Outlier detection for rail fares and second-hand cars dynamic price data

We are applying data cleaning techniques to web-provided and transaction data to remove out-of-scope observations and errors when calculating our consumer price indices.

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

Next release:
To be announced

Table of contents

1 . Main points

- We plan to introduce new methods using alternative data sources from 2023; the first categories we intend to transform are rail fares and second-hand cars.

- In this article we describe different outlier detection methods and assess their impact on second-hand cars and rail fares price indices.

- Based on our results we recommend use of relative-based outlier detection with user-defined fences for use with our new data sources.

- The chosen method does not change the overall indices for second-hand cars and rail fares by any great amount.

2 . Background to using data cleaning methods

Our programme of transformation across UK consumer price statistics involves identifying and using new, bigger data sources. We plan to introduce alternative data sources for calculating inflation for second-hand cars (web-provided data) and rail fares (transaction data) in 2023. For more details of these data and their impact on headline consumer price statistics see our Impact analysis on transformation of UK consumer price statistics: rail fares and second-hand cars, November 2022. The estimates in the impact analysis are indicative. There are still some minor improvements being made to the new production systems for these data, though these are expected to have a minor impact on the figures presented. This includes updating the data cleaning methodology currently in use to that recommended by this research.

To deal with the unprecedented quantity of transactions, we are developing and improving our methodologies to ensure the data are of sufficient quality. We are therefore adapting existing data cleaning methods for the newly available data sources.

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.

There are two underlying components to this strategy.

- Junk filtering uses variables in the dataset to determine observations that are not in scope and should therefore be removed prior to index production. The filters applied are pre-defined and are specific to a goods category. For example, we remove motorcycles when calculating a second-hand cars index.

- Outlier detection is used to identify products showing extreme, and potentially erroneous, prices or price movements. This article focuses on outlier detection; our junk filtering for rail fares and second-hand cars categories is described further in our Research and developments in the transformation of UK consumer price statistics: June 2022 article.

We are considering three main applications of outlier detection.

- Global outlier detection identifies extreme observation-level prices that are atypical for the whole consumption segment. For example, flagging any apple observation with a price above £8, because the price distribution for the remaining apples is between £0.40 and £0.80.

- Observation-level outlier detection identifies atypical values for an observation given a distribution of historic prices for that product. For example, flagging a specific observation for a £4 apple, because that same apple historically cost £0.40.

- Relative-based outlier detection identifies extreme month-on-month price movements in aggregated representative product prices. For example, flagging a 95% reduction in the average monthly price paid for a particular variety of apple.

Outlier detection is performed broadly in the same way for each of these approaches: defining lower and upper fences, identifying price observations (or price relatives) below and above these fences (respectively), then choosing whether to remove these observations from further processing.

The following methods are explored in this article.

- User-defined fence. For the user-defined fence method, the lower and upper fences are set by the user. This method heavily relies on user judgment to set the fences. The lower and upper fence values are user defined.

- Tukey (interquartile). For the Tukey (interquartile) method, the lower and upper fences are calculated by subtracting and adding a multiple "k" of the interquartile range from the first (Q1) and third (Q3) quartiles respectively. $LF = Q1 - k*(Q3-Q1)$ $UF = Q1 + k*(Q3-Q1)$

- Kimber. For the Kimber method, the lower and upper fences are calculated by subtracting and adding a multiple "k" of the semi-interquartile ranges to the lower and upper quartiles, better accounting for a skewed distribution compared with the Tukey method. $LF = Q1 - k*(quartile\ 2(Q2) - Q1)$ $UF = Q3 + k*(Q3 - Q2)$

- K-sigma. For the k-sigma method, the lower and upper fences are calculated by subtracting and adding a multiple "k" of the standard deviation (sd) to the mean. $LF = mean - k*sd$ $UF = mean + k*sd$

User-defined fences are set manually, whereas the other methods set fences algorithmically. We will sometimes collectively refer to these as "the algorithmic methods".

Any of the methods described identify potential outliers if their prices or price changes are outside the fences. The methods themselves do not remove outliers. However, because of the scale of data we receive through alternative data sources, the burden of manually validating each observation identified as a result of the method would be large and unmanageable.

As such, for the purposes of our analysis we do not include any observations identified as outliers by each method in our subsequent index calculations. While these observations are not used to construct the index, they remain in the data. We are currently working on a process to monitor and review the number and nature of outliers being detected and removed from index calculations each month.

For these reasons, the ideal method would flag the minimum number of transactions possible while still identifying any potential errors.

This article discusses the general data cleaning and outlier detection methods in Sections 2 and 3. Section 4 presents the two case studies for second-hand cars and rail fares, while the reasons for our recommendations are reported in Section 5. Section 6 details our future work.

3 . Examples of approaches to outlier detection

Within each dataset explored, unique products are defined as they are described in our Using transaction-level rail fares data to transform consumer price statistics, UK article and Using Auto Trader car listings data to transform consumer price statistics, UK article. Each month, a unique product can contain multiple price observations. The observation-level prices are averaged to calculate a representative price for the product within the month, that is then used for calculating price movements and inflation indices.

The main differences in the application of different outlier detection methods relate to the distributions and levels of aggregation that the methods are applied to. While global outlier detection is arguably the most straightforward, we encourage the use of observation-based and relative-based outlier detection in this article.

Observation-based outlier detection is applied to the price distribution of a specific product, as previously defined. Atypical price observations have the potential to skew the monthly representative price for any given product. Therefore, observation-level outlier detection can be applied to the price distribution of a specific product, to identify any erroneous observations that could be skewing the monthly price.

An extreme example is shown in Table 1 where the 15 January 2022 observation, a clear outlier, significantly increases the January representative price for this product.

Table 1: Observation-level outlier detection aims to remove observations to avoid the need for full elimination of the product because of outliers

Product	Date	Before outlier detection	After outlier detection
		Transaction price	Transaction price
Product A	01 January 2022	4	4
Product A	02 January 2022	4	4
Product A	04 January 2022	4	4
Product A	15 January 2022	30004	
Product A	25 January 2022	4	4
	Total	30020	16
	Representative price	6004	4

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

The goal of observation-based outlier detection is therefore to detect and potentially remove extreme observation-level prices without the need to eliminate the full representative monthly price for the product. This allows for more complete use of the data as a price for the month will still be calculable.

