ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Integration of scanner and web-scraped data into consumer price statistics: aggregation and weights

Status: Work in progress

Expected publication: For publication alongside minutes

Purpose

1. This paper outlines proposals and remaining challenges for the integration of scanner and web-scraped data into consumer price statistics, with a particular focus on aggregation of different data sources and their respective weights, as well as imputation for missing strata and consumption segments.

Actions

2. Members of the Panel are invited to:
   a) agree the proposed future aggregation structure for consumer price statistics integrating alternative data
   b) advise on the approach to introducing new consumption segments and discuss the proportion of market share required before realising the full potential of alternative data
   c) comment on the suitability of market share weighting using Annual Business Survey data
   d) discuss the appropriateness of the proposed imputation strategies for missing retailers/regions and missing consumption segments

Introduction

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

4. To date, most of our research that we have discussed with the Panel has considered methods for web-scraped and scanner data at the elementary index level of calculation. This paper discusses how we intend to integrate the new data sources with our traditional collection at higher levels of the aggregation structure, and some of our remaining challenges.

Proposed aggregation structure and sources of weights

5. Our proposed aggregation structure has been developed under the following considerations and principles:
   a. we have the flexibility to use alternative data in combination with traditional data (to ensure that we are still representing smaller/independent retailers and markets), weighted according to our best information on retailer market share
   b. we can more readily calculate regional consumer price statistics in future
   c. we can realise more potential from alternative data sources, while keeping the traditional collection as stable as possible to maintain RPI in its current form
   d. we enable transition towards the latest iteration of COICOP (2018)
   e. we realign our detailed (COICOP 6) level of the hierarchy coding with higher COICOP levels (item level index codes are currently aligned to the RPI hierarchy)

6. The current aggregation structure is provided in Figure 1a, and the proposed aggregation structure (for categories with alternative data sources) in Figure 1b.

Figure 1a: Current aggregation structure (ECOICOP)

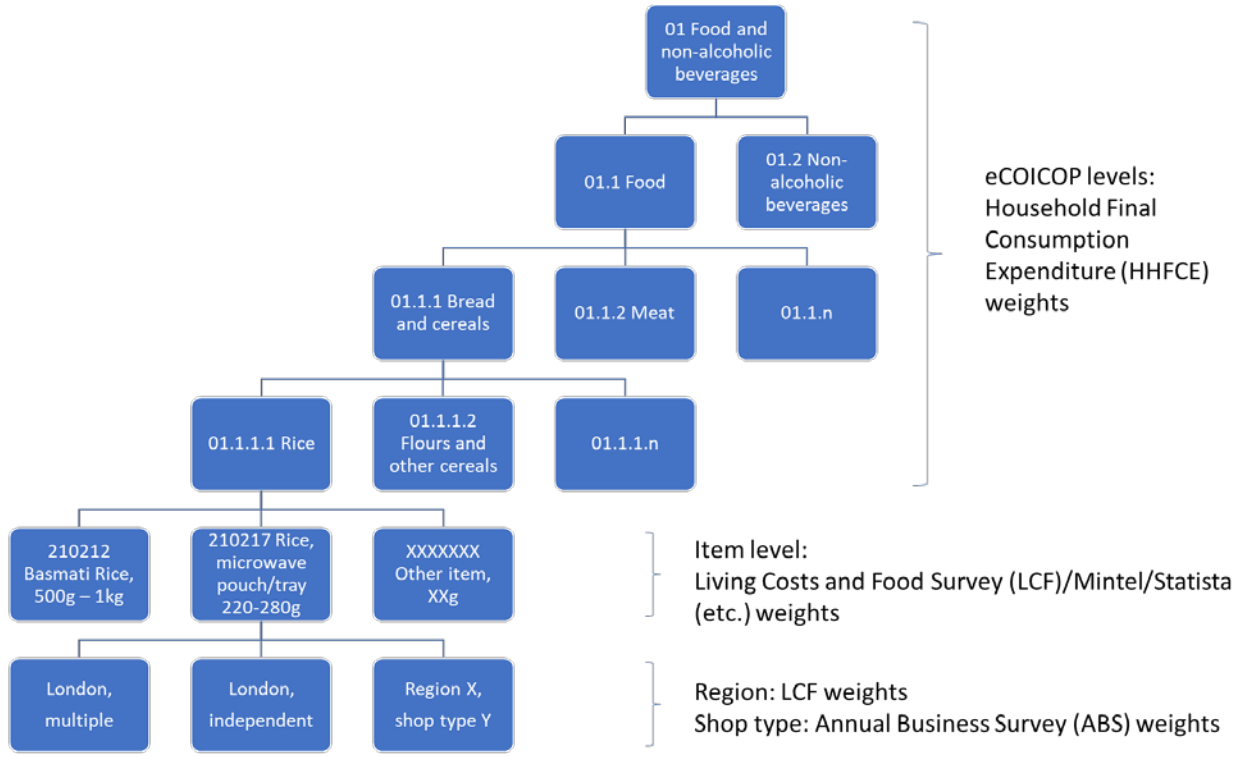


Figure 1a note: not all items are currently stratified by both region and shop type (see para 28, or section 8.3)

