Initial report on experiences with scanner data in ONS

Derek Bird, Robert Breton, Chris Payne and Ainslie Restieaux, Prices Division, Office for National Statistics

1. Overview

This report provides an initial assessment of using scanner data to calculate consumer price statistics in the UK. It identifies the opportunities and challenges of using scanner data from both a practical and statistical perspective. It is limited in its scope, using sample scanner data from one retailer for two items: shampoo and toothpaste. A number of basic metrics and filters are developed, which are assessed for their impact on price indices. The effect of discounts and the definition of a homogenous product are also considered.

The data manipulation, analysis and index creation have been programmed using the statistical analysis software, SAS. This exercise was limited in scope and is not intended to fully replicate the use of scanner data in a ‘live’ consumer prices index. Nevertheless, the research underpinning this report reflects a big step forwards in the understanding of the use of scanner data in the UK.

In addition to investigating the use of scanner data, ONS is also using computerised methods (known as scraping) to extract prices from a number of retailers’ websites as part of a Big Data project. That work will be used to assess whether scraping is a feasible and suitable alternative, or complement, to using scanner data as inputs for the measurement of consumer price inflation. These two projects also form part of the evidence on alternative data sources being considered by the review of the range of consumer price statistics, being led by Paul Johnson.

2. Introduction

2.1 Scanner data and its use in consumer price statistics

Every time you buy a product from a supermarket, the bleep of the scanned barcode creates a record. The barcode has an International Article Number (originally called European Article Number, the abbreviation EAN has been retained), which is unique for that item. The record created contains the price, quantity sold and EAN code. These records make up a valuable database of prices and quantities sold for unique items – information that can be used by National Statistics Institutes (NSIs) for the creation of consumer price indices. These databases are referred to as point of sale scanner data.
There is a growing recognition that scanner data offers real opportunities to increase the quality of outputs, and possibly reduce costs, for NSIs. However, change is not straightforward. NSIs cannot simply apply existing statistical methodology and quality adjustment techniques\(^1\) to scanner data and expect the same outcomes. Without significant innovation and adaptation there are many differences that make the traditional approach impractical. The challenge for producers is to identify the extent of that difference and assess its impact and implications for users of the indices.

Scanner data may be exploited in a number of different ways, from supporting traditional collection methods and research, to replacing existing price collections in stores and/or the collection of expenditure data. In addition, the data may be used to calculate price indices that might provide a benchmark for comparison against published measures.

### 2.2 Compliance with HICP legislation

The Harmonised Index of Consumer Prices (HICP, known in the UK as the Consumer Prices Index, or CPI) is a consumer prices index that is harmonised across European Union countries. The rules underlying the construction of the HICP are specified in a series of European Regulations. In particular, the HICP measures inflation with reference to the changing costs of a fixed basket of goods and services, with annual weights based on the consumption pattern from the previous year. Jörgen Dalén’s paper, ‘The use of unit values in scanner data’ (2014), investigates the use of scanner data in terms of compliance with HICP regulations. It concludes that, even though HICP regulations were intended to relate to traditional price collection, they do not rule out the use of scanner data, and that regulations should evolve alongside new data sources and research.

Another area for consideration is the concept of price. Unit prices (the total expenditure divided by the total quantity sold) which have been calculated using scanner data are not a true transaction price (the actual price paid to the seller by the buyer); rather, they are an average transaction price. For that matter, shelf prices (the price at which the good is advertised, i.e. on the ‘shelf tag’) are not necessarily transaction prices either and, as Dalén (2014) points out, ‘prices collected by price collectors in the traditional way are normally shelf prices.’ Debate remains around whether usage of shelf prices or unit prices is more appropriate. The debate also links to comparability issues between NSIs that still use traditional price collection methods and those that (at least in part) use scanner data.

Dalén (2014) also advocates the use of EAN codes within a chain of stores to indentify unique items. However, in practice the process of identifying and defining a unique item can be more difficult. He points out that retailers can change a product’s EAN code over time, even if there is no change in the product. Confounding this issue is the potential for the EAN code to then be used again for a different product. Should this occur, and an NSI is using the EAN code as the sole identifier for a product over time, there is a clear risk of using incorrect price information for an item. The extent of EAN switching is difficult to identify independently; it would require both in-store collection and scanner data, so would be expensive. As a result, NSIs will want to work with retailers to understand the extent of this practice.