The main difference between global outlier detection and observation outlier detection is that global outlier detection creates a single set of fences for the entire consumption segment, whereas observation-level outlier detection creates fences for each individual product. The aim is to better detect atypical prices for the individual product. Note that since fences will need to be set for each individual product, user-defined fences are not viable for this approach when there are a large number of products.

Observation-level outlier detection carries risks over the potential removal of genuine data in multi-modal distributions, and therefore needs to be applied with caution.

Relative-based outlier detection is applied to detect atypical price changes over time at the product level. The strategy applies outlier detection methods after observations have been averaged to product-level monthly representative prices, that are used in our index method calculations. Relative-based outlier detection will be applied to identify potentially erroneous price changes and subsequently remove them so they do not have an overt effect on our indices. If a price relative is considered an outlier, all observations for that product within a month are removed.

An example of this is shown in Table 2. An extreme price change in the Product C ticket has caused the index to fall substantially, despite increasing prices in the other tickets.

Table 2: An extreme price change in the Product C ticket has caused indices to fall despite rising prices in the other tickets

Product	January price	quantity	February price	quantity	price relative
Product A	3	8000	3.3	7950	1.1
Product B	5	30000	5.2	30050	1.04
Product C	40	4000	1	4000	0.025
Product D	4	3000	4.2	3050	1.05
				Index	0.426

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

4 . Case studies

In the previous sections, we described a variety of applications (global-, observation- and relative-based) and methods (user-defined fences, Tukey, Kimber, k-sigma) of outlier detection. We now look to apply these to the case studies of second-hand cars and rail fares. We look at how index calculation is affected when we remove observations identified by a range of different outlier detection methods from the data.

For our case studies, we compare our methods against a "no outlier detection removal" benchmark. For this exercise, we only consider one choice of parameter per method, but some exploration of parameter choice was performed.

Second-hand cars

Our first case study is second-hand cars. In Table 3 we present the methods and parameters we will explore in our analysis of outlier detection methods. Some methods flag more cases than others, and it is likely that better parameter optimisation would produce a more desirable level of outlier detection for each individual method. This demonstrates that National Statistical Offices must be careful about parameter choices.

Table 3: Methods of outlier detection to be explored with second-hand cars

Approach	Method	Parameters	Flagged, petrol (%)	Flagged, diesel (%)
Benchmark	No outlier detection removal	N/A	0	0%
Global	User-defined	LF, UF = 400, 60000	0.91%	0.29%
Observation	Tukey (interquartile)	k = 3	0.15%	0.10%
Observation	Kimber	k = 3	1.18%	0.89%
Observation	k-sigma	k = 3	0.21%	0.16%
Relative	User-defined	LF, UF = 1/3, 3	0.03%	0.04%
Relative	Tukey (interquartile)	k = 3	1.56%	0.96%
Relative	Kimber	k = 3	5.04%	3.41%
Relative	k-sigma	k = 3	0.90%	0.67%

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Figures 1 to 4 show price indices for petrol and diesel second-hand cars consumption segments following the removal of observations or products that have been flagged through global and observation-level and relative-based outlier detection procedures. All four figures show that outlier detection does not distort the general trend of the benchmark indices (where no outlier detection is performed). However, the Kimber method (with k = 3) removed more observations than other methods and in all four plots reduces the increase in the headline index more than others. This is likely a bias-towards-zero-inflation caused by removing too many genuine high-price observations.

Figure 1: Petrol second-hand cars, global and observation-level outlier detection

