UK Scanner Data Project – Update on Obtaining Scanner Data

Prices Division, Office for National Statistics, United Kingdom

Abstract

In 2011 the UK was awarded a Eurostat grant to undertake research on the exploitation of scanner data. The intention was to conclude the research project with the production of an end-to-end guide to the process of obtaining, implementing and using scanner data for the purpose of producing multipurpose consumer price statistics. The first of three intended volumes 'Stage 1 - Obtaining scanner data' has been written and delivered to Eurostat. However, despite significant effort, ONS was unable to secure support from retailers in obtaining scanner data, even for research purposes. This meant that it was not possible for us to produce the latter two volumes.

This paper explores the challenges in obtaining scanner data experienced during the project and the key lesson learnt by ONS that a lack of suitable legislation and a current data collection method that places little burden on retailers makes data acquisition from retailers more difficult. In continuing its own scanner work, ONS has looked at ways in which it can build the right relationships with retailers. More recently, ONS has successfully obtained two years of sample scanner data from a retailer for two items. This paper provides an update on the UK's experience of using this scanner data and includes some initial indices

We would welcome any comments on this paper, in particular answers to the following questions in the context of what is presented in the paper:

Q1. What can we deduce from the initial indices created?

Q2. What deductions, if any, can we make from the sensitivity of the indices to low quantity sales?

Q3. Do you have any general comments around cleaning scanner data?

Introduction

ONS has continued to expand its knowledge and understanding in obtaining, using and implementing scanner data. Despite initial issues faced in obtaining data, ONS has achieved some success in obtaining sample data from a key retailer. ONS has gained insights into scanner data from its literature review, continued investigation into methods used by the early adopting NSIs, and its direct experience in data collection. ONS is also seeking to obtain more data from retailers in order to further its knowledge and use of scanner data. The three topics covered by this paper are:

1. Key lessons learnt, both directly and indirectly, on obtaining data from retailers
2. Important considerations when using scanner data including a brief discussion of the methods adopted by other NSIs
3. Initial analyses conducted on the sample data acquired

Key Lessons in Obtaining Scanner Data

Legal requirement

The first key finding of the ONS, through its research and direct experience, was that the lack of relevant legislation was one of the key barriers that ONS faced in attempting to acquire scanner data from retailers.

Many European countries (though not the UK) have legal powers that require retailers to provide price and expenditure information to them for consumer price statistics, although these legal powers don't often prescribe that information must be provided in electronic form or in the form of scanner data specifically. In some cases, National Statistical Institutes (NSIs) who have implemented scanner data have found that some retailers have welcomed the provision of scanner data instead of submitting traditional paper forms, since this has reduced the burden on them. From discussions with other NSIs, it was clear that whilst having the right legislative framework in place was helpful, it often wasn't enough, and there was a strong emphasis on building strong working relationships with retailers.

Approaching retailers and gaining cooperation

The second key finding was the need to compliment any legal requirement with strong relationship building. The relationship building is multi-dimensional covering issues around building trust and confidence, maintaining relationships and data sharing.

Before contacting the retailers, it is essential to ensure that the scanner project has the support of the NSI's senior management. This will help demonstrate to retailers the importance the NSI places on obtaining scanner data. The majority of NSIs who receive scanner data from large retailers agreed that it was essential to contact the management of retail chains, since they will often be the ones who make the decision as to whether to provide the data. Previous experience by NSIs has highlighted that although less senior people may have better knowledge of the data, they are not able to make the decision to supply data. It is also important to meet the management face-to-face initially, and to conduct research into the retailers' corporate strategy, mission statement and key relationship managers prior to the initial meeting. The key to gaining agreement for the access of scanner data is dependent on demonstrating to retailers that providing scanner data will benefit them, by reducing the respondent burden on them in the longer term. There will be a need to present a strong argument to convince retailers to supply scanner data, by emphasising that scanner data is the future, and that retailers who cooperate at the early stages of the process have an opportunity to influence the design of the process.

During initial discussions with retailers, it may be important that there are IT representatives present, for both the retailer and the NSI. This is to ensure that consideration is given to the technical aspect of supplying scanner data, and retailers will want to know how the NSI intends to receive and store the data. They will also most likely need reassurance about the security of their data. Consideration should be given to setting up a secure file transfer system before initial discussions take place. It may also be appropriate to establish a working group with the management of several retailers, to provide a forum to openly discuss any barriers or challenges to providing scanner data.

To overcome initial resistance from retailers in providing scanner data, it is essential that they can be given assurance that the data they provide will be secure, and that it won't be used for any other purposes other than the purpose intended, since price information for large retailers can be commercially sensitive. If one or two retailers have agreed to provide scanner data, this can provide an incentive for others to get involved.

It is important to approach retailers with a clear idea of what information is required, but there will be a need to be very flexible to encourage retailer cooperation, based on how each retailer currently stores their data and how easily scanner data can be extracted from their systems and transferred.

Some of the key aspects that an NSI might establish when meeting with a retailer are:

- how the scanner data are housed by each retail outlet;
- understanding the structure of the existing scanner dataset(s);
- identify and understand the coding and classification structures used and understanding the level of detail in the data;
- whether the retailers change the unique coding, and if so, how often;
- whether the retailers currently supply data to other organisations;
- whether there are confidentiality issues associated with the data;
- how discounts are recorded;
- the timeliness of the potential future supply of data.

