Office for National Statistics

Article

Using statistical distributions to estimate weights for web-scraped price quotes in consumer price statistics

Feasibility of predicting sales quantities from product ranks, for potential use with web-scraped data in consumer price statistics.

Contact:
Helen Sands
cpi@ons.gov.uk
+44 (0)1633 456900

Release date:
1 September 2020

Next release:
To be announced

Table of contents

1. Main points

2. Introduction

3. Data and methods

4. Results

5. Discussion

6. Related links

1 . Main points

- Web-scraping is a way of collecting data that results in all products from retailers' websites being available for use in consumer price statistics, independent of whether those products have had any sales.
- Including products that have low to no sales in consumer price statistics could result in biased indices, so using a scanner dataset covering a 24-month period we demonstrate the feasibility of fitting statistical distributions to predict sales quantities from their ranks.
- Of the evaluated candidate statistical distributions, the log-normal distribution provided the closest approximation to observed quantities, with truncation reducing the extent of over-prediction among higher ranked products; the predictive performance of the Pareto distribution was notably inferior to that of the two log-normal variants.
- While these results are promising, they are dependent on having descriptive statistics on sales quantities from retailers as well as relying on retailers' websites ranking products in a meaningful way; therefore, more research is needed before these methods can be applied to web-scraped data in UK consumer price statistics.

2 . Introduction

Traditional methods of price data collection involve visiting or phoning physical outlets or visiting retailers' online websites and manually collecting the advertised prices. The selection of products within outlets is purposive. In each outlet, collectors choose one variety of a product that is representative of what people buy in their area from all available products. This ensures that only price movements for things that consumers purchase are included in the resulting inflation statistics.

For more information of how prices are collected and price indices are constructed in our current measures of consumer price inflation, please refer to our Consumer Prices Indices Technical Manual.

New data sources and methods are being introduced into the production of UK consumer price statistics from 2023. The new data sources, namely scanner and web-scraped data, provide many benefits compared with more traditional methods of data collection, including improved product coverage, higher frequency of collection and potential cost savings.

Scanner data are collected by retailers at the point of sale, whether online or in physical stores. Instead of making a judgement about the representativeness of the product, as is done in the traditional collection, scanner data can tell us the exact quantity and value sold, which can be used to weight price quotes by order of their economic importance. For example, if more people purchase a particular brand of bread than another, we can use the total sales value to give more weight to this product to ensure its price movements have more importance in the resulting inflation measure.

As scanner data are provided by retailers, we have been looking to acquire this data source largely where the market is dominated by a relatively small number of retailers, such as for groceries. In more saturated markets, such as clothing, there are numerous retailers that would need to be approached to cover a significant proportion of the market. Furthermore, scanner data can sometimes lack detailed product attributes that can be used to determine the quality of products. This information is useful when trying to appropriately account for quality changes in products over time and when trying to classify products.

For product categories where the market is heavily saturated or where we require a richer level of product detail, we have been investigating the use of web-scraped data as an alternative to traditional sources. [1] This is where automated tools collect prices from retailers' websites on a frequent basis. Web-scraping can be used across multiple websites and provides us with a richness of data that is difficult to obtain through other means.

However, web-scrapers are generally set up to collect all available prices from a retailer's website, without accounting for the popularity of products or the number that have been purchased. This means that when calculating price indices from web-scraped data, we may be including price movements for products that are infrequently or never purchased.

As can be seen in New index number methods in consumer price statistics, not using expenditure weights can lead to bias in the resulting price indices. This article investigates a method to approximate product expenditures based on their page rankings (that is, the order that products appear on a web page when sorted by popularity) via statistical distributions. This ensures that more popular products are given a greater weight in the resulting price index and adhere to the principle of price index movements being representative of general consumer spending.

This article sets out an initial investigation into the proposed method and is based on research presented to our Advisory Panel on Consumer Prices (Technical) in 2019. Because of data availability at the time of the research, the article is based on sales data for toothpaste by way of example.

Notes for: Introduction

1. Web-scraping is done in line with the Office for National Statistics (ONS) policy on web-scraping.

3 . Data and methods

Related literature

Studies around the globe have demonstrated the use of web-scraped data to estimate price indices. See Nygaard (2015), Polidoro et al. (2015), Bosch and Griffioen (2016), Van Loon and Roels (2018) and the Australia Bureau of Statistics (ABS, 2020) for a description of the Norwegian, Italian, Dutch, Belgian and Australian experiences, respectively.