UK, January 2020 to August 2022
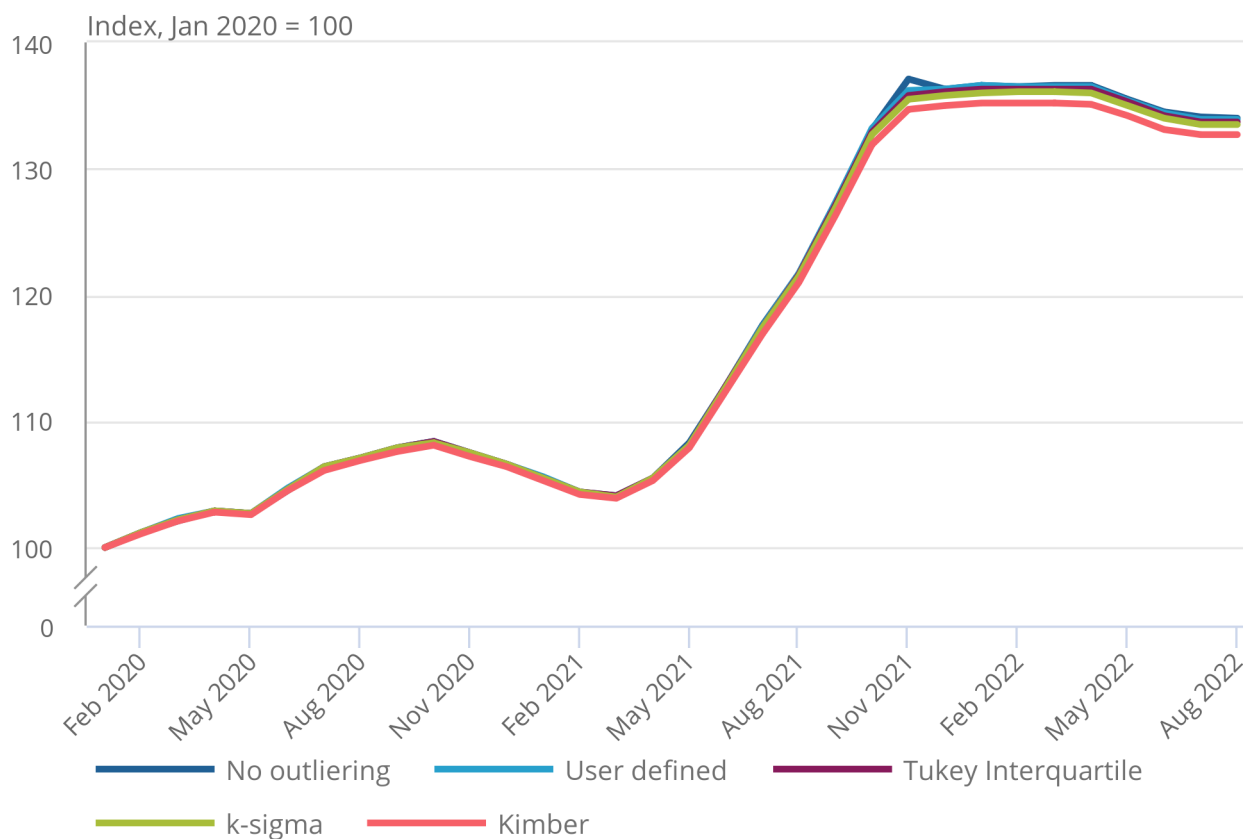
Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data
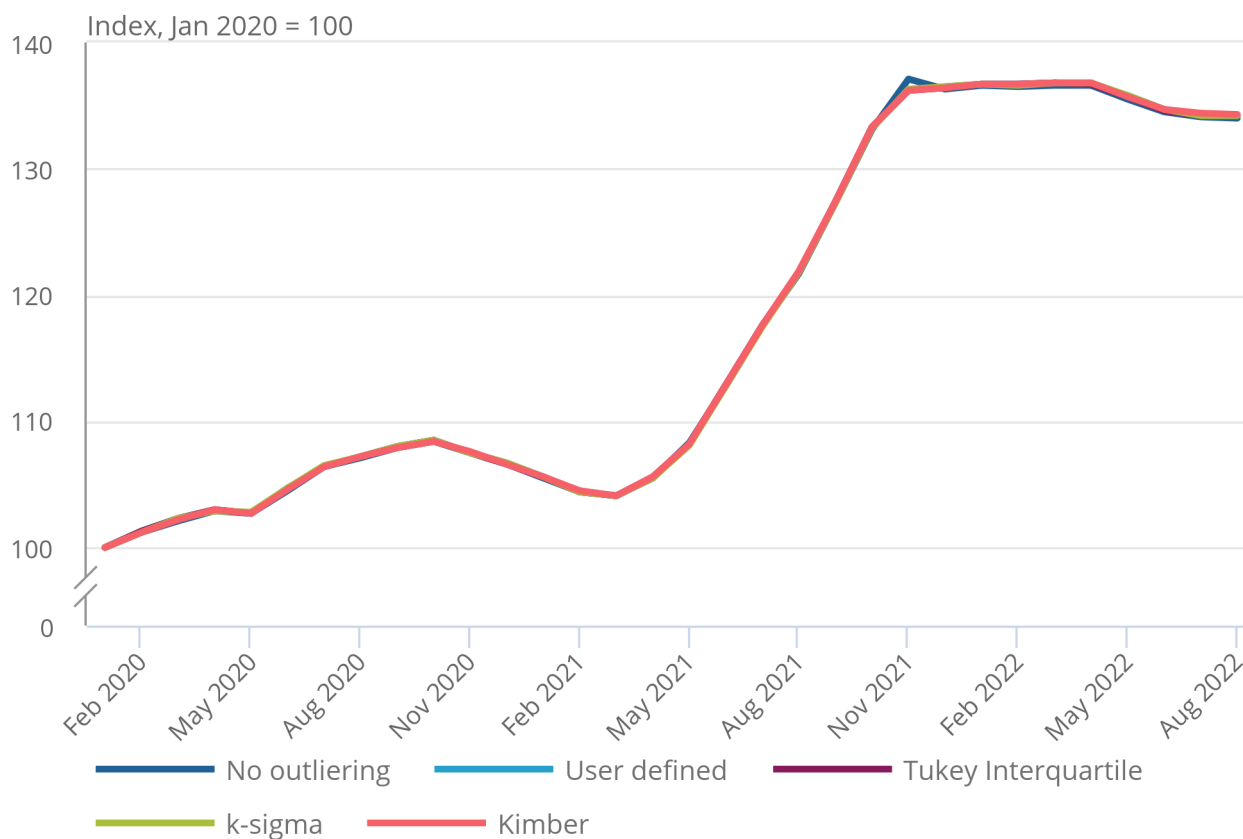
Notes:

1. User defined with lower fence (lf) = 400 upper fence (uf) = 60000, k = 3

Figure 2: Petrol second-hand cars, relative-based outlier detection

UK, January 2020 to August 2022

Figure 2: Petrol second-hand cars, relative-based outlier detection

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

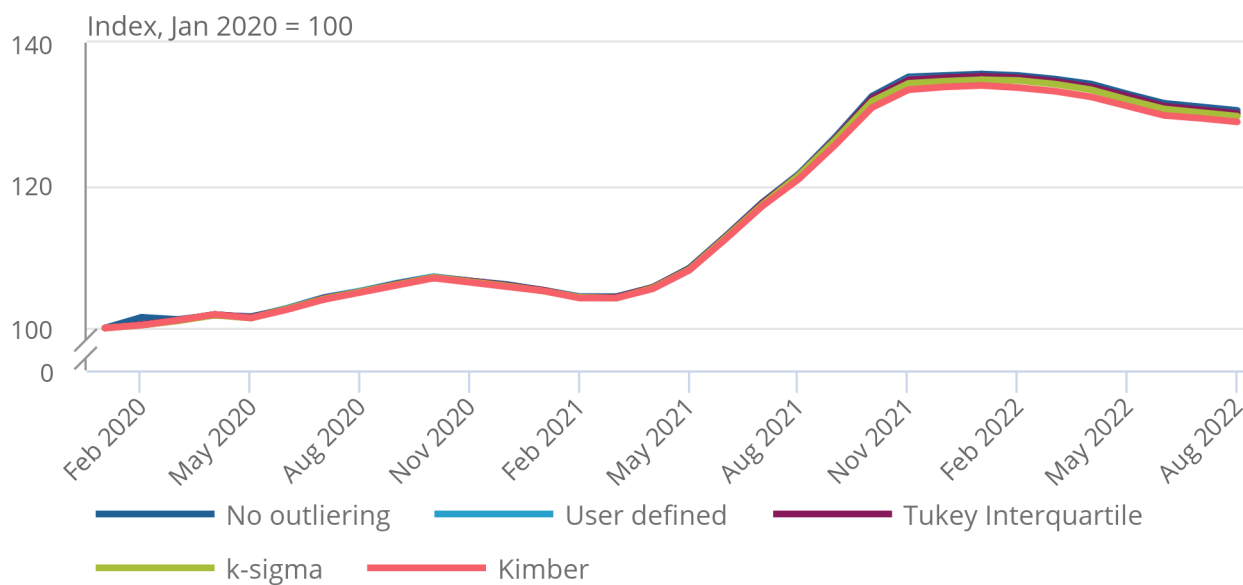1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

