ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

The use of approximate expenditure weights for web scraped data in consumer price indices

Status: final
Expected publication: alongside minutes

Purpose

1. As part of research into the potential introduction of web scraped data sources in consumer price index measurement, the ONS is conducting analysis into the feasibility and impact of incorporating these sources into the headline indices.
2. Expenditure and quantity information are not available at the product level in web scraped datasets. This paper is an update to research presented in APCP-T(18)14 in September 2018, and continues to investigate approximate weight allocation methods for the individual product quotes used in the calculation of an item's index.

Actions

3. Members of the Panel are invited to:
   a) comment on the results and methodologies presented in this paper
   b) advise on areas for further work and agree ONS's suggestions for future research

Introduction

4. This paper is a continuation of the analysis presented in paper APCP-T(18)14 at the September 2018 meeting, which investigates the use of approximate expenditure weights for web scraped data in consumer price indices.
5. Alternative data sources, such as web scraped data and scanner data, offer more frequent data collection, increased coverage and larger sample sizes than the current method. However, since all prices are scraped regardless of popularity, using an unweighted index at the lowest level of aggregation would mean that the more popular items would not have greater influence on the index. Indices calculated in this way may therefore not be representative of consumer spending, especially if the prices of less popular products behave differently to the more popular ones. Expenditure and quantity information are not available at the product level in web scraped datasets and must therefore be approximated.
6. One proposed indicator of popularity is the position of the product on the website, i.e. the page ranking. This assumes that the most popular products would be placed higher on the page and is a reasonable assumption since many websites provide the option to sort by popularity; however, the popularity ranking itself may not necessarily be reliable.
7. A major challenge with judging the quality of approximate weights produced from web scraped datasets is that the lack of sales information leaves nothing to compare the proposed weights against. Therefore, for this analysis a scanner data source for a single retailer, covering the year 2012 and the items toothpaste and shampoo, is used instead.
8. Using scanner data, with quantity and expenditure information available, allows approximate ranking weights to be compared to those assigned by calculating expenditure shares. No page rankings are available; therefore, the products can be ranked in order of quantity or expenditure

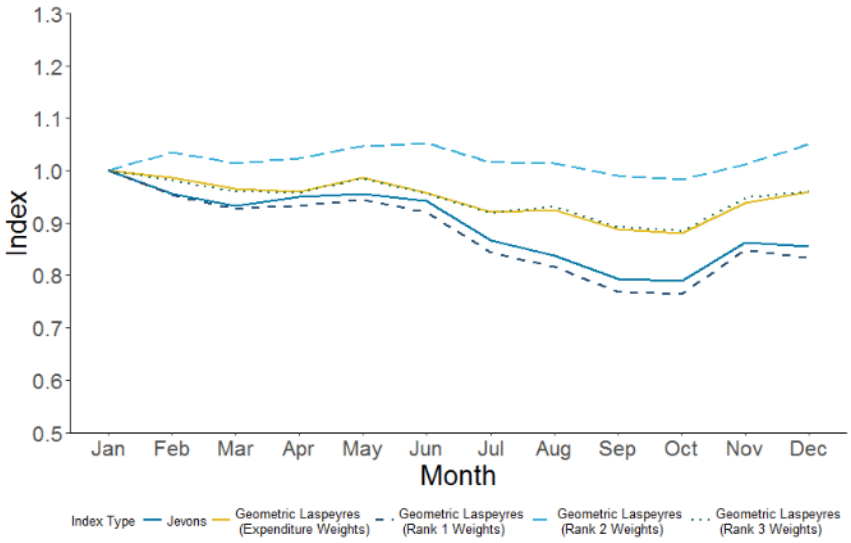
as a proxy - it is assumed that this is a good approximation to the page ranking that would be observed in a corresponding web scraped dataset.

Recap of previous research

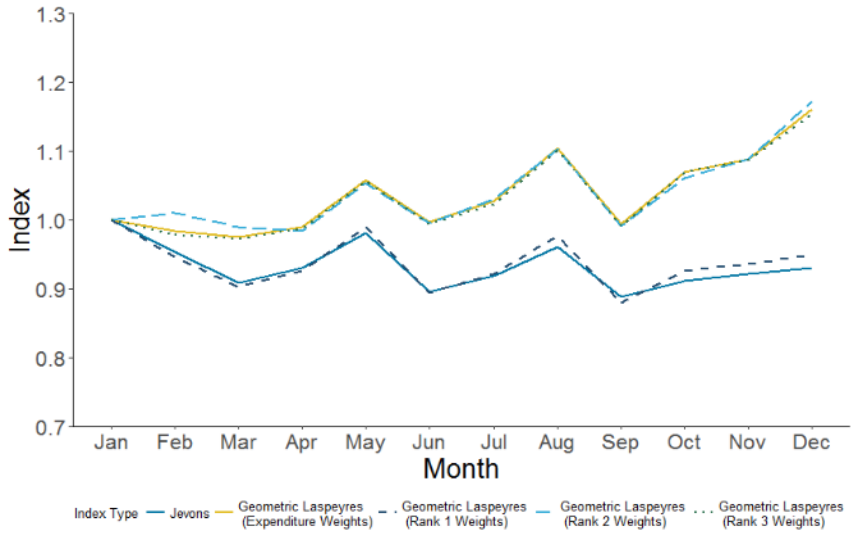
9. In the previous paper, three methods for transforming the assigned expenditure rankings to product weights were investigated. An effective transformation would have a linear relationship with the observed expenditure weights, with the resulting Geometric Laspeyres indices closely aligning to those calculated from the expenditure weights. A Geometric Laspeyres index was used for weighted indices, since the Jevons index is a Geometric Laspeyres index with equal weights, and therefore all differences in the index series can be entirely attributed to the change in the weights assigned to products. Thus, the expenditure-weighted Geometric Laspeyres index is the benchmark in this analysis.
10. The indices are monthly chained indices; the base period is the previous month. At the lowest level in the dataset, products were defined according to retailer-assigned product IDs and price relatives are taken to be 1 where no price exists in one of the months in question.
11. Figure 1 shows the resulting indices calculated for toothpaste and shampoo using the weights derived by the three evaluated methods in APCP-T(18)14, as well as a Jevons index and the benchmark expenditure-weighted Geometric Laspeyres index.

Figure 1: Toothpaste (a) and shampoo (b) indices with different weights applied

(a)



(b)



12. Although three methods were investigated, the results indicated that the use of the Equation 1 method with x = 6 to transform the expenditure rankings was most effective; the resulting weights were closer to those calculated from observed expenditure shares than the other methods, and resulting indices more closely aligned with the expenditure-weighted Geometric Laspeyres index, as shown in Figure 1. For both shampoo and toothpaste, x = 6 resulted in indices closer to the benchmark Geometric Laspeyres index than any other value for x. However, this particular value of x may only be optimal for the observed shampoo and toothpaste datasets; a more general method for selecting the optimal transformation for other items and in other datasets was not considered in paper APCP-T(18)14.

$$w_i^0 = \frac{(Rank\ share)^x}{\sum(Rank\ share)^x} \qquad \textbf{(Equation 1)}$$

$$where \quad Rank\ share_i = \frac{r_i}{\sum_{i=1}^{n} r_i}$$

$$r_i is\ the\ rank\ of\ product\ i\ in\ the\ base\ period\ (in\ ascending\ order\ according\ to\ popularity)$$

13. Based on APCP-T's advice, and on discussions subsequent to the presentation of the previous paper, the following objectives for future analysis were proposed:
    a. Carry out a literature review, summarising the possible methods and how to go about making recommendations/choosing the best.
    b. Repeat the previous analysis using the quantities observed in the scanner data set rather than the sales.
    c. Repeat the previous analysis on the top 5, 10, and 20 products in the scanner data set when sorted by quantity/expenditure.
    d. Calculate sample statistics such as the mean and standard deviation of quantities/expenditures across products using the available scanner data, and investigate the use of these parameters to recreate the observed distributions and translate observed ranks to estimated weights.

e. Repeat all analyses previously carried out using web scraped data from mySupermarket. This objective could make use of CD data, where both page rankings and expenditure are available.
f. Investigate the market share methodology discussed in Antoniades (2017) using the available mySupermarket data for printers and routers.
g. Depending on the availability of transaction data and web scraped data for the same retailer, investigate the use of web scraped rankings to approximate expenditure weights for the previously investigated methodology.
h. Previous research indicates that Equation 1 is sensitive to the choice of x. Investigate the choice of value for x across different items.

