ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Scanner data research – date trimming

Status: Draft of future publication

Purpose

1. Date trimming involves only using data from specific dates within a month to measure price indices. In this paper we outline how date trimming may be used in grocery scanner data.
2. In this initial iteration of the paper, we will present the theory behind date trimming, some preliminary analyses, and a provisional decision on usage of trimming, but impact on indices will be excluded. This is because pipeline development for scanner grocery index production is currently ongoing. We plan to follow up this paper with a second iteration containing these further analyses and a final proposed decision on trimming in October 2023.

Actions

3. Members of the panel are invited to:
    a. Advise on whether the outlined Scenario 1 is an appropriate (provisional) choice for date trimming
    b. Advise on any further issues that could be caused by implementing date trimming

Background

4. In traditional consumer price statistics, we typically measure inflation through point-in-time price collection, where prices are collected for most products once a month. An advantage of alternative data sources is to use information beyond a single day to give a better representation of the average transaction price paid by the consumer. For weighted data sources like scanner data, this means creating unit values (total expenditure sold divided by total units sold) for homogeneous products, and for unweighted data sources like web scraping, this means averaging the prices observed across the month. We will use "representative price" as a generic term for both unit values and price averaging. In both instances, we must decide on which days to use each month when calculating representative prices used in our monthly price indices.
5. From a methodological perspective, ideally, we would calculate representative prices using every day of the month. However, from a practical perspective this may not possible or preferred for a few reasons:
    a. While some datasets are provided at transaction level, for our grocery scanner data, transactions are typically aggregated daily or weekly. In weekly-aggregated datasets, some weeks can overlap two consecutive months. Since we do not have daily information, it is difficult to separate these weeks and so including an entire month's worth of data in every month is not possible.
    b. Within a monthly production round, it may be beneficial to only use an earlier portion of the month so that index compilation can begin earlier, giving more time for quality assurance.
    c. There may be rare instances where some days or weeks of data are missing due to issues with the data supply or ingest processes; where data from some portion of the month is of a lower quality and deemed unusable; or where the data is received too late to be included in the monthly calculations, and we may then need to use a smaller portion of the data from that month.
6. One method of handling the first and second reason could be date trimming. Date trimming involves filtering monthly datasets down to a timeframe where transactions within that

timeframe are in scope for measuring inflation, and where transactions outside of that timeframe are dropped from index calculations. For example, we could decide to use three full weeks of data for every month.

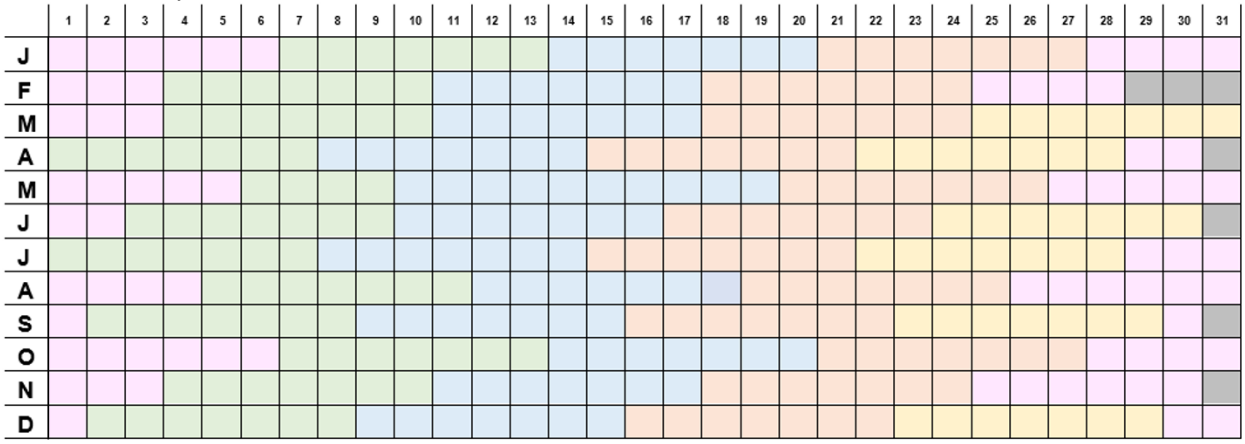
7. Previous literature has found conflicting results when comparing indices using different time aggregations. For example, time aggregation choices were found to lead to different price change estimates even for superlative indices, and it has been recommended that the unit value prices used for constructing CPI should be for the same period as the index to be constructed. This means that to produce CPI for a full month, the unit value price for the whole month should be used, rather than a representative shorter time period, to avoid upward bias of the index. On the other hand, a Luxembourg study (PDF 2.2MB) found little difference between using three and four weeks of data to construct price indices.

