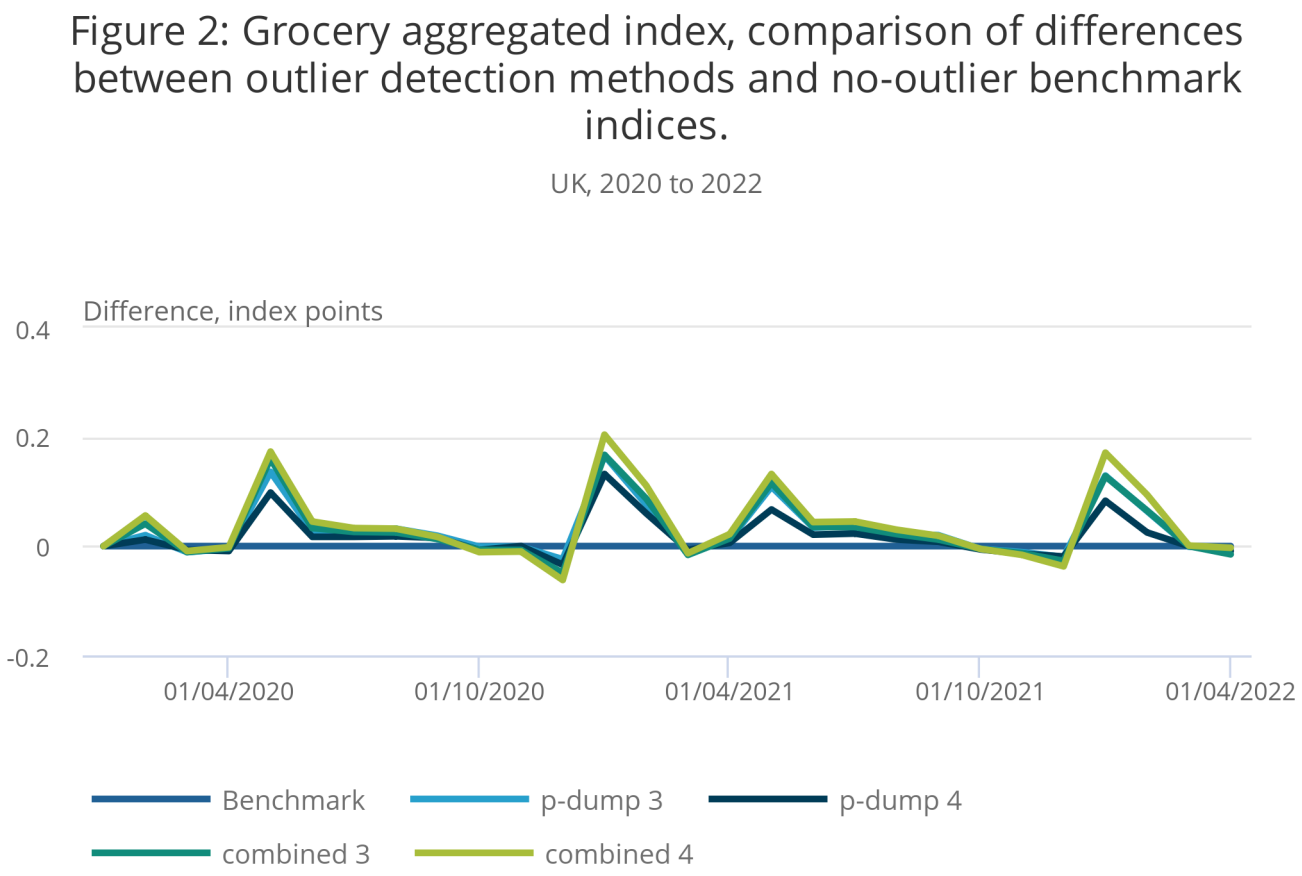


Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

This observation reinforces the idea that the outliers flagged in those months are likely to be dump prices. We prefer not considering the p-dump 3 filter, as it might reduce the impact of dump prices filters. The remainder of the analysis aims at selecting which of the three remaining strategies presented in Figures 1 and 2 is to be used, although, based on the results presented in the continuing of the article, we anticipate the use of the “combined 4” method.