14. This paper takes forward and builds on the conclusions of paper APCP-T(18)14. Specifically we consider three different approaches relating to the objectives above. For the first approach (objective b) we use website page rankings to estimate the quantity distribution, rather than the expenditure distribution, and use this to derive expenditure shares. For the second approach (objective c), we use the website page rankings to subset the data on only the top 5, 10 and 20 best-selling products and use these to construct an unweighted price index; this is broadly consistent with the idea of implicit weighting, whereby price collectors will aim to track the price development of a product that is representative of consumer's expenditure. For the third approach (objective d) we use sample statistics from the quantity and expenditure distributions to recreate the observed distributions using the website page rankings.
15. Objectives (e) to (f) are out of scope for this analysis until a sufficiently long time series accumulates for the mySupermarket web scraped data; objective (g) depends on the availability of transaction data and web scraped data from the same retailer; and objective (h) is subject to the availability of suitable data for more items. Previously mentioned research into the use of duplicates for approximating expenditure is also out of scope, as duplicates are not contained in the web scraped data sets currently available, and are unlikely to be so in future.

Characterising the analysed items

16. Since this analysis is limited to just two items, the results cannot be generalised to all items without further research. Differences in market behaviour between items is the main reason why a one-size-fits-all method is unlikely to be found; some items have market-leading products, while others exhibit the characteristics of perfect competition.
17. There are 692 products in the shampoo dataset in 2012, with each product available for an average of 9.7 months over the year. The most popular product, by sales, over the year makes up 2.7% of total sales (each product would make up approximately 0.14% of total sales in a market exhibiting perfect competition), whilst the top 50 products make up 44.8% of sales. There are 284 products in the toothpaste dataset in 2012, with each product available for an average of 9.4 months. The most popular product makes up 3.0% of sales (each product would make up approximately 0.35% of total sales in a market exhibiting perfect competition), with the top 10 products making up almost a quarter of all sales. Such data indicate that there exist shampoo and toothpaste market leaders and that their markets do not display signs of perfect competition.

Table 1: Summary statistics for shampoo and toothpaste, 2012

	Shampoo	Toothpaste
Number of products	692	284
% available in every month	59	60
Number of months a product is available on average	9.7	9.4
% available less than 6 months of the year	19.5	18.7
% of total sales made up by the highest expenditure product	2.7	3.0
% of total sales made up by the top 5 products	8.8	13.5
% of total sales made up by the top 10 products	14.2	24.6
% of total sales made up by the top 50 products	44.8	67.1

Literature Review

18. We conducted an online literature review to inform our work on estimating weights for products whose price information has been obtained via web scraping, the findings of which are summarised below and described in detail in Annex A.
19. Much of ONS's experience with web scraped data is summarised by Bhardwaj at al. (2017) and focusses on dealing with particular issues when working with web scraped data, such as product classification and churn. The paper recommends that further research be undertaken to assess the impact of not having expenditure weights for web scraped data, the first stage of which is the present study.
20. Many studies have demonstrated the use of web scraped data to estimate price indices, but without applying product-level weights (and generally making use of expenditure data from non-internet sources at higher levels of aggregation). For example, see Nygaard (2015), Polidoro et al. (2015), Bosch and Griffioen (2016), and Loon and Roels (2018) for a description of the Norwegian, Italian, Dutch, and Belgian experiences, respectively.
21. In the absence of online data, the present study makes use of scanner data to inform possible methods for constructing weights as and when web scraped data become available. Several published studies also combine, and in some cases compare, scanner and online data; see Krsinich (2015), Chessa and Griffioen (2017), and Cavallo (2017).
22. Numerous studies have investigated the statistical distribution of sales quantities for different product groups, which has informed our research into predicting quantities from their ranks. Of particular note are Chevalier and Goolsbee (2003), Hisano and Mizuno (2010), Touzani and Buskirk (2015), and Antoniades (2017), who between them made use of the exponential, Pareto, log-normal, and truncated log-normal distributions.

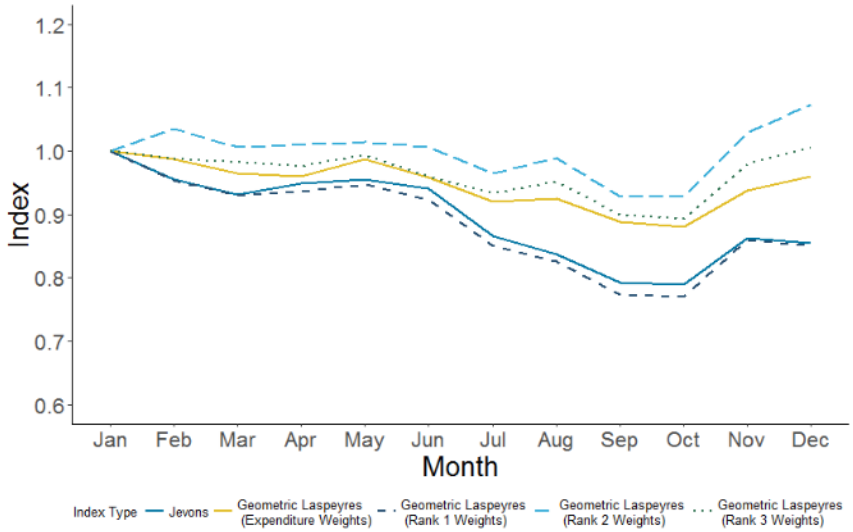
Using quantity- rather than expenditure-based rankings

23. Instead of total sales of products, the position of a product on a retailer's web page when sorted by popularity may instead be indicative of the quantity of sales of that product, i.e. a product placed first on the web page is likely to be the bestseller, with less popular products placed at the bottom of the page. Thus, the methods investigated in APCP-T(18)14 should also be tested using product rankings based on quantity as well as expenditure.

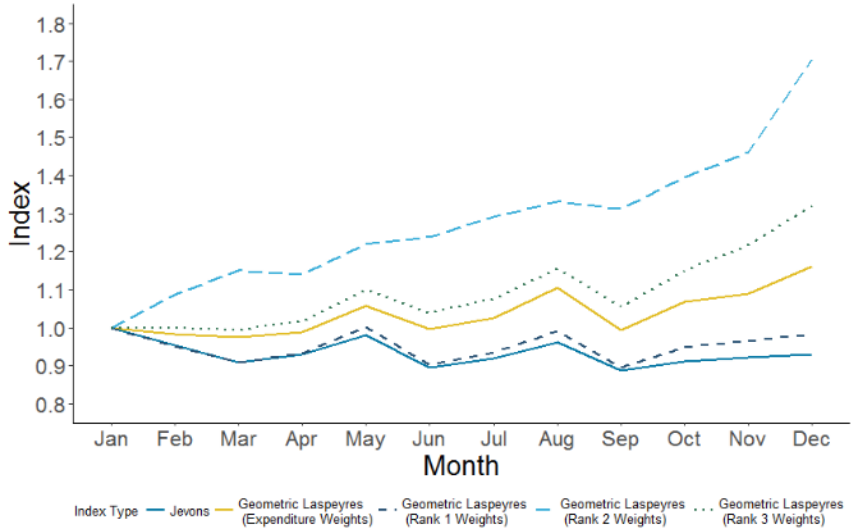
24. Although we investigate the idea that the page ranking is a proxy of quantity rather than expenditure, the benchmark Geometric Laspeyres index that we wish to compare our indices against is still expenditure-weighted for consistency with the previous analysis.

Figure 2: Toothpaste (a) and shampoo (b) indices with different weights applied, using quantity rankings compared with the expenditure-weighted index, 2012

(a)



(b)



25. Figure 2 shows that, as with expenditure-based rankings, the Rank 3 method (i.e. Equation 1 with x = 6) is the closest to the expenditure-weighted index when working with quantity rankings. For the remainder of the analysis we therefore discard the possibility of using the Rank 1 or Rank 2 methods as ranking transformations (see APCP-T(18)14 for details of these methods). Note that other values for x could be investigated as future work for quantity-based rankings.
26. Although the Rank 3 method of transformation is effective when quantities are used in place of expenditures, this is only an indication of the effectiveness of such a method in general; the method has only been tested for two CPI items, and the fact that the power-6 transformation was optimized on the observed sample means that it should not be applied to every item without further investigation. In the absence of product-level microdata on every item traded by

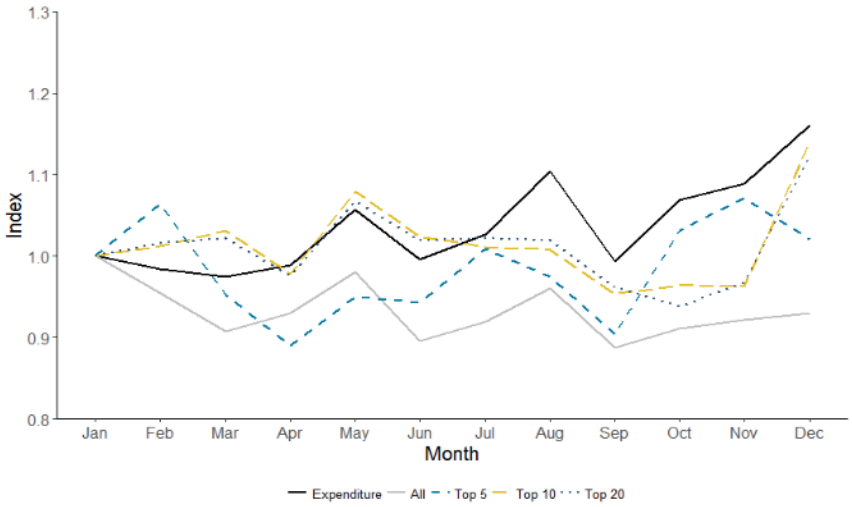
a retailer, Section 5 investigates fitting statistical distributions using known item-level sample statistics (rather than product-level microdata) as a potential method for estimating product weight that is applicable to any item.

