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

Heledd Thomas, Daniel Ayoubkhani

Abstract

As part of research into the introduction of web scraped data sources in consumer price index measurement, the United Kingdom's Office for National Statistics has conducted some research to investigate the potential impact of these data on item-level indices. This paper is an investigation into approximate weight allocation methods for the products used in the calculation of an item's index.

Alternative data sources, such as web scraped data and point of sale scanner data, are becoming more commonly available and have potential to improve consumer price indices through more frequent data collection, increased coverage and larger sample sizes. However, the main drawback with web scraped data is that, since all prices are scraped, calculating an unweighted index at the lowest level of aggregation means more popular items would not have enough influence on the index. Unweighted indices may not be representative of consumer spending if the prices of less popular products behave differently to those of more popular products.

In the absence of product-level expenditure or quantity information, we have investigated various methods for approximately assigning weights to products, including deriving weights from the product position on a web page or the market share. Geometric Laspeyres indices are calculated from these derived weights, such that their behaviour can be compared to the unweighted Jevons index calculated from the same data source. Further analysis considers how the difference in behaviour between items should be considered; does the market distribution indicate the existence of a market leader or is there perfect competition?

1. Introduction

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 have insufficient 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.

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.

This paper is an overview of the research conducted by the United Kingdom's (UK) Office for National Statistics (ONS). The ONS is researching the use of alternative data sources to replace the existing manual price collection, with both coverage and cost in mind.

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.

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.

1.1 Characterising the analysed items

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.

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

2. Methods for weighting

Given a dataset containing both prices and page rankings for all available products in an item category, this research looks at the derivation of proxy product-level expenditure weights and the resulting Geometric Laspeyres indices. A Geometric Laspeyres index is calculated so that the resulting indices are comparable with the Jevons indices for the dataset, i.e. the differences in the index series can be attributed to the weights applied, since the Jevons index is used in the current methodology.

Methods 1 and 2 (sections 2.1 and 2.2 respectively) aim to transform the page rankings into proxy expenditure weights, w_i, used in the calculation of the resulting Geometric Laspeyres index. Method 3 (section 2.3) looks at limiting the number of products used in the calculation of the index, applying equal weights to the selected products. i.e. a Jevons index is calculated on a subset of the data. The expenditure-weighted Geometric Laspeyres index is considered as a **benchmark** for other indices calculated.

2.1 Method 1 – Using formulae to transform rankings to weights

The aim is to find a formula for transforming product page rankings to product-level weights that closely align with weights calculated from expenditure shares.

Equations 1 and 2 make use of descending order ranks, i.e. rank 1 is assigned to the most popular product. Both methods for transforming the ranks ensure that a higher weight is allocated to the lowest ranked products, before normalising the weights to ensure that the sum of weights for products $1, \dots, n$ is equal to 1.

$$Rank\ Weight\ 1 \qquad w_i^0 = \frac{2}{n}\left(1 - \frac{r_i}{n+1}\right) \qquad (1)$$

$$Rank\ Weight\ 2 \qquad w_i^0 = \frac{1}{\sum_{i=1}^{n}\frac{1}{r_i}}\left(\frac{1}{r_i}\right) \qquad (2)$$

$w_i^0 is\ the\ weight\ assigned\ to\ product\ i\ in\ the\ base\ perio$

$r_i is\ the\ rank\ of\ product\ i\ in\ the\ base\ period$
$(in\ descending\ order\ according\ to\ popularity)$
$n\ is\ the\ number\ of\ products\ in\ the\ item's\ dataset$

Equation 3 makes use of ascending order ranks, i.e. rank 1 is assigned to the least popular product. Using rank shares alone (i.e. taking $x = 1$) does not give enough weight to the most popular products and gives too much weight to the least popular products. As the power is increased, higher weights are assigned to the higher ranked (most popular) products and lower weights are assigned to the lower ranked (least popular) products.

$$Rank\ Weight\ 3 \qquad w_i^0 = \frac{(Rank\ share)^x}{\sum(Rank\ share)^x} \qquad (3)$$