Figure 1b: Proposed aggregation structure (COICOP 2018)

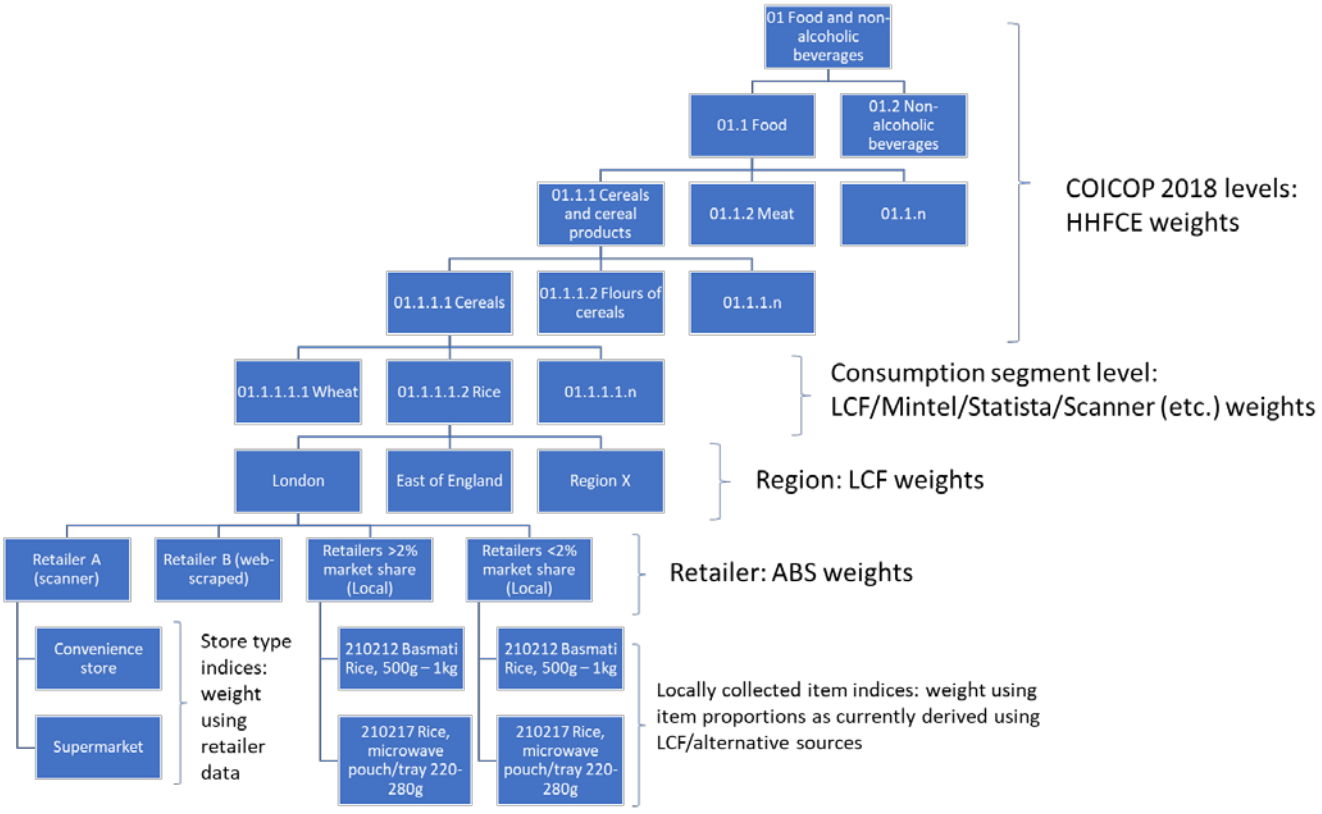


Figure 1b note: region and retailer weights will be more strictly imposed at lower levels of aggregation

7. Aside from the transition towards the latest iteration of COICOP, calculation of the indices and sources of weights at the COICOP 5 level and above remain unchanged. While the structure beneath the COICOP 5 level is made more complex with the introduction of alternative data, the principles of calculation remain the same; using a fixed basket Lowe-type index[1] to aggregate elementary aggregate indices to higher levels in the aggregation structure.
8. We will primarily publish indices at the COICOP 5 level, although index microdata could be made available for the consumption segment level indices (provided they are not disclosive of retailers), as they currently are for the item level indices.

