ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Scanner data research – data cleaning

Status: Draft of future publication

Purpose

1. In this paper we outline the data cleaning methods (primarily outlier detection and junk filtering) that we are exploring the use of with grocery scanner data.
2. This paper focuses on the theoretical approaches that we are planning to take to determine the best methods and gives some preliminary studies performed to determine our strategy on grocery data.
3. We anticipate releasing an updated version of this paper with the complete analysis on indices for all retailers available, and with our final recommendations for methods to use for outlier detection in grocery scanner data.

Actions

4. Members of the panel are invited to:
    a. Advise on the combined use of price filters with price-quantity dump filters
    b. Provide general feedback on this draft

Background

5. In our previous publication Outlier detection for rail fares and second-hand cars dynamic price data we presented a study of different outlier detection strategies applied to second-hand cars and rail fares transactions, where we indicated the price relative-based user-defined fences (UDF) with thresholds of 1/3 and 3 as our outlier detection strategy.
6. This document outlines the studies performed on grocery scanner data. It focuses on the theoretical approach to outlier detection, and it drafts the analysis we want to perform, together with presenting some early studies that will help us define our strategy.
7. Data cleaning - similarly to the case for rail fares and second-hand cars - is used to determine which observations within the data are in scope to calculate the indices. This approach removes observations that would introduce biases to the index calculation. The overall data cleaning is a two-step approach:
    a. **Junk filtering** uses variables on the dataset to determine observations that are not in scope and therefore should be removed prior to index production. The filters applied are pre-defined and are specific to a goods category.
    b. **Outlier detection** is used to identify and remove products showing extreme prices (quantity) and/or price (quantity) movements.
8. For grocery scanner data, we will use junk filters to remove the following types of observations:
    a. Products sold by weight where we do not have the weight of products sold
    b. Those with erroneous size information
    c. Those which are not genuine sales
    d. Those we cannot link to a UK region

e. Those without a suitable product identifier

f. Those without store type information

9. From preliminary studies performed only on one retailer, the junk filters described above remove approximately 1.545% of the transactions.

10. Based on the results from our previous publication, we are exploring only relative-based outlier detection methods. These methods are not designed to flag the individual transactions, but it looks for outliers after prices are aggregated at the monthly level.

11. Global outlier detection is ruled out because of the large number of products generating a broad price distribution. Observation-level outlier detection is not explored given the previous results, to avoid the risk of a poor fit or the introduction of zero-change bias.

12. Previously, our relative-based outlier detection was focussed on month-on-month changes in price relatives, by defining lower and upper fences, flagging price relatives outside these fences, then choosing whether to use these relatives for calculating the index.

13. We now extend price relative outlier detection to two additional forms of relative-based outlier detection, as summarised in Table 1:

    a. Quantity relative outlier detection, which focusses on extreme changes in quantities

    b. Price-quantity relative outlier detection, which focusses on simultaneous extreme changes in both prices and quantities change

Table 1. The fence-based methods to be explored in our research

Method	**Keep row if…**
Price relative fences	$L^p \leq$ price relative $\leq U^p$
Quantity relative fences ("q-dump")	$L^q \leq$ quantity relative $\leq U^q$
Price and quantity relative fences ("pq-dump")	$L^p \leq$ price relative $\leq U^p$ and $L^q \leq$ quantity relative $\leq U^q$

14. The goal of these new methods is to attempt to better capture and eliminate dump prices, which we discuss in the next section.

Why remove dump prices in groceries

15. Dump prices occur at the end of a product's lifecycle, when the retailer sells the remaining stock of a product at a very low clearance price, just before the product leaves the market.

16. This phenomenon is particularly common in the grocery market since retailers need to make shelf space for new product lines and need to sell perishable products prior to their sell-by date to avoid missing out on a sale completely. This encourages the retailer to drop prices rapidly to encourage a sale.

17. A limitation of the Törnqvist (and by extension, GEKS-Törnqvist) is that it is sensitive to dump prices (see page 34). An extreme example is given in Table 2. The second product drops substantially in price, but this sale price was not widely available – only one consumer was able

to benefit from it. Since the Törnqvist uses a mean of the base and current month weights (h), a non-trivial weight is allocated to the second product despite a trivial number of people benefitting from it, resulting in a very low index of 0.64.