These papers generally make use of unweighted index number techniques to construct elementary aggregates from web-scraped data, although highlighting that further research is required on the absence of expenditure data for these methods. To the authors' knowledge, nothing further has been published on this matter to date, although it has been touched upon by Chessa and Griffioen (2019). They compared web-scraped and scanner data from the same retailer and found that the number of product prices available on the website throughout a month was correlated with the number of products sold in the scanner dataset. They hypothesised that this may be traced back to the retailer's policy to promote items that are sold more often on their website.

Study data

We analysed monthly scanner data covering the 24-month period May 2011 to April 2013 for toothpaste sales, supplied by a single UK retailer. The dataset comprised 310 products appearing for a mean of 16 months each. Table 1 shows summary statistics for the dataset.

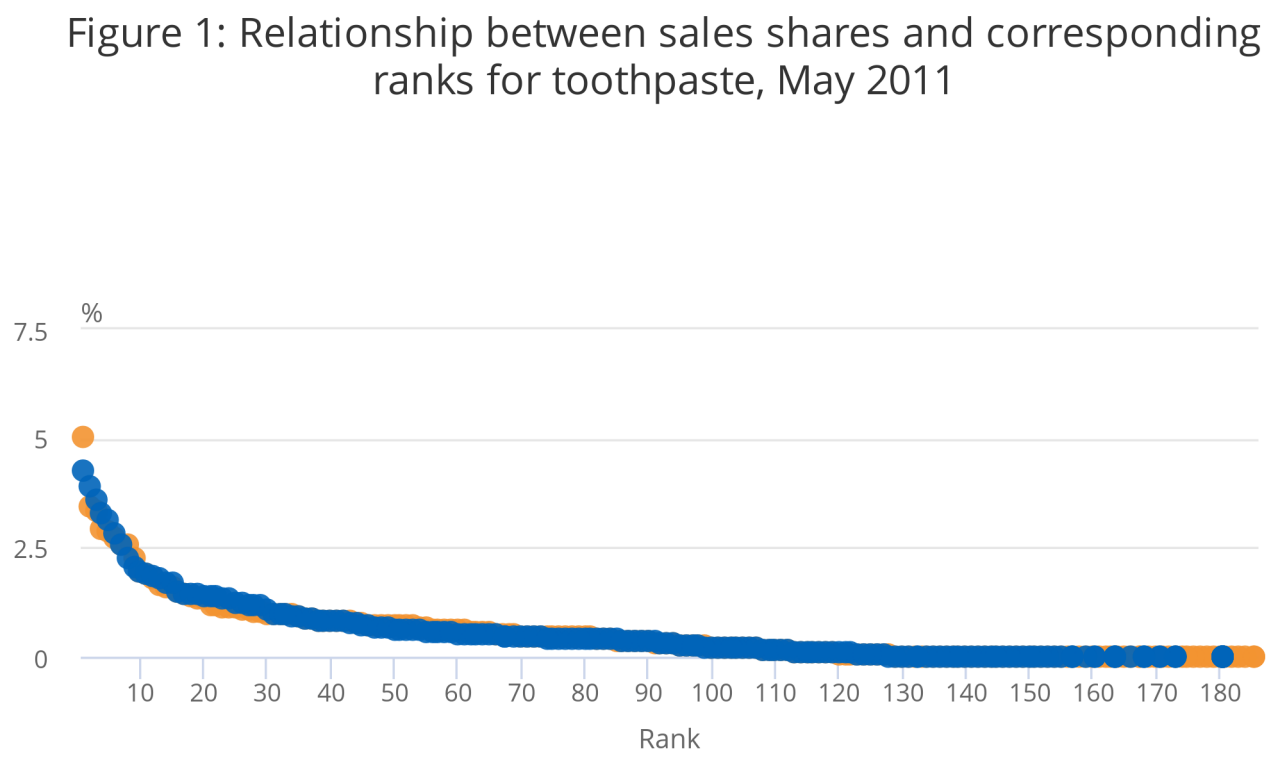
Table 1: Summary statistics for toothpaste sales, May 2011 to April 2013

Statistic	Value
Products available in every month	35%
Products available less than 6 months of the year	18%
Sales contributed by the highest expenditure product	3%
Sales contributed by the top 5 products	14%
Sales contributed by the top 10 products	24%
Sales contributed by the top 50 products	65%

Source: Office for National Statistics – Retailer data (anonymised)

The data are characterised by having a small number of market-leading products and a “long tail” consisting of many products with relatively low sales (Figure 1). The top-selling toothpaste product contributed 3.1% of sales value (under perfect competition, each product would contribute 0.3%), while the top 50 selling products made up nearly two-thirds of the sales value.