Assessing the performance of weighting methods on subsets of data

27. In the existing CPI price collection, price collectors will deliberately target items that they believe to be representative of consumers' expenditure based on retailer knowledge, shelf space and their own market knowledge (i.e. a broadly representative sample is selected). A Jevons index is calculated from the collected prices for each stratum.
28. The concern with the use of unweighted indices for web scraped data, as previously stated, is that less popular products in the dataset may have too much of an influence on the index, particularly where their behaviour differs to that of more popular products. We therefore attempt to replicate the representativeness of the existing CPI price collection by filtering the most popular products in terms of their expenditure or quantity. Using the scanner dataset, subsets are taken based on the top-ranking products, in total, over the year in terms of their expenditure or their quantity.
29. Figure 3 indicates that the Jevons indices for the top 5, 10 and 20 products are closer to the benchmark expenditure-weighted Geometric Laspeyres index than the Jevons index for all items in the dataset. Each subset shows a different pattern between months and none align closely to those evident in the benchmark Geometric Laspeyres index. A similar pattern was exhibited in Figure 4 for toothpaste.

Figure 3: Shampoo Jevons indices for different expenditure (a) and quantity (b) subsets, 2012

(a)



(b)

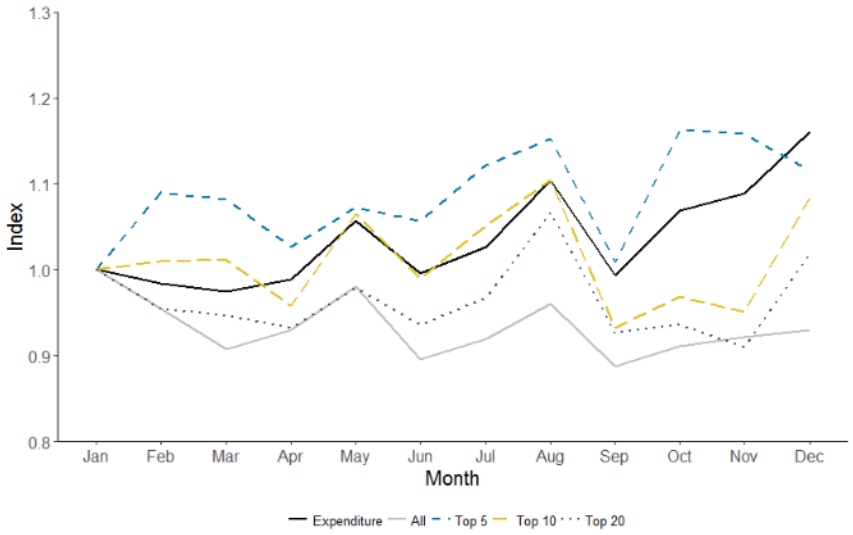
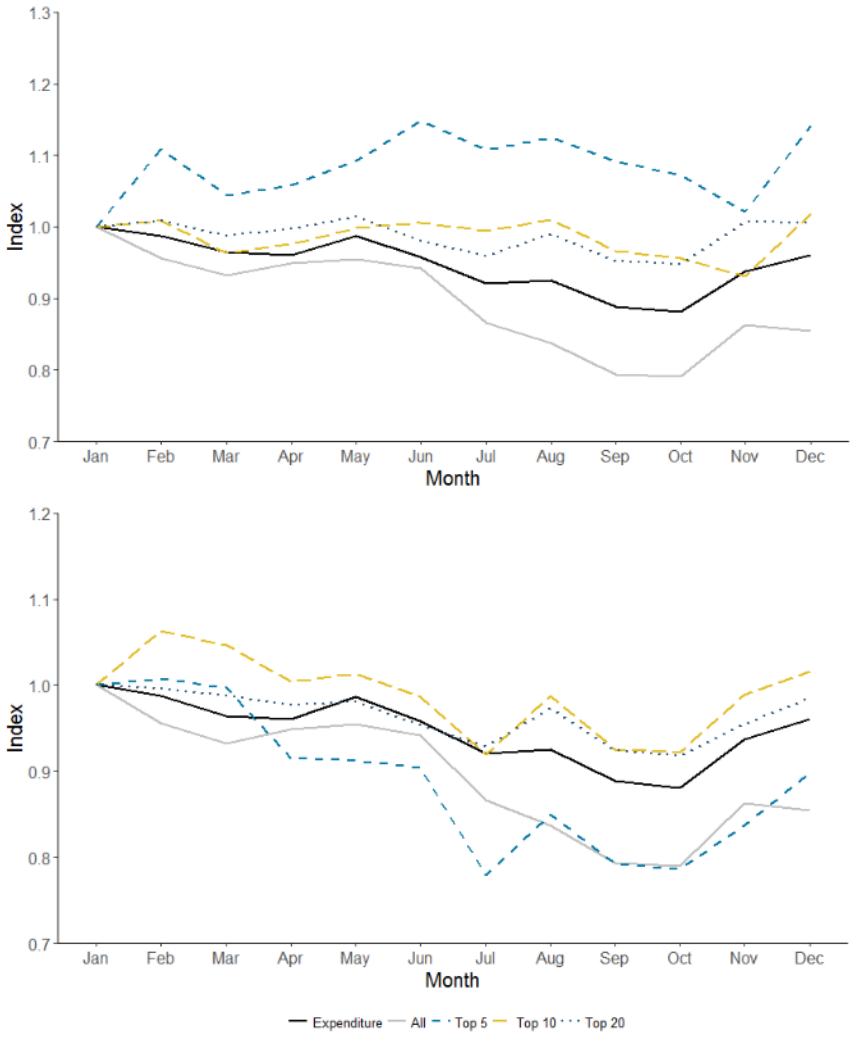


Figure 4: Toothpaste Jevons indices for different expenditure (a) and quantity (b) subsets, 2012

(a)

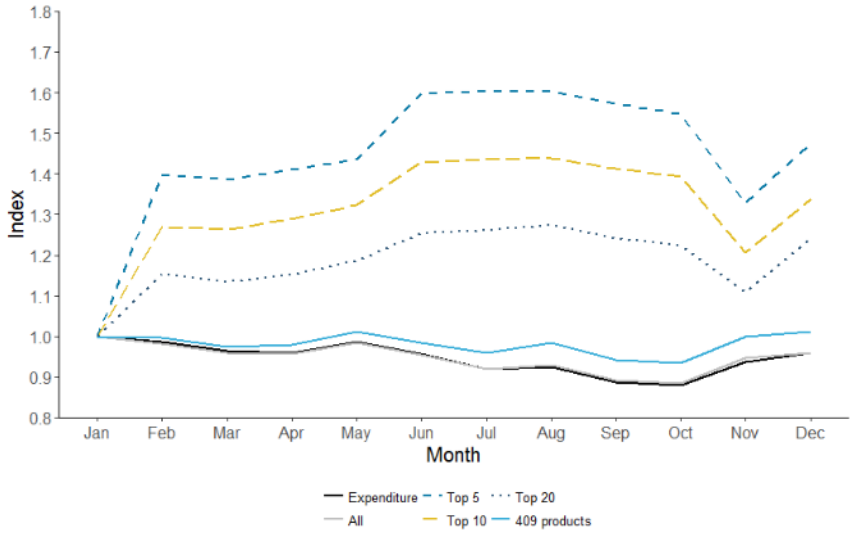
(b)