It is worth noting that as a minimum, scanner data must have an EAN identifier, expenditure and quantity information, price, discounts, as well as a classification code assigned by the retailer, for the data to be useful.

Maintaining Relationships

Once contact with retailers has been established, it is important to ensure that relationships are maintained. Some NSIs have their own account manager for retailers supplying scanner data, since there is significant value in providing a consistent contact over a number of years, to build and maintain relationships with retailers. It was also preferable to establish a single point of contact for each retailer, and possibly a secondary contact, at a working level. This helps to ensure that any issues with the transfer or use of the scanner data can be resolved swiftly, and maintaining good working relationships is generally the most important factor in ensuring the process is as smooth as possible. For those NSIs who regularly receive scanner data in their production processes, it is very rare for them not to receive data at all, and there are generally few issues with the data.

Most NSIs who regularly receive scanner data have drawn up written contracts, to ensure that there is a clear expectation of what data is provided to the NSI (and when), and to minimise any potential problems that could arise when there is a change in management structure of the retailer, for example. During the early stages of retailer cooperation, it may be more appropriate to put in place umbrella agreements with the retailers, to allow them to start sharing information with the NSI on data structure and allows for more detailed data sharing in the future.

Data sharing and its benefits

The format of scanner data should be whatever is easiest for the retailer to provide, although several NSIs have found it helpful to show the retailers an example of the data structure required, for example by specifying a minimum list of variables, or a template of the required information. Scanner data is usually already used by retailers for other purposes, for example for their own analysis purposes for marketing. Some NSIs have found it useful to request a 'reference price'[1] in the scanner data, as well as the minimum variables, against which the derived average price (expenditure/quantities sold) can be compared. This helps with the analysis of price information, and can reduce the need to contact the retailer to query the data.

During the initial stages of obtaining scanner data, it may be worth obtaining and analysing scanner data for only a sample of item groups. This can reduce retailer burden and increase the likelihood of their long term cooperation in the project, as well as ensure that the NSI is able to consider the many statistical aspects associated with only a defined group of items initially. A number of NSIs have also communicated the value retailers may get from having the analysis and quality assurance of the data being sent back to them by the NSI. This means the retailers can benefit from additional price analysis of the products they sell, or highlight whether there are any irregularities with their data systems. This benefit may be another way to encourage retailer cooperation.

It is also worth considering, and being flexible about, the coverage of the scanner data received, depending on retailer constraints in providing data for all their stores or data which represents all weeks in a given month. For example, some NSIs receive scanner data for a sample of stores for a retailer chain; scanner data could represent up to three weeks of transactions in a given month, or be based on an aggregate of weekly transactions, or be based on transactions covering the mid-week of a month.

In most cases, different retailers within a country provide scanner data to their NSI in different formats, although all are versions of flat text files, which are the most efficient from a data storage perspective. Data transfer is usually straightforward if a secure transfer mechanism has been set up, which can be used to transfer data in both directions. Some NSIs are considering the use of a common web portal for the supply of scanner data, in conjunction with other government departments, with the advantage that businesses should only have to provide data once.

Dependence on data suppliers: risks and contingencies

The access to scanner data initially depends on the cooperation of retailers to provide the information. A decision to incorporate scanner data into monthly production processes therefore results in a dependence on the retailer's ability and willingness to provide the monthly data deliveries on agreed dates. Should one or more retail chains suddenly be unable or unwilling to supply data, this could have serious consequences for the production of multipurpose price statistics. Therefore consideration should be given to the appropriateness of introducing written contracts, as well as a detailed contingency plan if one or more retailers are delayed in supplying information or unable to continue to provide information. Even if data can be provided, there is a

[1] For example, an advertised price at a point in the reference period

very real risk of being exposed to changes to the format or composition of the scanner data at short notice.

Important considerations when using scanner data

Price bouncing

A major advantage in the use of scanner data is that it simultaneously captures price and expenditure information. This simultaneity makes possible the creation of price indices that account for the substitution of consumers away from goods that have had price increases. Hence, this opens up the possibility of implementing superlative index formulas. Superlative indices account for substitution effects to a second order approximation – and with scanner data that can be implemented at all levels of aggregation - with virtually no time lag.

However, the application of superlative index formulas to scanner data introduces new problems for measurement, namely chain drift. Chain drift is the result of the interaction between pricing strategy and consumer behaviour, more specifically the effect of sales and discounts. Essentially, consumers stock up on items when the price is low or discounted, and when the price returns to its normal value they purchase less than prior to the discount (due to their inventories). The lack of symmetry in terms of price and expenditure movements result in a lower weighting of the post sales period price, and hence a downward drift in the superlative indices.

Fixed base indices, as opposed to the chain superlative indices, do not suffer from this issue; however a fixed base index has problems regarding new products and product discontinuation; and this is a key issue (for example supermarkets introduce thousands of new products every year). Hence, fixed base methods have significant issues regarding product matching through time.

Sensitivity to time aggregation

Higher frequency data has clear advantages when considering fast moving prices; however this has to be balanced against the effects of time aggregation on price volatility. Previous studies have considered this issue, and note substantial differences between weekly in comparison to monthly and quarterly data. In general, higher frequency data are more sensitive to discount and sales effects, and thus resulting in greater chain drift effects when superlative indices are used.