Figure 1: Relationship between sales shares and corresponding ranks for toothpaste, May 2011



Source: Office for National Statistics – Retailer data (anonymised)

Although expenditure approximates will be used with web-scraped data, the availability of both prices and quantities in the scanner dataset meant that estimated expenditure weights and the resulting indices could be compared to an index where products are weighted in accordance with their economic importance. In a real-world setting, product rankings would be scraped from retailers' websites alongside the associated prices. For this study, we assumed that quantity rankings (as observed in the analysed scanner dataset) would reasonably well approximate popularity rankings (as would be observed in a web-scraped dataset).

There are some limitations to this assumption that will need to be considered when applying these methods to actual web-scraped data. It relies on the popularity rankings that appear on retailers' websites being reflective of quantities sold; in reality, this will entirely depend on the algorithms used to define popularity by individual retailers, which may comprise facets other than, or in addition to, sales quantities (for example, the number of page views a product attracts or whether or not a product has been sponsored). In a sample of 15 of the "most popular" shampoo products scraped from a particular retailer's website, Auer and Boettcher (2017) found that only two were present in all nine months of the study period, and they suggest that this may be attributable to the popularity ranking being used by retailers as a marketing tool to promote certain products.

Price index formula

In our current headline consumer price statistics, price movements at the lowest level of aggregation (known as the elementary aggregate level) are typically combined using an unweighted geometric mean of price relatives. This index number method is referred to as the "Jevons" index:

$$P^{0,t}_{FB-Jevons} = \sqrt[n]{\prod_{i=1}^{n} \frac{p_i^t}{p_i^0}}$$

where: $P^{0,t}_{j}$ is the value of the Jevons index representing price changes among n products between base month 0 and current month t; and p^{0}_{i} and p^{t}_{i} are the price levels of product i in months 0 and t, respectively.

For this analysis, we constructed a weighted "benchmark" price index series using the geometric Laspeyres formula[1], with product weights equal to expenditure shares observed in the scanner dataset.

$$P^{0,t}_{FB-Geo\ Laspeyres} = \prod_{i=1}^{n} \frac{(p_i^t)^{w_i^0}}{(p_i^0)^{w_i^0}}$$

where: $P^{0,t}_{GL}$ is the value of the geometric Laspeyres index representing price changes among n products between base month 0 and current month t; p^{0}_{i} and p^{t}_{i} are the price levels of product i in months 0 and t, respectively; and w^{0}_{i} is the weight (expenditure share) of product i in month 0.

Different approaches to estimating expenditure weights could be assessed by replacing the observed w^{0}_{i} with its estimate and comparing the resulting price index series to the benchmark. To maximise use of the available data, the indices were chain-linked such that the base period and weights were updated each month using data relating to period t - 1.

Estimating product weights using statistical distributions

We estimated expenditure weights from observed product ranks solely by using distributional summary statistics for quantities, which could then be used alongside web scraped product-level price information. If operationalised, this would require retailers to provide summary statistics such as means and standard deviations for quantities sold. This is assumed to be more feasible than receiving more granular product-specific microdata from each retailer, but it is another possible limitation to productionising this method.

Sales quantity ranks were translated to quantiles of the cumulative distribution of sales quantities: $F(q_i) = 1 - r_i / n$. 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 was then to find a statistical distribution that suitably approximated the observed quantiles, and to use this distribution to predict sales quantities from their ranks. The log-normal, truncated log-normal and Pareto distributions were considered as candidates for predicting sales quantities. These distributions have previously been used to model the sales quantities of other consumer items, such as digital cameras and white goods.