\(^1\) Quality adjustment is a set of techniques used in price index construction to ensure the index measures like for like price change.
3. Key Opportunities and Challenges of Using Scanner Data

Scanner data presents many opportunities for improving the measurement of consumer price indices. For retail outlets that use the technology it can potentially provide a census of transactions over a continuous time horizon, which results in many possibilities for using the data. The opportunities are discussed in the literature that is primarily being produced by NSIs that use scanner data. Although using scanner data presents many opportunities, there will also be challenges to consider and overcome in accessing, implementing and using scanner data.

3.1 Opportunities

Census of transactions: Scanner data can provide a record of virtually all transactions within certain outlets, reducing sampling error and enhancing the representativeness of the indices. The use of scanner data could, therefore, reduce the need for NSIs to conduct regular surveys of consumer expenditure behaviour and patterns.

Continuous-time data: Scanner data can also improve the data collection process by including prices over an entire month. This improvement in temporal coverage would allow the consumer price indices to reflect price changes that are, perhaps, not captured by current price collection practices. For example, in the UK, most prices are captured on and around an ‘Index Day’ in each month. This means that potential price volatility across other weeks in the month may not be captured.

Surveillance for new goods: Since scanner data can produce information on transactions in real time, it presents the opportunity for observing the arrival of new products onto the market, new outlets (if data are sourced centrally for a retail chain) and market trends. In the current production of consumer price indices, the ‘fixed basket of goods and services’ is updated annually, meaning that expenditure on new products is included at annual intervals. This is among the most frequently updated baskets in the world; however, the use of scanner data could facilitate an even more frequent update. This would have implications on the traditional Laspeyres-type index used to calculate consumer price statistics.

Simultaneous observation of price and quantity information: One of the main advantages of scanner data is the access to both price and quantity information simultaneously. In current practices, these components are generally collected separately, often by different departments within NSIs. There is also a time lag between when expenditure (or ‘weight’) information becomes available and the period to which it refers. Using scanner data could reduce the burden on survey respondents, and create cost and time efficiencies by significantly reducing or eliminating the need to conduct monthly ‘field’ based price collection. The use of scanner data also presents an opportunity to calculate ‘superlative’ indices, as current period expenditure weights are available. Superlative indices account for substitution behaviour, whereby consumers economise by buying more of goods that become relatively less expensive and vice versa. However, this measure is conceptually different to that authorised by current EU legislation. It also presents some challenging statistical considerations, which are outlined in the next section.

Product attributes and measurement of quality: Scanner data not only provide price and quantity information but can also provide attribute information, which is an essential requirement for monitoring changes to attributes of existing goods. This information may be used to make
quality adjustments as well as for product replacement. However, this is not without its problems, as attribute information can be missing, incomplete or change over time.

3.2 Non-statistical considerations

Dependence on Information Technology (IT) systems: Retailers develop systems for their own sale and marketing purposes, resulting in databases that can differ significantly from company to company. Statistik Austria highlighted this in a recent paper: ‘data cleaning and data processing requirements before actual index calculations are considerable’ (Boettcher 2014). In the same paper Boettcher also notes the difference in processing required for ‘raw’ scanner data as opposed to relatively well ordered data from market research companies; the former requiring significantly more processing time.

Therefore, developing an IT system that can handle data in different shapes, sizes and formats and put them together to make a comprehensive database suitable for analysis and the creation of indices is a major challenge. Furthermore, scanner data sets are very large, given the detail and coverage of information available. As a result, this can present difficulties in ensuring that the supporting IT infrastructure is able to process this size and frequency of data on a regular basis, within a specific time frame. There are also data storage considerations, and associated issues around data protection and confidentiality. In addition, there is a need to clean the scanner data; some of this can be automated, but there could be unforeseen problems arising on a regular basis, which could be resource intensive. The paper ‘Scanner data in the Swiss CPI - an alternative to price collection in the field’ (Müller 2010) details how the Swiss NSI has invested in a generic tool suitable for scanner data price collection from any retailer in Switzerland thanks to a common interface definition. The NSI also purchases item allocation data from a market research institute, which it believes is the most efficient and inexpensive way of allocating data to the internationally recognised Classification Of Individual Consumption According to Purpose (COICOP).