Introduction of consumption segments

9. The first notable difference in the aggregation structure comes with the introduction of consumption segment level indices. We refer to "consumption segments" as this is commonly used international terminology when discussing a 6th, or more detailed, level in the COICOP hierarchy.
10. Consumption segments are slightly broader than the current item definitions, though still defined based on a 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 "women's full-length leggings"; by broadening the consumption segment definition to "women's leggings" we use more of the alternative data.
11. However, to enable continuation of the RPI in its current form, we intend to maintain the more granular item level elementary aggregates when collecting from physical stores. In the CPIH and CPI, these items will then be treated as representative of the consumption segment and will aggregate together with alternative data for the broader consumption segment definition to form the higher-level index (see Figure 1b).
12. A further benefit of having broader definitions is that it can improve the potential of our classification. For example, it is easier to identify from text descriptions whether a garment is a "legging" than a "full-length legging", the length of a garment is not always specified.
13. Adding in a consumption segment level also enables us to realign our detailed CPIH and CPI indices with the COICOP hierarchy, improving user navigation. Currently our item codes are aligned with the RPI hierarchy which cannot be easily read across into the COICOP hierarchy without a mapping file, for example basmati rice has the item code 210212 within the COICOP5 category "Cereals" 01.1.1.1.
14. Some NSOs have modified their hierarchy so that the retailer hierarchy is used beneath the COICOP 5 level. We have chosen not to do this for the following reasons:
    a. the inconsistencies in how retailers organise their hierarchies would prevent us from having consistent sub-aggregates to calculate contributions. For example, one retailer may organise their fruit by country of origin, whereas another may organise by type. The total retailer contribution towards fruit would be calculated, rather than contributions from specific products.
    b. the retailer hierarchy is not always intuitive. Often, the lowest level in the retailer hierarchy maps to multiple consumption segments across the COICOP hierarchy (e.g.

[1] Multilateral methods will be used at the elementary aggregate index level for scanner and web-scraped data sources, but these will be aggregated with traditional sources using Lowe-type index methodology

"cake decorations" contains icing, chocolate chips, cupcake cases and candles) and some products can be misclassified by the retailers (e.g. pears in the apples category)

Classifying individual products in the alternative data to consumption segments is therefore the best chance we have of ensuring maintained accuracy, homogeneity, and interpretability of our elementary aggregate indices.

Realising the full potential of alternative data sources: how many consumption segments to include?

15. Items are traditionally chosen for the consumer price statistics basket to be representative of their broader group. For example, we may collect lemons as being representative of citrus fruit, or garden spades as being representative of all garden tools. With alternative data, we have access to prices for a near-census of items, so by maintaining our current item samples within the consumer prices basket, we may not realise the full potential of these new data.
16. We currently collect prices from physical stores for approximately 215 distinct items in COICOP divisions 1 and 2 (food and non-alcoholic beverages, alcoholic beverages, and tobacco). With our scanner groceries data, we have been classifying to over 650 consumption segments within COICOP 1 and 2, more than three times the number of categories[2].
17. The question of how many consumption segments to include therefore becomes of utmost importance. The conservative option would be to introduce consumption segments in line with the current sample of items. This maintains consistency with our current sampling approach and, as all retailers and data sources should be available for all consumption segments, allows a statistically controlled influence from different retailers and data sources. However, much of the alternative data would remain unused. An estimate of the amount of data that become unused at each stage in the process is provided in Annex A.
18. The more liberal approach is to introduce a full complement of consumption segments (under any given resource constraints) to try and realise the full potential of the alternative data. This will provide better coverage of product categories and potentially reduce some volatility in the aggregate index. However, without bolstering our collection in physical stores[3] to ensure that we have items representative of each consumption segment, we run the risk of alternative data retailers dominating the index. In the groceries example above, we would have ~1/3 of the indices with a full complement of retailers from all data sources, but ~2/3 of the indices would be based only on retailers for which we have alternative data.
19. For groceries, we currently have scanner data from retailers that together account for less than 50% of the food and drink market. We are hoping in the coming months we can complete this process for some further retailers, increasing our market share coverage using alternative data alone to ~75%. **We need to decide what proportion of the market needs to be covered if we are to take a more liberal approach to introducing new consumption segments, to ensure that we do not have a small number of retailers unduly influencing category indices, and the consumption segments introduced are not disclosive to individual retailers.**

[2] Note that, even then, we are not making use of all the data, some of which does not fit in COICOPs 1 or 2, or our current consumption segment definitions within these divisions.