Date Trimming Options

8. Before introducing date trimming, we should research the impact of different scenarios on our indices to ensure that we are not producing biases. There are three scenarios to be investigated:

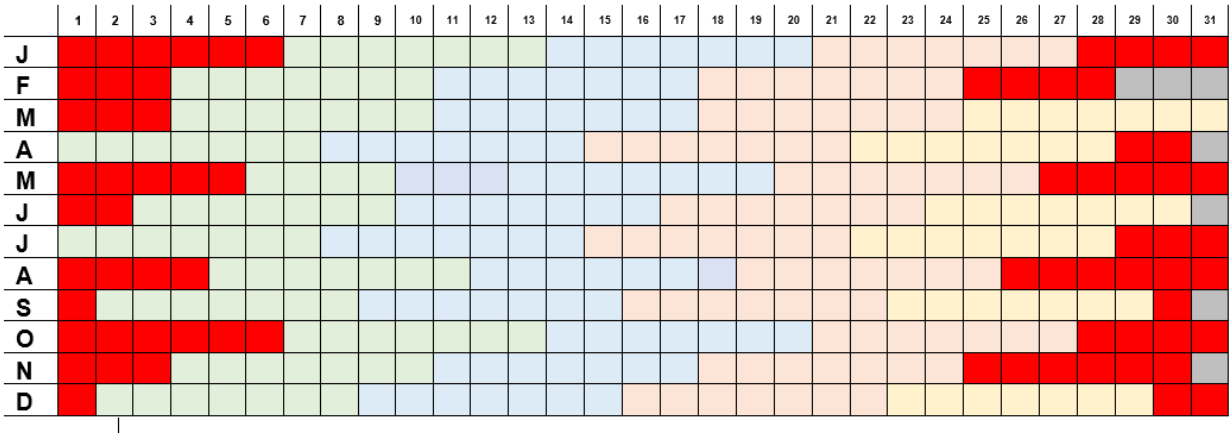
    a. **Scenario 1** is to use all data available to calculate the indices. This would mean using all days in the month for retailers where we receive daily data (Figure 1), and either three or four weeks which fall fully into each month for retailer where we receive data on a weekly basis (Figure 2).

Figure 1. Visualisation of days used and lost when employing Scenario 1 in 2023, for retailers where we receive daily data.



    b. **Scenario 2** is to use all weeks which fall fully into each month regardless of whether we receive weekly or daily data. This would mean using three or four weeks for each month dependent on how the days fall in each month (Figure 2).

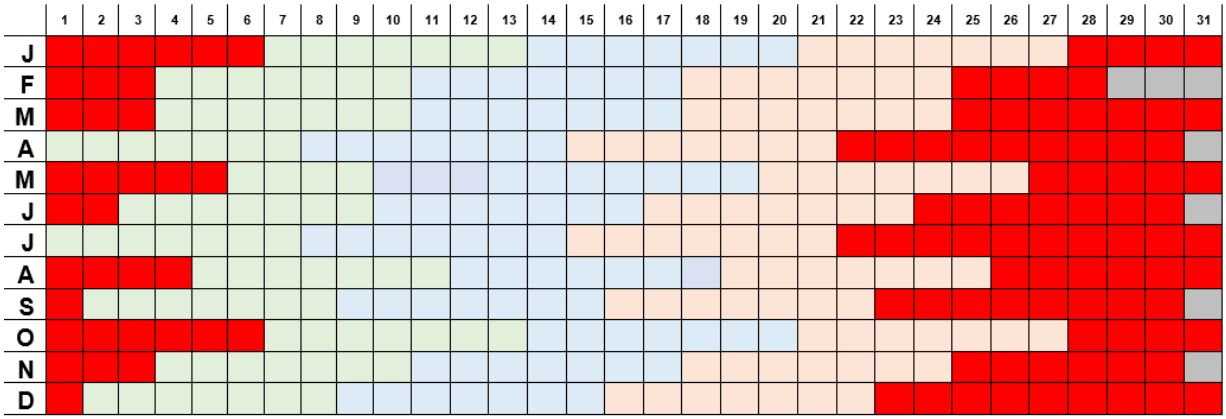
Figure 2. Visualisation of days used and lost when employing Scenario 2 (or Scenario 1 for weekly data) in 2023.