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

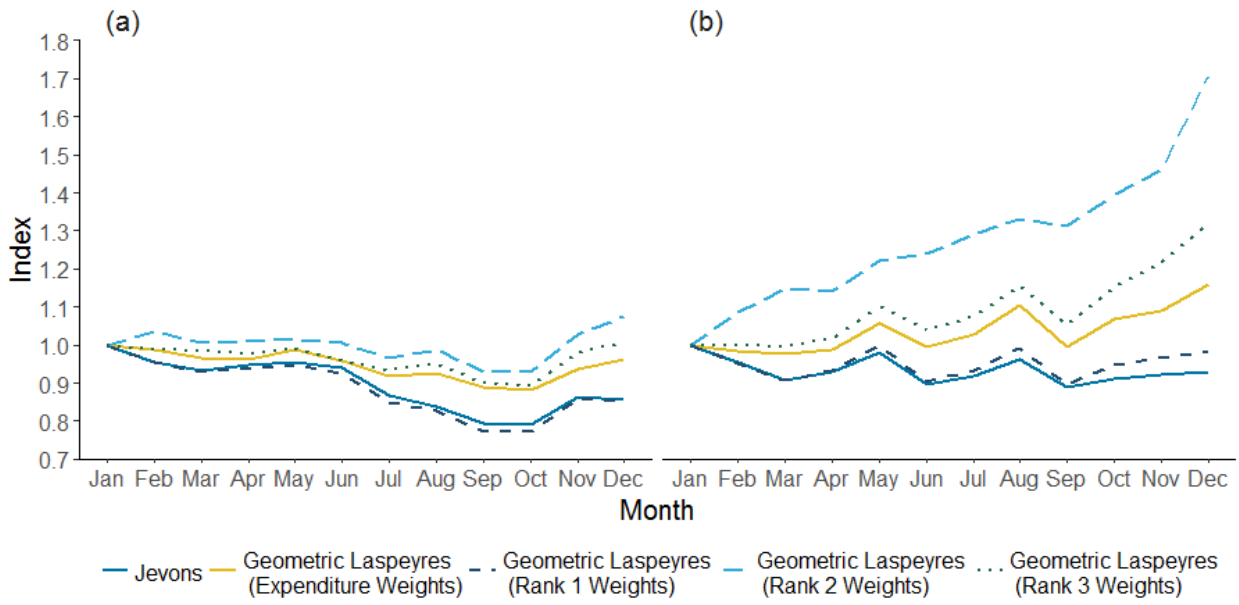
$r_i is\ the\ rank\ of\ product\ i\ in\ the\ base\ period$

$(in\ ascending\ order\ according\ to\ popularity)$

$x\ is\ an\ integer$

Figure 1 shows that 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.

Figure 1: Toothpaste (a) and shampoo (b) indices with different weights applied, using quantity rankings compared with the expenditure-weighted index, 2012



The method has only been tested for two CPI items, and the fact that the value for x was selected by inspection of the index series calculated using different values for x means that it should not, therefore, be assumed as the optimal solution for other items for which this methodology has not yet been tested. In the absence of product-level microdata on every item traded by a retailer, Method 2 (section 2.2) 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.

2.2 Method 2 – Using distributions to estimate quantities from ranks

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, which they may be unable or unwilling to supply.

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.

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.

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).

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.

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.

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.

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 2 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.

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 3 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 4).

Figure 2: Quantity vs. rank (log scale), simulated and observed quantities, January 2012