Figure 2: Grocery aggregated index, comparison of differences between outlier detection methods and no-outlier benchmark indices.

UK, 2020 to 2022

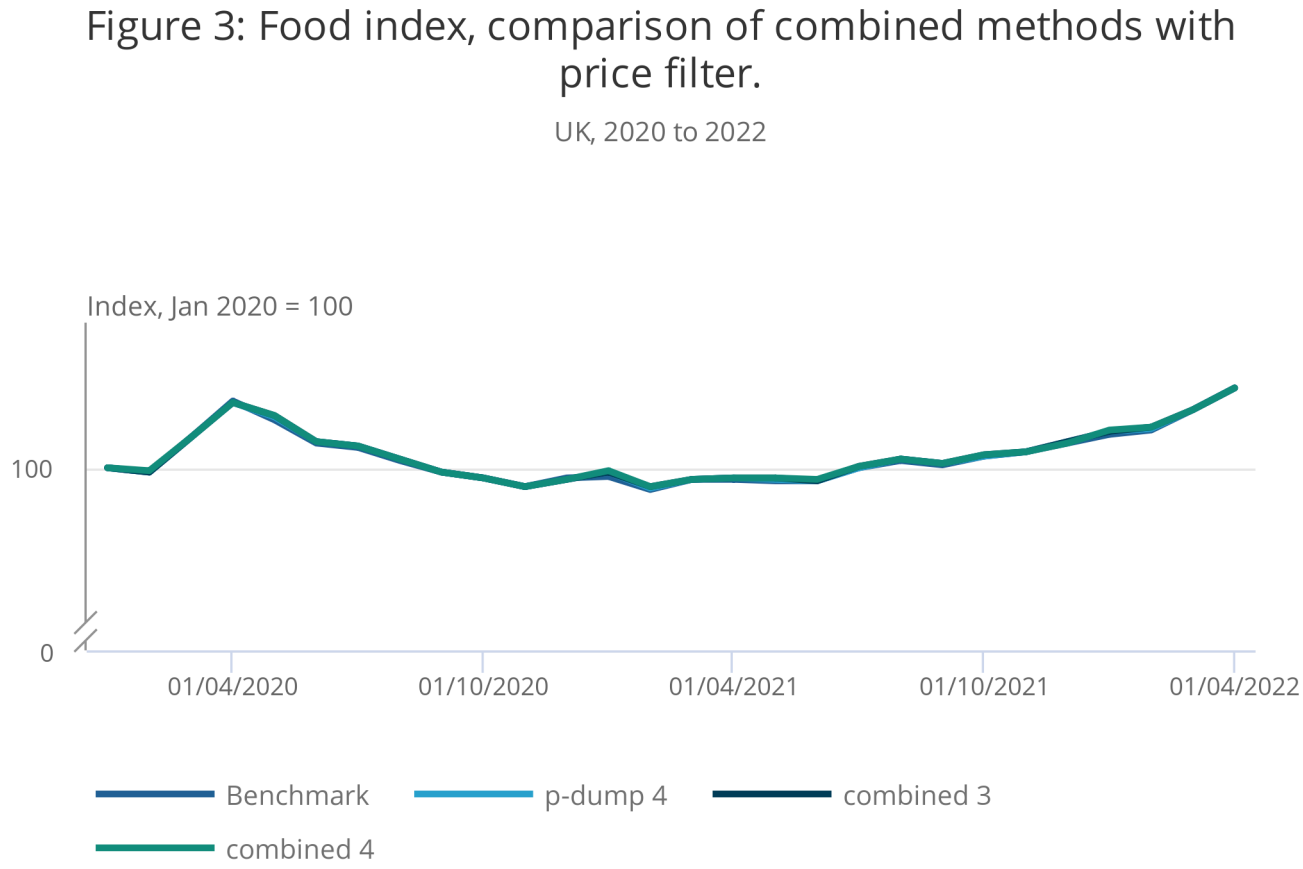


Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

The grocery category consists of over 130 consumption segments across four Classification of individual consumption according to purpose (COICOP3) categories (food, alcoholic beverages, non-alcoholic beverages, tobacco). Looking at the impact of each method for every consumption segment would be time-consuming. To optimise the decision process, we investigated only those categories and consumption segments for which the difference between the outlier index and the benchmark index is relatively large, that is, larger than the 0.1 index points. Following this definition, Figure 3 and Figure 4 show the food COICOP3 category, which presents the largest difference among the four COICOP3 categories.

