Introducing grocery scanner data into consumer price statistics

From 2026, we intend to introduce grocery scanner data into our consumer price statistics. This methodology article gives an overview of the data and methods we will use.

Contact:
Consumer Prices Methods Transformation
cpi@ons.gov.uk

Release date:
29 April 2025

Next release:
To be announced

Table of contents

1 . Overview

On 29 April 2025, we published the first indicative estimates of the impact of introducing groceries scanner data into UK consumer price statistics. We plan to parallel run grocery scanner data for a year before incorporating them into live production in 2026. These datasets cover over a billion units of products per month. This article summarises the methods used to integrate data of this size into our statistics.

2 . Grocery scanner data: introduction

From March 2026, we intend to introduce grocery scanner data into consumer price statistics. These datasets are created as consumers purchase goods at a supermarket checkout (or online), giving information on the prices that customers pay. (Note that personal customer data are not included in data feeds.)

Scanner data offer advantages compared with traditional data. We can:

- increase automation of price collection
- expand on what our inflation figures cover
- use many more product prices in our inflation statistics
- reflect which specific products drive inflation more
- more accurately reflect the average price paid by the consumer for each product

We currently have regular scanner data feeds, which cover the sale of over a billion units of products per month and make up around 50% of the grocery market. We intend to enter production with these retailers in 2026 and look to expand our coverage in future years. We will continue to reflect the remainder of the grocery market with our traditional price collection practices.

Scanner data also present challenges compared with traditional methods of inflation measurement. Our traditional methods involve considerable manual scrutiny. It would be too difficult to scale these methods to be used on the much bigger scanner datasets, and they would fail to make full use of the scanner data. Therefore, new methods and systems are needed to process scanner data. These methods were first presented in our Research into the use of scanner data for constructing UK consumer price statistics article, and have been developed over the course of several years, with feedback from our Advisory Panels and the wider international prices community.

This article describes the scanner data acquisition process, along with the methods that are being introduced to transform scanner data into price indices. These methods have been developed into a statistical pipeline, making use of cloud infrastructure and expanding the system built for rail fares, discussed in our paper Developing reproducible analytical pipelines for the transformation of consumer price statistics: rail fares (PDF, 400KB).

3 . Grocery scanner data: acquisition and quality

Scanner data are provided to us from grocery retailers, containing aggregated transactional data on products sold in-store or online. The data we receive do not contain customer information but are rather an aggregated view of sales regarding a product within a particular store and time range.

A (synthetic) example is given in Table 1, where the first row shows the total expenditure and number of units sold of a pack of Jazz apples within the Rotherham outlet between 11 and 17 January.

Table 1: Example fields from a (synthetic) grocery scanner dataset

From	To	Outlet	Product	Weight	Expenditure	Quantity sold
11 January 2026	**17 January 2026**	Rotherham	Jazz apples (6-pack)	Each	7000	3000
18 January 2026	**24 January 2026**	Boston	Baked Beans	400g	8020.52	15000

Source: Office for National Statistics

Acquiring a stable grocery scanner data feed requires a considerable setup process as we build a partnership with the retailer. This starts with initial meetings to clarify the data we need, and how we will use them. We then work with the supplier's data team on iteratively analysing and improving on a smaller supply of test data. This process allows us to design the final data feed, ensuring the content and format of the data supply are fit for purpose.

Where possible, we look to use automated data transfer pipelines to reduce human dependency and ongoing burden to the supplier. After both parties are content with the proposed data feed, any decisions, expectations and security details are documented formally in a data sharing agreement. Once the agreement is signed off, regular data feeds (potentially along with historical data) can be sent.

Each time we receive data, we conduct a variety of checks to ensure the data are of a sufficient quality before we use them in further statistical pipelines. If the data pass the checks, then we standardise the raw data using processes developed by our data engineers. This may involve renaming variables to make each scanner data feed follow a consistent data schema, deriving new variables to simplify further processing of the data, standardising weight values (for example, converting kilograms to grams) and linking on geographic markers. This pre-processing limits the need for hardcoded retailer-specific processing within our statistical pipeline.