Transactions can become more sporadic when using very short time aggregations. For example, if we were considering a particular brand of shampoo, over a sample of a few stores, taking daily data may result in a volatile series: with the potential for some days not to have any sales. This would create gaps in the series.

Item grouping and unit prices

Traditional index number theory states that the value of transactions must be preserved. The complexity occurs when we have multiple prices over the period, this could be a result of discounts or other pricing activity. Hence a solution is to calculate the unit price (where the unit price is the transaction value divided by quantity), but this raises questions as to how we group prices across multiple dimensions (place, items and time). Differing assumptions will impact the indices.

Methods (data, analysis and recommended index types) from other NSIs

GEKS

Price bouncing and resulting chain drift are issues that have been previously considered by the academic literature and other NSIs. A key solution is to combine the benefits of more frequent data, but use each month as a new base. This would then maximise product matches (accounting for new and discontinued products) and at the same time avoid the issues of price drift. The method utilised by the Statistics Norway is an adapted version of Gini (1931) Elteto and Koves (1964) and Szulc (1964), it is known as the GEKS method. It is adapted as a rolling year, RYGEKS, to avoid the continuous recalculation of previous periods (which is not considered desirable for price indices).

The 'Dutch Method'

Statistics Netherlands had an expansion of scanner data in January 2010, this accompanied their development of a new index construction method, at the lowest level of aggregation. They do not directly use the expenditure data to form the indices, but use it indirectly to select a sample of items (using cut off sampling). Monthly chained unweighted geometric (Jevons) price index numbers are calculated on the selected sample. The exclusion of expenditure data avoids the chain drift issue and the monthly chaining avoids the attrition rate issue.

Initial analysis

This section presents an overview of the initial analysis that has been conducted on the sample data that ONS has acquired. The analysis is experimental and the results presented below will be subject to revisions as ONS refines its methods over the coming months. Although the results provide insight into scanner data and the construction of indices, they are not necessarily indicative of the actual inflation rate for these items, especially considering that this is a single retailer.

The data

A leading retailer gave the ONS access to a sample dataset of two products: namely, shampoo and toothpaste. The data is aggregated over months; and contains several key variables such as transaction values, quantities, product descriptions and unique identifiers.

In constructing initial indices unit prices for non discounted sales have been used. The data has been restricted to include those items which span the full 24 month time period. In the coming months ONS will refine its methods to test the impact of including discounted sales and new or discontinued items. The indices are chained with January as the base periods.

Initial indices

Shampoo

Figures (1 – 3) below present initial indices for Shampoo using the scanner data, acquired by ONS. Figure 1 shows the results from using a Paasche, Lasperyes and Fisher index formulae. Figures 2 and 3 present the results of using elementary aggregate formulae Jevons, Dutot and Carli. Figure 2 excludes sales with a quantity less than 100 whereas figure 3 includes all sales. The results show the elementary aggregate indices are very sensitive to this and change from falling over time in figure 3

to rising in figure 2. In addition, the results in figure 2 are much more comparable to the indices presented in Figure 1 (Paasche, Laypeyres and Fisher). Finally, figure 4 presents, a chained version of the index for shampoo used in the CPI (the UK's HICP) - which has been more volatile and grown much more quickly over the two year period compared the indices calculated using the scanner data.

Toothpaste

Similarly, figures (5 – 7) below present initial indices for toothpaste using the scanner data. Figure 5 shows the results from using a Paasche, Lasperyes and Fisher index formulae. Figures 6 and 7 present the results of using elementary aggregate formulae Jevons, Dutot and Carli. Figure 6 excludes sales with a quantity less than 100 whereas figure 7 includes all sales. As with shampoo, the results show the elementary aggregate indices are very sensitive to low level sales items. Finally, figure 8 presents a chained version of the index for toothpaste used in the CPI (the UK's HICP) – it demonstrates a slightly more volatile series, which is falling over time, in comparison to the scanner data indices.

Figure 1: Shampoo, Scanner data, Indices - Paasche, Laspeyres and Fisher
Excluding low level sales