Dependence on data suppliers: Gaining access to scanner data is dependent on the cooperation of retailers. However, for the most part, retailers in the UK have been unwilling to cooperate. Even with cooperation, building up the necessary relationship to gain access can take a long time to develop. According to advice from the Central Bureau of Statistics (CBS) in the Netherlands, it can take anything from 6 months to several years to negotiate the sharing of scanner data. Data acquisition is a major risk for any scanner data project.

Furthermore, a decision to incorporate scanner data into monthly production processes results in a dependency on the retailer’s ability and willingness to provide the monthly deliveries on agreed dates. Should one or more retail chains suddenly be unable, or unwilling, to supply its data, it would have serious consequences for the production of the index. Therefore, consideration should be given to the appropriateness of introducing service level agreements, as well as a detailed contingency plan if one or more retailers are delayed in, or are unable to continue, supplying information. Even if data can be provided, there is a very real risk of changes to the format or composition of the data occurring at very short notice.

3.3 Statistical considerations

Data Continuity and EAN matching: Products continually disappear (sometimes only on a temporary basis) and new ones appear with new EAN codes. Work conducted by other NSIs demonstrates major challenges relating to matching EAN codes for the same product over time,
since a minor change (for example, the colour of packaging) can result in a new EAN code being used. These changes are termed re-launches by some NSIs. This creates time consuming problems when trying to identify and link the same products over time.

In addition, it is possible for identical store-specific EANs (for example, those applied to seasonal goods only sold by the store) being irregularly assigned to different products in different months. CBS addresses this problem by purchasing additional data from a market research company to aid the data mapping process, in particular to map EAN codes to COICOP classes (‘The use of supermarket scanner data in the Dutch CPI’, De Haan J, van der Grient H 2010 ). It is clear that any time saved by accessing scanner data, instead of using traditional price collection methods, can be offset by a significant increase in resource dedicated to data processing and cleaning.

Using scanner data with a fixed basket: The Swiss Federal Statistical Office currently uses scanner data to improve its price collection process. Due to the many methodological considerations associated with scanner data, it made no changes to its statistical methodology when scanner data was implemented. The scanner data are, therefore, aggregated at an elementary level using a geometric mean of price relatives. The higher aggregates are calculated using a fixed basket approach (or Lowe) and are reweighted on an annual basis using outputs from the Swiss Household Budget Survey. The Australian Bureau of Statistics follows a similar approach, although it is investigating the use of more sophisticated methodology.

The use of superlative indices: By introducing more frequent and timely price and expenditure information, scanner data presents the opportunity to reconsider the common approach of using a Laspeyres-type formula (with weights fixed in the base period). Instead, there are several statistical possibilities to compute a superlative index, updating quantities at the same time as price information is collected. There are three well known superlative indices which produce similar results to each other: these are the Fisher, Törnqvist and Walsh.

Superlative indices are the theoretical method for calculating cost of living price indices (although, in practice, NSIs are unable to produce timely superlative indices) - the International Labour Organization’s (ILO) CPI manual (ILO 2004) describes superlative price indices as the best choice and an ideal framework when both detailed price and quantity information are available. A superlative index would allow the basket of goods to follow market dynamics, capturing new goods that become available and enhancing the representativeness of the basket.

However, the issue is broader than this; whilst a cost of living index, which accounts for substitution behaviours, is desirable from an economic point of view, there are practical considerations that should also be taken into account. Laspeyres-type indices have been shown to act as an upper bound for index numbers, so switching to a superlative index, such as the Törnqvist or the Fisher, would mean that the level of inflation would be lower. This would have the implication that substitution behaviour would be necessary to keep track with inflation. Moreover, there remains debate around whether substitution behaviour is even an appropriate assumption to make.