4 . Grocery scanner data: methods

The following subsections discuss the methods we will be using to both calculate price indices from grocery scanner data and then integrate into our consumer price inflation statistics. Most of these methods have been discussed in more detail in previous articles we have released, and we will link to these articles where relevant. In this article, we will summarise and confirm our final methods, highlighting any changes we have made since we first discussed the topic.

Date trimming

When using scanner data, we need to decide which days or weeks of data to use to represent a month. There are a variety of challenges, such as how to treat weeks that overlap two months, and whether to treat daily and weekly datasets differently. We discussed this topic at length in our Date trimming for consumer prices alternative data sources article.

Previously we had considered maximising the amount of data available, using three weeks to represent some months and four weeks for others, depending on how many full weeks fall within a month. However, after further work and timetable planning, our final decision is to instead use the first three full weeks that fall fully within the month. To give a few examples, this would mean using:

- 4 to 24 January 2026
- 1 to 21 February 2026
- 1 to 21 March 2026
- 5 to 25 April 2026

Note that we would use the same dates regardless of whether the data are provided to us daily- or weekly-aggregated.

Empirically we found a negligible difference in indices between using three and four weeks of data. However, had we used the full four weeks, we would have needed to calculate then quality assure our indices within two working days of the final receipt of data to meet production deadlines. This lack of time would have created too much risk for us to sustain, and so using three weeks is preferred.

Strata definition

Our groceries strata definitions follow the aggregation structure outlined in more detail in Section 4 of our Introducing alternative data into consumer price statistics: aggregation and weights article.

In the previous version of this article, we had planned to stratify scanner data by groceries outlet type - for example, having different strata for supermarket and convenience stores. However, we will no longer be including this level of stratification, and instead control for homogeneity by using the outlet type within the product definition (as described in the following "Product definition" subsection).

Classification

After the strata are defined, we need to classify each product in each scanner dataset to one of the consumption segments in our aggregation hierarchy, allowing the scanner data to then be split into strata according to their consumption segment, region and retailer. Retailer hierarchies do not cleanly map onto our hierarchy, and therefore we need a bespoke method to ensure each product is appropriately classified.

To perform this classification task, we are using a method described as "machine-assisted manual classification". In summary, each product is manually assigned a classification by a labeller by using a bespoke labelling dashboard. Although this task is done manually, and therefore each product is manually scrutinised, the labeller also has access to a shortlist of machine-recommended options, allowing the labeller to find the appropriate label faster than they would on their own. A full description of the method can be found in Section 3 of our Classification of new data in UK consumer price statistics article.

Defining invalid strata

Since there are thousands of scanner data strata because of the various permutations of consumption segment, region and retailer, there may be some strata that would not produce reliable indices for a full production year because of low product count. To avoid an overreliance on imputing these strata, we instead seek to remove them.

To do this, we run a test to check whether the stratum could have been calculated over the two years leading up to the production year. We calculate a 25-month GEKS-Törnqvist window covering from the January of two years ago to the current January, and if a lack of product matches means we cannot form a mathematically valid index in any of the 25 months within the window, we describe it as an "invalid stratum". Invalid strata are then removed from our aggregation structure for the following year and the retailer would then be represented by traditional data instead.

For example, suppose a scanner retailer stratum had a 10% market share and was set to invalid, then we would exclude this stratum from our aggregation structure and instead transfer the 10% weight to the corresponding traditional stratum, which would now be considered to represent the retailer. The invalid strata process generally affects very few strata.

Calculating unit values for products

Scanner data contain multiple rows per product. For example, suppose a retailer provides daily data and has 10 outlets within London, and we are using a three-week month in line with our date- trimming strategy. If a product is sold in all 10 outlets on all 21 days, then this would result in 210 rows of data representing that one product in London in the month. We call these rows "observed expenditure and quantities".

Our index methods require that each product is represented by a singular price and quantity each month. Therefore, we sum the "observed expenditure and quantities", giving the total expenditure and quantity for the product across the entire region within the month. We then divide expenditure by quantity to give a "representative price", which is then used in our index methods. Representative prices are often described as a "unit value".

A product may be sold at different prices in different outlets and at various times of the month. A representative price will generally not match any single one of these individual prices exactly, but instead represents the average price paid by consumers for the month.