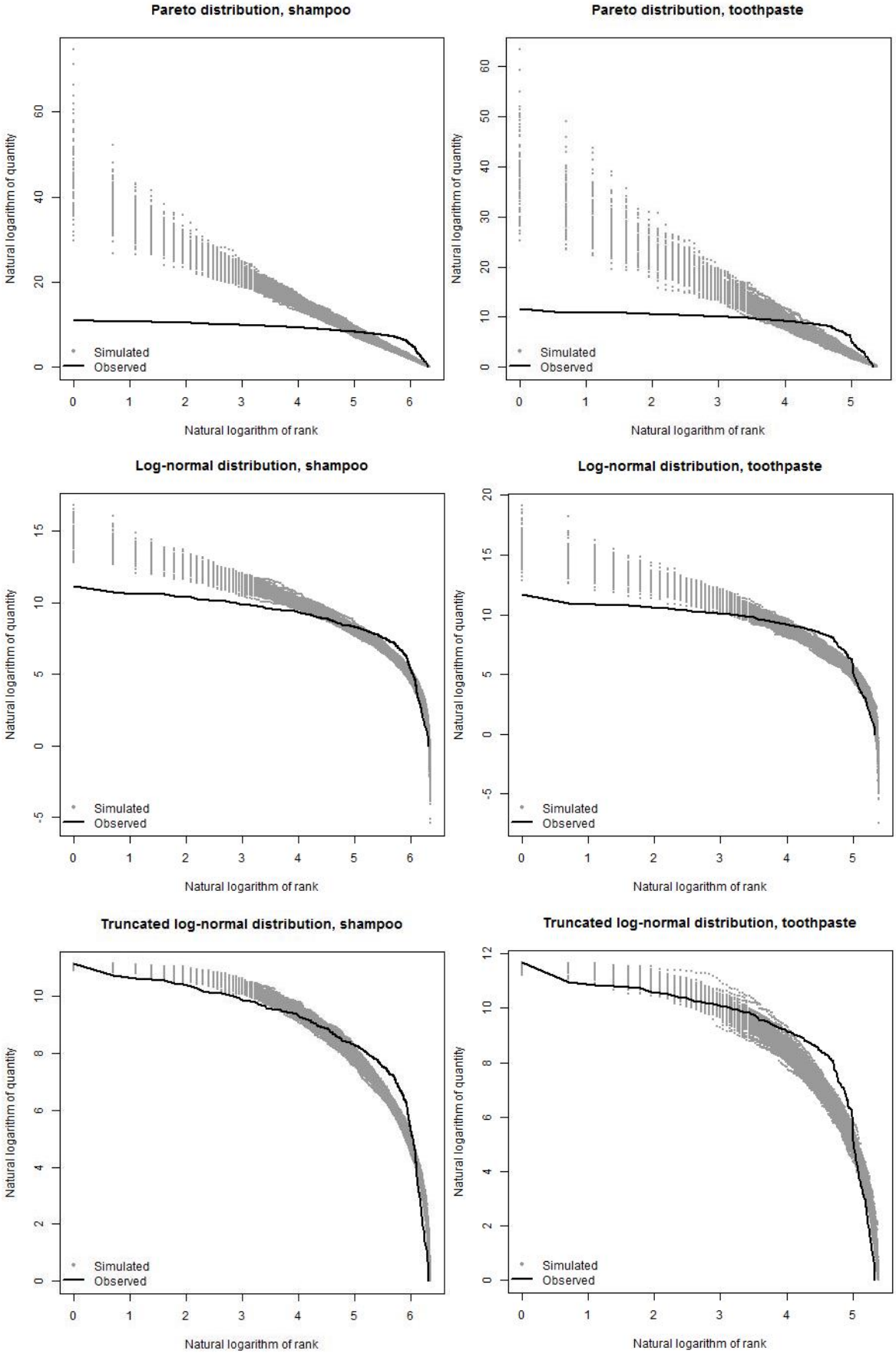


Figure 3: Quantity vs. rank (log scale), fitted and observed quantities, January 2012