Figure 3: Food index, comparison of combined methods with price filter.

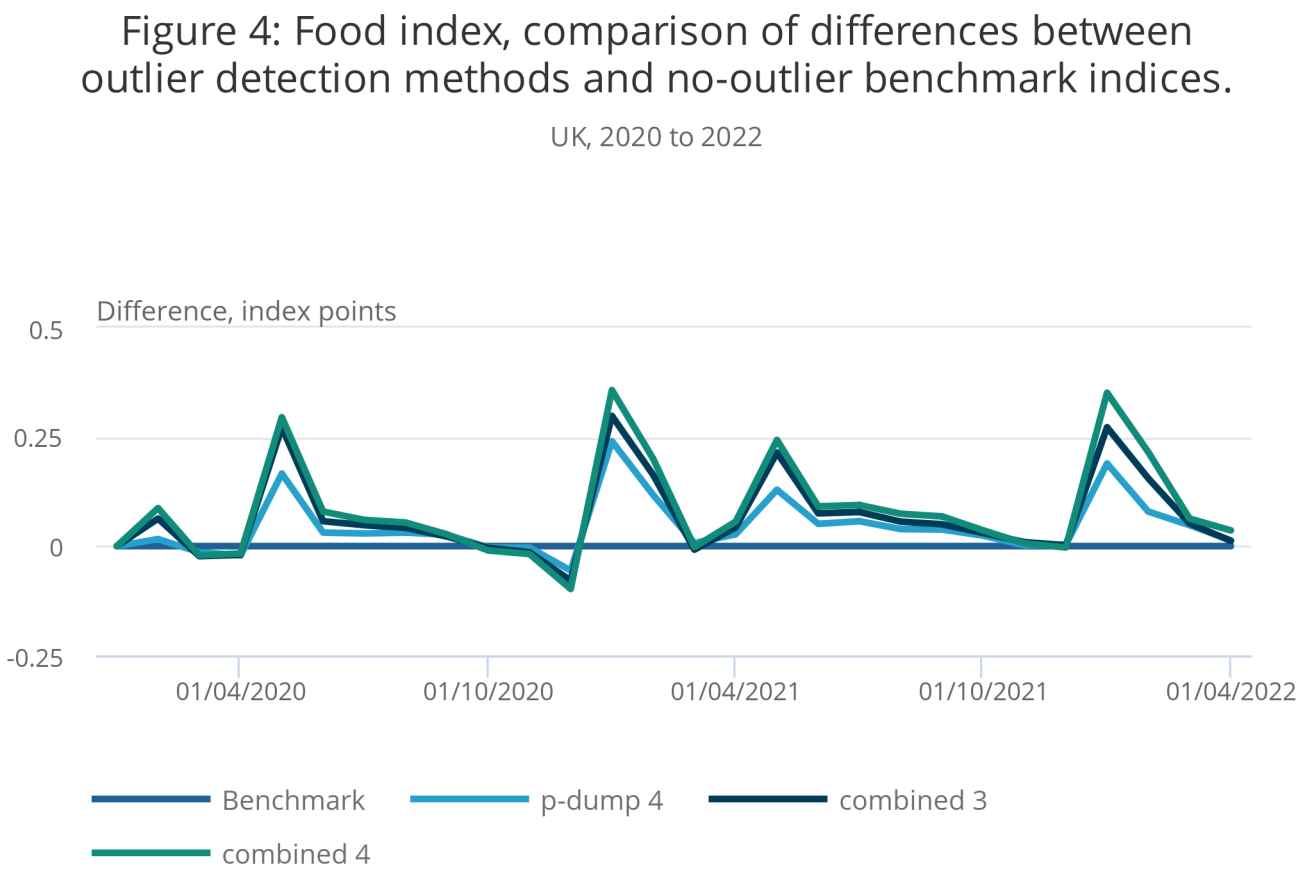
UK, 2020 to 2022



Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Figure 4: Food index, comparison of differences between outlier detection methods and no-outlier benchmark indices.