Once representative prices are calculated, we go further and perform a size-based quality adjustment to the prices and quantities. For products with a fixed size, this transformation involves dividing the price, and multiplying the quantity, by the product weight.

For example, if a 500 gramme tin of tomatoes is sold for £0.50 and 700 people buy it in the month, then the size-adjusted price is £0.001 (£0.50 divided by 500 gramme), and the size-adjusted quantity is 350,000 (700 multiplied by 500). This can be interpreted as consumers paying £0.001 per gramme of tomatoes, with 350,000 grammes of tomatoes sold.

This transformation allows us to quality-adjust for changes in weight, since price-per-gramme is affected by both price changes and weight changes. If the product weight is expressed in kilogrammes or litres, the weights will be standardised to grammes or millilitres, respectively, so we can compare on a like-for-like basis.

Where products are sold loose by unit of weight, for example, loose carrots sold per kilogramme, the product's total expenditure is divided by the total weight sold, in grammes, to similarly obtain a price per gramme. If the total weight sold is not available, the product is excluded from our aggregates as we are not able to calculate a meaningful price for the product.

Product definition

As described in the last section, derivation of representative prices (unit values) involves averaging many observations. We need to perform this averaging over products which are "homogeneous in quality". If, for example, we were to create a product that encompasses a 60 pence Pink Lady apple and a 40 pence Granny Smith apple, and both of their prices stayed constant, then the representative price may shift solely because of the proportion of people buying each apple. This is known as a unit value bias - where changes in price are not because of pure price change, but rather because of compositional effects. In the example given, the Granny Smith and Pink Lady apples are not homogeneous in quality, since they are distinct types of apples.

To be certain our products are homogeneous, we need to use a product definition that encompasses observations of a consistent quality. To do this, we use the SKU (Stock Keeping Unit), a retailer-derived product identification variable, as part of the product definition to ensure we are comparing products on a like-for-like basis across time. This will help avoid us doing something like grouping a Pink Lady and Granny Smith apple into one product. We also break down products by outlet type (supermarket or convenience store) to avoid unit value bias stemming from the different prices a retailer may set at different types of outlets. We also use the weight type in the product definition to ensure the size-adjustment to prices (described in the "calculating representative prices section") is done on a like-for-like basis.

For example, a product ID of "20736312_G_supermarket" would represent the rows where the SKU is 20736312, where the weight is given in grammes, and where the sales are in supermarkets (rather than, for example, convenience stores).

Discounts and refunds

Historically, discounts have been difficult to account for when measuring product price change because of a lack of information on discount take-up rates. For example, a box of grapes may cost £2 in January, and continue to typically cost £2 in February, but may also be available at a lower price of £1 for members of a loyalty scheme.

Without knowing the proportion of consumers who are members of the scheme and thus able to take advantage of the lower price, it would not be possible to know the true average price paid by the consumer. In traditional practices, we would therefore ignore the loyalty scheme price and compare only pre-discounted prices.

In scanner data, the unit values calculated group together both consumers who bought the product at a discounted price, and consumers who bought it at full price. For example, if one person bought the box of grapes at the full £2, and one person bought the box of grapes at the discounted £1, then our data would reflect this as £3 of expenditure and two units sold, giving a unit value price of £1.50. As a result, a true average price paid by the consumer is measured.

This means that scanner data typically account for price promotions, multibuy offers and loyalty scheme discounts - expanding the range of discounting behaviour we can account for compared with traditional methods.

Refunds relate to products that have been returned to the retailer by the consumer. The product is not consumed, and therefore considered not in scope for inflation measurement. Some retailers provide refunds as separate rows, and where this is true, we can aggregate sales with refunds to remove refunded products. However, note that for food and drink categories (where we are focusing our initial use of scanner data), refunds make up less than 1% of total expenditure (see our Research into the use of scanner data for constructing UK consumer price statistics article) and their effect on indices are negligible. This suggests that even where we cannot account for refunds, the impact of including some sales, which are subsequently refunded, is expected to be very small.