Chain drift and item level superlative indices: Chain drift is a potential problem when using superlative indices at the item level. The paper, ‘The use of supermarket scanner data in the Dutch CPI’ (De Haan J, van der Grient H 2010), describes how a monthly chained matched-item index at the elementary level was initially deemed the most desirable approach to maximise the use of scanner data. However, although using a superlative index seems the most obvious elementary
aggregate to apply to the price indices, the paper details how superlative indices can suffer from chain drift. Chain drift occurs if a chained index ‘does not return to unity when prices in the current period return to their levels in the base period’ (ILO 2004). A stylised example of this is shown in figure 1. Note that this is not necessarily representative of what happens in practice. Prices and quantities are randomly generated, but prices in months one and seven are constrained to return to the base price. A Fisher price index is then calculated. The index demonstrates how, over time, the chained monthly index drifts further away from unity at the constraints.

Chain drift can arise because prices and quantities ‘bounce’ as a result of promotional sales. Households sometimes ‘stock up’ during sales periods and then consume from their inventories when the item is not on sale. In extreme situations, as explained in the paper, the quantities sold on sale can rise a hundred-fold on quantities sold when the item is not on sale. This is particularly true of non-perishable goods. For example, toilet paper might be an item that consumers might stock up on when the good is on sale.

**Figure 1: Example of chain drift in a Fisher price index**

![Graph showing chain drift in a Fisher price index]

**Solutions to chain drift:** Due to the problem of chain drift associated with applying superlative index formulae, CBS Netherlands (De Haan J, van der Grient H 2009) chose to apply an unweighted geometric mean, or Jevons formula, at the elementary aggregate level. The paper then describes how the expenditure information is used in applying cut-off sampling to overcome the problem with the Jevons index not using product weights. This is especially important since, in CBS Netherlands’ experience, expenditure across items within a product category is highly skewed – for example, c.40-50% of items (volume) counts for less than 10% of expenditure. The cut-off is applied so that approximately 50% of the items (volume) are selected, representing 80-85% of expenditure.

In 2009, a paper was published (Diewert WE, Fox KJ, Ivancic L 2009) which detailed an approach to aggregating scanner data using current weights, without incurring the problem of chain drift. The proposed approach is to use a Rolling Year Gini (1931), Eltető and Köves (1964), and Szulc
(1964) method (RYGEKS). This method was originally developed to make cross county comparisons but has been repurposed for use with scanner data. This is calculated as the geometric mean of the ratios of all bilateral Fisher indexes, where each entity is taken in turn as the base. This method makes use of fixed base comparisons but uses each month in turn as the fixed base and then averages the resulting comparisons.

Furthermore, the paper ‘Eliminating chain drift in price indexes based on scanner data’ (De Haan J, van der Grient H 2009), and more recently, ‘Scanner data price indexes: the “Dutch Method” versus rolling year GEKS’ (De Haan J, van der Grient H 2011), details CBS Netherlands’ application of the GEKS method and compares the results to the ‘Dutch method’ (described in the paragraph before last) currently adopted in their monthly CPI calculation. An assessment of the two methods found that, theoretically, the RYGEKS is an improved method since it makes use of all available price and weight information without suffering the effects of chain drift. At the all items level of aggregation, it was found that both methods produced similar indices; however, the differences were more marked at lower levels of aggregation. Application of the RYGEKS method might also result in increased volatility of the time series, although this is most likely going to reflect real world market dynamics. Although the clear benefits of trialling the RYGEKS methods are acknowledged by CBS Netherlands, they are keen to monitor the trial application over a longer time period, as well as monitor the development in international best practice and guidance.

**Sensitivity to time aggregation:** The paper, ‘Scanner data, time aggregation and the construction of price indexes’ (Lorraine Ivancic, Kevin J. Fox and W. Erwin Diewert 2009), also assesses the impact of aggregating weekly, monthly and quarterly estimates of price change. The tentative conclusion is that unit values averaged over months and quarters are preferred to those averaged over a week, due to the large volatility seen in the weekly aggregates. Dalén’s paper, ‘The use of unit values in scanner data’ (Dalén 2014), concludes that ‘a unit value over a period of up to a month is a reasonable approach’ and that this should be based on data from ‘the same chain of stores during the first three full weeks of a calendar month,’ since timeliness considerations make it impractical to use data from the full month.