UK, 2020 to 2022



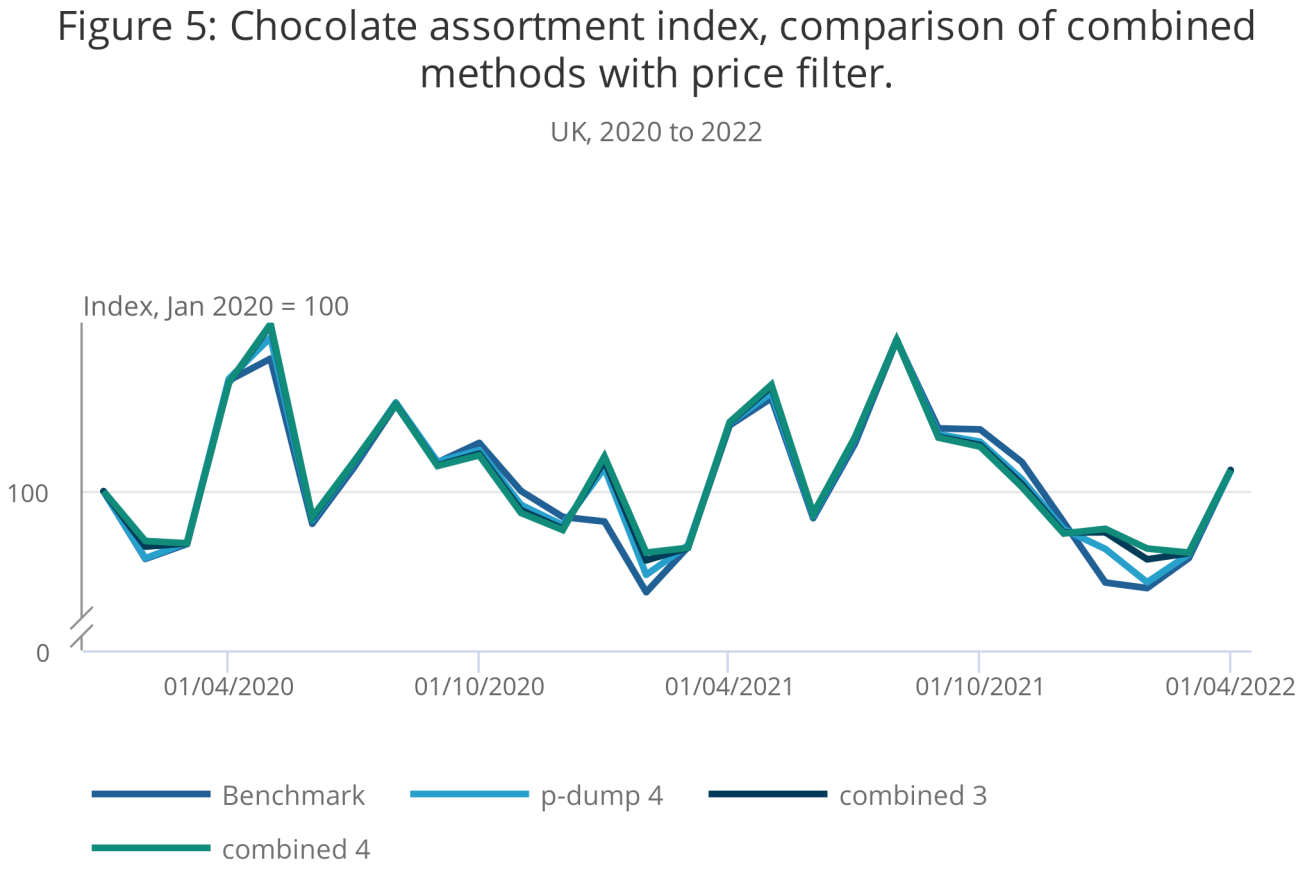
Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Figure 3 and Figure 4 show the same seasonal trends as Figures 1 and 2, but with larger differences, leading to the conclusion that the food category is the largest contributor to the differences observed. This is expected, given that the food category consists of 102 consumption segments.