Figure 3: Diesel second-hand cars, global and observation-level outlier detection

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data
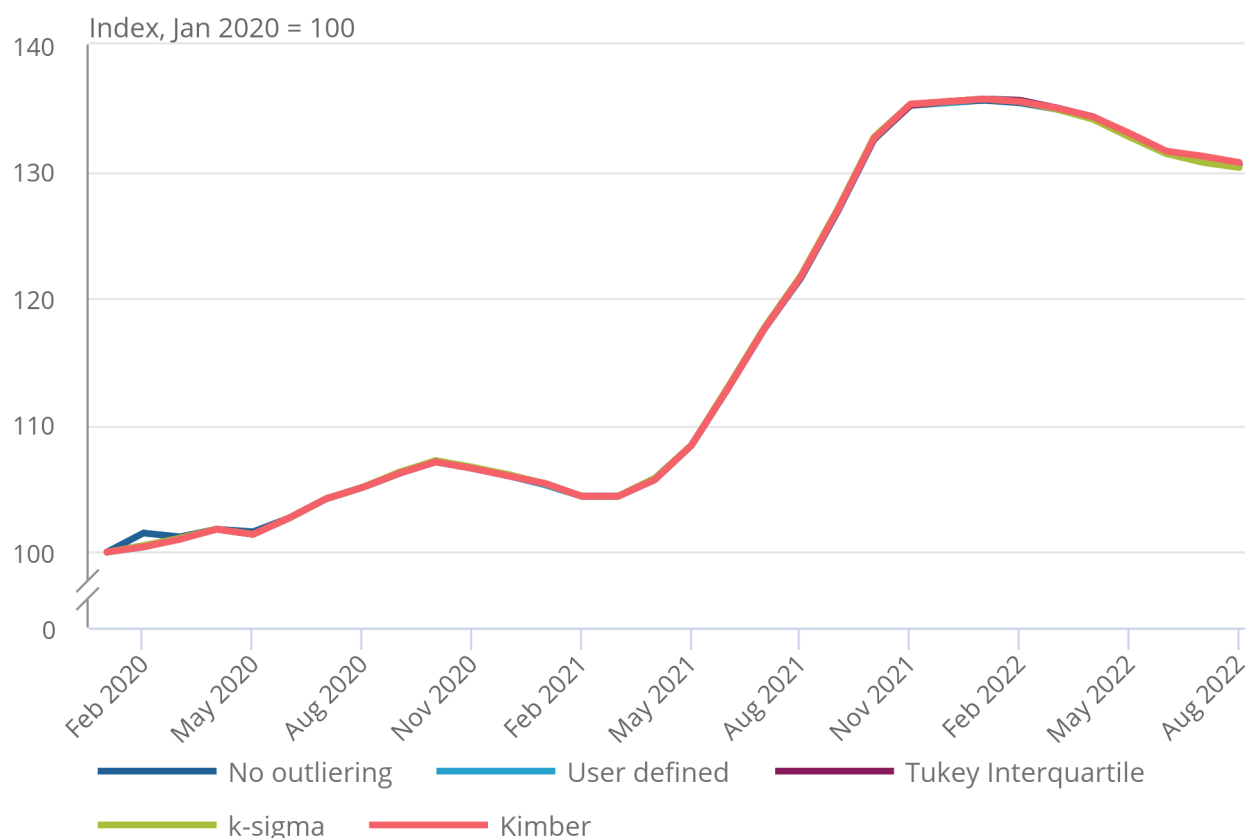
Notes:

1. User defined with lower fence (lf) = 400 upper fence (uf) = 60000, k = 3

Figure 4: Diesel second-hand cars, relative-based detection

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

There are two particularly interesting months. The benchmark indices contain mild spikes in November 2021 for petrol (Figures 1 to 2) and in February 2020 for diesel (Figures 3 to 4) cars.

The November spike was mostly driven by a single, highly weighted car with a month-on-month price relative larger than 16. In this example, a car that typically has a value of approximately £100,000 (in November 2021), also had a single price quote around £10 million, which considerably distorted the average price.

The February spike contained highly contributory month-on-month relatives of 100, 37, 7 and 3.7. Every form of outlier detection we have explored appears to identify these observations, and removing them eliminates this volatility. This shows that even the mildest forms of outlier detection can help us avoid potential errors introduced by extreme values, while removing few genuine observations.

In Figures 5 to 8 we present, for the same indices as in Figures 1 to 4, the difference in indices between each method and the benchmark, allowing a more perceptible comparison of the methods. We first consider global and observation-based outlier detection in Figures 5 and 6.

