Integration of alternative data into consumer price statistics: the UK approach

Helen Sands, Office for National Statistics, UK

May 2022

Introduction

New, alternative data sources, namely scanner and web-scraped data, and methods to best use these data sources, are being introduced into the production of UK consumer price statistics from 2023.

These new data sources will result in millions more prices being processed each month; therefore, for UK consumer price statistics, changes are required at the lowest level of aggregation to integrate these new data while ensuring that they are appropriately represented within our price indices.

This article details our proposed hierarchy and methods that are to be implemented as we begin to incorporate new, alternative data; details on our existing hierarchy and methods can be found in Consumer Prices Indices Technical Manual, 2019.

New data

From 2023 we plan to transform the measurement of UK consumer price statistics for rail fares and second-hand cars.

For rail fares we have recently acquired access to transaction-level data for rail sales in Great Britain, sourced from the rail industry's Latest Earnings Networked Nationally Overnight (LENNON) ticketing and revenue system. These are provided to us daily by the UK's Rail Delivery Group, dating back to January 2019, and are transaction-level data, so explicit information is available on both cost and quantity of tickets within each transaction. There is also a wealth of information on price-defining attributes, such as origin and destination station, ticket type and class.

For second-hand cars, we have been working with the largest and most visited vehicle advertising website in the UK, Auto Trader. We now have access to daily feeds of their vehicle data, including advertised prices along with a wealth of vehicle attributes, dating back to January 2018. As these data are for advertised vehicle listings, they do not include explicit sales revenue information.

From 2024 we plan to transform the measurement of UK consumer price statistics for groceries using retail scanner data. We have engaged directly with the UK's largest retailers to gain access to these data. Retailers provide product sales totals for each store, for each day, week, or month (dependent on retailer) along with additional information about the products being sold. Our current retailers account for nearly 50% of the UK grocery market, though we are seeking to acquire data from additional retailers to further improve our coverage.

This is just the start of a continuous programme of improvements to UK consumer price statistics as we intend to gradually increase our use of new data sources in future years.

New aggregation structure

To help enable the introduction of new data sources we will move to a new aggregation structure (Figure 1). This has been developed under four key considerations:

1. We have the flexibility to use alternative data in combination with, or in place of, traditionally collected data, weighted according to our best information on retailer market share
2. We can realise more potential from alternative data sources, while continuing existing practices for our traditional collection
3. We can more readily calculate regional consumer price statistics, for which there has been growing demand in the UK
4. We enable a smooth transition towards the latest iteration of Classification of Individual Consumption According to Purpose (COICOP 2018), while also realigning our numerical system for our detailed (COICOP 6) level of the hierarchy coding with higher COICOP levels

The example in Figure 1 is illustrative, there are consumption segments for which we do not have new data and will continue to use traditionally collected data to produce price indices. There are also consumption segments that we will not stratify into groups based on market share, as large retailers dominate the market for some products. Furthermore, sometimes there may be greater or fewer locally collected item indices to represent each consumption segment, depending on the range of product varieties in each consumption segment and the amount of expenditure each consumption segment accounts for. There may also be consumption segments where we choose to rely entirely on alternative data sources, such as used cars or rail fares.