To explore our methods in lower-level detail, we ranked the 102 consumption segments within the food category by the largest differences caused by outlier detection, and found that 80 have a difference larger than 0.1 index points. In Figure 5 and Figure 6 we show the index for the consumption segment with the largest difference, "Chocolate, assortment (for example, selection box)".

Figure 5: Chocolate assortment index, comparison of combined methods with price filter.

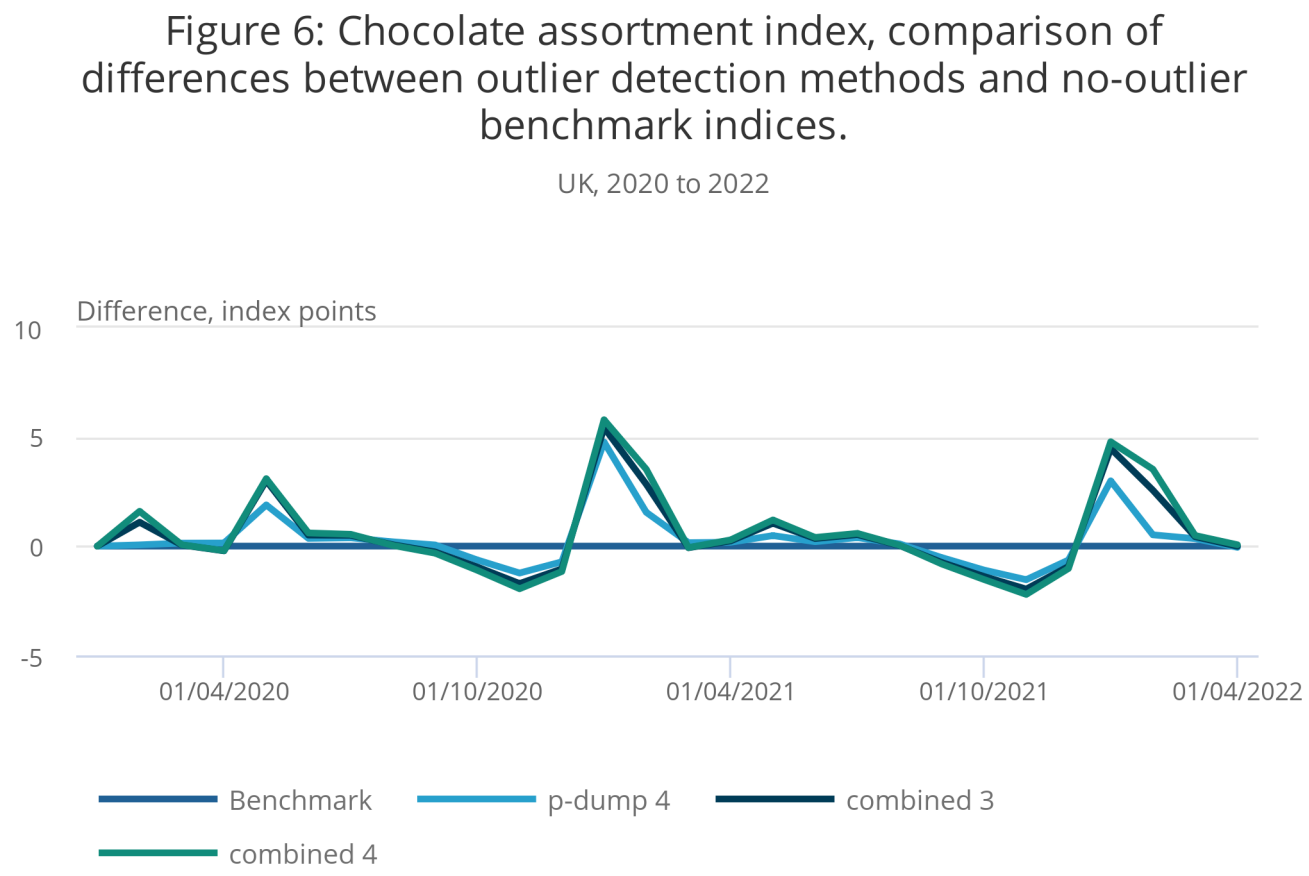
UK, 2020 to 2022



Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Figure 6: Chocolate assortment index, comparison of differences between outlier detection methods and no-outlier benchmark indices.