4. Sample scanner data

4.1 Sample Scanner Data

The data used in this analysis is from a single retailer for two products: toothpaste and shampoo. The data are aggregated by month and cover a two year period from May 2011 to April 2013. There are several key variables in the data as detailed in table 1.

**Table 1: Data variables**

<table>
<thead>
<tr>
<th>ITEM CODE</th>
<th>ITEM DESCRIPTION</th>
<th>EAN CODE</th>
<th>CATEGORY</th>
<th>YEAR</th>
<th>MONTH</th>
<th>UNITS ON DEAL</th>
<th>UNITS OFF DEAL</th>
<th>SALES ON DEAL</th>
<th>SALES OFF DEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234567</td>
<td>Brand X shampoo 225ml</td>
<td>123456789</td>
<td>Shampoo</td>
<td>2012</td>
<td>APR</td>
<td>200</td>
<td>200</td>
<td>£400</td>
<td>£400</td>
</tr>
</tbody>
</table>
The first four variables are identification variables; the next two capture the time dimension and the final four capture units and sales. Unit and sales are split by those sold at full price and those sold 'on deal'. On deal sales includes a range of discounting methods that are not necessarily mutually exclusive, including:

- price reduction
- Y for £X
- multi-buy
- extra free
- gift with purchase
- coupon
- meal deals
- other promotions.

4.2 Definition of price

Price is defined to be the item (according to its EAN code) unit value; which is the item sales revenues divided by the quantity of the item sold in the corresponding month. For example, if sales revenues for Aquafresh Complete Care 100ml toothpaste are £10 and 10 tubes are sold in November, then the 'price' is £1.00.

5. Analysis

This section provides analysis of the data and resulting prices indices. Several validation metrics have been applied, such as the Tukey algorithm (which identifies and invalidates price movements which differ significantly from the norm for a particular item) and the price ratio test, to gain a better understanding of the data. Three forms of indices are presented:

1) A Jevons unweighted index (a Jevons formula is currently used to calculate elementary aggregates in the UK's CPI).

2) A Laspeyres base-weighted index (higher level CPI aggregates are Laspeyres-type indices, where the weights reflect expenditure from an earlier period).

3) A Törnqvist index, which is a superlative index.

The sensitivity of these indices to various data restrictions is tested by applying cleaning filters. The results show that scanner data indices that use the Jevons formula are particularly sensitive to the basic restrictions applied, but the Törnqvist index is much less sensitive to these restrictions.

The indices presented are based on items that are continually available across the year (based on EAN codes). EAN codes which either appeared or disappeared are removed from the scanner data sample. No attempt has been made to match products or to impute for missing products. Unless otherwise specified, on- and off-deal data have been combined. As January is the base month for price comparison used in the UK consumer price statistics, the analysis is presented from January 2012 onwards and chained in February 2013, giving 16 months of data. Although the
results provide insight into scanner data and the construction of indices, they are not indicative of the actual inflation rate for these items, especially considering that this is a single retailer.

Index results using only discounted or full price items are also considered using the on-deal off-deal split available in the data supplied. Finally, the use of EAN codes to define homogenous items is also considered.

5.1 Comparison with CPI item indices
Initially, un-cleaned summary scanner data, as received from the retailer, were used to create a price index with no filters applied. Monthly unit values (with on-deal and off-deal sales combined) were calculated for all products based on matching EAN codes. The resulting price indices for shampoo and toothpaste are presented in figure 2 below, alongside the item indices used in the CPI which have been chained together at the item level for comparison.

The scanner data indices are more volatile and there are also distinct differences in the growth rates compared to the CPI item indices. The results are not expected to match for a number of reasons. Most importantly, the scanner data represents prices from one retailer, versus the CPI price collection that covers a wide range of retailers. The two data sets are formed in very different ways. CPI data are collected locally and validated on entry (for example by using a price ratio test, for example). Further data checks are applied, including use of the Tukey algorithm. The validation process is laid out explicitly in table 2. This process will identify extreme price movements and allows decisions to be reached on their inclusion in the index. None of these checks have been applied to the scanner data (and possibly more importantly, they cannot be applied to individual prices as the data are aggregated over a month).