[3] To enable consistent production of the RPI in its current form, we have decided to not substantially refocus our field collection to improve coverage of new consumption segments

20. While we can account for a large proportion of the market with a small number of retailers for grocery categories, this is not as achievable for other product categories such as clothing. Therefore, there may need to be a different approach to introducing consumption segments, or the threshold for which we decide to take a more liberal approach, dependent on the class of product or market dynamics.

21. If we do take a more liberal approach to introducing consumption segments, we will likely be left with empty stratum cells in the hierarchy where we have not collected prices for some consumption segments in physical stores. Here we could either use an implicit imputation (under the assumption that the remaining stores will have similar price movements to those that we have ADS for), or an explicit imputation based on a parent or neighbouring index. This is explored further in paragraphs 41:43.

Use of the Annual Business Survey to produce market share weights

22. The Annual Business Survey (ABS) publishes financial information from businesses representing the UK non-financial business economy (about two-thirds of the UK economy). The ABS surveys businesses in the UK.

23. Our current method of weighting elementary aggregate indices together typically involves using ABS data to give weight to stores that are either multiples (more than 10 stores) or independents (less than 10 stores). The main problems with this current shop-type stratification are that:

    a. it gives equal weight to retailers that are quite different, e.g. small newsagent chains, discount stores and large supermarkets

    b. it misrepresents predominantly online retailers which may have few physical outlets (and therefore classed as independents) but still have a large market share

    c. it doesn't provide an easy means to integrate indices from scanner and web-scraped data with the local collection

24. Furthermore, other NSOs appear to typically opt for weighting retailers by market share (see Annex B). We therefore propose to use the ABS reported turnover to calculate the market share data needed to weight retailer indices using scanner or web-scraped data together with the retailers where prices are collected from physical outlets (Figure 1b).

25. Although we currently use the ABS to calculate shop-type weights, weighting individual retailers or groups of retailers by their market share requires using the data at a more granular level. Inspection of the data at this level has shown some outliers, so we are working with data collection teams to improve quality assurance and year-on-year checks at this more granular level moving forward.

26. We have chosen the ABS as the most suitable potential data source because:

    a. the commodity categories that turnover is reported within are the most granular we have found of available data sources. This helps prevent the weighting of consumption segments being affected by the sales of unrelated products

    b. all retailers with >250 employees are surveyed, meaning there is good coverage of retailers with large market share

27. The limitations of the ABS are that:

    a. the data are only available with a significant lag, for example, revised estimates for 2020 are not available until May 2022, so could not be used in consumer price statistics until the start of 2023. If there are any big shifts in market share over this 3-year period, they will not be appropriately reflected in our indices

    b. the commodity categories are still broader than item or consumption segment. As a result, some retailers will not sell every item that maps onto a commodity category. For example, a large bakery chain may have a high market share in the cereals and bread commodity category in some years but does not sell some of the items that fall into that category, such as pasta

    c. the reporting unit is national, so the data do not give information about regional turnover, so we cannot calculate region-specific market shares. This would have been helpful where certain retailers are more predominant in certain regions. One way this could be approximated would be by weighting the national turnover by number of local units or employees in a region, with the assumption that the location of outlets, staff and spending is positively correlated. However, this would be an approximation, and subject to bias due to the location of non-retailer locations, such as distribution hubs and headquarters. Instead we plan to apply national market share data to retailers within each region, and weight each regions indices together using another data source

Strata and market share weighting

28. Currently, price quotes are aggregated into strata of either shop-type, region, both shop-type and region, or not stratified at all. Two types of shop are identified for the stratum weights, multiples and independents, and there are 12 regions (including Northern Ireland, Wales and Scotland). This means that price quotes are typically aggregated into either 2, 12, 24 or 1 strata/stratum. For food and non-alcoholic beverages, approximately 3/4 of items are currently stratified by region, and 1/4 are stratified by both region and shop-type.

29. If we were to calculate consumer price indices giving every retailer an individual weight based on their market share, this would significantly increase the number of strata and we would run into issues with sample sizes and availability. Therefore, we only intend to use explicit market share weighting for retailers whom we have alternative data sources for (which have been chosen due to their market dominance and who we are confident we will have sufficient sample sizes/ongoing availability of data).

30. For the remaining retailers in the local collection, we propose aggregating price quotes into two market share strata within each region; one stratum for retailers with large market share (more than 2%) in the relevant commodity category, and one stratum for retailers with low market share (less than 2%) – see earlier Figure 1b. This enables us to better account for online-only retailers and make a clear distinction between small and large retailers.