UK, 2020 to 2022



Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Similar to our headline grocery comparisons, Figure 5 and Figure 6 show that "Chocolate, assortment" follows similar seasonal patterns, with large differences caused by outlier detection occurring post-Christmas and Easter. However, in this case the difference is much larger, going up to six index points. This is unsurprising since it is very common for retailers to stock seasonal chocolates during Christmas or Easter, which are steeply discounted and cleared post-season. It is interesting to note that the two combined indices behave roughly in the same way across the 28-month period explored.

Consumption segment analysis

In the following we present some low-level analysis based on the consumption segment "Chocolate assortment" and the products therein. This analysis aims at determining which of the combined methods is to be preferred.

Figure 7 shows the distribution of quantity relatives (x-axis) and price relatives (y-axis) from the selected consumption segment covering January 2020 to April 2022. The dark pink lines represent the "combined 4" strategy from Table 3, where there is a combination between the price filter with fences of [0.25, 4] and the price-quantity filter with fences $r^{p}_{t-1,t}$ 0.5, $r^{q}_{t-1,t}$ 0.1. The transactions outside those lines are not used for calculating the inflation index, as they are flagged as outliers. The dark green vertical and horizontal lines show the boundaries of the two individual components of the combined filter. Note that the figure has logarithmic axes.

Figure 7: Two-dimensional distribution of quantity and price relatives with flagging selection

Source: Office for National Statistics – Outlier detection for grocery scanner data in consumer price statistics

Since Figure 7 shows a very large number of points, the observations identified as outliers might seem more abundant than they really are. To avoid this, the black contour lines give a sense of the densities of price and quantity relative points:

- the innermost contour contains 20% of the data
- the second innermost contour contains 50% of the data
- the third innermost contour contains 80% of the data
- the final contour contains 95% of the data

Figure 7 shows how the price relative fences method alone can remove a lot of the suspected dump prices transactions: those with a larger price change. The combination of the price and price-quantity filters allows us to select more dump prices, located in the rectangular region between the pink and dark green horizontal lines in the bottom left of the plot.

As mentioned, this consumption segment is the one with the largest difference between any outlier detection method and the benchmark. This can also be observed by the large percentage of flagged transactions, reported in Table 4, compared with the aggregated figures reported in Table 3. For example, for the "combined 4" filter considered in Figure 7, the percentage of expenditure removed is approximately 10 times larger than for the aggregated grocery case. In any case, even at the consumption segment level, more than 99% of expenditure and 95% of rows are kept for index calculation.

Table 4: Impact of different fence-based methods on the percentage of expenditure and rows flagged for the chocolate assortment consumption segment index

Fencing method	Abbreviation	Keep row if…	% removed	
			expenditure	rows
Price	p-dump 4	0.25 RP 4	0.0181%	1.9491%
Price-quantity	pq-dump 0.1	0.5 RP OR 0.1 RQ	0.0121%	3.1202%
Price or price-quantity	combined 4	p-dump 4 AND PQ-dump 0.1	0.0290%	3.7516%

Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

As Figure 7 shows, most of the outliers have a price relative below 1, indicating products with a price reduction. This observation further justifies that the index calculated including outlier detection, for example in Figure 5, is in general higher than the benchmark index, where no outlier detection methods are applied.