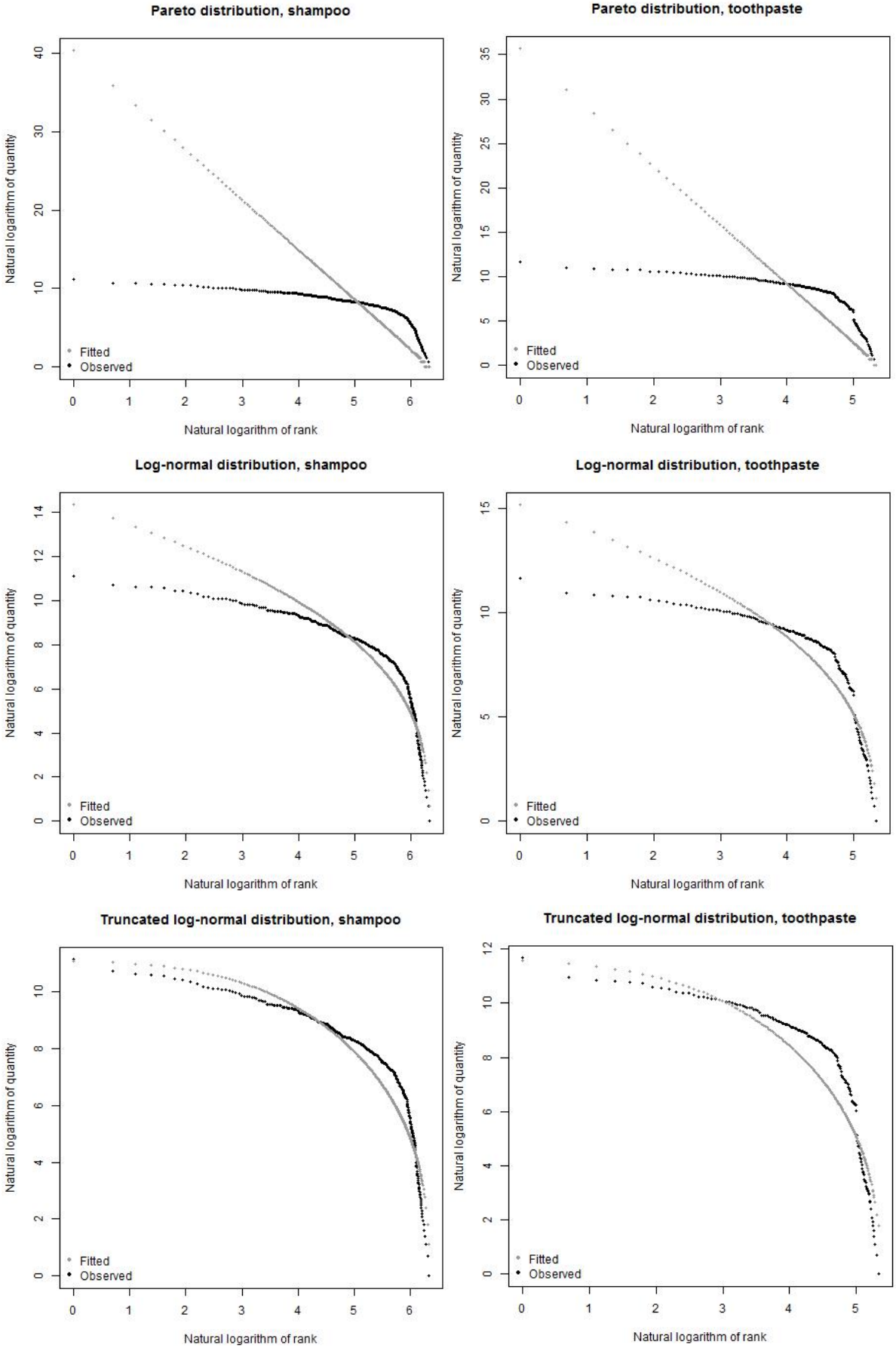
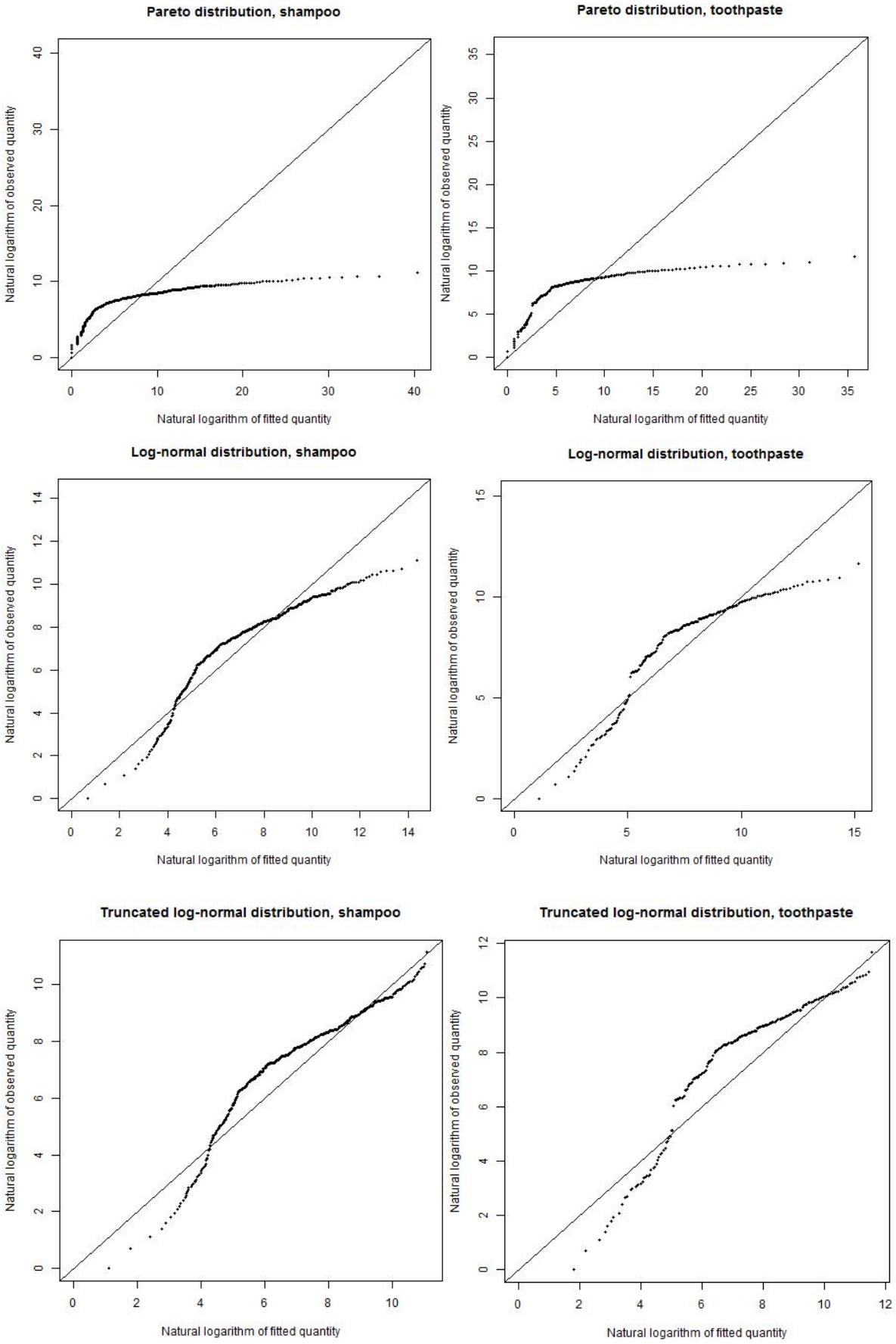