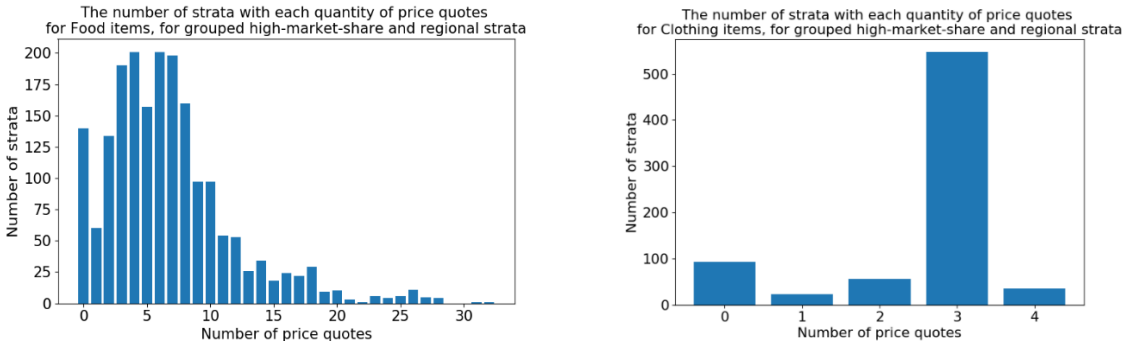
31. We suggest 2% as the threshold for market share strata as this typically provides a good split between the larger and small retailers. A higher threshold would result in different retailer types being grouped into the ‘low market share’ category, for example, discount supermarkets and independent butchers; a lower than 2% threshold would also result in this.

32. As not every item is weighted by both shop-type and region currently, strictly imposing a market share strata split within each region will either increase the number of strata that price quotes are spread across or change the distribution of price quotes across strata.

33. After removing price quotes for the retailers for whom we have alternative data for, the distribution of sample sizes in the remainder of the traditional (local) collection was investigated when imposing 12 regional x 2 market share (< or > 2%) strata on all current items within COICOP 1 (food and non-alcoholic beverages) and COICOP 3 (clothing and footwear). This analysis uses a single month's data from 2019.

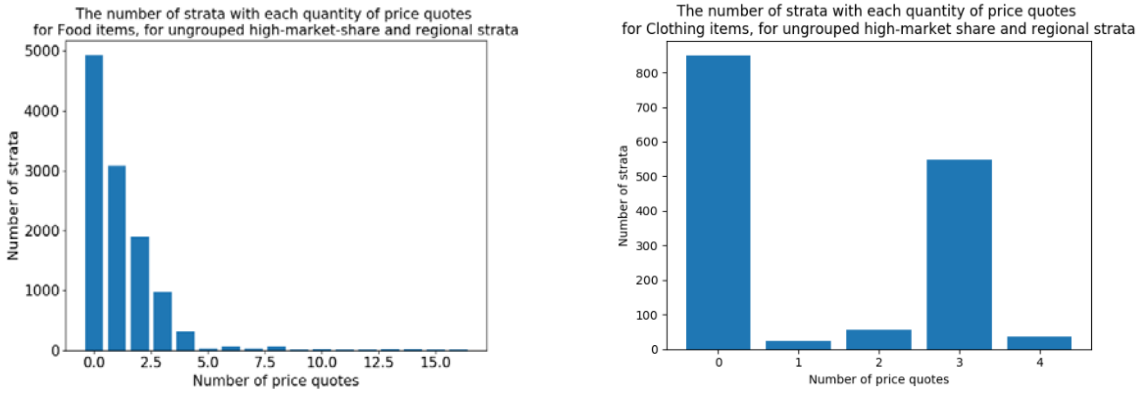
34. For retailers with above 2% market share within each region (Figure 2), the proposed aggregation results in some empty strata, particularly for food. This may be because some items are not available in every region, as well as the removal of 3 large retailers whom we have alternative data for.

Figure 2: Distribution of sample sizes in imposed strata for food and clothing in each region, retailers with >2% market share grouped



35. For comparison, price quotes were also aggregated weighting each unique retailer within each region above the 2% market share threshold using their respective market share (Figure 3). E.g. a single strata would be dessert apples, from Retailer A, in London. However, this left far more strata empty, and tended to give smaller sample sizes per strata as not all retailers have prices for all items in all regions.