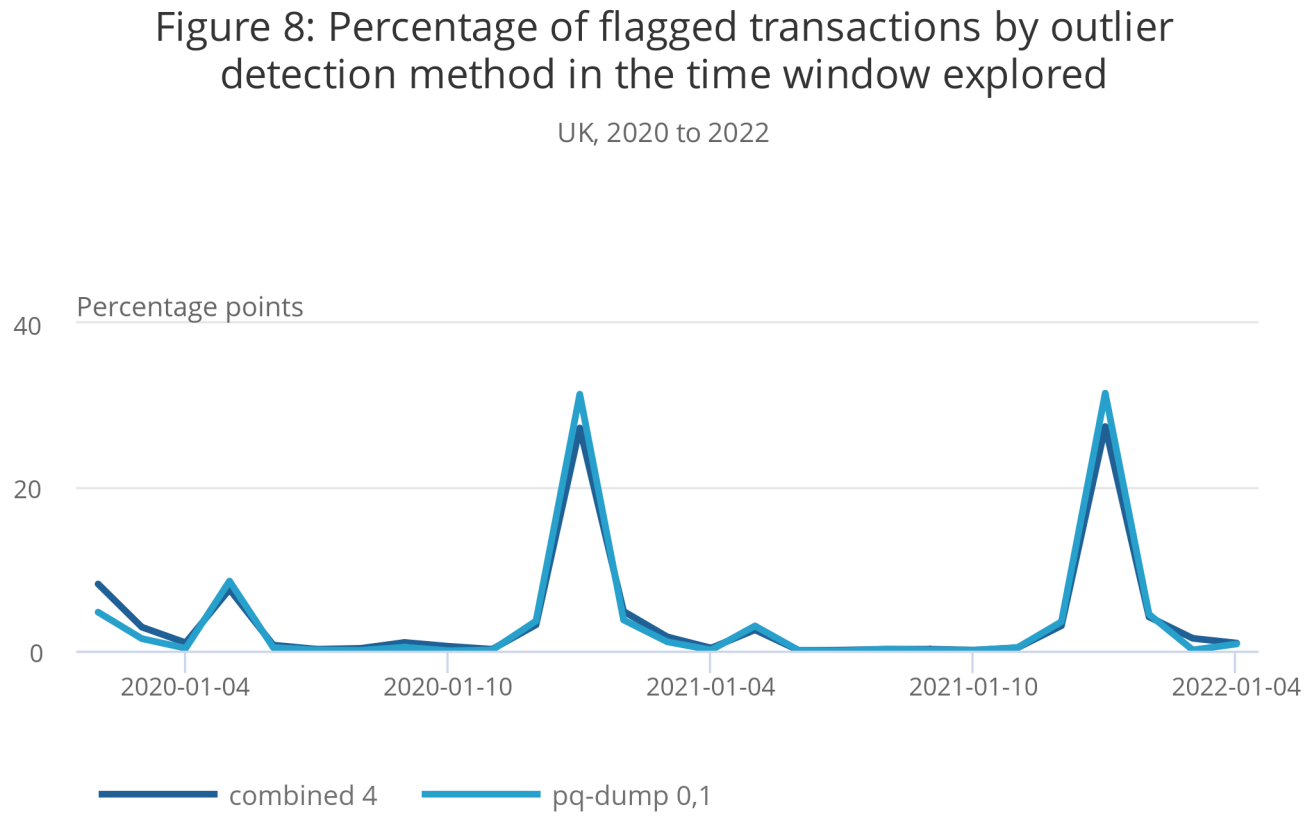
Seasonal dump prices analysis

The scope of this part of the analysis is to evaluate the impact of dump prices on the total flagged outliers. If outliers were to be mostly caused by price errors, we might not expect to find seasonal patterns in the frequency of outliers.

However, Figure 8 shows extreme seasonality when outliers are observed using the pq-dump 0.1 and combined 4 outlier detection methods over the 28 months studied. Around 60% of outliers are found in January or May. This observation suggests that most of the outliers identified by our strategy are because of seasonal dump prices rather than random errors.

Figure 8: Percentage of flagged transactions by outlier detection method in the time window explored

UK, 2020 to 2022



Source: Outlier detection for grocery scanner data in consumer price statistics from the Office for National Statistics

Notes:

1. pq-dump 0,1 in the label corresponds to pq-dump 0.1 in the text.