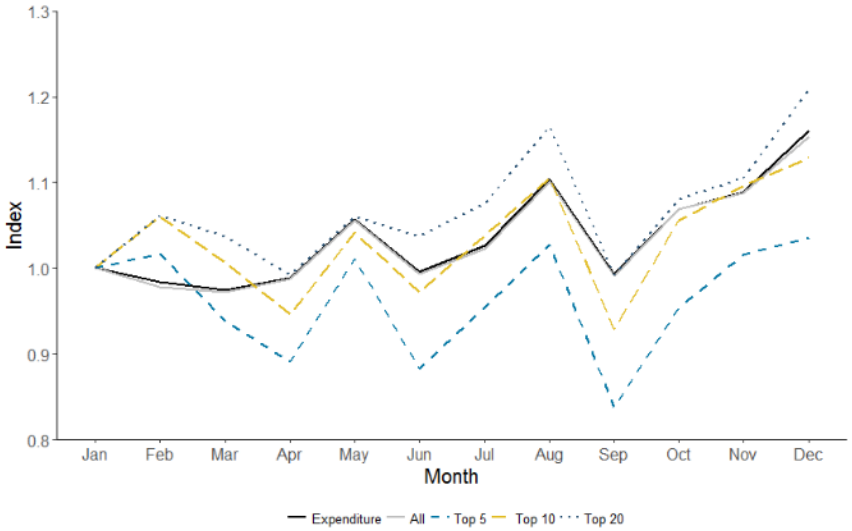
30. We can also replicate the above analysis for the Rank 3 method. Figure 5 shows that for toothpaste, the indices calculated using this method on the expenditure subsets of data are far from the benchmark expenditure-weighted Geometric Laspeyres index series and seemingly, the larger the subset of products taken, the closer the series become. The toothpaste indices for all products available throughout every month of the year (409 products - the blue line) are still not as close to the benchmark Geometric Laspeyres indices as those calculated from all products in the dataset. The patterns shown for the shampoo indices are not as clear as those exhibited for toothpaste, with no notable association between the number of products included in the subset and closeness to the benchmark Geometric Laspeyres index series. This finding highlights the importance of exercising caution when choosing a weighting method across items and generalizing results.

Figure 5: Toothpaste (a) and shampoo (b) indices calculated using the Rank 3 weighting method for various expenditure subsets, 2012

(a)



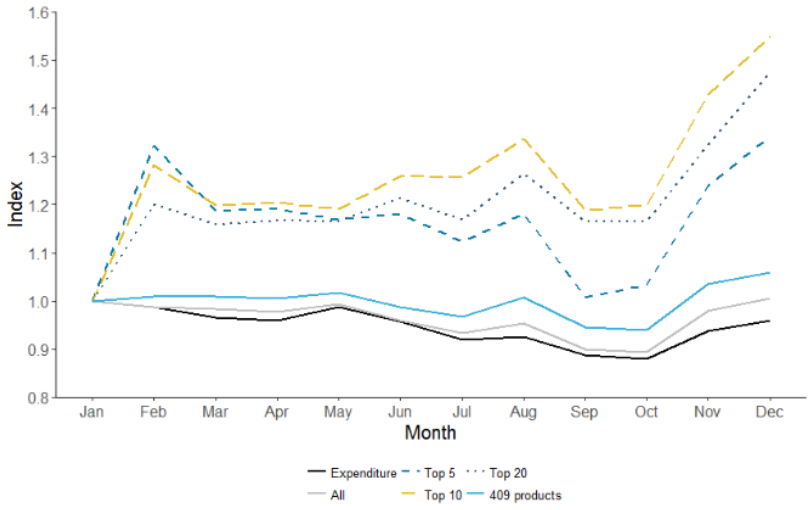
(b)



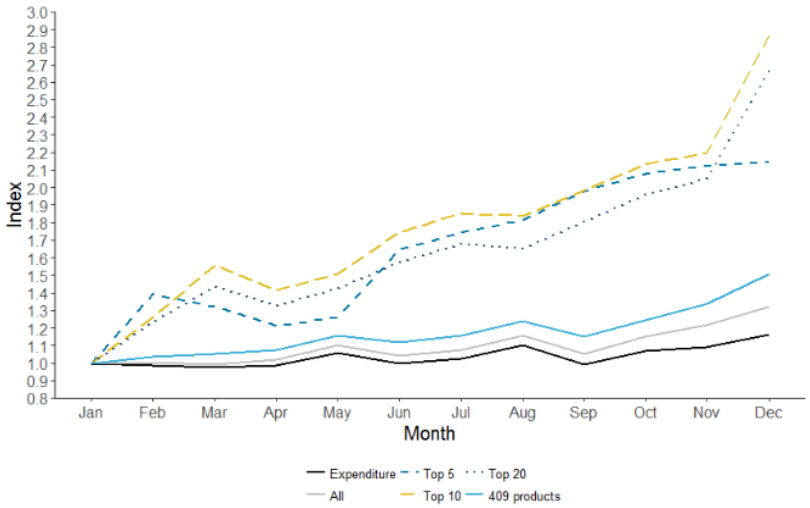
31. For the quantity subsets, the resulting indices are also volatile and far away from the benchmark expenditure-weighted Geometric Laspeyres index. Figure 6 shows that for the shampoo indices in (b), the Top 10 index series reaches a level of 2.9, displaying signs of chain drift. Analysis of the price relatives for the subsets shows that the most popular product (by quantity) in the shampoo dataset for February has a price relative of 1.48. This suggests that the product was on sale in January and returned to full price in February. The weights used in February are those for January and thus, while the product was on offer in January, quantities purchased were high. Therefore, for items such as shampoo and toothpaste that often have significant price decreases when on offer and due to consumers tending towards products that are on offer, taking subsets as small as 5, 10 or 20 products is not suitable, as the resulting indices are volatile and prone to chain drift.

Figure 6: Toothpaste (a) and shampoo (b) indices calculated using the Rank 3 weighting method for various quantity subsets, 2012

(a)



(b)



Estimating weights from ranks using statistical distributions

32. The aforementioned Rank 3 method, whereby product rank shares are raised to the sixth power in the calculation of weights, was found to result in monthly price indices that most closely resemble those obtained when weighting by observed expenditure shares. However, this analysis was only conducted on two items (shampoo and toothpaste) and for one year (2012). It is unclear whether the Rank 3 method would be optimal for other items and other years because, at present, ONS does not have access to sufficiently detailed product-level quantity/expenditure data to verify this.
33. The aim of the analysis presented in this section is therefore to estimate expenditure-based weights from observed product ranks solely by using distributional summary statistics for quantities within items (which could then be used alongside web scraped product-level price information). If operationalised, this would require retailers to simply provide ONS with summary statistics such as means and standard deviations for quantities, rather than more granular product-specific quantity/expenditure microdata.
34. It should be noted that, as with the previously described analysis, this research was conducted with scanner data (for both prices and quantities) rather than web scraped information. It is therefore assumed throughout that observed product ranks based on quantities sold are a good approximation to those that are yet to be observed based on webpage ordering. Note that the analysis was conducted on expenditures as well as quantities but little difference in predictive performance was observed, so the focus of this paper is on the quantity distributions.

Methods

35. The research dataset is the same as that used in the previously described analysis: shampoo and toothpaste sales for each of the months in the calendar year 2012. Products with zero sales in a particular month (for example, due to being out of stock or discontinued by the retailer) do not contribute to the analysis in that month.
36. For each product group, sales quantity ranks are translated to quantiles of the cumulative distribution of sales quantities as follows $F(q_i) = 1 - r_i/n$. This formulation may be interpreted as there being r_i products with sales quantities greater than or equal to that of product i (i.e. q_i). The goal of the analysis is then to find a statistical distribution that suitably approximates the observed quantiles, and to use this distribution to predict sales quantities from their ranks.
37. The observed frequency distributions of both shampoo and toothpaste quantities exhibit long tails, with a very small number of products having very large sales quantities, and the majority of the products making up the rest of the distribution. The log-normal, truncated log-normal and Pareto (power-law) distributions are therefore considered as candidates for predicting sales quantities. These distributions have previously been successfully fitted to retail sales of books, consumer electronics and household appliances by Chevalier and Goolsbee (2003), Hisano and Mizuno (2010) and Touzani and Buskirk (2015), respectively. Note that these distributions are all continuous rather than discrete; it is assumed that the discrete rank data are sufficiently well approximated by continuous statistical distributions due to the relatively large number of observations in each of the samples (692 products for shampoo and 284 products for toothpaste across all months of 2012).
38. The parameters of the log-normal distribution are the mean and standard deviation of the natural logarithm of sales quantities; for each product group, these are estimated using the

corresponding sample statistics calculated on the observed dataset (i.e. the maximum likelihood estimates of these parameters). The truncated log-normal distribution additionally requires pre-specification of the truncation points; these were set to the minimum and maximum quantities observed in the dataset for each item. The scale and shape parameters of the Pareto distribution are estimated by their maximum likelihood estimates, calculated from the observed data for each item: $\min(q_i)$ for scale and $n \times (\sum_{i=1}^{n} \ln[q_i/\min(q_i)])^{-1}$ for shape. Each item's distributional parameters are estimated for each month separately, rather than estimating a single set of parameters by pooling the data over the year.