Figure 3: Distribution of sample sizes in imposed strata for food and clothing in each region, retailers with >2% market share ungrouped

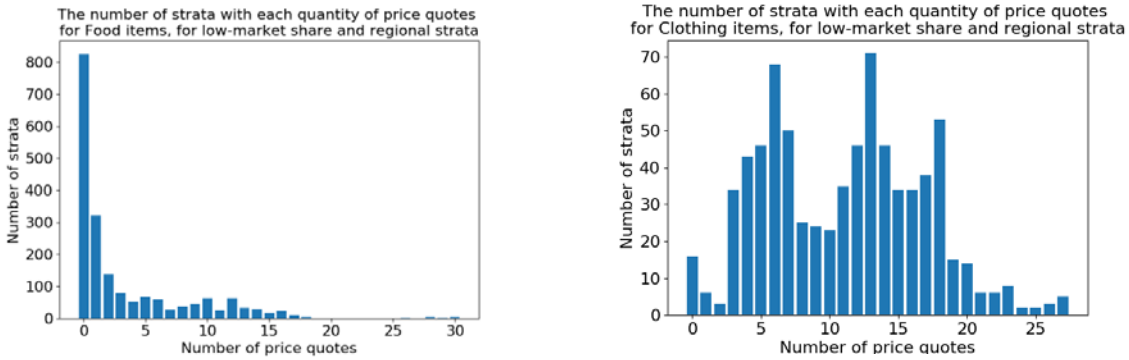


36. Grouping together retailers over a given threshold still loses some specificity by grouping together multiple retailers. However, the retailers with more than 2% market share tend to be of similar types. On balance, we decided this is preferable to having several missing strata.

In addition, using market share rather than number of physical outlets means that this is still based on spending patterns, as opposed to the current shop-type strata.

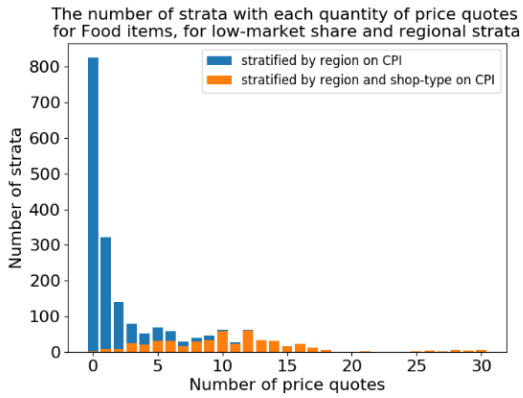
37. Figure 4 shows the distribution of sample sizes for retailers with less than 2% market share within each region. See that, for food, a large proportion of strata are either empty or have few price quotes, this is because many items are not stratified by shop-type currently, and are therefore only collected from larger stores (multiples and/or those with >2% market share).

Figure 4: Distribution of sample sizes in imposed strata for food and clothing in each region, retailers with <2% market share grouped



38. Figure 5 shows that, in the food category, items that are currently stratified by both shop type and region tend to have a greater sample size, in strata grouping retailers with less than 2% market share, than those currently only stratified only by region. In items currently only stratified by region, the collection is dominated by supermarkets, leaving very few price quotes in strata when grouping retailers with less than 2% market share. This is, therefore, not a result of imposing the new strata methodology but is also present currently.

Figure 5: Distribution of sample sizes in imposed strata for food in each region, retailers with <2% market share grouped, broken down by current stratification



39. In summary, we propose a below-item-level aggregation structure that uses ABS data to:

    a. give retailers we have alternative data for weights corresponding to their respective market shares

    b. group remaining retailers that have a greater than 2% market share, and weight accordingly

    c. group remaining retailers that have a less than 2% market share, and weight accordingly

40. We do not have web-scraped or scanner data for all categories within the consumer prices basket, so for some items there will not be any elementary aggregate indices calculated from web-scraped or scanner data that get given the retailer's market share as a weight. For these categories we propose still adopting the new market share strata, this will be roughly equivalent to current shop-type stratification but will better differentiate different store types and account for online-only retailers.

Handling missing strata

41. Based on our new stratification proposals and the potential to introduce more consumption segments based solely on retailers for whom we have alternative data, it is likely we will have some missing stratum level indices. For these cases there are a few options that could be used.

42. If we are missing price quotes for a group of retailers (e.g. retailers with <2% market share) within a region (e.g. London) within a consumption segment (e.g. apples), we could use:

    a. **implicit imputation**: basing the index solely on the retailers whom we do have data for (e.g. ADS retailers and retailers with >2% market share). This would imply the assumption that we think price movements are item driven and the price movements for retailers who we do have data for are effective approximations of the retailers who we don't have data for

    b. **nearest neighbour imputation**: imputing the price movement based on price movements of the consumption segment (apples) for the same retailer group (<2% market share) within a different region (e.g. South East) or based on a different consumption segment (pears) for the same retailer group (<2% market share) within the same region (London)

    c. **nearest parent imputation**: imputing the price movement based on price movements of the remaining consumption segment (apples in all regions from all available retailers), incorporating price movements from different retailers and regions as the best approximate for the price movement of the missing strata

43. We have considered these three options (see workbook sent as an addendum). While the nearest neighbour imputation may offer a closer approximation in some instances, it is somewhat subjective as to what neighbour index should be chosen. The nearest neighbour also could change over time. We therefore consider the **nearest parent imputation as the most appropriate way of handling missing strata**. This option accounts for price movements within the same retailer or group of retailers within different regions as well as the price movements of other retailers, or groups of retailers, within the same region.