We performed several spot checks of products near the border of the pq-dump 0.1 filter shown in Figure 7. These checks confirmed that most of the products were indeed seasonal items being dumped out of market, such as advent calendars or Christmas chocolate boxes. This observation, together with all the other discussed, gives us confidence that the combined 4 strategy is the best one for grocery scanner data.

5 . Final outlier detection method and impact

Our preferred outlier detection method when introducing grocery scanner data is price relative fences of [0.25, 4], which will identify products where the representative monthly price increases more than four times and decreases to less than a quarter month-on-month. We propose using this filter in combination with a price-quantity filter with price relative fence of $r^{p}_{t-1,t}$ 0.5 and quantity relative fence of $r^{q}_{t-1,t}$ 0.1, which will identify additional dump prices caused by products reaching the end of their lifecycle and being removed from the market. These observations would then not be used for calculating our grocery indices.

We have several reasons to prefer this approach:

- it performed well in our case studies; we have given evidence to show that most of these observations appear to be dump prices
- compared with the previous results for second-had cars and rail fares, the thresholds for the price filter have widened, since we prefer price relative fences of [0.25,4] instead of [0.3334,3], because of the wider price change distributions that can occur within groceries
- the percentage of expenditure removed by the combined filter (0.00344%) is in line with what was observed for second-hand cars and rail fares, based on a price relative filter with fences [0.3334, 3]
- it is a straightforward method that is easy to explain and understand
- because of minimal calculations and aggregations, it also scales well with data size, limiting costs, environmental impacts, and ensuring faster runtimes

The impact of the strategy on the aggregated grocery index was shown previously in Figure 1 and Figure 2. Outlier detection seems to be driven (mostly) by the removal of dump prices, therefore correcting for a mild downward bias caused by dump prices, with particular impacts in January and May, when a lot of dumping behaviour may be expected post-Christmas and post-Easter. At the most aggregated level, the method does not change the index by more than 0.2 index points.

The strategy can have larger impacts on some lower-level aggregates. For the Classification of individual consumption according to purpose (COICOP3) categories, the largest difference is observed in the food category. Figure 3 and Figure 4 show that in this case the method does not change the index by more than 0.35 index points, while showing a clear seasonal structure because of the removal of seasonal dump prices.

As expected, there are some consumption segments with much larger impacts, particularly those with a lot of seasonal products. Out of 102 consumption segments for food categories, 80 show a difference between the benchmark index and the index obtained by applying the outlier detection strategy larger than 0.1 index points. The largest difference, of up to six index points, is observed for the "Chocolate, assortment (for example, selection box)", which is also used to perform the seasonality study.

The seasonality studies presented in this article shows a link between the month and the percentage of outlier flagged, reinforcing the fact that outliers are mostly originating from dump prices. The case study presented in Figure 8 shows that the outlier detection strategy proposed can identify and remove from the index calculation seasonal dump prices.

6 . Future developments

We plan to introduce alternative data sources for groceries in 2025, according to our programme of transformation across UK consumer price statistics. We plan to publish an analysis of the impact of including this data in our indices at the end of 2024.

In the future, we may explore applying outlier detection at the transaction price level, rather than at the representative price level. This may allow us to only remove the outlier transaction(s) without completely removing the product within the month. Note that since our outlier detection methods do not remove many observations, it is likely that this will only offer a mild improvement.

7 . Related links

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).

Outlier detection for rail fares and second-hand cars dynamic price data
Methodology article | Released 28 November 2022
We are applying data cleaning techniques to web-provided and transactio data to remove out-of-scope observations and errors when calculating our consumer price indices.

Guide on Multilateral Methods in the Harmonised Index of Consumer Prices
Manuals and Guidelines | Eurostat | 2022 edition

Introducing multilateral index methods into consumer price statistics
Methodology article | Released 28 November 2022
How we will use the GEKS-Törnqvist to introduce alternative data into our consumer price statistics, including how the method works and its advantages.

8 . Cite this methodology

Office for National Statistics (ONS), released 1 December 2023, ONS website, methodology, Outlier detection for grocery scanner data