39. For each candidate distribution in each month, goodness-of-fit is assessed using R^2 (the proportion of variation in observed quantities explained by fitted quantities) and mean absolute percentage error (MAPE, a measure of the accuracy of the fitted quantities) across all products within each item.
40. Fitted quantities are multiplied by observed prices to estimate product-level expenditures and, in turn, product weights are calculated using estimated expenditure shares. The resulting Geometric Laspeyres price index series (spanning January to December 2012) can then be compared to that obtained using observed rather than estimated expenditures.

Results

41. Of the three candidate statistical distributions, the observed data are mostly in accordance with simulated draws from the truncated log-normal distribution, as illustrated in Figure 7 for January 2012. The observed quantity-rank pairs are generally within the range of those simulated by the truncated log-normal distribution, but lie below the range simulated by the Pareto and log-normal distributions for higher ranked products.
42. The observed log-quantity versus log-rank relationships do not follow the "signature" linear trend that would be expected if the data followed a power-law distribution such as the Pareto distribution (illustrated in Figure 8 for January 2012), whilst the log-normal distribution tends to over-predict quantities for higher ranking products. This over-prediction is somewhat (but not completely) remedied by truncating the log-normal prediction, and there remains a tendency to under-predict for medium-low ranking products (Figure 9).

Figure 7: Quantity vs. rank (log scale), simulated and observed quantities, January 2012

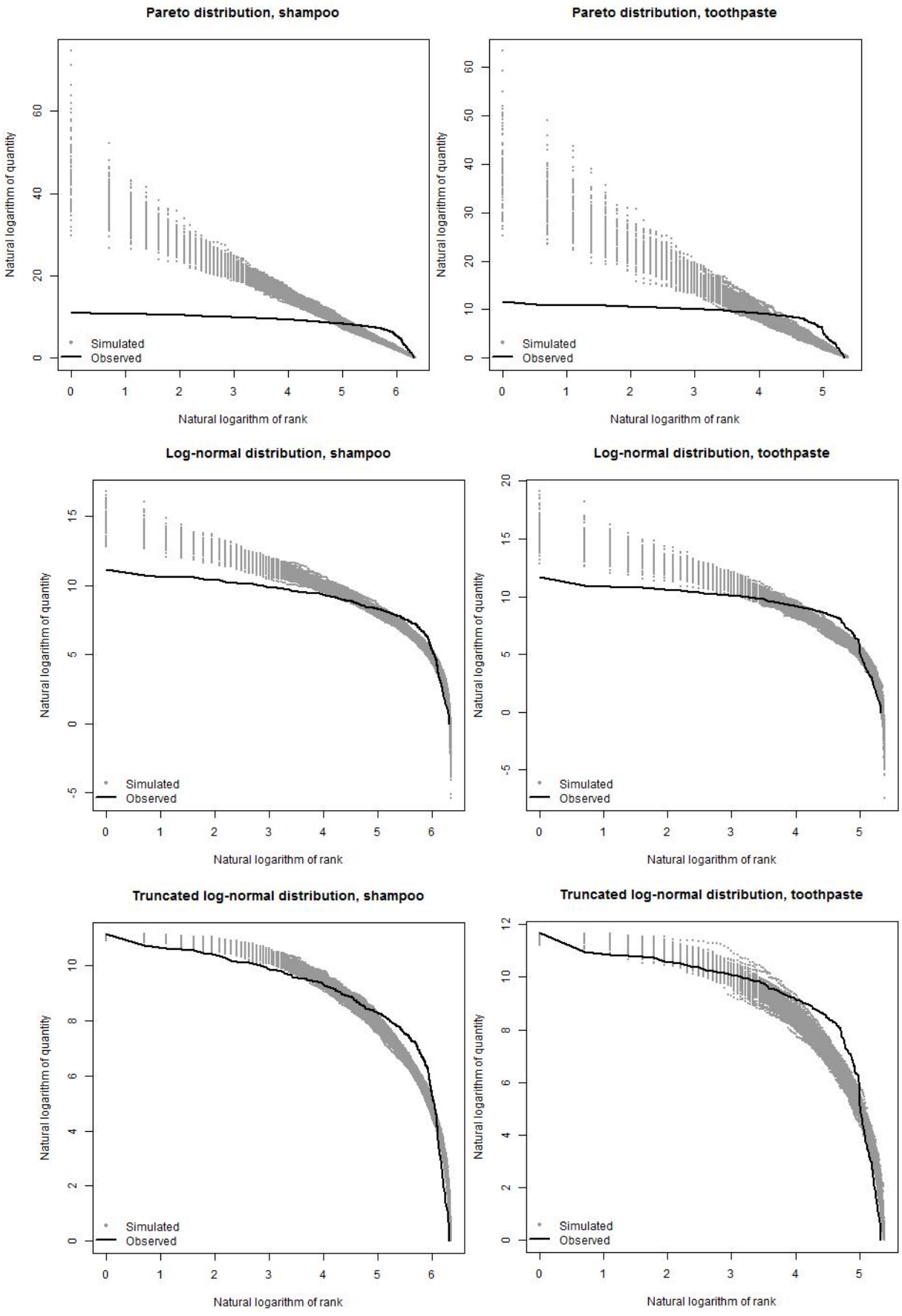


Figure 8: Quantity vs. rank (log scale), fitted and observed quantities, January 2012