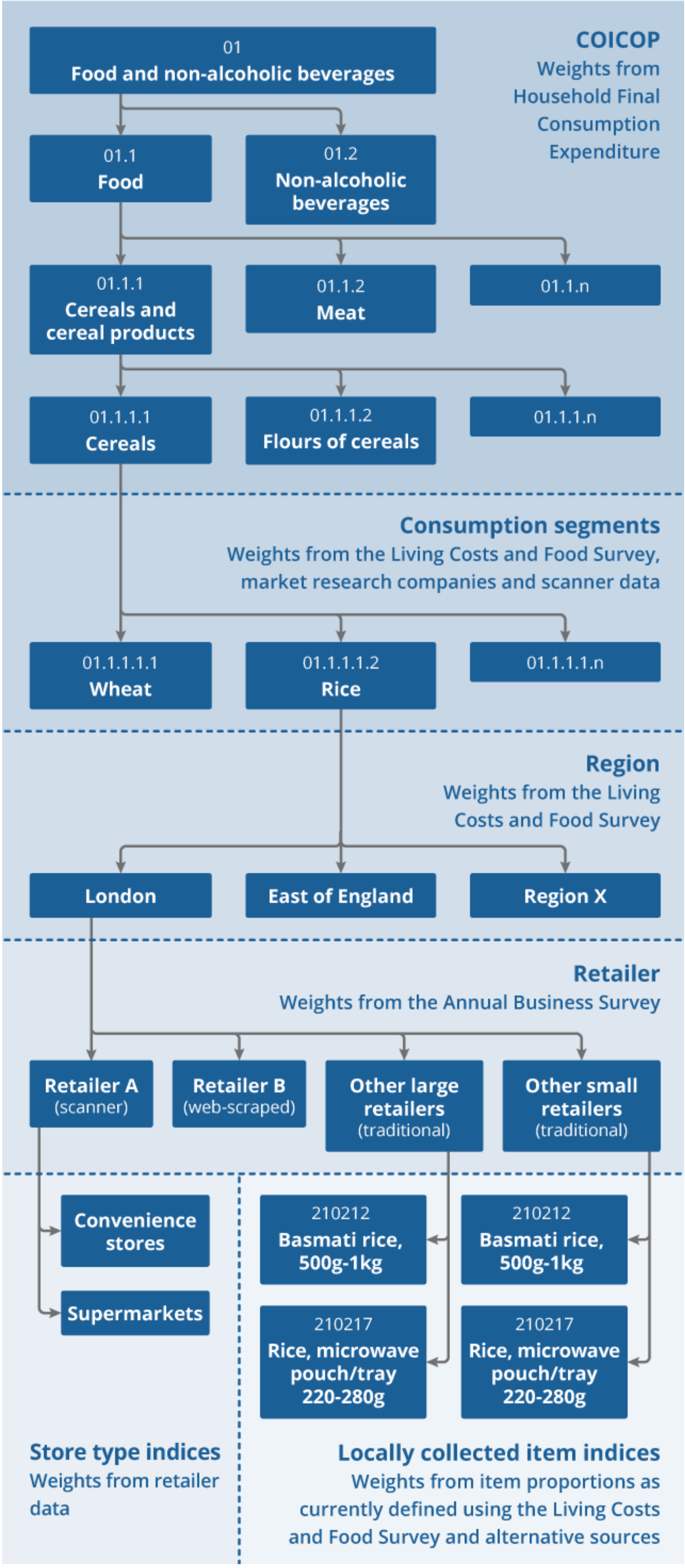
In Figure 1 we define other large retailers as those with greater than 2% market share within a COICOP class, and other small retailers as those with less than 2% market share within a COICOP class. This means a retailer can be classed as a large retailer for one COICOP class, such as food and drink, but a small retailer for another COICOP class, such as toys and games, where other retailers may dominate the market. Each retailer for whom we have alternative data sources for receives a weight according to their individual market share. We group non-ADS retailers to maintain adequate sample sizes within each strata.

Introduction of consumption segments

A key change in the proposed price aggregation structure comes with the introduction of "consumption segments". Currently, UK consumer price indices produced at this detailed level in the hierarchy are referred to as "item indices".

Items are selected for the consumer price statistics basket to be representative of a broader group. For example, we collect peanuts to represent price movements for nuts, and garden spades to represent price movements for garden tools. For retailers for whom we have alternative data, we have access to prices for a near-census of products. We are therefore introducing consumption segment level indices to realise greater potential from these new data, for example including all nuts instead of just peanuts.