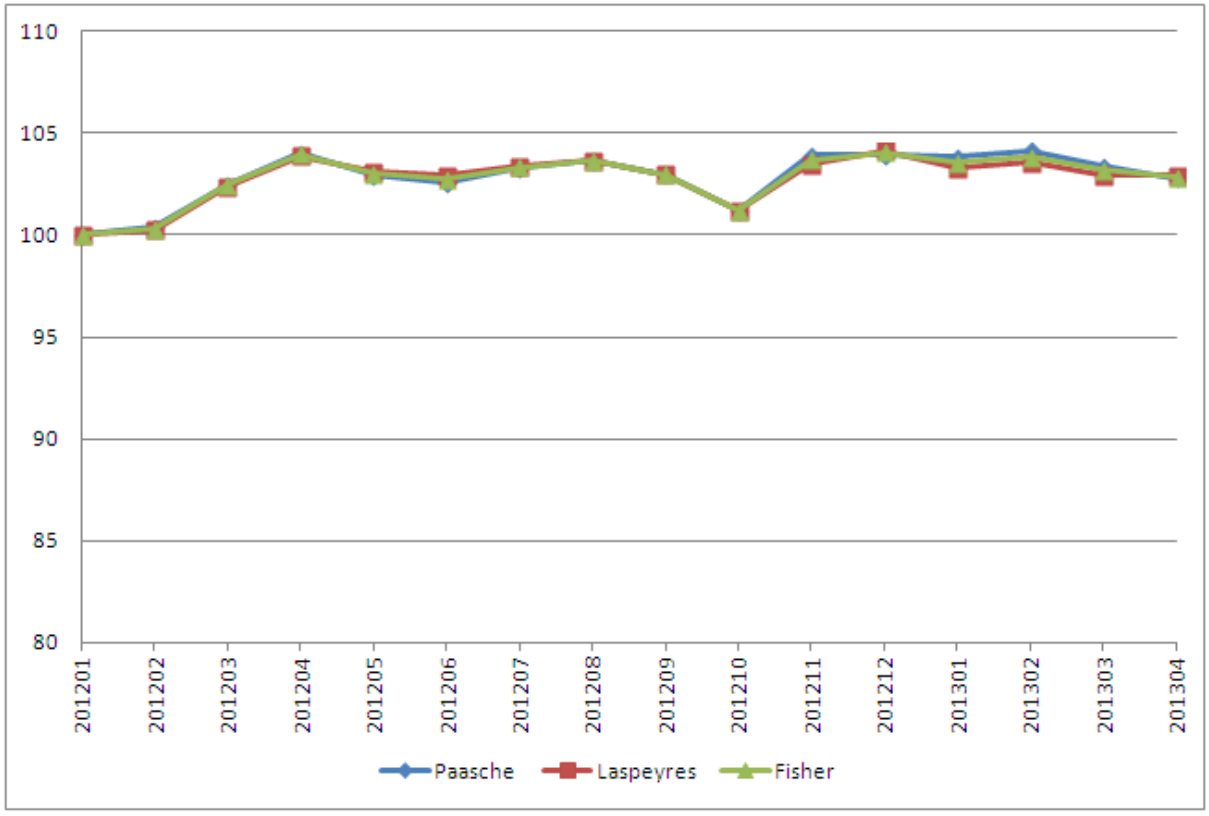


Figure 2: Shampoo, Scanner Data, Indices – Jevons, Dutot, Carli

Excluding low level sales

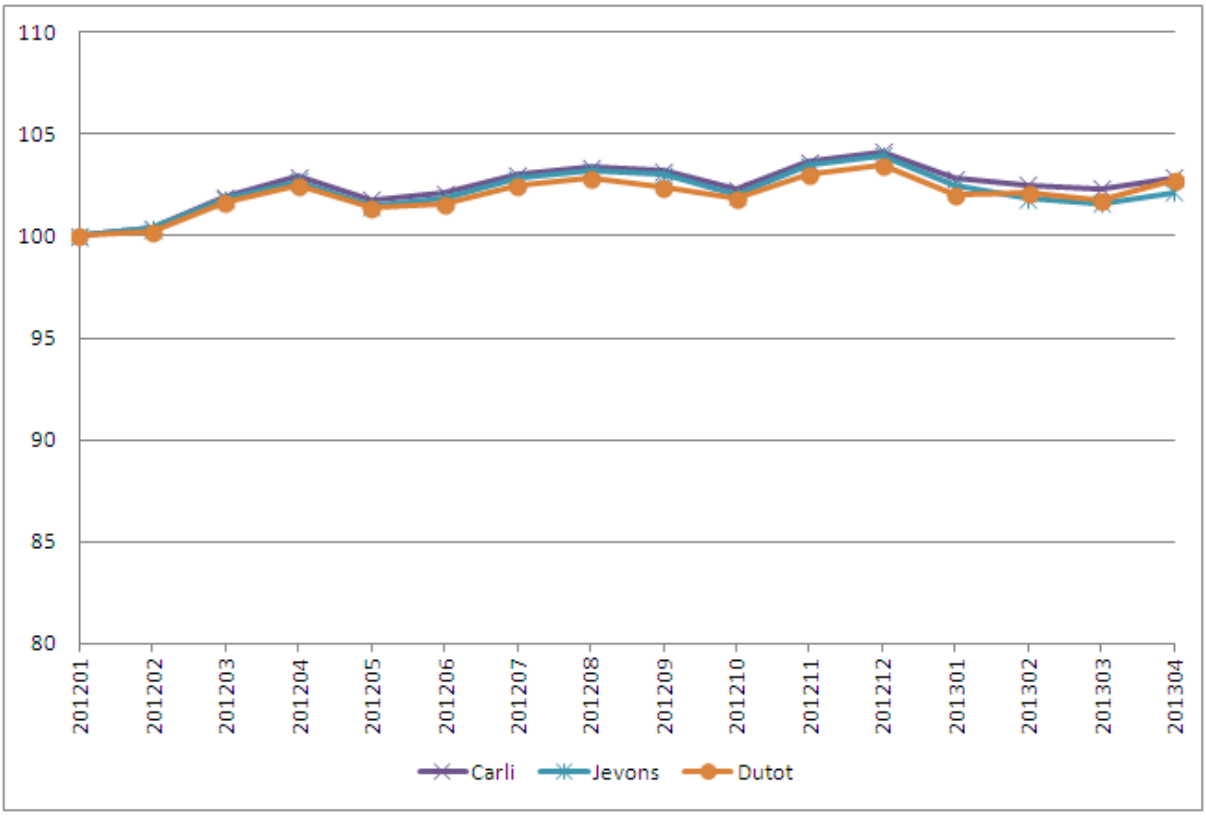


Figure 3: Shampoo, Scanner Data, Indices – Jevons, Dutot, Carli

All Sales included