Handling missing consumption segments

44. There also some points in time where entire consumption segments are not available within a month, or consecutive months. This is most likely the case when considering seasonal items that are only available for part of the year, though throughout the COVID period there have been several occasions where consumption segments have been entirely unavailable too.

45. While there may be availability for seasonal items in out-of-season months in the alternative data, we propose to not make use of these data as low demand in out-of-season months could cause unusual price behaviours. We do however want to capture the effect of price

movements for seasonal products when they are in-season, to be truly representative of consumer expenditure. Carrying forward the last seen price for seasonal or unavailable products is also undesirable, as it could bias the index towards being stable.

46. There are currently conflicting methods for dealing with seasonal items and unavailable items. We therefore intend to make our approach more consistent, aligned with our approach with the fixed-basket principle underlying our consumer price indices. Although annual fixed weights will not be representative of the monthly consumption pattern, it is conceptually consistent with a fixed-basket index and ensures weight changes are not reflected in the monthly price change.

47. Using a fixed-basket approach, we see two potential options for the imputation of seasonal and unavailable items:

    a. **parent imputation,** if the product is out-of-season in the base period, use the last observed price to be used as a base price[4]. Calculate the imputation factor for out-of-season items based on the monthly index movement of the COICOP 4[5] level index that the consumption segment belongs to. This is based on the current approach for dealing with seasonal items (see section 9.5.1)

    b. **all items imputation,** if the product is out of season or unavailable, impute the missing price based on the all-items index to ensure the missing category is not impacting on the headline rate (for non-seasonal items, use the monthly price index of all available items; for seasonal items, use the annual price index of all available items). This is based on the current approach for dealing with unavailable items that have occurred due to COVID-19

Summary

48. We are restructuring the consumer prices hierarchy beneath the COICOP 5 level to allow us to effectively integrate alternative data sources, maintain the RPI in its current form, more readily produce regional indices, and provide a more effective means of weighting different retailers together.

49. Our new approach to weighting retailers together improves some of the shortcomings of our current approach but results in imposing strata that have not existed until now, some of which are likely to remain empty. The introduction of new consumption segments to realise the full potential of alternative data sources is likely to further increase the number of missing strata, if we are to take a more liberal approach to their introduction. Therefore, there are several key questions we are yet to answer as addressed throughout this paper.

Helen Sands and Annabel Summerfield
Prices and Methodology Division, ONS
July 2021

[4] Is this last observed price going to cause a downward bias?

[5] COICOP 4 level has been chosen as some consumption segments are unique within a COICOP 5 level, using the COICOP 4 level provides consistency across calculation, but goes against our approach for imputing missing strata based on the nearest parent.

List of Annexes

Annex A	Sankey diagram showing data loss through classification and choice of consumption segments
Annex B	Some other NSOs approach to weighting retailers

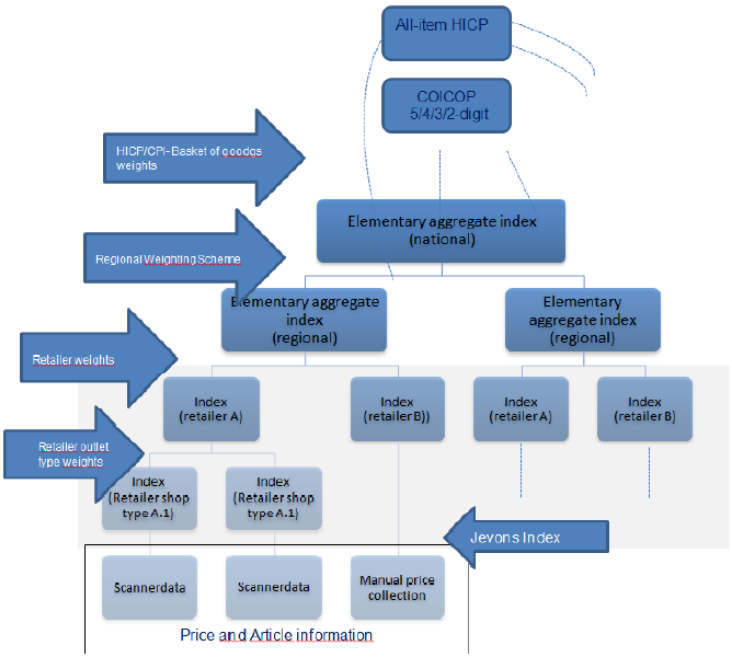
Annex A: Sankey diagram showing data loss through classification and choice of consumption segments