Figure 1: Illustrative example of how the future aggregation structure could look for the "Rice" consumption segment



COICOP
Weights from Household Final Consumption Expenditure
01
Food and non-alcoholic beverages
01.1
Food
01.2
Non-alcoholic beverages
01.1.1
Cereals and cereal products
01.1.2
Meat
01.1.n
01.1.1.1
Cereals
01.1.1.2
Flours of cereals
01.1.1.n
Consumption segments
Weights from the Living Costs and Food Survey, market research companies and scanner data
01.1.1.1.1
Wheat
01.1.1.1.2
Rice
01.1.1.1.n
Region
Weights from the Living Costs and Food Survey
London
East of England
Region X
Retailer
Weights from the Annual Business Survey
Retailer A
(scanner)
Retailer B
(web-scraped)
Other large retailers
(traditional)
Other small retailers
(traditional)
Convenience stores
Supermarkets
210212
Basmati rice, 500g-1kg
210212
Basmati rice, 500g-1kg
210217
Rice, microwave pouch/tray 220-280g
210217
Rice, microwave pouch/tray 220-280g
Store type indices
Weights from retailer data
Locally collected item indices
Weights from item proportions as currently defined using the Living Costs and Food Survey and alternative sources

Consumption segments are broader than the current item definitions, though still defined based on a relatively homogeneous set of products. Broadening the definitions allows us to make better use of the alternative data. For example, one current item definition is for "Basmati rice 500g-1kg"; by broadening the consumption segment definition to "Rice" we will be able to use data for long-grain, short-grain, white, brown, and flavoured rice, weighted according to popularity.

However, when continuing to collect prices using traditional methods, the current item level definitions will be maintained. This is because sample sizes are smaller and index methods may be more sensitive to heterogeneous prices. These item indices will be treated explicitly as being representative of the broader consumption segment, for example, "Basmati rice 500g-1kg" and "Rice, microwave pouch/tray 220-280g" will be aggregated together to form a rice index for traditionally collected data within each region. These will then be aggregated with "Rice" indices from retailers for whom we have alternative data, based on their respective market shares, to form regional and higher-level rice indices (see Figure 1).

New consumption segments will only be included in UK consumer price statistics if there are corresponding items in the sample of representative goods from our traditional collection. This ensures that retailers for whom we have alternative data are proportionately represented within our consumer price indices. This is particularly important when we haven’t been able to reach a high market share coverage using alternative data within a COICOP class; if we were to include consumption segments where the remaining retailers were not represented, price movements for the retailers who we do have data for would dominate the higher-level aggregates.

New index number methods

The index methods used at the elementary aggregate (EA) level will vary, even within a single consumption segment index, due to the mix of new and traditionally collected data at the lowest level. Price indices constructed using new, alternative data sources will begin to make use of multilateral methods.

Fox et al. (2022) recently reviewed our work (to date) on determining the most appropriate index method for producing elementary aggregate indices using new data. They discussed and evaluated the properties of different multilateral index numbers for measuring high frequency price changes, drawing on household scanner data. They concluded that use of the Caves-Christensen-Diewert-Inklaar (CCDI, also understood as the GEKS-Törnqvist) index, updated using the mean splice over consecutive 25-month windows, is to be preferred for both theoretical and empirical reasons.

While we are continuing our own programme of work to better understand the nuances in different multilateral index methods, we currently plan to use the CCDI in our initial introduction for these new data sources from 2023.

The CCDI index will only be used for EA indices where we are making use of new, alternative data such as scanner or transaction data. EA price indices constructed using traditionally collected data will continue to primarily be formulated using the fixed-base Jevons index (though some of our EA indices also make use of other methods including Dutot and Lowe indices).

Once EA indices are calculated for all data sources, these indices will be aggregated together using existing aggregation methods: weighting the indices together according to our most recent, accurate information on expenditure shares in the base period (using a Lowe index).

As multilateral indices are no longer calculated within-year, and no longer have a coherent fixed base period, prior to aggregation they are re-referenced to the base period to be consistent with EA indices based on traditional data. For example, in March 2023 the (UK) base period for the traditional collection would be January 2023. But for our multilateral indices that span a longer period they could be referenced to January 2021 = 100. To ensure all indices are consistent for aggregation we would rereference our March 2023 multilateral index to January 2023 = 100.