Table 2. The Törnqvist index method can be sensitive to dump prices

	p1	p2	q1	q2	e1	e2	s1	s2	r	h	r^h
Prod 1	3	3	10000	10000	30000	30000	0.5	0.999983	1	0.74999	1
Prod 2	3	0.5	10000	1	30000	0.5	0.5	1.667E-05	0.1667	0.25001	0.6389
				sum	60000	30000.5				Törnqvist	0.6389

18. Since product 2 has largely ceased to exist as a mass-produced and -sold product in month 2, had we removed product 2 prior to calculating the index, then we would have obtained an index of 1, since there had been no change in price in product 1. This seems a more realistic index since most consumers are paying the same price in both months.

19. Typically, dump prices can be identified when a large price drop occurs, and rather than observing a substantial increase in quantity sold as consumers substitute to this product, we instead observe a large drop in quantity at the same time.

20. Note that with a price relative of 0.1667, product 2 would have already been removed by our price relative check, independently of changes in quantity. Therefore, our current approach already provides protection against the most extreme dump prices, which may be sufficient alone. We will show how we may look to combine a price filter with a price-quantity dump filter.

Preliminary results

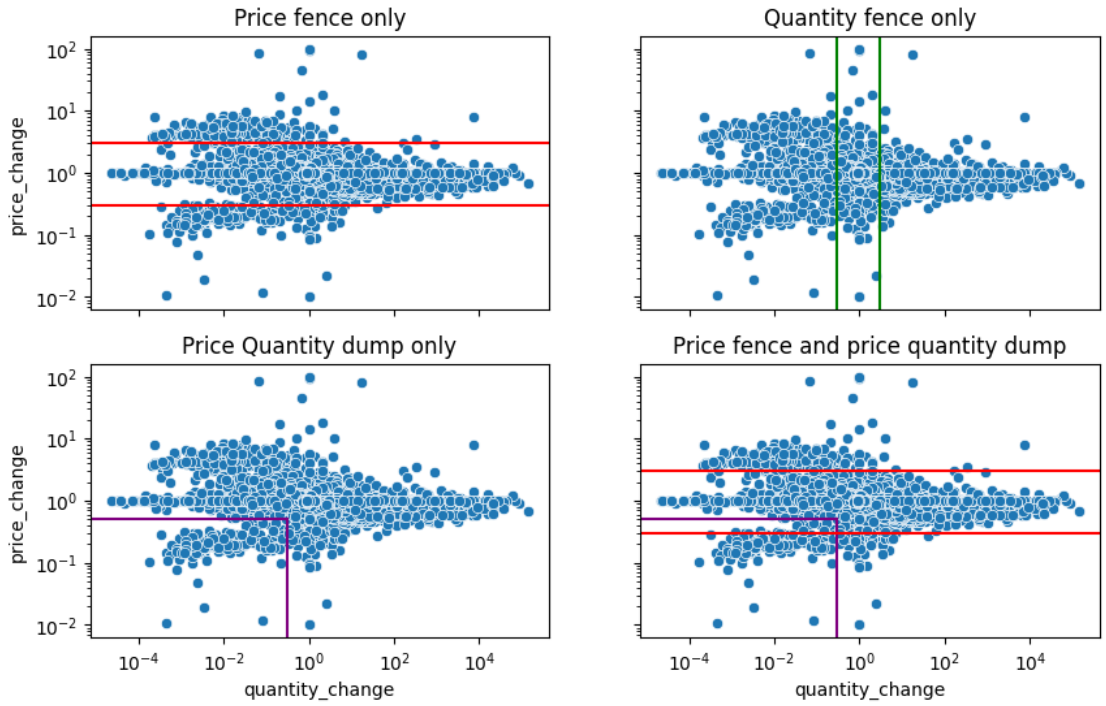
21. The preliminary analysis presented below is performed on transactions obtained from one single retailer, covering a six-month window from January to June 2019. The analysis is not based on index calculation, but it looks at two-dimensional distributions of month-on-month price and quantity changes. The analysis aims at visualising the impact of several fencing strategies and calculating the percentage of flagged transactions. Note that the fences explored are not yet optimised and are to be updated once the full analysis is completed. Given the limited time window and the fact that transactions from only one retailer are considered the results are not generalisable, but they nevertheless provide useful insights that we aim to test on all retailers and in longer time windows.

22. In Figure 1 we show some distributions of quantity relatives (x-axis) and price relatives (y-axis) from a single retailer covering January 2019 to June 2019. In the four subplots the transactions distribution is always the same, but we overlay cut lines corresponding to different data cleaning methods. We show how price fences create “horizontal cuts” (top left), quantity fences create “vertical cuts” (top right), and price-quantity fences create “rectangular (or corner) cuts” (bottom left). We show how we can combine (for example) price and price-quantity fences to create more-refined cuts on the data (bottom right). The parameters used to create these fences are shown in Table 3.