5.2 Descriptive metrics

Continuity
This metric assesses the percentage of items (as defined by EAN code) that exist across all months in 2012. There is a high appearance and disappearance of products (also known as product churn) in the sample data, as shown in figure 3 below. Around 60% of the products are in

---

2 Indices in the CPI production system are chained at class level and above. For the purposes of this report, in year indices at the item level have been chained for comparison
the sample for 12 months for both shampoo and toothpaste, and the remaining 40% of products either go out of stock or are new items. This proportion is not dissimilar to the proportion of products which were replaced with comparable or non-comparable items over the same year in the CPI (21.88% for shampoo and 44.8% for toothpaste). The high churn rate is an issue in scanner data because, whereas in the CPI we are dealing with approximately 150 price quotes a month for a particular elementary aggregate, in scanner data we are potentially dealing with thousands. This makes it far more difficult to deal with finding replacements and similar issues.

Table 2: Current checks used in CPI

<table>
<thead>
<tr>
<th>Stage</th>
<th>Current CPI Checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIELD collection</td>
<td>Price changes of plus or minus 40% on the previous (or last available) month prompt a warning message.</td>
</tr>
<tr>
<td>FIELD collection</td>
<td>Prices outside of minimum – maximum range prompt a warning message.</td>
</tr>
<tr>
<td>ONS Phase 1</td>
<td>FIELD checks are rerun and failures flagged for investigation.</td>
</tr>
<tr>
<td>ONS Phase 2</td>
<td>The Tukey algorithm is used to identify extreme price movements.</td>
</tr>
<tr>
<td>ONS Phase 3</td>
<td>Price quotations where the collector has entered additional uncategorised information are investigated.</td>
</tr>
<tr>
<td>ONS Phase 3</td>
<td>Base price imputation procedures are implemented when a product has been missing or invalid for three consecutive months.</td>
</tr>
<tr>
<td>ONS Phase 3</td>
<td>Individual price quote indices that are greater than 180 or less than 60 are scrutinised.</td>
</tr>
<tr>
<td>ONS Phase 3</td>
<td>The Tukey algorithm is used to identify extreme price movements.</td>
</tr>
<tr>
<td>FIELD audit</td>
<td>Price collectors are accompanied on collection.</td>
</tr>
<tr>
<td>FIELD audit</td>
<td>Collected prices are back checked.</td>
</tr>
</tbody>
</table>

In the current analysis, if an item goes out of stock or its EAN code changes mid-year then it is excluded from the entire index. In the CPI collection, replacement products are identified and missing products are imputed for. Finally, scanner data captures all discounts. Therefore, sales for discounts such as multi-buy, which are not accounted for in the CPI price collection, are captured here.

According to a recent Eurostat report³, around 30% of products churn over a 12 month period in scanner datasets. This property partly reflects the dynamism in the market place, specifically the births, deaths and re-launches of products. The remaining diagnostics are applied to products that existed across 2012 (based on matching EAN codes) in the sample scanner data.

**Extreme changes in prices through time**

Methods such as the price ratio test and the Tukey algorithm can be used to identify extreme price changes and remove these from the data. Under the Tukey algorithm, price changes which lie outside of a specified range are flagged up for investigation. The mean of the price relatives is used to separate the data into two halves, upper and lower. Any price changes greater than 2.5

---

³ Eurostat (Network of experts) September 2013 – Scanner Data Workshop
times the mean of the upper half of the data, and less than 2.5 times the mean of the lower half of the data, are then flagged. Note that price relatives of 1 are excluded from the calculations.

As shown in figure 3, around 22% of toothpaste unit value prices and 16% of shampoo unit value prices fail the Tukey test. Two price ratio filters have also been tested, removing prices that move by more than +/- 33% and +/-40%, compared to their previous value. These filters are based on the threshold set for the price change check in the local collection. The expected failure rate for the local collection is between 1 and 2%. Figure 3 shows that the number of failures for the +/- 33% filter are much higher than this, at 8% for shampoo and 9% for toothpaste.

The inclusion of items coming to the end of their lifecycle could be an explanation for the higher rate of failure in the scanner data. The CPI is compiled in accordance with the European Regulation that states that ‘only goods and services which are expected or likely to be available again at standard prices should be included’. This presents a rationale for price collectors to identify items that may be coming to the end of their lifecycle (a closing shop sale, or items being discontinued, for example) and exclude these from their collection. Additionally, scanner data unit values capture multi-buy deals, which could increase the volatility of unit prices.