[REDACTED]

Annex B: Some other NSOs approach to weighting retailers

Belgium	
Integrating Big Data in the Belgian CPI – Ken Van Loon, Dorien Roels (May 2018)	'To combine scanner data and classical price collection data, expenditures by retailer and their specific outlet types are used to calculate the weights for these new strata' (Annex 2 p. 16) Indices at the ECOICOP 5-digit level are combined with other data (web scraping, manual price collection, …) using a stratification model in which each stratum and retailer gets a weight based on expenditure or market share.
Sweden	
Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket – Can Tongur (July 2018)	'…observations, relative prices, within each retail chain (=stratum) are averaged and summarised to the product group by weight in with the average market share of each retailer to result in the equation for the complete product group.' (p. 40) 'weights are normalised so that depending on the number of retail chains within each product group, the retailers' average relative price is assigned to a priori known weight'
Finland	
Chain Error as a function of Seasonal Variation - Kristiina Nieminen, Yrjö Varti a, Antti Suoperä, Satu Montonen (May 2019)	'The scanner data elementary aggregates are integrated together with the traditionally collected and processed elementary aggregates using enterprise-specific weights' (p. 15)
Austria	
From price collection to price data analytics - Josef Auer, Ingolf Boettcher (2017)	Currently price data is not weighted until the regional level. With scanner data, additional levels of aggregation are intended to be used primarily at the retailer level. 'To combine scanner data and classical price collection data, expenditures by retailer and their specific outlet types are used to calculate the weights for these new strata'

Chart 1 - *overall price index compilation for the Austrian HICP with scanner data*



(Annex 2 p. 16)

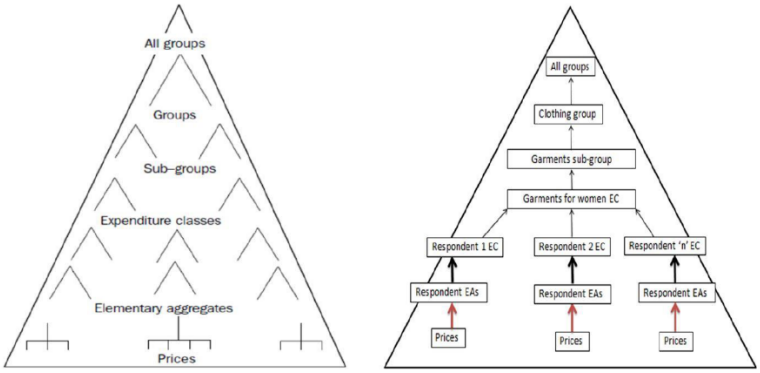
Australia

Experimental clothing indexes using Australian web scraped data – Andrew Glassock, Michael Holt (May 2019)

ABS now apply a modified aggregation structure when using scanner data below the expenditure class level in order to capture possible retailer-specific effects.

'Instead of using a single elementary aggregate, prices collected through scanner data are aggregated to retailer or respondent elementary aggregates and expenditure classes. Respondent expenditure classes are then weighted by the retailer's market share to obtain the expenditure class indexes.'

Figure 2: Current and Proposed Aggregation Structures



Expenditure class indexes are put together by aggregating the retailer elementary aggregate indexes to the retailer expenditure class level and averaging the movement across retailers.

'In the first stage, expenditure class indexes for each retailer are calculated by weighting each elementary aggregate by expenditure shares obtained by weighting each elementary aggregate by expenditure shares obtained from the Household Expenditure

	Survey (HES). However, aggregation across retailers is more problematic since information about expenditures on each retailer expenditure class is required. The ABS Retail Trade Survey provides a natural starting point for obtaining expenditure weights for each retailer, although not all retailers represented in out sample are included in the survey. Expenditure information from supplementary data sources is therefore also required to attain expenditure weights for these retailers.' (p. 8) 'They have identified alternative assumptions for weighting individual products for clustered product definitions as an area of possible investigation for the future.' (p. 17)
Eurostat	
Elementary aggregation: A not so elementary story! – Claude Lamboray (May 2019)	First elementary aggregates are defined based on available expenditure data. In addition to the product dimension, it is common practice to stratify according to the outlet or the regional dimensions. Some additional information, such as market shares, can be used to explicitly or implicitly weight the sampled prices during this step of aggregation.
Italy	
Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP – Polidoro, Giannini, Lo Conte, Mosca and Rossetti (Istat, 2015)	Aggregation of micro indexes by geometric means (elementary level) and by weighted arithmetic means (upper levels). Weights (here available) are proportional to the market shares of each brand and each segment. p. 169