N.B. This would also be the case for retailers who deliver weekly data in Scenario 1.

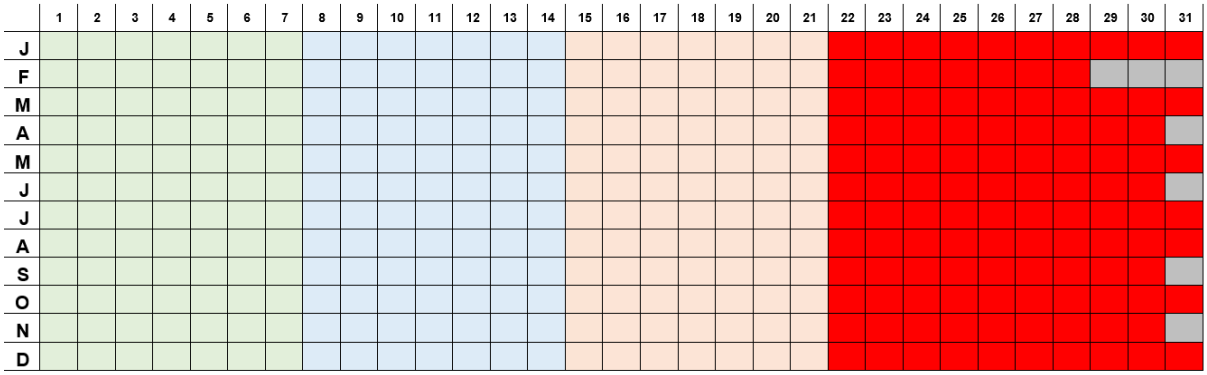
c. **Scenario 3a** is to use a consistent, fixed timeframe every month. This would entail using the first three full weeks in each month to calculate indices and disregarding the rest of the data (Figure 3).

Figure 3. Visualisation of days used and lost when employing Scenario 3a in 2023.



d. There may be scenarios where we receive a monthly delivery of data that covers a specific timeframe that is not represented by the other scenarios. For example, the retailer may only be able to provide the first 21 days in the month. **Scenario 3b** shows the data that would be kept or removed according to this example, but we would likely also use the other scenarios for data sources where data is provided on a more frequent basis.

Figure 4. Visualisation of days used and lost when employing Scenario 3b for daily aggregated data in 2023.



9. Each of the scenarios has various advantages and disadvantages. We have described these below with reference to the five quality dimensions.

10. **Accuracy.** All three scenarios result in some measure of data loss. Scenario 1 uses all data available, meaning a full month of data for those retailers who deliver daily data, and up to four weeks of data for those who deliver weekly data, resulting in a ~19% data loss for some retailers. Scenario 2 involves an equal data loss of ~19% for all retailers since we only use full weeks for all retailers. Scenario 3 results in the most data loss, as we lose approximately 31% of data over the year by using a fixed time period. The comparative scale of data loss can be seen in Figures 1-4. Furthermore, the scenarios differ in the amount of time that they allow for quality assurance. Scenario 1 gives us reduced, and variable, time to scrutinise our indices before publication. Because we receive the data on the same day of the week, we may have to wait multiple days between the end of the month and receiving the data. This means that the time we have to calculate and scrutinise the indices is less than in Scenarios 2 and 3, which may increase the risk of less accurate indices. We can mitigate this risk substantially by producing and scrutinising interim indices as we accumulate data throughout the month, giving additional time to quality assure the data. This is already something we implement elsewhere (for example, in rail fares), so is likely to be implemented for groceries. Scenario 2 also, in some cases, gives us reduced time to scrutinise, depending on where the last day of the month falls, whereas Scenario 3 provides us with the most time in between receiving the data and publication.

11. **Relevance.** There is potential for Scenarios 2 and 3 to be slightly less relevant or interpretable to the end user. This is because the user requires a monthly index, but Scenarios 2 and 3 will, in some cases, provide them with an index constructed on a smaller timeframe. However, since we will ensure that the chosen scenario has the least impact possible on the final index, this is not anticipated to be a problem. We also note that all of these scenarios provide greater time coverage than our traditional data sources.