**Products coming to the end of their lifecycle**

Products coming to the end of their lifecycle typically have low sales revenues, which indicate they are significantly discounted and are sold in low numbers due to limited stock. Thus, a simple test is to exclude sales below a certain threshold. In the descriptive statistics low sales are defined to be below £1000\(^4\) although, clearly, this figure would vary according to the size of the retailer and on the product in question. Excluding sales below this threshold results in 25% of toothpaste products and 17% of shampoo products being removed from the data.

**Minimum market share**

Collectors focus on items that are typical of what consumers buy. A way to proxy for this in scanner data is to produce a metric that assesses market share. With a single retailer it is only assessing the relative sales by that retailer, rather than market share, so it is only an approximation. The metric assesses those items which make up more than 0.1% of total sales revenues. Around 55% of toothpaste and 66% of shampoo prices would pass this metric.

---

\(^4\) The critical values used here are estimates: quantity or a combination of price change and quantity could be used; the acquisition of more data may lead to firmer conclusions as to which metrics are the most important.
### Figure 3: Descriptive metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Shampoo test failures</th>
<th>Toothpaste test failures</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item discontinuity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Price quotes for all months in 2012)</td>
<td>Fail 41%</td>
<td>Fail 40%</td>
</tr>
<tr>
<td></td>
<td>Pass 59%</td>
<td>Pass 60%</td>
</tr>
<tr>
<td><strong>Tukey test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Price changes &lt; 2.5 x upper midmean and &gt; 2.5 x lower midmean)</td>
<td>Fail 16%</td>
<td>Fail 20%</td>
</tr>
<tr>
<td></td>
<td>Pass 84%</td>
<td>Pass 80%</td>
</tr>
<tr>
<td><strong>Price ratio test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Price changes &lt; 33% and &gt; -33%)</td>
<td>Fail 8%</td>
<td>Fail 9%</td>
</tr>
<tr>
<td></td>
<td>Pass 92%</td>
<td>Pass 91%</td>
</tr>
<tr>
<td><strong>Low sales revenue test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Item sales &gt; £1,000)</td>
<td>Fail 17%</td>
<td>Fail 25%</td>
</tr>
<tr>
<td></td>
<td>Pass 83%</td>
<td>Pass 75%</td>
</tr>
<tr>
<td><strong>Minimum market share test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Item sales &gt;0.1% of total sales revenue)</td>
<td>Fail 45%</td>
<td>Fail 34%</td>
</tr>
<tr>
<td></td>
<td>Pass 55%</td>
<td>Pass 66%</td>
</tr>
</tbody>
</table>
5.3 Scenario testing: applying cleaning filters

Price ratio filter

The price ratio restrictions applied on the Jevons index for shampoo have the biggest impact from September 2012 onwards. Both the +/-33% filter and +/-40% filter increase the level of the index. Conversely, when the restrictions are applied on the Laspeyres index, the filters decrease the level of the index throughout the series. The price ratio restrictions do not appear to have a significant impact on the Törnqvist price indices for shampoo. The impacts are even more pronounced on toothpaste (as shown in figure 4). It is likely that the impact of the filter on the Jevons index is upwards due to the exclusion of heavily discounted items, which have a low weight in the Törnqvist index.

Figure 4: Price ratio filter
Low sales revenue filter

Excluding items with low sales revenues (including only sales greater than £500 or greater than £1000) leads to an upward shift in the Jevons indices for both shampoo and toothpaste. It has less of an effect on the superlative indices. Again, the upward movement in the Jevons indices is probably due to the exclusion of heavily discounted items at the end of their product life cycle. There is also an upward shift in the Laspeyres and Törnqvist indices for toothpaste, although this effect is negligible for shampoo.

Figure 5: Low sales revenue filter
Minimum market share filter

Removing items with a low market share (as defined by less than 0.1% and 0.05% of sales) shifts the Jevons index upwards for both shampoo and toothpaste, similar to the results seen for the low revenue sales filter. This restriction has a negligible effect on the shampoo Laspeyres and Törnqvist indices, although it does impact on the toothpaste index.