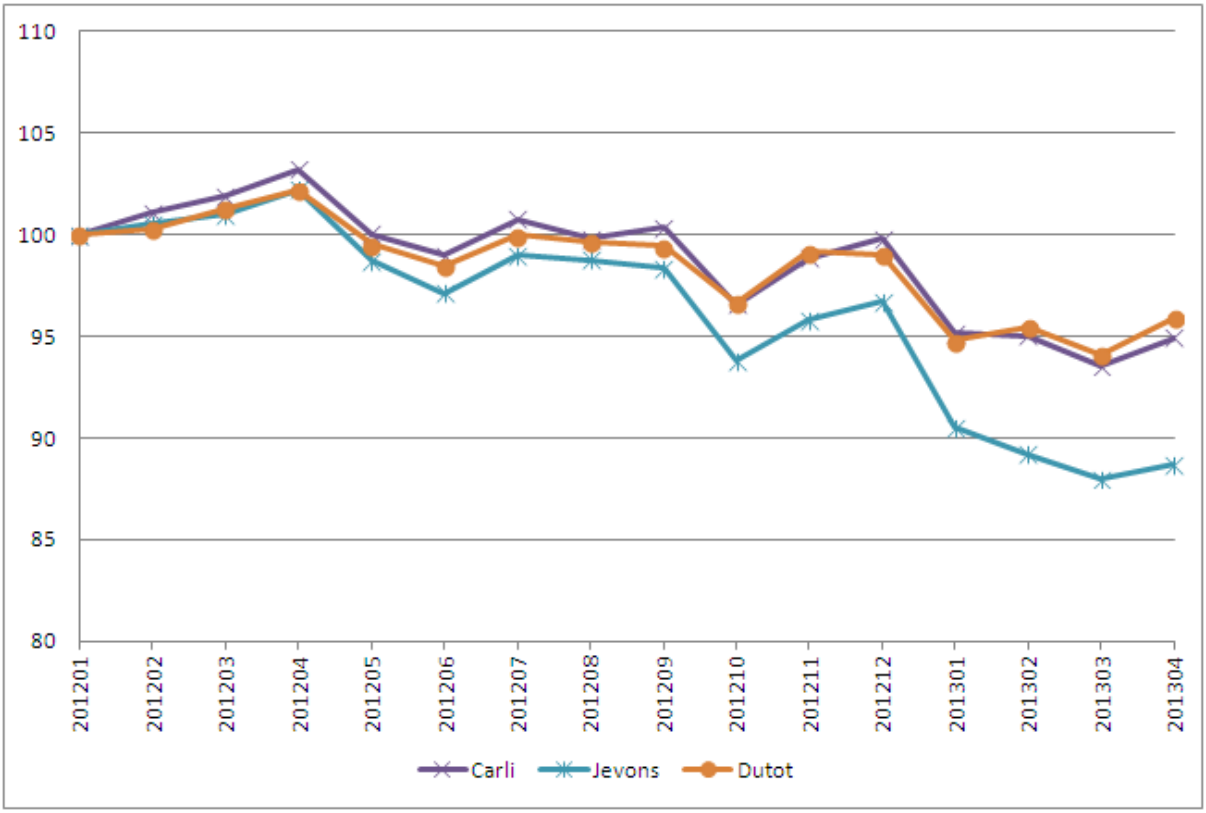


Figure 4: Shampoo, CPI, Index – Jevons

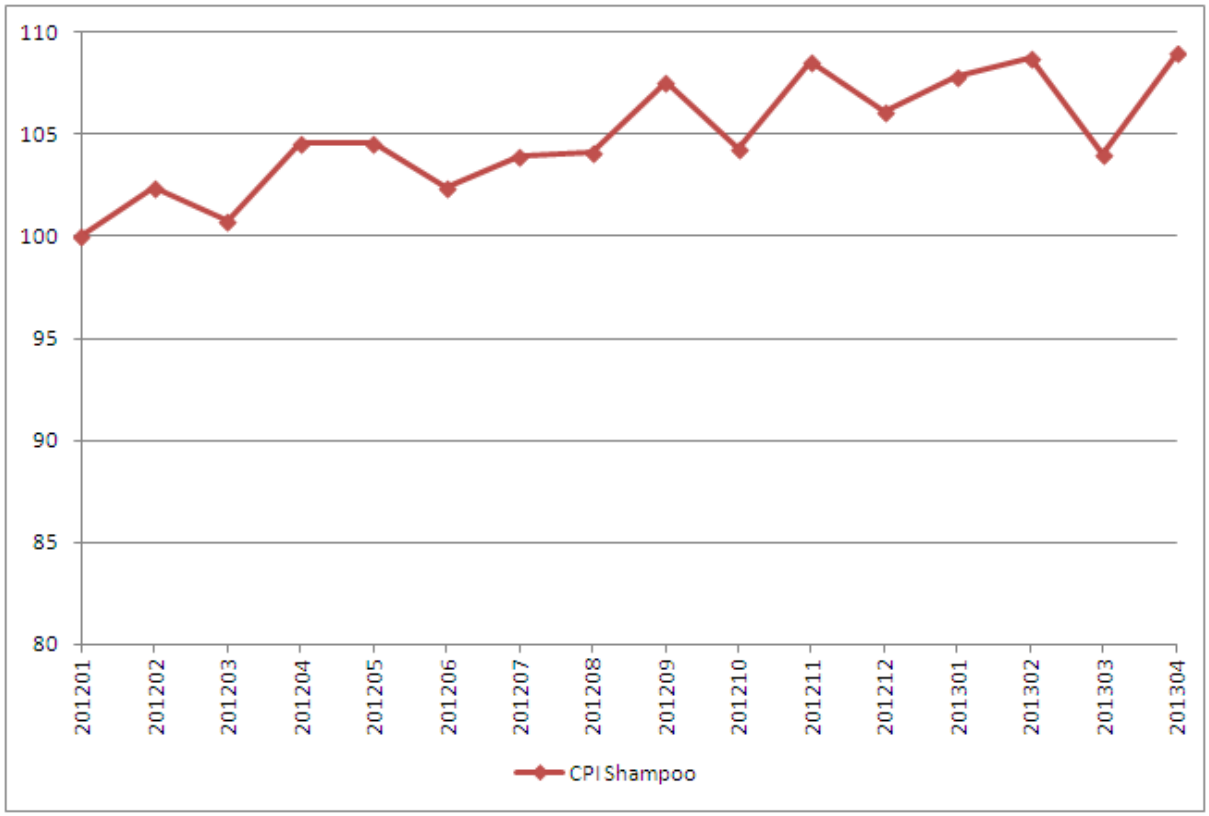


Figure 5: Toothpaste, Scanner data, Indices - Paasche, Laspeyres and Fisher

Excluding low level sales

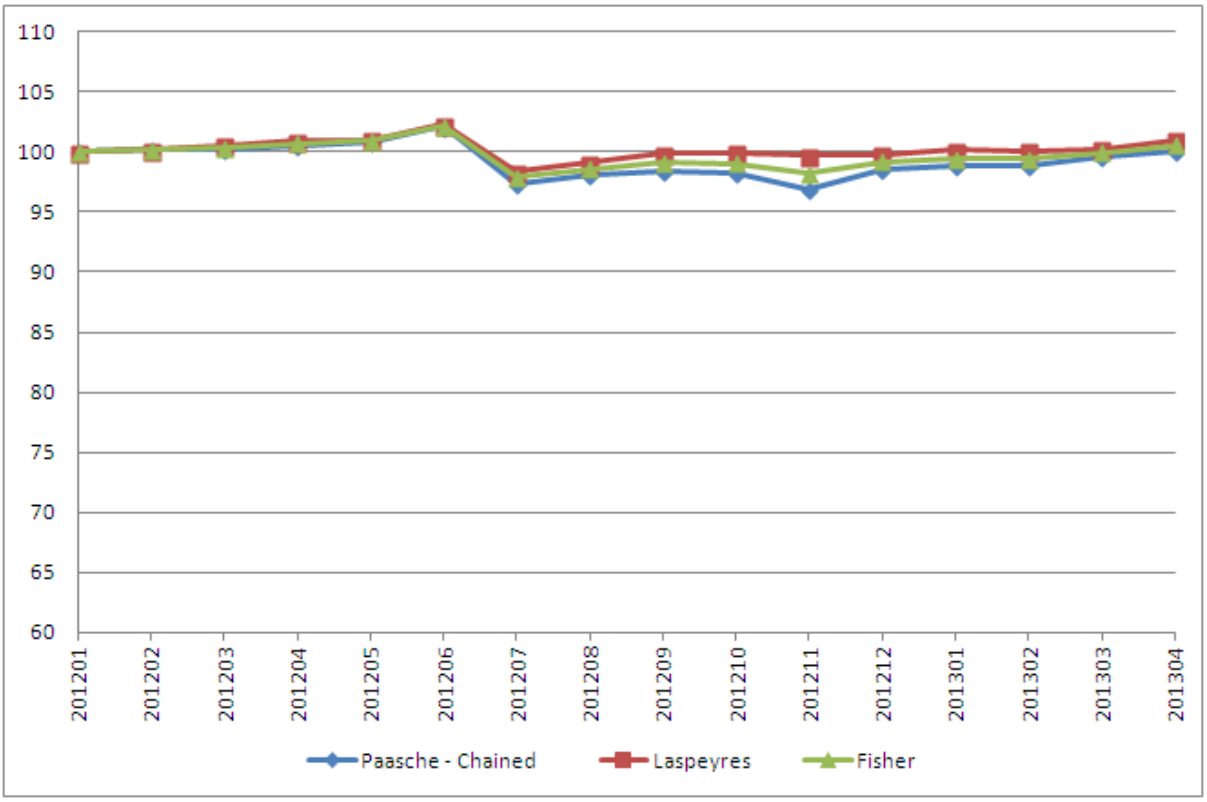