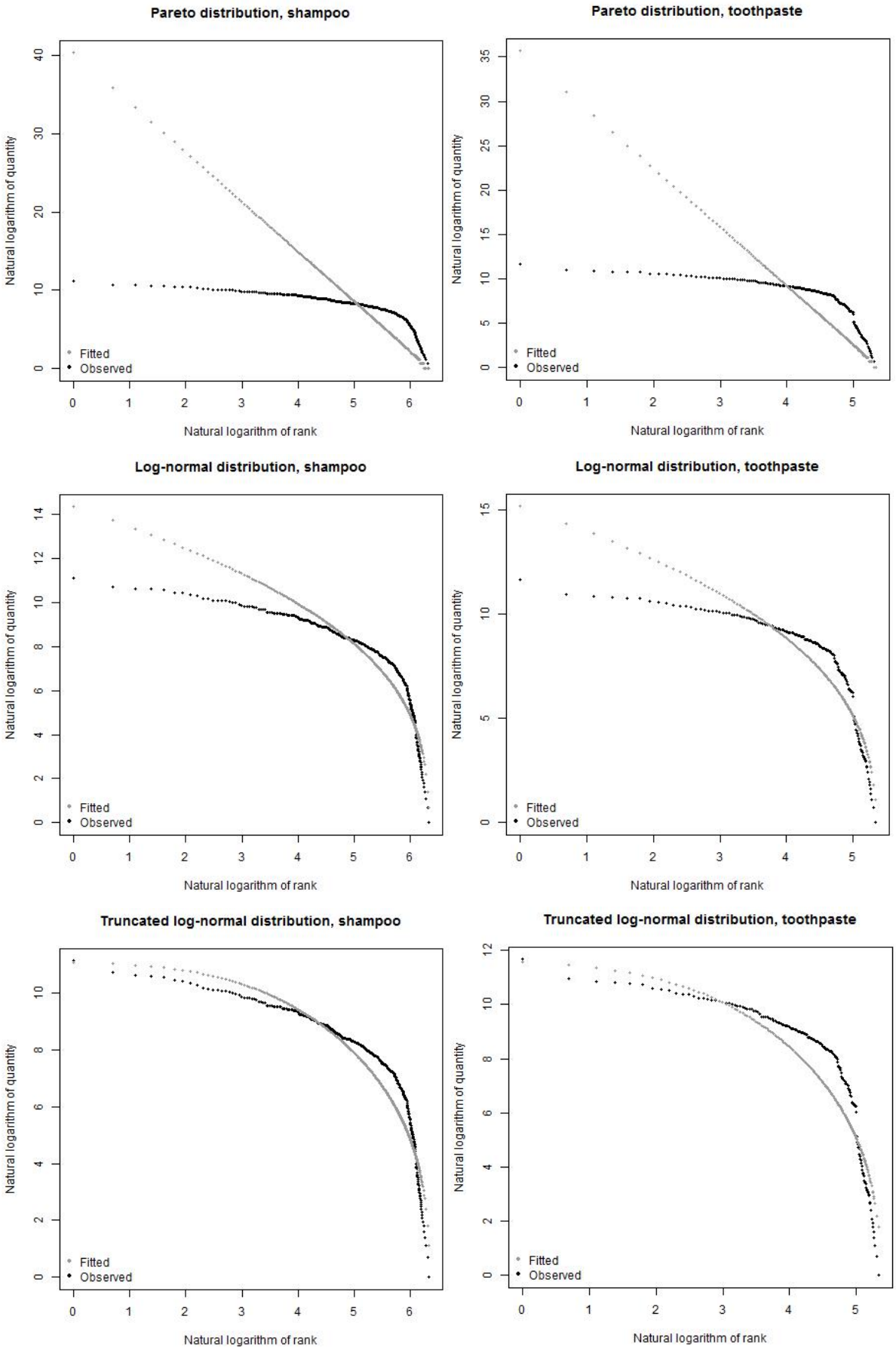
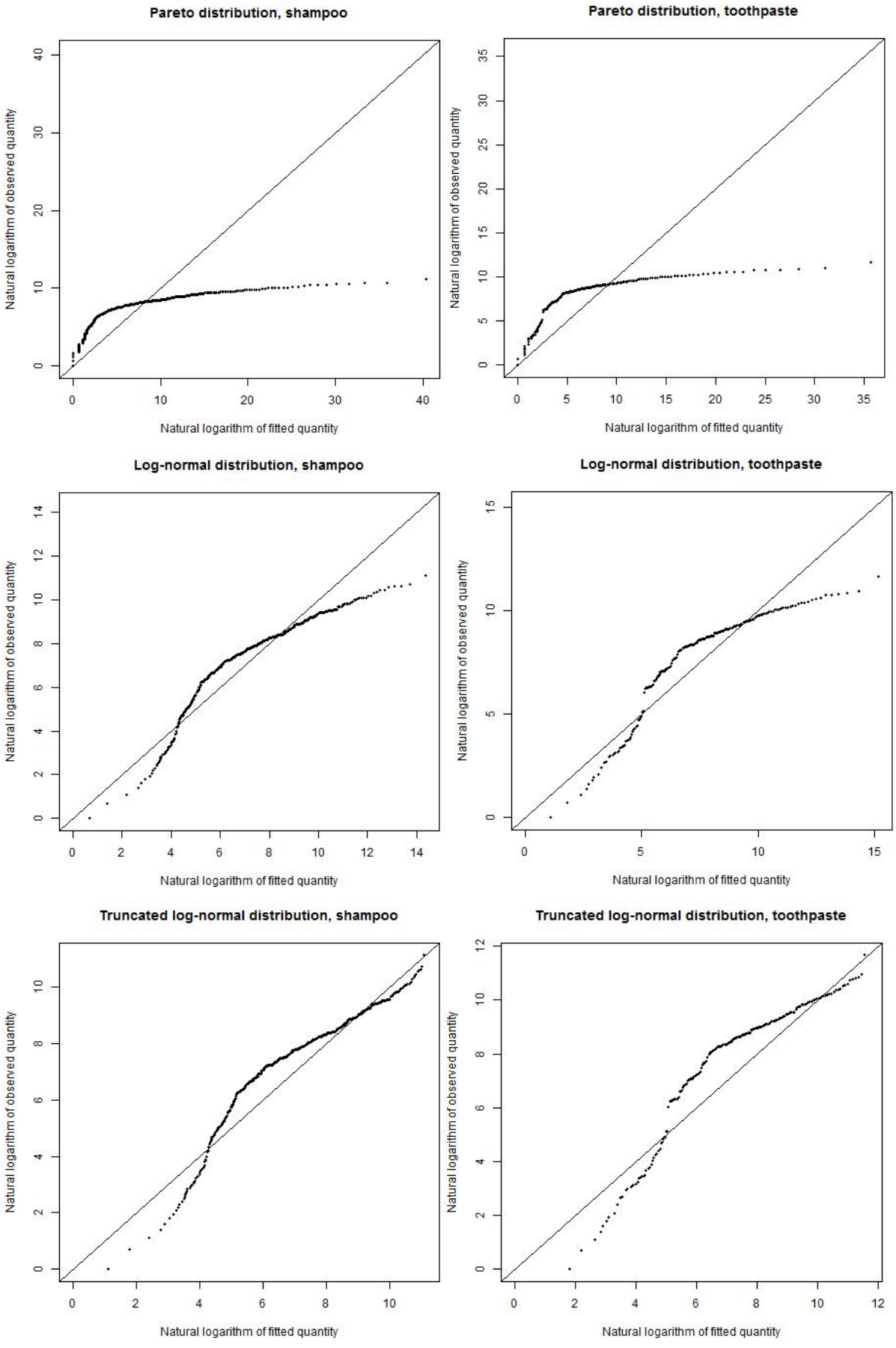


Figure 9: Observed vs. fitted quantities (log scale), January 2012



43. Across all of 2012, fitted quantities from the truncated log-normal distribution explain the majority of variation in observed quantities for both shampoo and toothpaste (Table 2), achieving an R^2 ranging from 86.5% (December) to 91.8% (January) for shampoo, and from 84.6% (February) to 90.5% (August) for toothpaste. In terms of the predictive accuracy of the truncated log-normal distribution, MAPEs range from 19.9% (January) to 31.8% (August) for shampoo quantities, and from 17.6% (August) to 29.4% (June) for toothpaste quantities.
44. The preceding results reported in this section, focussing solely on January 2012, are not atypical of the goodness-of-fit of the truncated log-normal distribution throughout 2012 in general (Table 2); however, it should be noted that the R^2 is maximised and the MAPE is minimised in January for shampoo quantities.
45. Fitting the truncated log-normal distribution to expenditure rather than quantity does not result in any notable improvement in goodness-of-fit (Table 2). For toothpaste, the R^2 is greater for seven months and the MAPE is lower for six months when the distribution is fitted to expenditure rather than quantity. For shampoo, although the R^2 is greater for all 12 months and the MAPE is lower for 10 months when the distribution is fitted to expenditure rather than quantity, the differences in goodness-of-fit are generally small in absolute terms.

Table 2: Goodness-of-fit statistics, truncated log-normal distribution, 2012

Variable	Month	Shampoo		Toothpaste	
		R^2	MAPE	R^2	MAPE
Quantity	January	91.8	19.9	89.3	23.5
	February	89.7	24.3	84.6	24.9
	March	88.2	23.9	86.4	23.3
	April	88.8	24.6	84.9	20.6
	May	88.3	27.9	85.5	21.6
	June	87.7	27.3	86.4	29.4
	July	87.0	25.8	87.3	22.5
	August	87.6	31.8	90.5	21.2
	September	87.2	27.5	90.2	17.6
	October	87.2	25.5	89.7	22.8
	November	86.8	30.9	89.3	28.2
	December	86.5	22.1	86.9	28.0
Expenditure	January	92.1	17.9	88.4	23.2
	February	91.5	23.5	83.6	27.0
	March	90.6	19.8	87.3	32.5
	April	90.4	20.7	86.1	26.3
	May	89.3	21.9	85.2	31.1
	June	89.5	23.2	85.2	23.1
	July	88.4	26.2	86.2	25.9
	August	89.6	21.7	90.9	21.1
	September	89.8	25.5	91.4	50.0
	October	88.3	25.3	90.2	22.3
	November	88.3	27.4	89.8	25.8
	December	88.3	26.6	87.2	23.9

46. After multiplying the fitted quantities from the truncated log-normal distribution by the corresponding observed prices to derive expenditure weights, the resulting price index series for shampoo closely tracks that constructed using the aforementioned Rank 3 method (Figure 10). However, the levels of both index series are consistently above that of the benchmark Geometric Laspeyres series utilising weights constructed from observed expenditure, with the difference increasing with time from the reference period.
47. As with shampoo, the toothpaste price index series resulting from use of the fitted quantities is consistently above the benchmark Geometric Laspeyres index series (Figure 11). However, it is also consistently above the index series constructed using the Rank 3 method, which provides greater accuracy in reproducing the benchmark series.