23. Note that in the two bottom plots of Figure 1 we are only considering lower price-quantity fences, corresponding to asymmetric fences. This choice causes the presence of only a corner area being flagged by the price-quantity dump filter.

24. Note also that when combining price fences with price and quantity fences, using the same price outlier parameters would cause the price-quantity fence to be redundant.

Figure 1. Flagging of outliers depending on different methods of outlier detection.



25. Figure 1 might appear misleading, as with well over 20,00 products it is difficult to get a sense of scale of the proportion of data removed by each method due to overlapping data points in the centre, making peripheral data points appear more frequently than they are. Therefore, to avoid misinterpretations, in Table 3 we present the parameters used and proportion of data removed by each method.

Table 3. Percentage of rows removed by each approach within Figure 1.

		% of rows removed:		
Fences used	**Row kept if…**	**Lower**	**Upper**	**Total**
Price	$0.3334 \leq p \leq 3$	0.278%	0.245%	0.523%
Quantity	$0.3334 \leq q \leq 3$	5.266%	3.823%	9.089%
Price-quantity	$(1/2 \leq p \leq 100000$ & $0.3334 \leq q \leq 100000)$			0.219%
Price and price-quantity	$(0.3334 \leq p \leq 3)$ & $(1/2 \leq p \leq 100000$ & $0.3334 \leq q \leq 100000)$			0.539%

26. Table 3 suggests that we are flagging more cases as potential outliers using these price relative fences within groceries (0.523%) than we did with rail fares (0.02%) or second-hand cars (0.03% for petrol, 0.04% for diesel), as discussed in our previous paper. Although some caution should be taken when performing this comparison, especially due to the different time windows

explored in the two studies, promotional offers within groceries can cause much steeper price changes to occur. Given a 50% reduction in price seems viable (either as a price promotion, or as a buy-one-get-one-free offer), it may be that the [0.3334, 3] fences are too narrow for groceries, and we may seek to widen to [0.25, 4] (or potentially further). To identify the correct fences, we will follow up this work with more low-level product inspections to determine whether price changes at this scale seem viable.

27. Note also that care should be given in use of quantity fences. Threefold changes in quantity are not uncommon, causing the method to flag 9.089% of transactions when looking at month-on-month changes. In continued investigations we will widen these parameters much further, as well as considering different dump filters.

28. In Figure 1 we identified two key "extreme clusters" of data points:

    a. In the top-left, where a rapid price increase and a rapid quantity decrease occur simultaneously. These potentially make sense as a price increase on this scale is likely to cause a large degree of substitution away from the product. We may wish to keep (more of) these transactions.

    b. In the bottom-left, where a rapid price decrease and a rapid quantity decrease occur simultaneously. Since consumers typically substitute towards products that are decreasing in price, it is likely that the reduction in quantity on this scale is caused by end-of-lifecycle dump prices. We may wish to remove (more of) these transactions.

29. A potential way of accounting for the two 'extreme clusters' as discussed is to firstly broaden the price filter (for example, exploring fences of [0.25, 4]) and pair it with an asymmetric price-quantity filter to only remove the dump prices in the bottom-left corner. These refined fences will allow us to keep the top-left cluster whilst dropping the bottom-left cluster.

Future work and provisional strategy

30. In line with previous research, we continue to favour use of price relative UDFs applied at the consumption segment level as a base method to remove unrealistic price movements since they are easy to fit, monitor and generalise, allowing us to avoid removing too many data points due to a poor fitting algorithm, as discussed previously in APCP-T(22)13. The chosen outlier detection strategy is to be applied at the whole 'groceries' category.

31. In line with our preliminary results, we will explore widening the choice of parameters within these price relative fences.

32. As discussed, the price relative filter already provides some protection against dump prices and may be viable to use on its own. However, broadening the parameters of this filter may lead to more dump prices being kept. In this case, we will explore using a price-quantity filter to target these dump prices, to work in conjunction with our standard price filter.

33. Our next steps will be to extend these analyses to longer timeframes and more retailers and try different parameterisations of the price and price-quantity filters, performing:

    a. Low-level investigations of the products flagged to determine the viability of the price and price-quantity changes;

    b. High-level analyses of the impacts on indices.

34. We anticipate releasing a second draft of this paper with further information on the outcome of this planned work, along with a final proposed on the decision for methods to use for outlier detection in grocery scanner data.

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

List of Annexes

No annexes