Figure 6: Minimum market share filter
5.4 Indices for discounted and non-discounted products
The UK’s CPI data collection captures single product discounts, but excludes multi-buy deals. In addition to all transactions, scanner data includes most forms of discounts. The data provided contain an ‘on-deal’ flag (see the data section for a list of discounts included). However, no distinction can be made between the different forms of discounts, we just have the overall weight and quantity sold. Most shampoo items are discounted regularly in the sample (although this does not mean they are discounted continually). Over the course of each month most (98%) shampoos will have had some sort of discount applied (even if it accounts for only a small amount of items sold). Sales of discounted and full price shampoos generate similar levels of revenue in the sample.

Figure 7: Discount and non-discount
Indices using only discounted and full price (or ‘non-discounted’) items are shown above in figure 7. Products are allowed to move between indices; that is, a product which is on sale in a particular month and full price the next will be a component of the discounted index in the first month, and a component of the full price index in the next. The discounted indices for both shampoo and toothpaste are more volatile compared to the full price equivalents. This is expected due to the different types of discounts which can be offered over time.

5.5 Using EAN codes to define homogenous items

Using EAN codes to define a homogenous item is an assumption often applied to scanner data and has been used in the above analysis. However, this might not always accurately replicate decisions made in the field and this could have implications for the use of scanner data. It is possible that more than one EAN code could be used to define a homogenous item, or a single EAN code may be used to define different items. The examples below provide unit values by brand or sub-brand for items which might be considered ‘comparable’ by collectors in the field and where a continuous price chain is maintained.

Figure 8: Brand pricing

In most instances the unit prices for the shampoo brands follow each other almost identically, except in instances where a specific products appears to be discontinued and its unit price drops sharply. The toothpaste items are less consistent; for example, one brand of whitening toothpaste is consistently higher than the unit price for the other two products of that brand. The change from a non-whitening to a whitening toothpaste in the field is likely to be considered a non-comparable replacement and so deemed to be a quality change.
6. Conclusion

Scanner data expands the options for developing our understanding of inflationary pressures in the consumer sector, but is not a panacea. There are competing costs and benefits, risks, resource constraints and access issues, all of which have to be assessed to determine a way forward. Additionally, competitors to scanner data, such as web scraping, may be equally as attractive and so any strategy for use of ‘big data’ needs to reflect a range of alternative data sources.

This paper has demonstrated that the approach taken to index construction, even in two products that are reasonably homogenous, makes a real difference. Consideration of discounts, product life cycles, data matching, chaining and aggregation have all demonstrated that, even when a single methodological approach is taken, different paths for indices can be generated. We have not used this paper to look at the full range of methodological issues that will have to be addressed if a decision to use scanner data were planned, but note that the work is central to progress. Recent Eurostat research (Dalén 2014) concludes that there is nothing in the HICP regulations which ‘prevents the use of scanner data in general’ and that they should ‘evolve with time and especially with new types of data and research emerging.’ It is also important to consider the impact of a change in data sources on the outputs, and that must be undertaken with the users of the statistics, not in isolation from them.

This paper has also looked at a number of aspects of the limited range of data available to ONS using some elementary filters, which all produce similar results. There has, however, been no assessment of which filter(s) might be most appropriate. Similarly, no attempt has been made to match EAN codes over time to capture re-launches of products, or to impute for products which may be temporarily out-of-stock; aspects that would be important if scanner data were to be used in a production environment.

The acquisition of scanner data in the UK has been a difficult and slow process. There is reluctance on the part of the retail industry to share data. These data are necessary, however, to complete the research that will be required if progress is to be made. The experiences that have been developed with this limited data set have provided early insights into the challenges of handling multiple data sources with different structures. Knowledge around the ways the data might be stored and processed has increased, albeit to a limited extent. These steps, while small, are important in the development process. A much wider research programme is crucial to fully understand the benefits and limitations of using scanner data in price index production.

With drivers from the EU and domestic priorities around reducing collection costs it seems prudent to invest more resource into the investigation and understanding of this topic. Therefore, we will continue to investigate the potential for the use of scanner data and will use the opportunity afforded by EU task force meetings and other similar forums to report on progress.
7. References