Figure 4: Observed vs. fitted quantities (log scale), January 2012



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.

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.

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.

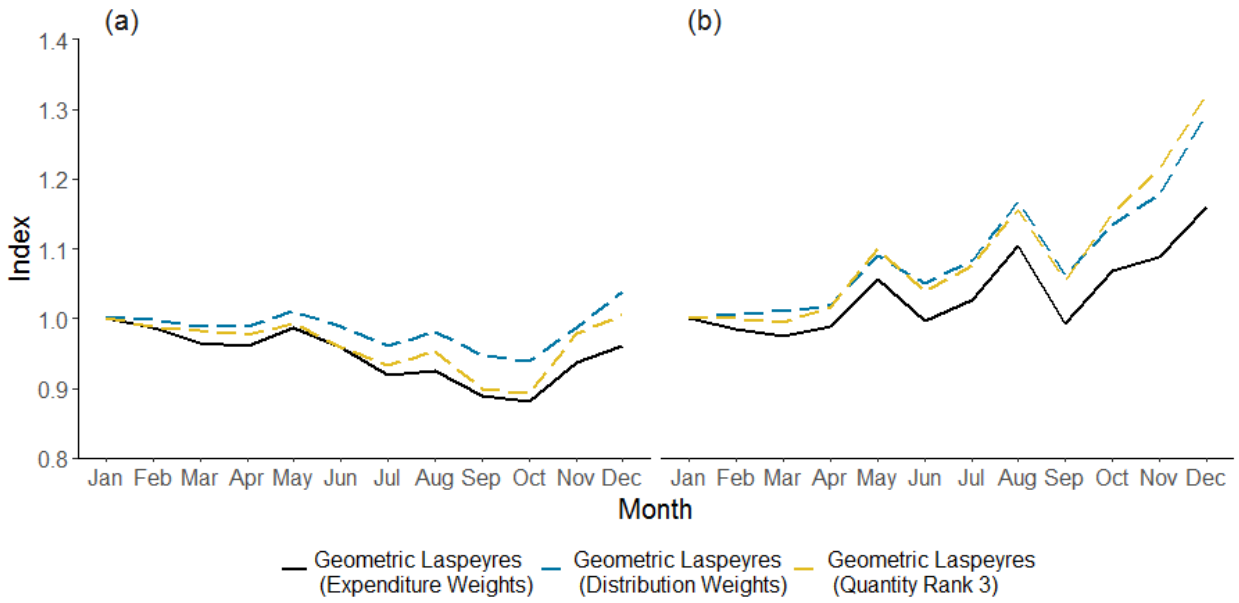
Table 2: Goodness-of-fit statistics, truncated log-normal distribution, 2012

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

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 5(b)). 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.