The log-normal distribution assumes that natural logarithms of sales quantities follow a normal distribution, reflecting the skewed nature of the distribution of toothpaste sales between products; truncation of the log-normal distribution restricts the range of quantities covered by the distribution to a given range (in this analysis, the theoretical range was determined based on the minimum and maximum quantities observed in the toothpaste dataset), and the Pareto distribution is suitable when a small number of the most popular products account for a large share of the market, as is the case for the observed toothpaste dataset. The parameters of the distributions were estimated by maximum likelihood estimation (for each month separately, rather than by pooling the data through time).

Fitted quantities were multiplied by observed prices to estimate product-level expenditures and, in turn, product weights were calculated using estimated expenditure shares. The resulting geometric Laspeyres price index series could then be compared to that obtained using observed rather than estimated expenditures.

Assessing predictive performance

The performance of each candidate distribution in each month was assessed using the mean absolute percentage error (MAPE) of predicted log-quantities log across all products:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} |[\log(q_i) - \log(\hat{q}_i)] / \log(q_i) \times 100|$$

To avoid overstating the accuracy of the statistical distributions that might be observed in their real-world application, performance was evaluated out-of-sample by fitting the distributions to the first 12 months of the data (the “training” set, May 2011 to April 2012) and then calculating the MAPE of their predictions over the next 12 months (the “holdout” set, May 2012 to April 2013). This simulates the situation whereby the models would be periodically re-estimated using a given year of data from a retailer, and the fitted models would then be applied to make predictions for one or more subsequent years.

Notes for: Data and methods

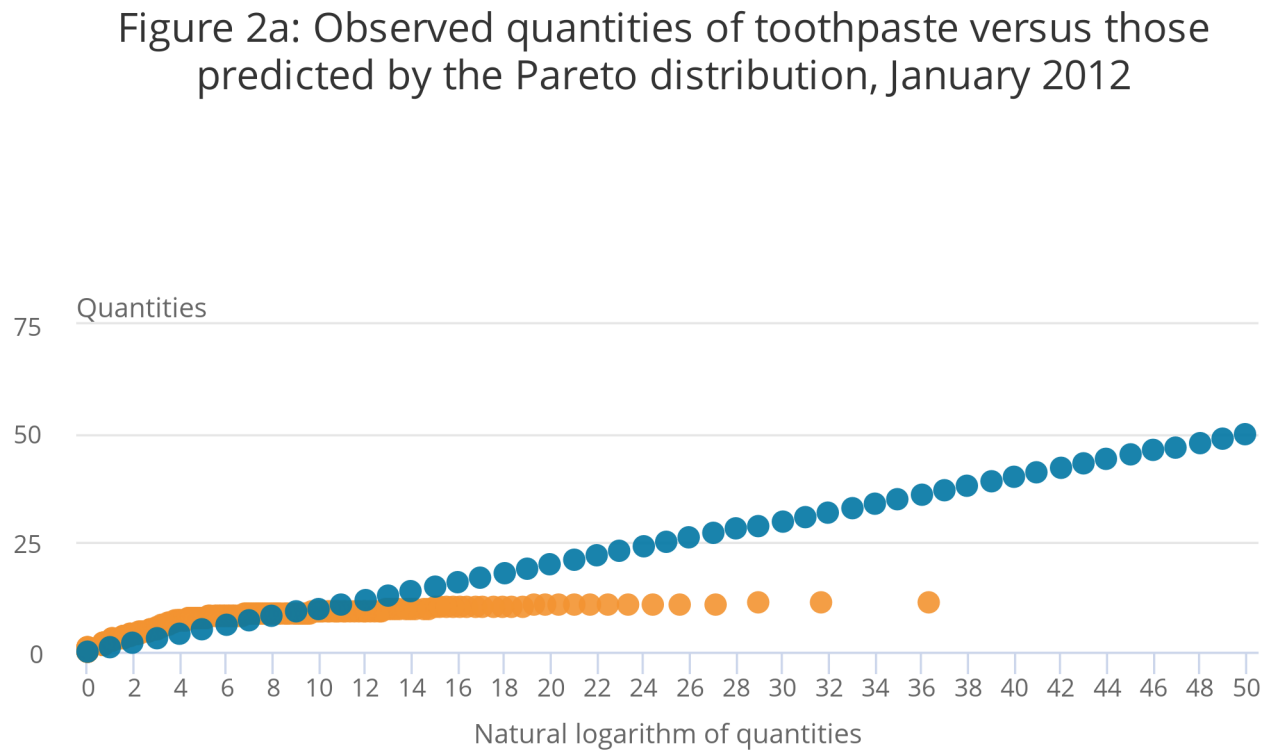
1. The geometric Laspeyres formula was used as it is conceptually similar to the current methods used in consumer price statistics. Research into suitable index number methods for use when expenditure information is available is ongoing.