Figure 6: Toothpaste, Scanner Data, Indices – Jevons, Dutot, Carli

Excluding low level sales

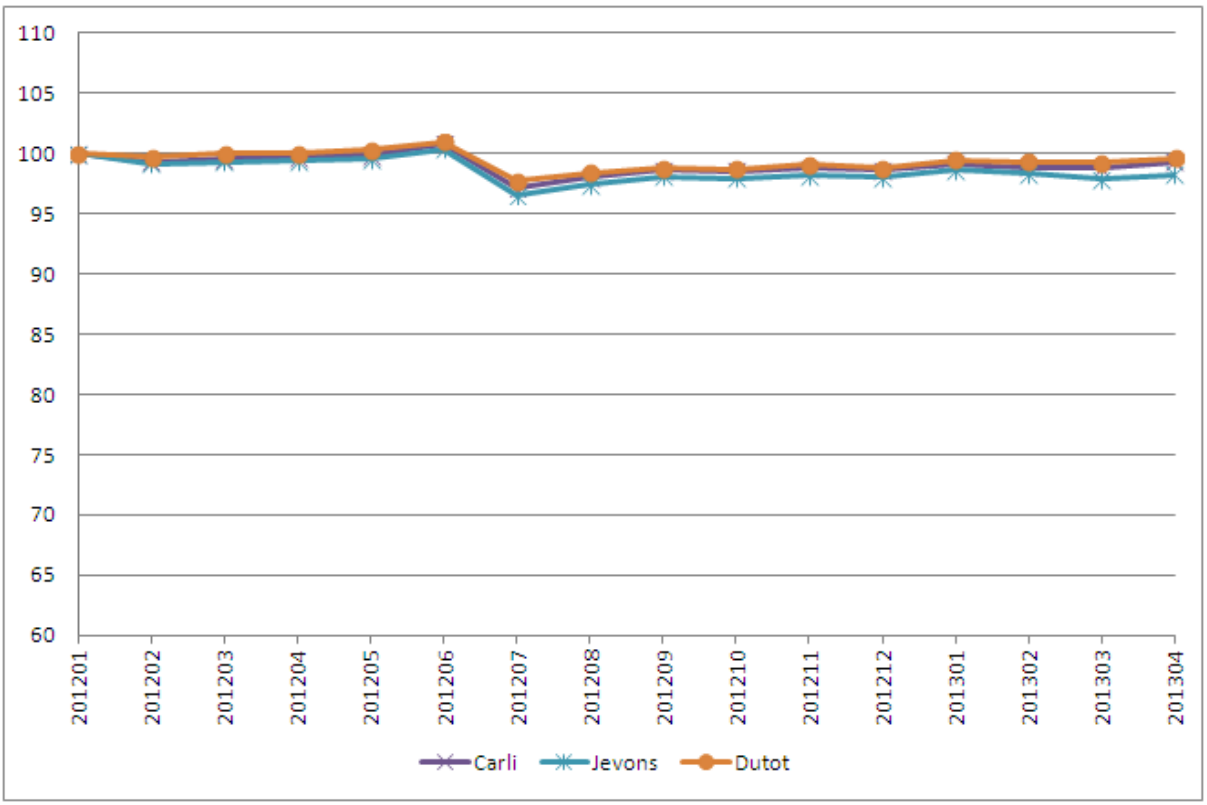


Figure 7: Toothpaste, Scanner Data, Indices – Jevons, Dutot, Carli

All Sales included

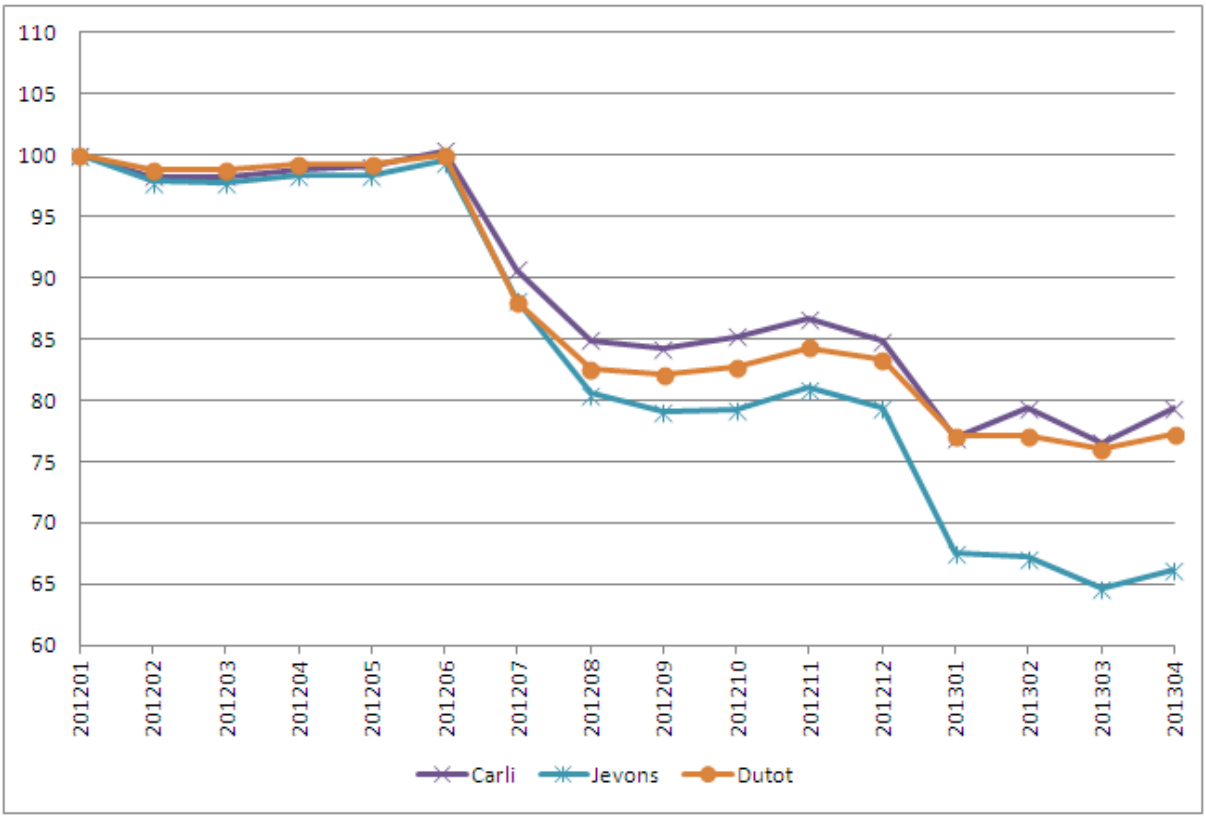