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 5(a)). 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 5: Toothpaste (a) and shampoo (b) price index series, 2012



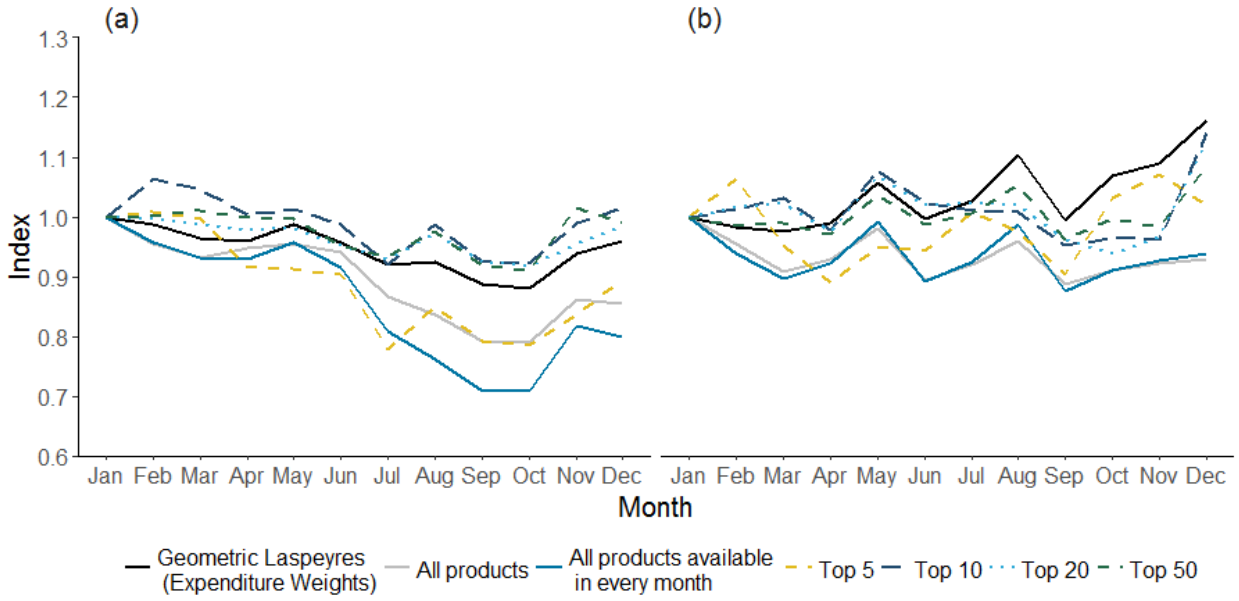
2.3 Method 3 – Using subsets of the data

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.

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 quantity.

Figure 6 indicates that the Jevons indices for the top 10, 20 and 50 products by quantity are closer to the benchmark expenditure-weighted Geometric Laspeyres index than the Jevons index for all products in the dataset and all products available in every month. The results are more variable when the top 5 products are taken. Each subset shows a different pattern between months and none align closely to those evident in the benchmark Geometric Laspeyres index.

Figure 6: Toothpaste (a) and shampoo (b) Jevons indices for various quantity subsets, 2012



3. Conclusions and future work

3.1 Conclusions

The key findings from the preceding analysis for each of the three considered methods can be summarised as follows:

- **Method 1 – Rank weights based on quantity rankings:** The Rank 3 method for transforming the rankings is closest to the benchmark expenditure-weighted Geometric Laspeyres index in 2012, although not as close.
- **Method 2 – Estimating quantities from ranks:** The truncated log-normal distribution provides the best approximation to the quantities of toothpaste and shampoo. The resulting indices display similar period-on-period movements to the benchmark expenditure-weighted Geometric Laspeyres index but at a consistently higher level.
- **Method 3 – Using quantity subsets:** The Jevons indices calculated from quantity subsets are volatile but closer to the benchmark expenditure-weighted Geometric Laspeyres index than the Jevons indices calculated using all data.

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.

Taking a subset of the products in the datasets 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.

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.

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.

3.2 Future work

If statistical distributions 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

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. 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.

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.

4. References

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.

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.

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.