4 . Results

In-sample exploratory analysis

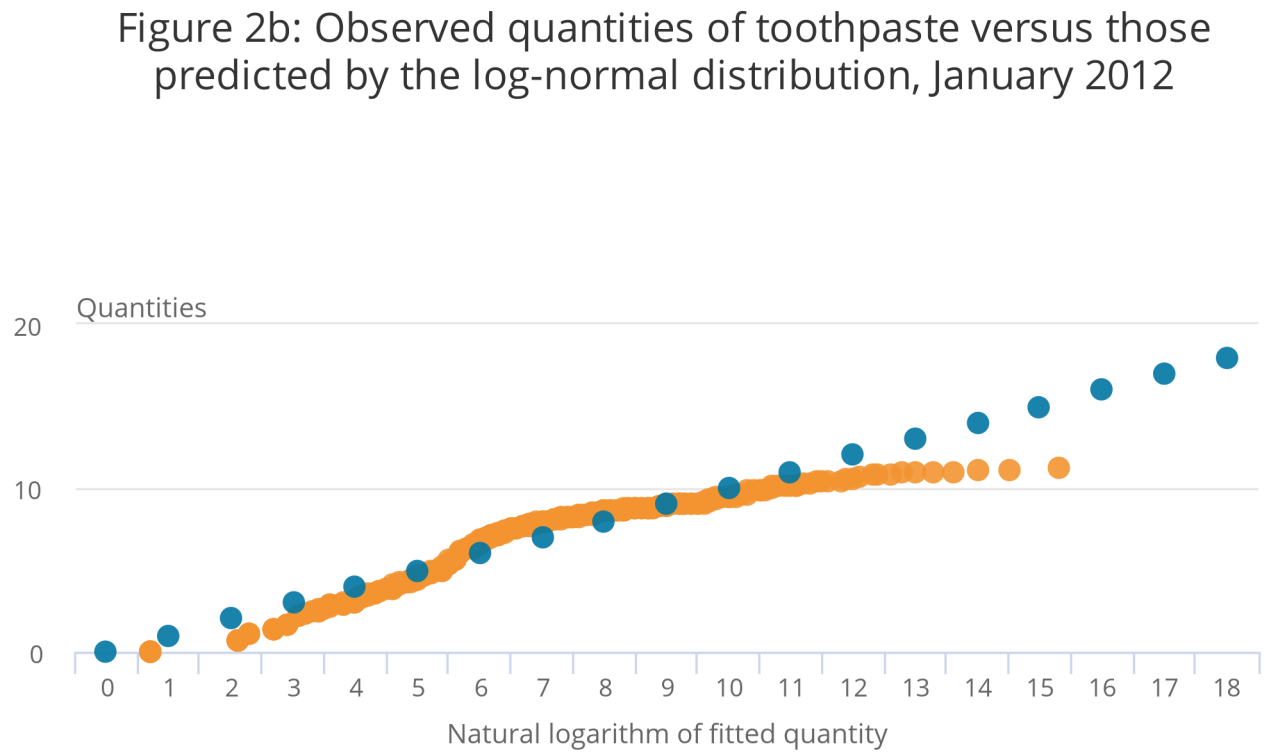
We conducted in-sample exploratory analysis of a single month, January 2012, by comparing observed and fitted log-quantities of toothpaste sales (Figure 2), which demonstrates that most of the improvement in goodness-of-fit that arises by truncating the log-normal distribution is a result of a reduced tendency to over-predict sales at the top end of the quantity distribution. However, truncation does not appear to remedy inaccuracies in the predictions in the middle and bottom end of the quantity distribution. The Pareto distribution does not appear to be suitable for predicting sales quantities of toothpaste.