Relaunch linking

Sometimes manufacturers remove a product from the market and then "relaunch" it. A relaunched product may differ to the original variant in price, quality, or both. The product may be relaunched for a variety of reasons, such as because of recipe changes, changes to packaging, and most pertinently because of a change in weight. Products can either increase in weight (especially because of promotional offers) or decrease in weight as part of a cost-cutting strategy (often described as "shrinkflation").

When a relaunch occurs, some retailers group together both the original and relaunched product under the same SKU, resulting in both the original and relaunched product having the same product ID and allowing us to capture any quality-adjusted price changes in the product. However, some retailers assign the original and relaunch different SKUs, and if this is not corrected, our indices would not capture any (quality-adjusted) price change associated with the relaunch.

Relaunch linking describes our process of linking together the original product SKU and the relaunch SKU. The methods are described in detail in Section 3 of our Research into the use of scanner data for constructing UK consumer price statistics article. (Although note that we have made one change to no longer use price as a means of identifying relaunches compared with when this article was first released.)

In theory indices can be both upwardly and downwardly biased by not accounting for weight changes. However, empirical evidence suggests indices tend to be biased down by not accounting for relaunching behaviour - as shown in our own work in the article linked in the previous paragraph.

Data cleaning

The data cleaning process involves three main levels of data cleaning.

Firstly, this is the stage where any rows pertaining to "invalid strata" (as described in the previous subsection) are removed from the data.

Secondly, we remove out-of-scope observations (which we have sometimes referred to as "junk filtering"). The filtering we perform includes removing:

- observations that cannot be assigned a valid product ID or stratum
- observations covering products sold by weight, where we lack weight information
- observations that do not represent sales to consumers
- observations with invalid sales or quantity information (such as negative prices)

Thirdly, we perform two types of outlier detection. The first removes extreme price changes, where a representative price goes up or down by a factor of four. This gives enough room to allow significant promotional offers (such as buy one get one free) to be accounted for, without implausible price changes distorting our indices.

The second removes dump pricing behaviour. Dump prices occur when a product is reduced in price by more than 50%, but quantity sold has fallen by more than 90%. They are often associated with end-of-lifecycle products, where remaining stock is sold at extremely low prices, but the average customer is unable to benefit from this sale as only very limited stock are sold at these clearance prices. These methods are described in more detail in our Outlier detection for grocery scanner data in consumer price statistics article.

Indices and aggregation

The index methods used for elementary aggregate calculation within scanner data will be different to those used for traditional data. For traditional data, in most places, we will continue to use Jevons and Dutot for Consumer Prices Index and Consumer Prices Index including owner occupiers' housing costs (CPI and CPIH) and Carli and Dutot for Retail Prices Index (RPI) construction.

For scanner data, consistent with methods used for introducing alternative data sources for rail fares in 2023 and second-hand cars in 2024, we will be using the GEKS-Törnqvist with a 25-month window and a mean splice on published extension method. This method is described in more detail in our Introducing multilateral index methods into consumer price statistics article, and is summarised in our video.

We will continue to use "Laspeyres-like" aggregation to aggregate from our (scanner and traditional data) elementary aggregates to higher levels of our aggregation structure. Full guidance on how our aggregation and weighting approach works can be found in our Introducing alternative data into consumer price statistics: aggregation and weights article.

5 . Future developments

We aim to publish final impacts of this transformation at the end of 2025. Following our publication, a decision will be made as to whether we move these new data and methods into use in live production. If we are satisfied that our data, methods and systems are ready for live monthly production of these indices, the first time they will be introduced is in the figures for February 2026, published in March 2026. The existing published series will not be revised.

Our broader plans to transform UK consumer price statistics by including new improved data sources and developing our methods and systems are discussed in our Transformation of consumer price statistics: August 2024 article.

6 . Related links

Impact analysis on transformation of UK consumer price statistics
Article series
This series demonstrates the indicative impacts on our headline consumer price statistics of introducing new data and methods into our consumer price inflation statistics.

Introducing alternative data into consumer price statistics: aggregation and weights
Methodology article | Released 29 April 2025
This article describes the aggregation structure used to aggregate alternative and traditional data together.

7 . Cite this methodology

Office for National Statistics (ONS), released 29 April 2025, ONS website, methodology article, Introducing grocery scanner data into consumer price statistics