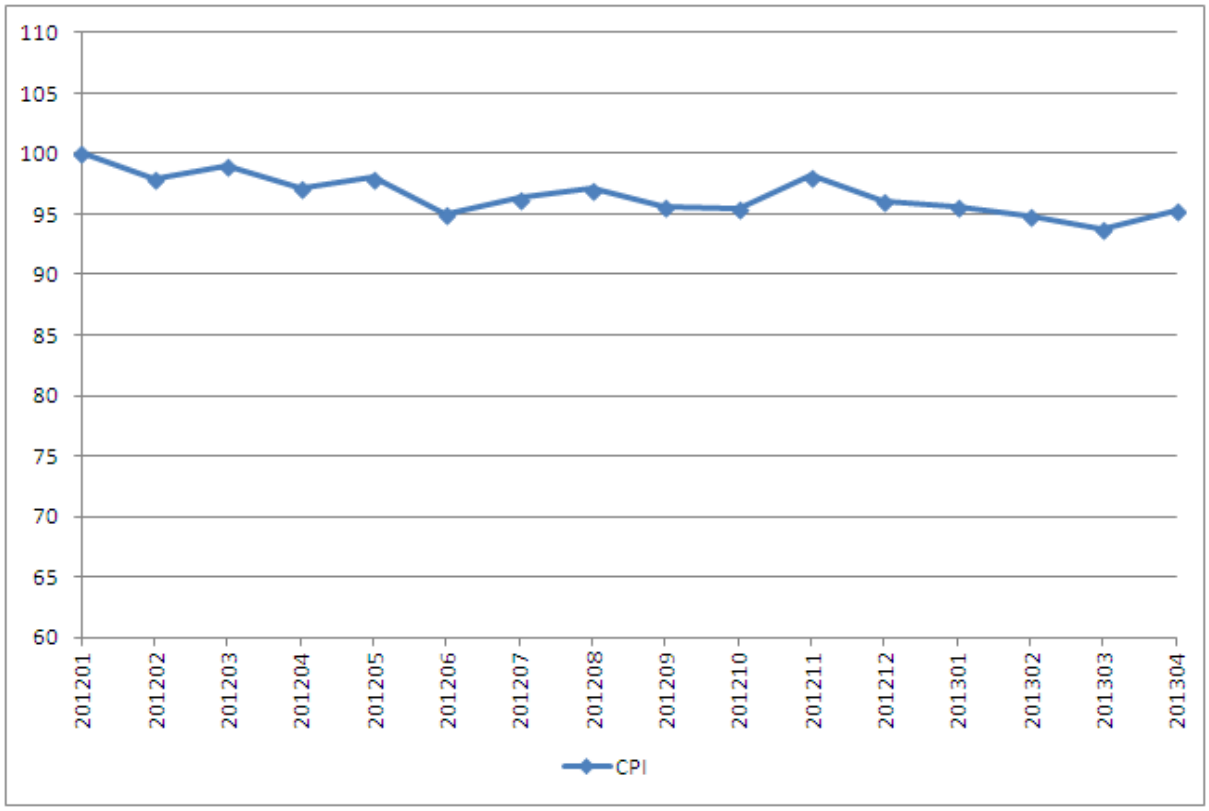


Figure 8: Toothpaste, CPI, Index – Jevons



Next Steps

ONS plans to refine its methods further which may result in revisions to the results presented above. ONS will also extend its analysis of the scanner data obtained to test the impact of including discounted sales, including new and/or discontinued items over time and producing alternative indices such as RYGEKS. The results will help ONS assess the future possibility of using scanner data in consumer price statistics.