Figure 2a: Observed quantities of toothpaste versus those predicted by the Pareto distribution, January 2012



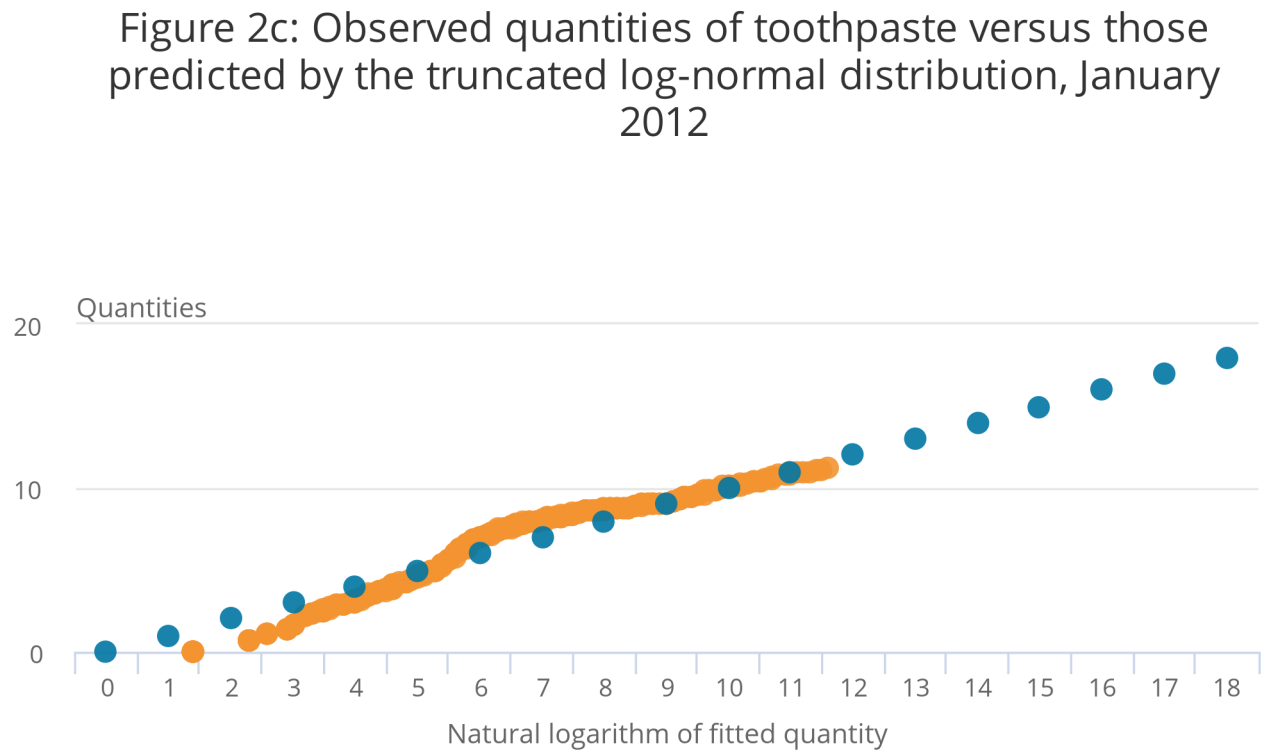
Source: Office for National Statistics – Retailer data (anonymised)

Figure 2b: Observed quantities of toothpaste versus those predicted by the log-normal distribution, January 2012



Source: Office for National Statistics – Retailer data (anonymised)

Figure 2c: Observed quantities of toothpaste versus those predicted by the truncated log-normal distribution, January 2012



Source: Office for National Statistics – Retailer data (anonymised)

Out-of-sample performance

The preceding exploratory analysis suggests that of the three considered distributions, the truncated log-normal distribution provided the best fit to the observed quantity data; Table 2 shows that in terms of out-of-sample predictions, it achieved holdout-set mean absolute percentage errors (MAPEs) ranging from 19.5% to 44.1%. However, truncating the log-normal distribution did not appear to result in a systematic improvement in predictive accuracy over the untruncated variant, with differences in holdout-set MAPE ranging from negative 1.6 to positive 2.5 percentage points.