Annual update process for new data sources

Given we will restrict the number of consumption segments to only include those that have representative items in the traditional collection, we will also need to continue the process of annually updating our basket of goods and services to include new consumption segments to reflect changing consumption patterns.

We will only introduce new consumption segments (or new alternative data sources retailers) where we already have a 25-month period of data so we can consistently use a 25-month window for all alternative data based elementary aggregate indices. This means that every year the consumption segments will essentially be reset, the data re-classified to the new consumption segments over a historic 25-month period, and then the new index for the following 12 months would splice onto this recalculated historic index.

As an example, say we introduce a new consumption segment “Plant-based milk” in 2023. We then classify 25 months (Jan 2021 – Jan 2023) of historic scanner data to plant-based milk and calculate an initial unrevised index for this period. We then begin to use a mean splice in the 26th – 37th months (Feb 2023 – Jan 2024) to continue the index and re-reference this index to month 1 (Jan 2023) = 100 so it can be aggregated with the new fixed-base “plant-based milk” index from the traditional collection. This process is then repeated annually to onboard any new consumption segments or new retailers providing alternative data sources.

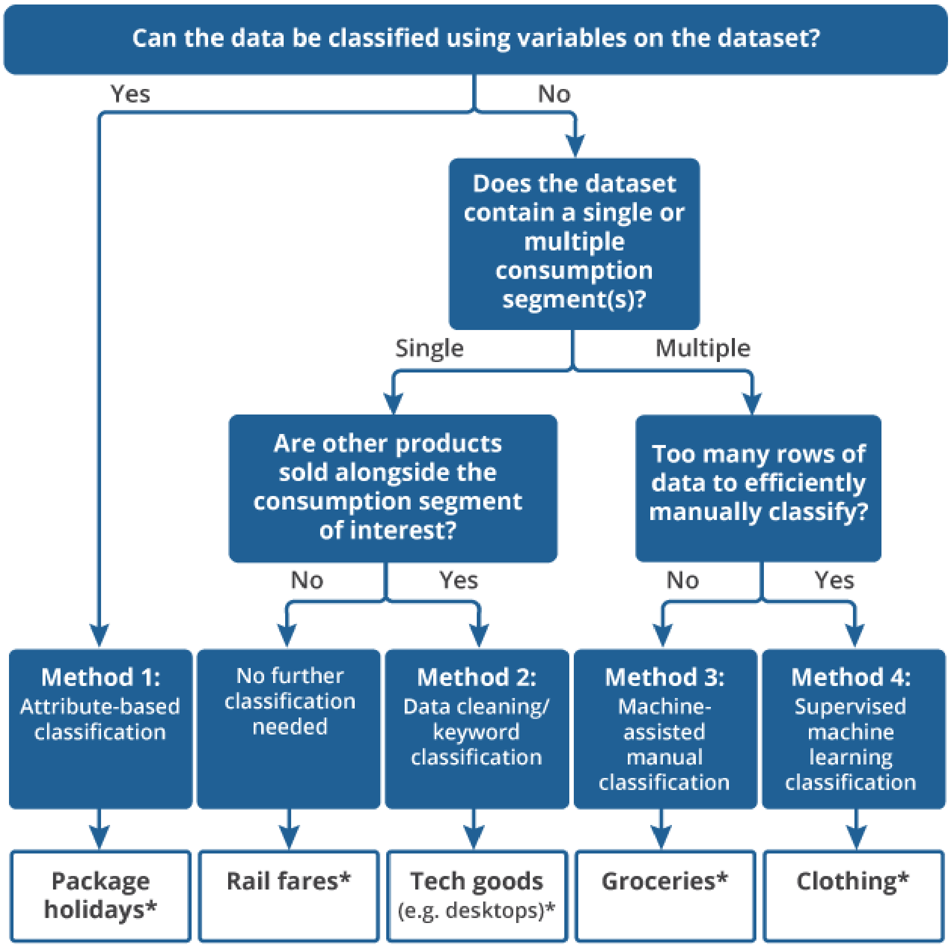
Quality assurance of new data sources

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

Data filtering and classification of new data sources

Different methods are required to ensure the data that we use in the EA index are relevant to the consumption segment that we're trying to measure. In ONS (2021a) we outlined our strategy for classification of different data types and categories, Figure 2 shows the suite of methods we intend to use.