12. **Timeliness.** None of the three scenarios will affect the punctuality or timeliness of the published indices.

13. **Clarity.** All three scenarios offer similar levels of ease of explanation because the methods are not complex to explain and the format of the indices outputted from each scenario will remain the same.

14. **Coherence/consistency.** Scenario 1 involves using an inconsistent amount of data **across** retailers, since retailers providing daily data will have more data feeding into their indices than retailers providing weekly data. This is not considered a methodological issue since indices are stratified by retailer, and although it could affect comparability of our indices between retailers, it is not anticipated to be an issue because we do not publish indices to this level. Similarly, Scenarios 1 and 2 involve using an inconsistent amount of data **within** retailers, since each retailer can be represented by a different number of days/weeks each month. This could affect the comparability of our indices across different months and years, because in one year a month could have three full weeks falling into it, and in another year have four full weeks. However, assuming that the three- and four-weekly price and quantity distributions are similar, there is a question of whether the scaled difference in quantities causes any measurement changes. For example, does a three-weekly price and quantity of £2 and 30 units sold give the same result as the four-weekly price and quantity of £2 and 40 units sold? This is a question of whether the multilateral index method with window length T is invariant to proportional changes in quantities in one (or more) months, or in other words, does the following equality hold for any generic month i and scaling factor λ:

$$P(\{\boldsymbol{p_1}, \dots, \boldsymbol{p_i}, \dots, \boldsymbol{p_T}\}, \{\boldsymbol{q_1}, \dots, \boldsymbol{q_i}, \dots, \boldsymbol{q_T}\}) = P(\{\boldsymbol{p_1}, \dots, \boldsymbol{p_i}, \dots, \boldsymbol{p_T}\}, \{\boldsymbol{q_1}, \dots, \lambda\boldsymbol{q_i}, \dots, \boldsymbol{q_T}\})$$

15. In the case of the GEKS-Törnqvist, this equality clearly holds since the underlying Törnqvists use expenditure shares as weights, which are unchanged regardless of quantity scaling in any given month. Since none of the underlying Törnqvists are changed, neither is the GEKS-Törnqvist. Therefore, quantity scaling changes due to changes in the amount of time used each month is not considered a concern. Scenario 3 is the most consistent and comparable of the scenarios outlined here, since it uses the same amount of data for each retailer and each month.

16. A summary of these various advantages and disadvantages of the three strategies is given in Table 1.

Table 1. Summary of Scenarios and impact on amount of data used.

Scenario	Weekly aggregated data summary	Daily aggregated data summary	Data Loss	Pros	Cons
Scenario 1	Use all weeks falling fully in month (figure 2)	Use all data available (figure 1)	Daily: None Weekly: ~19%	Minimises data loss	Puts pressure on publication timings Inconsistent use of time **within** and **across** retailers
Scenario 2	Use all weeks falling fully in month (figure 2)	Use all weeks falling fully in month (figure 2)	~19%	Consistent use of time **across** retailers	Data loss of c. 19% which could lead to biased indices Inconsistent use of time **within** retailers
Scenario 3	Use three weeks falling fully in month (figures 3 and 4)	Use three weeks falling fully in month (figures 3 and 4)	~31%	Consistent use of time **within** and **across** retailers Increased time for scrutiny	Data loss of c. 31% which could lead to biased indices