Table 2: Mean absolute percentage error (MAPE) of toothpaste quantity predictions from various statistical distributions, 2012 to 2013

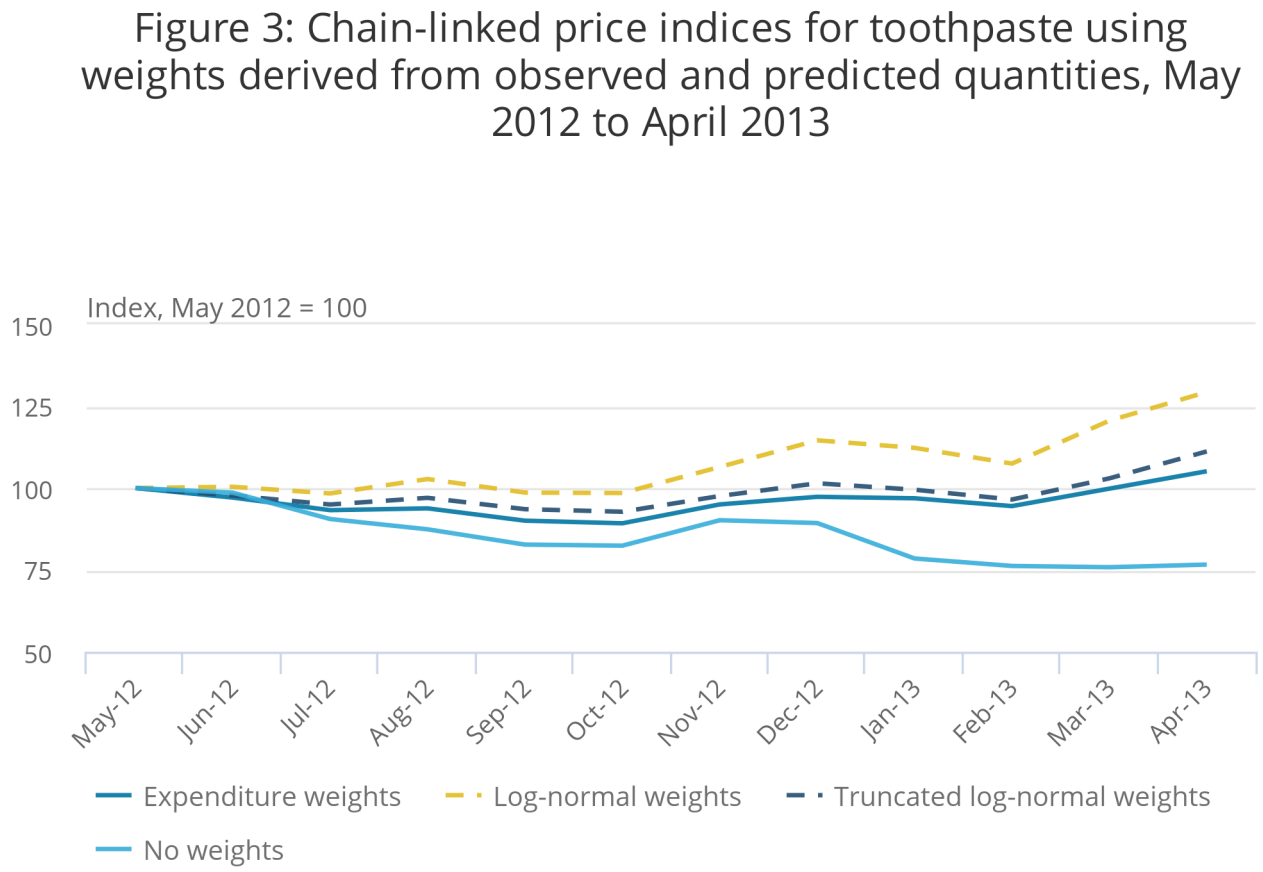
Month	Log-normal	Truncated log-normal	Pareto
May	18.9	19.5	52.9
June	26.1	28.6	49.4
July	19.9	20.1	54.1
August	20.8	21.0	51.0
September	22.3	20.7	48.1
October	31.4	30.6	45.6
November	34.0	34.1	42.4
December	36.6	35.7	44.3
January	23.9	25.6	46.3
February	32.1	31.5	47.4
March	31.2	32.8	45.4
April	44.4	44.1	39.0

Source: Office for National Statistics – Retailer data (anonymised)

Price indices

Despite the apparent similarity in out-of-sample predictive performance between the log-normal and truncated log-normal distributions, using expenditure weights derived from the truncated version resulted in an index series that was notably closer to the benchmark expenditure-weighted geometric Laspeyres index (Figure 3). However, both distributions lead to index series that are consistently above the benchmark in terms of their levels, and the differences between the series increase through time, in part because of the chain-linked nature of the indices. There is clear evidence of bias in the unweighted index series, which contrary to any of the weighted index series suggests a downward trend in prices over the period.