UK, January 2020 to August 2022

Figure 5: Petrol second-hand cars indices: global and observation-based method minus benchmark

UK, January 2020 to August 2022
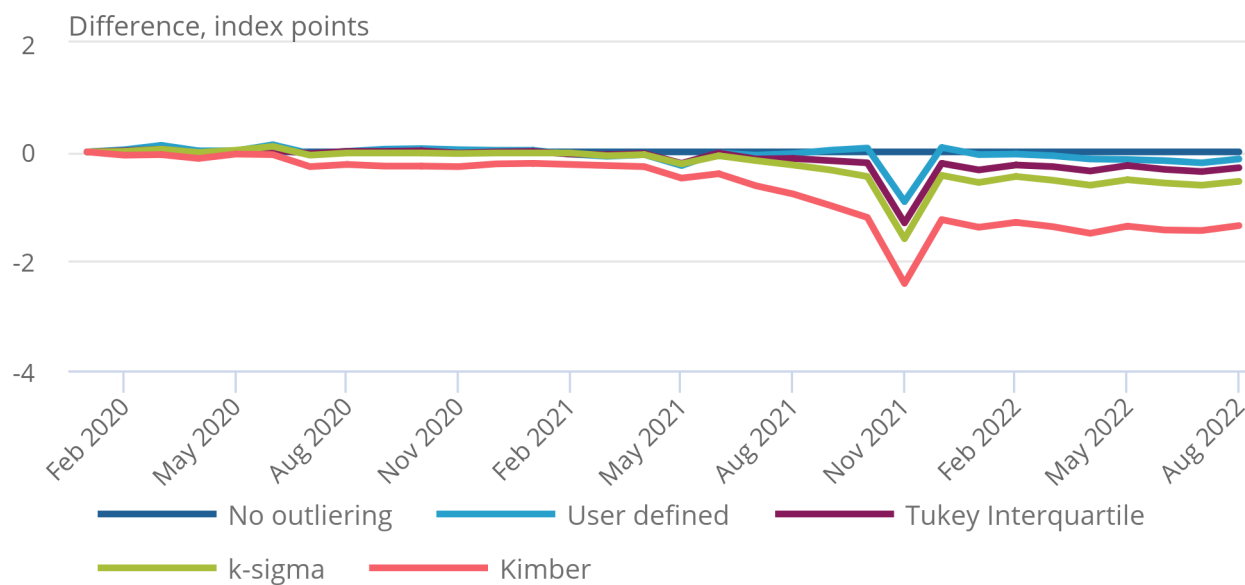
Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data
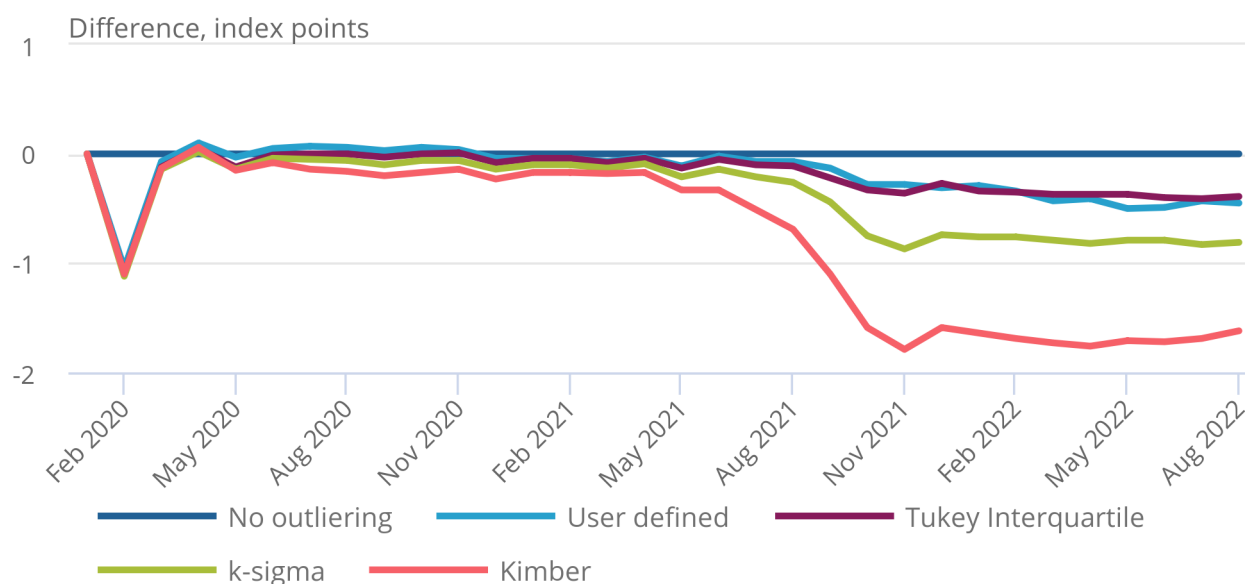
Notes:

1. User defined with lower fence (lf) = 400 upper fence (uf) = 60000, k = 3

Figure 6: Diesel second-hand cars indices: global and observation-based method minus benchmark

UK, January 2020 to August 2022

Figure 6: Diesel second-hand cars indices: global and observation-based method minus benchmark

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

1. User defined with lower fence (lf) = 400 upper fence (uf) = 60000, k = 3

Figures 5 and 6 may suggest a bias is introduced by our implementation of observation-based outlier detection and provide a warning against further use. We lacked sufficient sample sizes to perform the method monthly and instead performed the check longitudinally over the entire time frame. As prices of second-hand cars increased rapidly between May 2021 and November 2021 (as shown in Figures 1 to 4), differences with the benchmark became wider. This may be because setting a single non-updating fence over a longitudinal period of increasing prices meant an increasing number of high-value flags and a decreasing number of low-value flags, causing a potential downwards bias in the index.

We therefore do not believe observation-based outlier detection to be appropriate unless there are sufficient observations to perform it over single-month distributions. This might be more appropriate for other goods categories, such as groceries.

Global outlier detection with non-updating fences may also be open to a similar criticism for similar reasons. Note, however, that removing quotes flagged using user-defined fences (global-based approach) eliminates only slightly fewer observations than the Kimber (observation-based approach), but the impact on the index is much smaller. This may be because global outlier detection behaves differently, fully flagging all observations of more expensive or cheaper products and only partially flagging some observations where products are close to the fence boundaries. By contrast, observation-based outlier detection is more likely to partially flag observations across many products. This more widespread pattern of partially flagging observations may carry a greater risk of introducing this potential bias more widely.