Figure 2: The classification of new data sources requires a suite of methods


Can the data be classified using variables on the dataset?
Yes
No
Does the dataset contain a single or multiple consumption segment(s)?
Single
Multiple
Are other products sold alongside the consumption segment of interest?
Too many rows of data to efficiently manually classify?
No
Yes
No
Yes
Method 1: Attribute-based classification
No further classification needed
Method 2: Data cleaning/ keyword classification
Method 3: Machine-assisted manual classification
Method 4: Supervised machine learning classification
Package holidays*
Rail fares*
Tech goods (e.g. desktops)*
Groceries*
Clothing*

*Recommendations for these categories are subject to change due to ongoing research. These categories are given as examples and are not an exhaustive list of all categories being explored.

For rail fares and second-hand cars, we have found that other products are in fact sold alongside the consumption segment of interest, and therefore we are using a data cleaning/filtering approach to remove any unwanted data. For example, the second-hand cars data contain prices for vans and motorbikes and the rail fares data contain data for bus travel and car parking tickets; we use product attributes to filter out unwanted data.

For groceries we have continued to use manual classification to date but are exploring how we can reduce the amount of manual classification using machine-assisted classification. More information on our data processing specifically for grocery data can be found in ONS (2021b).

Defining a unique product

One of the most important decisions when constructing indices is determining what our product definition should be. In some cases, defining a product is relatively simple. For groceries we are provided numerical indicators as to what constitutes a unique product, via the Global Trade Item Number (GTIN) or the Stock Keeping Unit (SKU). The GTIN is consistent between retailers and unique to each product, but the SKU is sometimes better at capturing small changes in say packaging or ingredients which we would consider as constituting the same product. Therefore, where available, we use the SKU to define a unique product for groceries, though we may use the GTINs to help link and classify products across multiple retailers.

For other categories, such as second-hand cars, defining a unique product is more challenging if we want the product to persist through time at a constant quality. That's because every car registration is unique and, even for each unique car registration, the quality of the car may depreciate month on month. For second-hand cars we are therefore using several quality-defining attributes, including the cars age at time of listing, to define a product. This results in groups of cars of similar quality, and we can then follow the price of each constant-quality group through time.

To ensure the product definition we have chosen appropriately balances the need for homogeneity with the need for these products to persist through time, we use Chessa (2019) proposed MARS method that balances these components and produces a score to judge the success of the product definition.

As with second-hand cars, for rail fares we use attributes within the data to define a product, based on a subset of price-determining quality characteristics, such as the origin and destination stations, the ticket type, and the travel class.

Ongoing work and next steps

We are currently still considering the best data cleaning methods for use with these data, to identify and remove erroneous values. We are also still investigating nuances between some of the highly rated multilateral methods and the impact of imputation in the GEKS formula, as well as the interaction between the multilateral indices and higher-level chaining.

We are still considering the use of web-scraped data and how we might be able to approximate expenditures to ensure bias from infrequently sold products and the use of unweighted indices is not exacerbated, as well as improving some of our existing work in classification and product grouping using machine learning.

We publish bi-annual research articles updating on our research progress in these areas. Our next publication in late June 2022 will show the methods we have considered for rail fares and second-hand cars and the resulting indices. A full impact analysis will be published in November 2022 and the new data will be live in the indices from February 2023 (published in March 2023).

In 2023 we will also commence a parallel run of new data and methods for our grocery indices, again with publications in June and November 2023.