Figure 10: Shampoo price index series, 2012

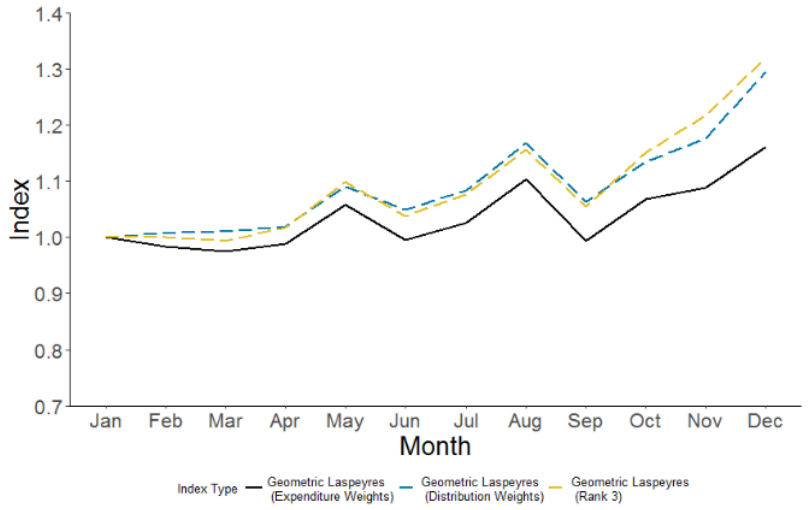
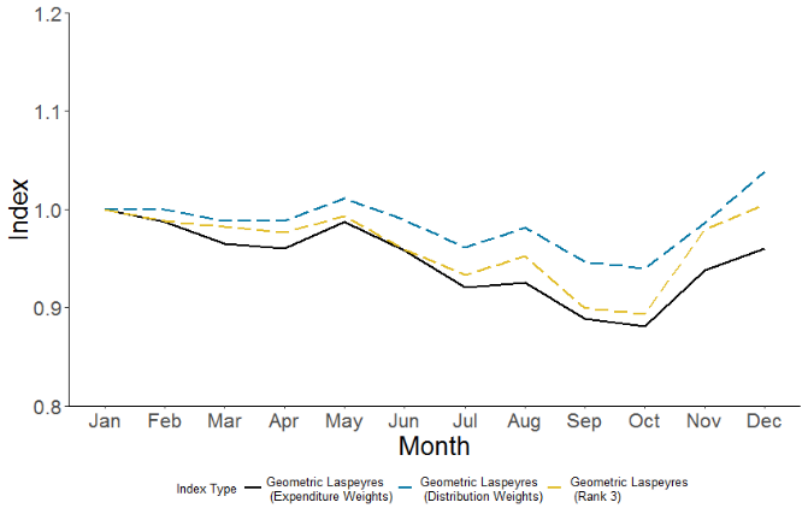


Figure 11: Toothpaste price index series, 2012



Conclusions and further work

48. Changing the assumption of the analysis carried out in paper APCP-T(18)14, such that a page ranking is considered a proxy of *quantity* in place of *expenditure*, does not change the conclusions reached previously, with the Rank 3 method for transforming the rankings still resulting in indices closest to the benchmark expenditure-weighted Geometric Laspeyres index in 2012. Although this method provided a reasonable approach to approximating expenditures using a scanner dataset, the method is yet to be tested on web scraped data (since web scraped and scanner data for the same retailer are not available) and a suitable method for selecting a value for x for different items is yet to be determined.
49. Taking a subset of the products in the datasets, either by expenditure or quantity, and calculating either a Jevons or Rank 3 method weighted index does not lead to indices that are close to the benchmark Geometric Laspeyres index series in 2012. The Jevons indices are volatile and it is difficult to decide on a subset that results in the closest index, although all subsets are closer to the benchmark Geometric Laspeyres index than the Jevons index for the entire product set. The Rank 3 method indices experience chain drift due to a small number of particularly high price relatives dominating the index for smaller subsets.
50. It is thus not advised to take subsets of the dataset for the Rank 3 method, as indices calculated using this method for the entire product set are closer to the benchmark Geometric Laspeyres index. This analysis addresses the possibility of lower ranked products impacting the index, i.e. this is true for the Jevons index, as the products are all given equal weight, but including the lower ranked products when using the Rank 3 method works well since these products are weighted appropriately.
51. Of the three candidate statistical distributions, the truncated log-normal distribution provides the best approximation to the observed quantities of both shampoo and toothpaste in 2012. Truncation of the distribution reduces the propensity for over-prediction amongst higher ranking products compared to the standard log-normal distribution, while the data do not exhibit the linear log-quantity versus log-rank relationship that is characteristic of a power-law distribution.
52. Using expenditure weights derived from predicted quantities results in shampoo and toothpaste price index series that exhibit similar period-on-period movements to their benchmark Geometric Laspeyres series, but that are consistently greater in terms of their levels. Furthermore, the Rank 3 method provides a somewhat closer representation of the benchmark Geometric Laspeyres index series for toothpaste.
53. If implemented in a production environment, data providers would need to supply ONS with the following parameters for each item (calculated across all products within the item) to fit the truncated log-normal distribution: the minimum quantity; the maximum quantity; the arithmetic mean of the natural logarithm of quantities; and the standard deviation of the natural logarithm of quantities. Whilst relatively trivial to calculate, in practice these quantities may not be provided to ONS on an ongoing monthly basis. Future work may therefore seek to explore:
    - the impact on goodness-of-fit (and the resulting index series) of fitting the truncated log-normal distribution using parameter estimates obtained from annualised rather than monthly quantity data

- the out-of-sample predictive performance of the fitted truncated log-normal distribution, by estimating the parameters of the distribution on a training dataset (e.g. 2011) and then assessing goodness-of-fit (and the resulting index series) on a holdout dataset (e.g. 2012) - thereby simulating an annual delivery of parameter estimates from the data provider to ONS

54. The next steps in this research are those indicated above, as well as those outlined in objectives (e) to (h), although the latter is dependent on the availability of data on which to perform the analyses.
55. Furthermore, the use of a *chained* Geometric Laspeyres index appears to have resulted in chain drift, particularly when restricting the dataset to subsets based on quantity/expenditure. The analysis on subsets of data could therefore be repeated, taking a sample of the top 5, 10, etc. products in only January and following these products throughout the year, eliminating the impact of chain drift. Further to this, taking subsets consisting of a greater number of products (e.g. 50) could be tested.
56. The conclusions of this research are limited by the caveat that the dataset used is a scanner dataset, and the hypothesis that the ranking of a product on a web page indicates popularity has not yet been tested; throughout the analysis presented in this paper, we have assessed the best treatment of product popularity rankings *once they are known*. Future research will test this hypothesis where both transaction data and web scraped data are available.

References

57. Antoniades A (2017). 'Distribution as expenditure', *not yet published*.
58. Auer J and Boettcher I (2017). 'From price collection to price data analytics: How new large data sources require price statisticians to re-think their index compilation procedures. Experiences from web-scraped and scanner data'. In: International Working Group on Price Statistics, Fifteenth Meeting of the Ottawa Group, 10-12 May 2017, Eltville am Rhein: Germany.
59. Bhardwaj H, Flower T, Lee P and Mayhew M (2017). 'Research indices using web scraped price data: August 2017 update', Office for National Statistics.
60. Bosch O T and Griffioen R (2016). 'On the use of internet data for the Dutch CPI'. In: UNECE/Eurostat/OECD, Meeting of the Group of Experts on Consumer Price Indices (2016), 2-4 May 2016, Geneva: Switzerland.
61. Bosch O T, Windmeijer D, Delden A V and Huevel G V D (2018). 'Web scraping meets survey design: Combining forces. In BigSurv 2018, Exploring new statistical frontiers at the intersection of survey science and big data, 25-27 October 2018, Barcelona: Spain.
62. Cavallo A (2012). 'Online and official price indexes: Measuring Argentina's inflation', Journal of Monetary Economics, Volume 60, Issue 2, pages 152 to 165.
63. Cavallo A (2017). 'Are online and offline prices similar? Evidence from large multi-channel retailers', American Economic Review, Volume 207, Issue 1, pages 283 to 303.
64. Cavallo A and Rigobon R (2016). 'The Billion Prices Project: Using online prices for measurement and research', Journal of Economic Perspectives, Volume 30, Issue 2, pages 151 to 178.
65. Chessa A G and Griffioen R (2017). 'Comparing price indices of clothing and footwear for scanner data and web scraped data', *not yet published*.
66. Chevalier J and Goolsbee A (2003). 'Measuring prices and price competition online: Amazon.com and BarnesandNoble.com', Quantitative Marketing and Economics, Volume 1, Issue 2, pages 203 to 222.