As discussed, because of the data size, we have to remove all flagged observations without manual scrutiny to check if they are genuine errors. Therefore, it may be necessary to widen the user-defined fences to avoid such a bias from removing too many genuine highly priced observations. The fences set would need to be monitored to avoid such a potential bias being re-introduced.
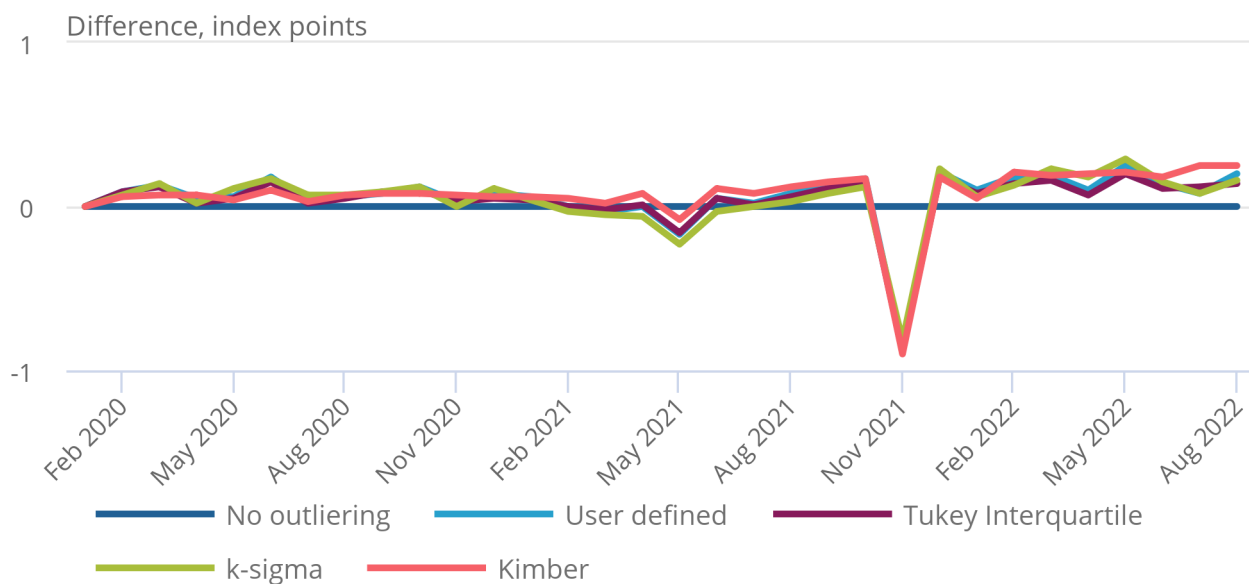
We consider relative-based outlier detection in Figures 7 and 8.

Figure 7: Petrol second-hand cars indices: relative-based method minus benchmark

UK, January 2020 to August 2022

Figure 7: Petrol second-hand cars indices: relative-based method minus benchmark

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

UK, January 2020 to August 2022

Figure 8: Diesel second-hand cars indices: relative-based method minus benchmark

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data
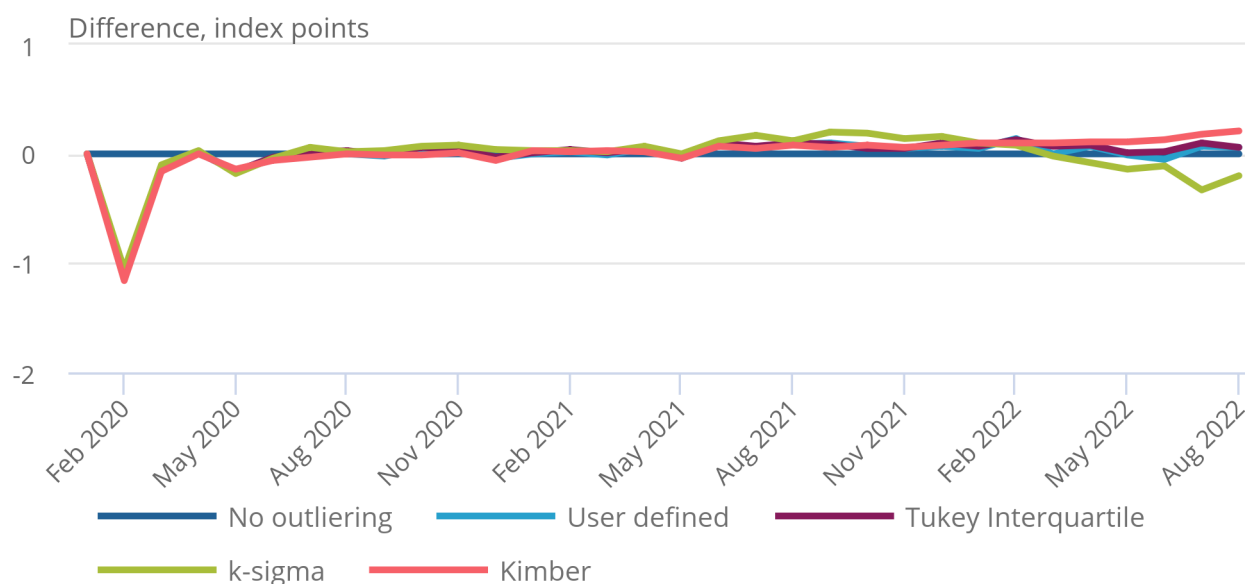
Notes:

1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

Many of the algorithmic methods set fences symmetrically, setting upper and lower fences by adding or subtracting a value to a measure of central tendency. For example, a method may set fences by adding or removing 0.99 to an average price relative of 1, giving a lower fence of 0.01 and an upper fence of 1.99. Prices would have to go down by 99% to be detected by the lower fence, but would only have to double to be detected by the upper fence. Price relative distributions are positively skewed and so symmetric methods (Tukey, k-sigma) may not be preferred. However, Figures 7 and 8 show that this does not seem to be a major issue, with similar results given to the benchmark.

The Kimber method sets asymmetric fences, better accounting for skewed distributions. However, Figures 7 and 8 show the method to be further away from the benchmark. This is likely because the Kimber is detecting substantially more cases than Tukey and k-sigma, perhaps because of a suboptimal choice of k-value.