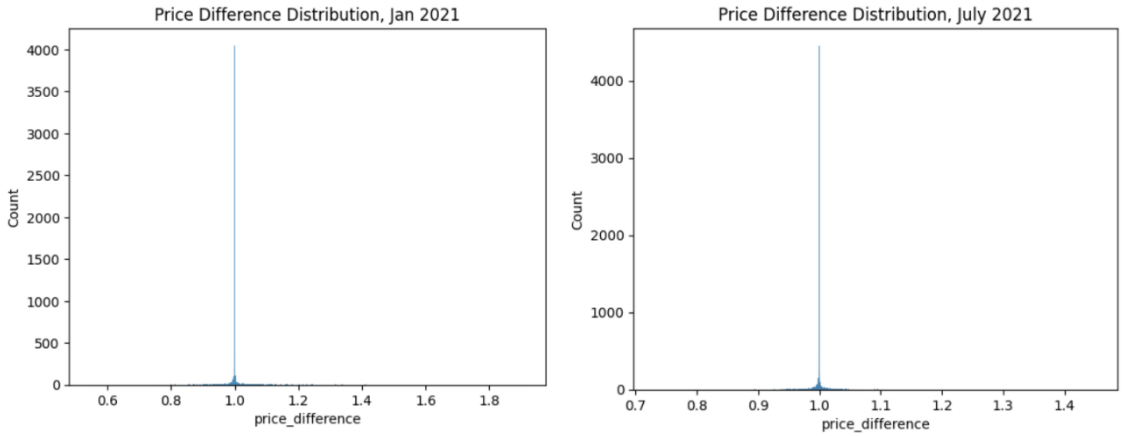
17. Disregarding any amount of data could potentially introduce some bias to our indices. This could be particularly prevalent in months where large seasonal events such as Easter occur, or where weather events affect the weight of different products in the market.

Preliminary analyses

18. To examine whether representative prices calculated using three weeks of data would be similar to those calculated using four weeks of data, we calculated a ratio between these two prices by dividing the three-weekly price by the four-weekly price. A price difference ratio of 1 would indicate no difference in price, and a price difference ratio of 0.75 would indicate a 25% reduction in average price when calculating over three weeks compared to four. We plotted the distribution of price difference ratios for all grocery products in one retailer in January and July 2021. The results are shown in Figure 4.

19. We recognise that the results shown in this section are not generalisable as they do not represent all the data available. We analysed data from one retailer in January and July of 2021. This is for simplicity, and because we needed to choose months which contained four full weeks of data to compare three-weekly and four-weekly prices. January and July fit this criterion and are at different times of year, so we chose them as relevant months for the analysis. Some additional summary information is provided in Table 2.

Figure 4. Distribution of price difference between three and four weeks of data for grocery products in January 2021 and July 2021, for retailer a



N.B. Price difference is calculated as three-weekly price / four-weekly price

Table 2. Descriptive statistics for the price difference between three-weekly and four-weekly prices for grocery items in 2021.

Measure	Jan 2021	July 2021
Count	13,545	13,638
Mean	1.002	0.9995
S.D.	0.038	0.021
1st percentile	0.909	0.938
10th percentile	0.982	0.985
Median	1.000	1.000
90th percentile	1.021	1.013
99th percentile	1.132	1.059

20. The 10th and 90th percentiles in Table 2 show that 80% of products do not differ by more than around 2% when calculating their prices using three weeks or four weeks of data. This may lead us to consider Scenarios 2 and 3 which give us more time to scrutinise our indices.

Provisional decision and future work

21. The default position should be to maximise use of the data using Scenario 1 unless there is a compelling reason not to. There are two potential drawbacks to consider in this light.

22. The first potential drawback is the inconsistency both within and across retailers in each month. As previously discussed, this is not likely to cause issues methodologically – even so, consistency should still be seen as a target. However, we consider mitigating the risk of bias from using smaller portions of the month as being preferred to improving this consistency in this way.

23. The second potential drawback is that using this scenario reduces the amount of time for scrutiny. This practical consideration is a bigger concern for us. However, we can mitigate this risk by producing weekly indices as the data is received, allowing us to scrutinise the data on an ongoing basis throughout the month. We are still in the process of finalising

timescales for a monthly production round within the context of scanner grocery data and so it is currently unknown how much risk each scenario would introduce.

24. We therefore recommend the following provisional decision:
    a. Our ideal option is to use Scenario 1 to maximise the use of data, provided that doing so does not introduce an unacceptable level of risk due to the reduction in scrutiny time in the monthly production round.
    b. Scenarios 2 and 3 will be considered if monthly production round timescales are considered a concern.
25. For future work, we plan to finalise our timescales for the monthly production round and study the impact on indices of using the various scenarios, considering the interaction of this timing with our classification and relaunch linking processes each month. We then plan to return to the panel with these analyses and a final proposed decision.
26. Finally, it is worth noting that, although we anticipate being able to use Scenario 1 for most prices projects, we will review this date trimming decision for each goods category, since it could affect some categories more than others.

Laura Christen and Liam Greenhough
Consumer Prices Methods Transformation, ONS
July 2023

List of Annexes

No annexes