Figure 3: Chain-linked price indices for toothpaste using weights derived from observed and predicted quantities, May 2012 to April 2013



Source: Office for National Statistics – Retailer data (anonymised)

5 . Discussion

Conclusions

Web-scraped data can provide many benefits compared with more traditional methods of data collection. However, because of the nature of web-scraping, calculating price indices from web-scraped data may result in the inclusion of price movements for products that are infrequently or never purchased. The lack of weighting information at the product level when using web-scraped data can bias price indices, as discussed in New index number methods in consumer price statistics and seen in Figure 3.

This article presents initial analysis into an alternative method for calculating approximate expenditure weights for web-scraped price quotes. It demonstrates the potential for bias in an unweighted index series and the feasibility of fitting statistical distributions to predict sales quantities from their ranks to overcome this bias. Of the evaluated candidate statistical distributions, the log-normal distribution provided the closest approximation to observed quantities, with truncation reducing the extent of over-prediction among higher-ranked products. Using the truncated log-normal weights was shown to reduce the bias present from the lack of weighting information that we would likely experience when using web-scraped data without approximating product-level expenditures. The predictive performance of the Pareto distribution was notably inferior to that of the two log-normal variants.

Limitations

There are several limitations that need to be overcome before this work can be operationalised.

The choice of statistical distributions included in our analysis was driven not only by the characteristics of the data but also by the requirement for the parameter estimates of the distributions to be readily calculable by retailers (that is, estimated using straightforward summary statistics with closed-form solutions rather than numerical optimisation techniques). This approach means that the Office for National Statistics (ONS) could estimate product-level expenditure weights alongside web-scraped price quotes, while the retailer need not supply product-level microdata, which may involve practical hurdles. However, it still relies on retailers providing summary information, which may be infeasible in practice.

Other statistical distributions may provide greater predictive accuracy than the three we assessed but at the cost of increased complexity to parameterise them. A longer span of data than that available to us in this study may lead to reduced error associated with using year-old data to parameterise the distributions (for example, by using time series models to “nowcast” the upper truncation point of the truncated log-normal distribution).

Our out-of-sample validation paints a reasonably realistic picture of the prediction error that might be faced in a production setting, when timeliness constraints mean that retailers may only be able to supply monthly summary statistics relating to the previous year. However, it should be noted that we only analysed two years of historical data (divided equally into training and holdout sets) from a single retailer for one item. It is unlikely that our results can reliably be generalised to other time periods, retailers and items, and certainly further research would need to take place before the methods described in this paper could credibly be implemented in the Consumer Prices Index including owner occupiers’ housing costs (CPIH).

A further possible limitation, already discussed in Section 3: Data and methods, is the reliability of websites’ popularity rankings. If retailers use factors other than, or in addition to, the number of units sold to determine a product’s page ranking, this will limit the plausibility of using this measure as an approximation to the product’s quantity rank.

Next steps

As we continue our research into approximating expenditure weights for web-scraped price quotes, our next steps involve applying these methods to web-scraped data for which we have a direct comparison with a scanner dataset. This will allow us to see how well the results from this study generalise to actual page rankings from web-scraped data.

We have also modified our approach to web-scraping to allow for the collection of multiple observations for the same product from a website (for example, if the product appears in the categories “Women’s Dresses” and “New in for Summer”, then this will result in count of two, whereas we were previously removing these duplicates). This approach will allow us to test the findings of Chessa and Griffioen (2019).

We will also investigate how reliably distributions fitted to data provided by one retailer can be applied to the page rankings of another; again, this will require access to both web-scraped and scanner data for the same retailer so the assumptions can be appropriately tested.

We will continue to contribute to the international literature in this field and investigate other methods of approximating expenditure weights for web-scraped data when they have shown promising results for other national statistical institutes.

6 . Related links

Research and developments in the transformation of UK consumer price statistics: September 2020
Article | Released 1 September 2020
The first in a series of biannual articles to update users on our research to modernise the measurement of consumer price inflation in the UK.

Automated classification of web-scraped clothing data in consumer price statistics
Article | Released 1 September 2020
Research into using supervised machine learning algorithms to efficiently classify web-scraped clothing data, for use in consumer price statistics.

New index number methods in consumer price statistics
Article | Released 1 September 2020
Research into the use of new index number methods to calculate price indices using web-scraped and scanner data.