User-defined fences on price relatives avoids this issue through manual selection, where fences can be set to respect distributional skew. In our case we are using ratio-3, where ratios between consecutive months are flagged if more than 1:3 (trebling of prices) or less than 3:1 (thirding of prices). This ensures consistency over time and seems to work well, identifying few observations while still correcting the spikes previously mentioned.

Rail fares

We consider similar analyses for our rail fares data. As previously mentioned, observation-level outlier detection may not work well when observation distributions are multi-modal. We observed such distributions when exploring rail fares because of a scheme where child tickets cost £1 when bought alongside an adult ticket. This resulted in bimodal distributions with peaks at £1 and the normal ticket price. Fences on prices typically cut out one of these genuine prices. We therefore avoid using observation-level outlier detection with rail fares, though multimodal fences could be considered in future.

We also considered global outlier detection with lower and upper fences of £0.50 and £5,000 respectively. However, but the impact of removing these observations was so negligible (the difference with the benchmark was less than 0.01 in every month) that no further results are presented.

Therefore, we focus on relative-based outlier detection. The methods we explore are presented in Table 4. Despite using the same k-values that worked reasonably well in the second-hand cars case study, the Kimber and Tukey methods are flagging an unreasonable proportion of cases. This may be caused by the lower and upper fences being so close together that the methods are setting extremely tight fences. This shows the importance of distributional shape when using algorithmic methods.

Table 4: Methods of outlier detection to be explored with rail fares

Type	Method	Parameters	Flagged	Percent
Benchmark	No outlier detection	N/A	0	0%
Relative	User-defined	LF, UF = 1/3, 3	132,796	0.02%
Relative	Kimber	k = 3	182,006,519	29.91%
Relative	k-sigma	k = 3	5,751,068	0.95%
Relative	Tukey (interquartile)	k = 3	145,194,524	23.85%

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Price changes in rail fares are unlike other categories. Most ticket prices change annually in March, when the price cap of regulated rail fares is increased in line with the Retail Prices Index. This behaviour is observed in the index, where a large uplift in the index occurs each year in March (since 2021, before this this the uplift was applied each January).
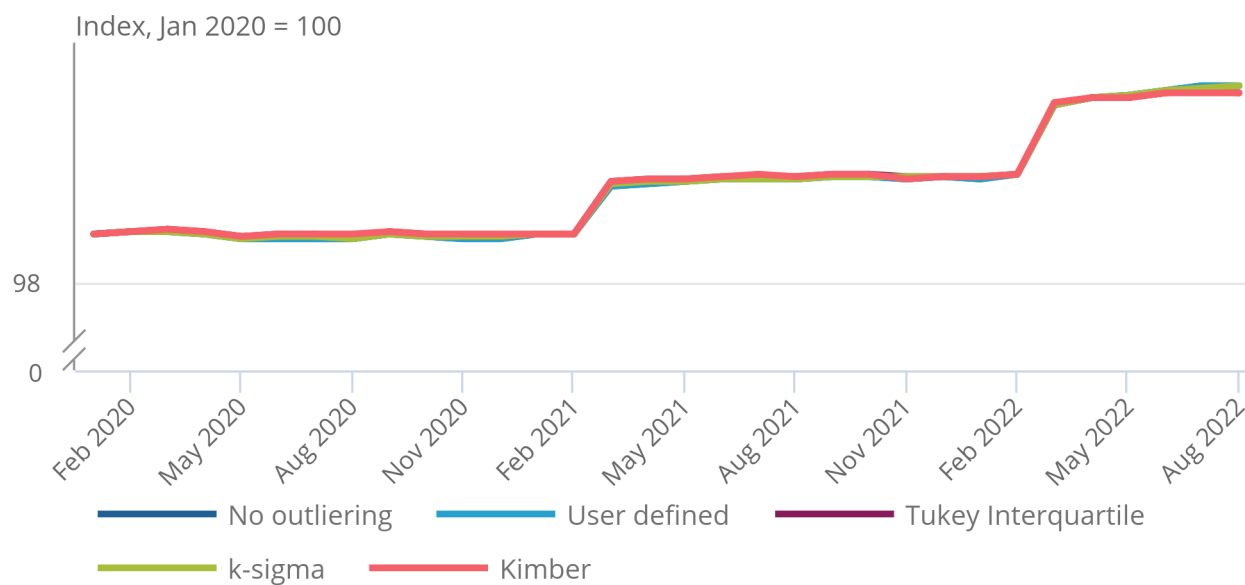
Figure 9 shows the indices for the different outlier detection methods, whereas Figure 10 shows the difference between these indices and the benchmark.

Figure 9: Rail fares indices for different data cleaning methods

UK, January 2020 to August 2022

Figure 9: Rail fares indices for different data cleaning methods

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

Figure 10: Percentage point difference between indices and the benchmark

UK, January 2020 to August 2022

Source: Office for National Statistics – Outlier detection for rail fares and second-hand cars dynamic price data

Notes:

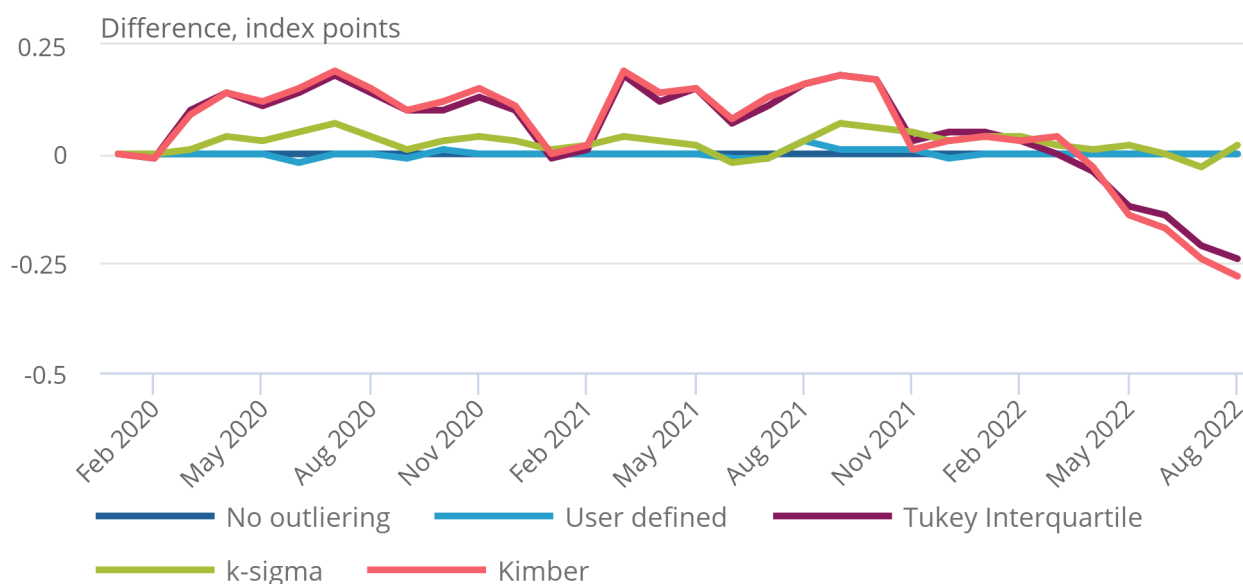1. User defined with lower fence (lf) = one-third upper fence (uf) = 3, k = 3