67. Hisano R and Mizuno T (2010). 'Sales distribution of consumer electronics', Physica A: Statistical Mechanics and its Applications, Volume 390, Issue 2, pages 309 to 318.
68. Hull I, Lof M and Tibblin M (2017). 'Price information collected online and short-term inflation forecasts'. In: ISI Regional Statistics Conference, IFC – Bank Indonesia satellite seminar on big data, 20-24 March 2017, Bali: Indonesia.
69. Krsinich F (2015). 'Price indexes from online data using the fixed-effects window-splice (FEWS) index'. In: International Working Group on Price Statistics, Fourteenth Meeting of the Ottawa Group, 20-22 May 2015, Urayasu City: Japan.
70. Loon K V and Roels D (2018). 'Integrating big data in the Belgian CPI'. In: UNECE/Eurostat/OECD, Meeting of the Group of Experts on Consumer Price Indices (2018), 7-9 May 2018, Geneva: Switzerland.
71. Nygaard R (2015). 'The use of online prices in the Norwegian Consumer Price Index'. In: International Working Group on Price Statistics, Fourteenth Meeting of the Ottawa Group, 20-22 May 2015, Urayasu City: Japan.
72. Polidoro F, Giannini R, Conte R L, Mosca S and Rossetti F (2015). 'Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation', Statistical Journal of the IAOS, Volume 31, Issue 2, pages 165 to 176.
73. Stanley M H R, Buldyrev S V, Havlin S, Mantegna R N, Salinger M A and Stanley H E (1995). 'Zipf plots and the size distribution of firms', Economics Letters, Volume 49, Issue 4, pages 453 to 457.
74. Touzani S and Buskirk R V (2015). 'Estimating sales and sales market share from sales rank data for consumer appliances', Physica A: Statistical Mechanics and its Applications, Volume 451, Issue 1, pages 266 to 276.
75. Willenborg L (2017). 'Elementary price indices for internet data', Statistics Netherlands.

Heledd Thomas and Daniel Ayoubkhani
Methodology, ONS
January, 2019

List of Annexes

Annex A	Findings from the literature review

Annex A – Findings from the literature review

1. An online literature review was conducted to inform our work on estimating weights for products whose price information has been obtained via web scraping. The findings of the literature review are described below.

Use of web scraped data for price indices, but without product-level weights

2. There exists a substantial and growing body of research relating to the use of internet price data to compile consumer price index series. As demonstrated by the literature summarised below, this research has taken place both within and outside of national statistical institutes, and across numerous countries and product groups. However, the issue of product-level weighting has been largely overlooked; investigators have generally made use of unweighted index number techniques at the elementary aggregate level, as per the traditional approach to compiling price statistics.
3. Cavallo (2012) compared official Argentine consumer price inflation estimates with those derived from web scraped data, while a similar exercise was undertaken by Cavallo and Rigobon (2016) for the inflation experience of numerous other countries; in both instances, online price data were collected via MIT's Billion Prices Project. When constructing price indices using web scraped data, an unweighted geometric mean was used to aggregate product-level price relatives by product group, with a weighted arithmetic mean being used for aggregation above this level; weights were taken directly from the published consumer price indices of the countries under analysis (i.e. not based on web scraped data).
4. Research into the use of online prices of personal care products in Norway has been described by Nygaard (2015). The authors attempted to mitigate the issue of not having quantity/expenditure data with which to construct weights by including only the most popular products (when sorted by popularity on the scraped webpages) from the most popular retailers (according to registered turnover) in the price index. In one method explored in the paper applied to daily web scraped shampoo data, price relatives for homogenous product strata were weighted together in proportion to frequencies of price quotes, with the resulting index series generally being above that obtained from an unweighted index formula.
5. The future use of internet data in the Dutch consumer price index was summarised by Bosch and Griffioen (2016), proposing that prices for product groups are calculated as an unweighted arithmetic mean of each retailer's web scraped prices across products within the group within each month. Weights would only be applied above the elementary index level using sales data from other (non-internet) sources, as per the existing CPI compilation process.
6. Polidoro et al. (2015) analysed internet prices for consumer electronics in Italy, calculating unweighted geometric mean prices at the elementary aggregate level and weighting in proportion to known market shares to aggregate indices at higher levels. A similar approach was taken by Hull et al. (2017), who collected web scraped prices for a number of fruit and vegetable products in Sweden, and compared the resulting item-level price indices to those published in the Swedish consumer price index. Unweighted geometric mean prices were calculated at the product level, while item-level price indices made use of pre-existing (non-web scraped) expenditure weights from the Swedish CPI.
7. Loon and Roels (2018) present various case studies of research into using web scraped data in the Belgian consumer price index. The most relevant of these to the present study is footwear,

for which products were classified into groups using scraped product-level information, and an (unweighted) Jevons price index was calculated at this group level.

8. Various methods for constructing price indices from internet data are outlined and evaluated by Willenborg (2017). The methods are grouped according to whether they are based on product matching (through time) or product classification (into items). However, the author concentrates solely on calculating elementary (unweighted) price indices, with weighted aggregation to higher levels remaining beyond the scope of the review.
9. A possible high-level workflow for using web scraped data in the context of survey methodology was outlined by Bosch et al. (2018), focussing on remedying errors and biases when deriving inferences from combined internet and survey data. However, this framework does not cover the possibility of deriving product-level weights from the web scraped data itself.

Research involving both scanner and internet data

10. In the absence of online data, the present study makes use of scanner data to inform possible methods for constructing weights as and when web scraped data become available. Several published studies also combine, and in some cases compare, scanner and online data – these are reviewed below.
11. Chessa and Griffioen (2017) compared clothing and footwear prices, and the resulting price indices, derived from scanner and web scraped data obtained from the same Dutch retailer. Point-of-sale and internet prices were found to exhibit a high degree of correlation, as too were the resulting Geary-Khamis price indices (weighted by quantities for point-of-sale data and number of web scraped prices for internet data). This research was expanded on by Cavallo (2017) who, as part of MIT's Billion Prices Project, simultaneously collected internet and physical store prices for 24,000 products sold by 56 multi-channel retailers across 10 different countries. Price levels were found to be identical between the two sources most of the time (though the match rate varied considerably by country, product group and retailer), with price changes exhibiting similar frequencies and average magnitudes (though not necessarily timing).
12. Krsinich (2015) discussed the application of the FEWS index method, developed to perform quality adjustment when faced with products entering and leaving the market, to web scraped information. The impact of not having access to quantity data was simulated by producing weighted and unweighted index series using scanner data for a range of consumer electronic and grocery products in New Zealand. A relatively small impact was observed for groceries (with the effect of weighting being most noticeable on the seasonal pattern of prices changes rather than the general trend), but the impact for electronics was more variable.

Using statistical distributions to estimate sales quantities from their ranks

13. Although internet data does not routinely include sales quantities, on many websites it is possible to sort products by popularity, which may be considered a reasonable approximation to quantity or expenditure ranking. Several papers, most notably those summarised below, have investigated the statistical distribution of sales quantities for different product groups; this information has informed our research into predicting quantities from their ranks.
14. Stanley et al. (1995) demonstrated that manufacturing firms' sales can be well approximated by the log-normal distribution. However, a Zipf plot of log-sales against log-ranks illustrated that

the upper tail of the empirical distribution was too thin relative to the theoretical log-normal distribution.

15. Online book sales for around 26,000 titles available on Amazon.com and BN.com during 2001 were analysed by Chevalier and Goolsbee (2003), who obtained both prices and ranks for each of these products (note that the ranks were themselves based on sales, as reported by the retailers). The authors modelled sales quantities as being Pareto (power-law) distributed, and used this distribution to translate observed sales ranks into predicted sales quantities. The resulting sales-weighted average book prices were notably different to the corresponding unweighted average prices for both retailers.
16. Daily data on the sales of digital cameras in Japan between 2004 and 2008 were analysed Hisano and Mizuno (2010). The authors found that the observed sales distribution could be well approximated by the log-normal distribution for some periods of the analysed timespan, but a power-law distribution provided a better fit to the data in other periods.
17. Touzani and Buskirk (2015) used scanner data from a sample of US retailers between 2004 and 2011 to model the quantity distributions of refrigerators, freezers and washing machines (over 200,000 observations in total). The authors found that sales quantities could be predicted from their ranks with reasonable accuracy by modelling them with a log-normal distribution. Furthermore, double truncation of the log-normal distribution led to further improvements in goodness-of-fit, particularly amongst the top selling products for each item.
18. Antoniades (2017) demonstrates a method whereby importance weights may be estimated from prices alone (i.e. without quantities): firstly, a measure of retail distribution is calculated as the number of retailers carrying a particular product divided by the total number of retailers in the sample (weighted according to the number of price quotes observed); and secondly, weights are estimated by assuming an exponential relationship between market share and retail distribution. Through simulation, the author finds that incorporating these estimated weights reduces the bias associated with measures of price inflation by 73% compared to those obtained from an unweighted index.

Other issues

19. Auer and Boettcher (2017) demonstrate the challenges of incorporating web scraped data in the Austrian consumer price index. Although they pay most attention to the technical hurdles, and do not mention the statistical challenge of estimating expenditure weights, the authors do provide some insight the appropriateness of using “most popular” page ranking as a proxy for quantity ranking. In a sample of 15 of the “most popular” shampoo products scraped from a particular retailer’s website over a nine-month period, only two were present in all nine months, and the average duration in the “most popular” list is 3.8 months; the authors suggest that this may be attributable to the “most popular” ranking being used as a marketing tool by retailers to promote the launch of certain products, and it’s reliability as a proxy for sales quantity should therefore be questioned.