Even though the Kimber and Tukey identify an unreasonable proportion of data, when those data are removed the resultant indices surprisingly reflect the overall rail fares trend reasonably well (as shown in Figure 9). However, it is unrealistic to expect this proportion of outliers, and so the volatility in the comparison to the benchmark (Figure 10) is expected to be undesirable.

Removing observations based on quotes flagged through user-defined fences give results extremely close to the benchmark showing that the most extreme potential outliers have a very mild effect, perhaps because of the size and quality of the data.

5 . Final outlier detection method and impact

Our preferred outlier detection method for use with rail fares and second-hand cars in production is relative-based outlier detection with a user-defined lower fence of one-third and upper fence of 3. This will identify products where the representative monthly triples in price within a month or is reduced to more than a third of the price within a month. We subsequently will not use any observations for the flagged product in that specified month in our downstream calculations, these observations are however identified in a separate dataset so that they can be further scrutinised.

We have several reasons to prefer this approach, including:

- it performed well in our case studies removing few observations and did not unrealistically distort any of the benchmarks and corrected the likely erroneous (albeit mild) spikes in second-hand cars

- it carries less risk of using outdated fences, limiting the need for manual updates; a trebling of the month-on-month price may always be considered extreme, whereas setting a £60,000 upper fence on prices for second-hand cars becomes increasingly less extreme over time because of inflation

- we may be able to use these parameters consistently across different types of goods, our case studies have shown that this is likely not possible for the other methods

- by not being dependent on distributional shape, it carries less risk of using poorly-defined fences as a result of a poor fit with the shape of the distribution

- 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 on second-hand cars of using this method was shown previously in Figure 7 (for petrol cars) and Figure 8 (for diesel cars). Outlier detection removal corrects the (likely) erroneous mild spikes observed, but otherwise the method does not change the index measured between January 2020 and August 2022 significantly. The largest difference was observed in May 2022, when the benchmark index was 135.5 and the index with the user defined fences was 135.8, resulting in a difference of 0.25 index points on the second-hand cars index.

The impact on rail fares of using this method and subsequently removing prices was shown previously in Figure 10. Over the 33 months measured, indices are substantially unchanged by applying outlier detection, with a larger absolute difference observed in August 2021, when the benchmark index was 102.4, while the index with the user defined fences was 102.3, resulting in a difference of 0.03 index points on the rail fares index.

In production we will look to monitor the number of outliers detected by the lower and upper fences and compare indices before and after outlier detection to ensure that outlier detection does not cause any unusual behaviour in the indices.

There may be rare situations where many products drop in price by 66% or increase by 200% within a single month for genuine reasons, for example if the government were to introduce a subsidy. Our monitoring approaches will allow us to detect these situations, and we can respond by adjusting the fence values appropriately. This makes the adopted method simple to monitor and adjust.

6 . Future developments

Our chosen method works well with goods categories presented in this article, but we will need to research its continued use, and alternatives, when considering the application of other goods categories. One of the alternatives we have previously considered (and presented to the Technical Advisory Panel on Consumer Prices in April 2021) is DBScan, a clustering algorithm. This is much more complex to productionise, maintain and interpret so has been deprioritised in favour of simpler approaches, and may be reconsidered if these methods are deemed unsuitable for future goods categories.

We will also look to explore another form of data cleaning, "dump price removal". This may be particularly relevant when considering categories such as groceries, clothing, or technological goods. Dump price removal involves identifying products where both the prices and quantities fall by a large factor, having entered a clearance period where all remaining stock is being cleared from the market. Since rail fares and second-hand car sales generally do not exhibit these clearance patterns, this is considered as a lesser concern for these categories.

7 . Related Links

Research and developments in the transformation of UK consumer price statistics: November 2022
Article | Released 28 November 2022
Research to modernise the measurement of consumer price inflation in the UK: fourth in a series of biannual articles to update users.

GEKS-Törnqvist: introducing multilateral index methods into consumer price statistics
Methodology article | Released 28 November 2022
We are planning to use the GEKS-Törnqvist for introducing alternative data into our consumer price statistics. In this article we describe how the method works and its advantages.

Impact analysis on transformation of UK consumer price statistics: rail fares and second-hand cars, November 2022
Article | Released 28 November 2022
Research into the use of new index number methods to calculate price indices using web-scraped and scanner data.

Transformation of consumer price statistics: April 2022
Article | Released 27 April 2022
Our plans to transform UK consumer price statistics by including new and improved data sources and developing our methods and systems for production from 2023.

Consumer Prices Indices Technical Manual, 2019
Methodology | Last revised 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.

Consumer price inflation, UK: October 2022
Bulletin | Released 16 November 2022
Price indices, percentage changes, and weights for the different measures of consumer price inflation.

8 . Cite this methodology

Office for National Statistics (ONS), released 28 November 2022, ONS website, methodology, Outlier detection for rail fares and second-hand cars dynamic price data