ADVISORY PANEL ON CONSUMER PRICES – STAKEHOLDER

Consumer Prices Data Collection Strategy

Status: (final)
Expected publication: (alongside minutes)

Purpose

1. In line with the Better Statistics Better Decisions Strategy there will be a move from collecting prices manually, either on the high street or by trawling websites, to electronic means wherever feasible and efficient. The paper in Annex A summarises the future vision for Consumer Prices data and outlines a future work plan.

Actions

2. Members of the Stakeholder Panel are invited to:
   a) provide feedback on the research to date
   b) comment on the future work plan for the Consumer Prices data collection strategy

Jack Phillips
Prices Division, ONS
January, 2018

List of Annexes

Annex A	Consumer Prices Data Collection Strategy

Annex A – Consumer Prices Data Collection Strategy

1. Introduction

Price data is currently collected through a combination of central (within ONS) and local (contracted out) collection. Around 105,000 price quotes, from around 140 locations, are collected by contracted price collectors visiting shops and other outlets. The prices are collected once a month, at the same time each month. These are supplemented by prices collected centrally by ONS staff, from emails, web sites, telephone calls, catalogues and brochures.

In line with the Better Statistics Better Decisions Strategy the vision for the Consumer Prices data in the future is that it will:

- Support current and future analysis needs
- Consider alternative data sources first
- Be based on resilient systems and processes
- Provide value for money
- Keep pace with how items are purchased in practice

This vision will be achieved through a move from collecting prices manually, either on the high street or by trawling websites, to electronic means wherever feasible and efficient. This will be via:

1. Point of sale scanner data for the largest retailers (including online transactions)
2. Web scraped data for other products which are dominated by a few smaller retailers or for which attribute data is also required.
3. Manual collection, as now, where point of sale and web scraped data are not available/efficient

Analysis of the current collection methods show that there will always be elements of the current collection that will have to be collected locally by price collectors. However, there is considerable scope to investigate the use of alternative data sources for both the central collection and elements of the local collection. These data sources may also provide a more efficient way to capture the increase in online expenditure that has occurred over the last decade. We believe there is potential to collect around 25% of the basket from scanner data and 20% from web scraping which will reduce the amount of locally collected prices as well as the use of internet and CD/Brochure for centrally collected prices.

2. Constraints on collection

Scanner data (covering both local and online retail transactions) requires the agreement of retailers to supply the data on an ongoing basis and up to this point we have not been able to secure this data. Web scraped data requires the terms & conditions of websites to allow for this scraping, and we have found this not to be the case for a number of important retailers that are included in our basket. There are also legal issues to consider relating to any changes to the Retail Price Index, and compliance issues with European statistics. Evidence of the robustness and impact of the new data sources will therefore be required prior to using the new data collections in the production of the

consumer price statistics, which may require several years of parallel run. We have begun the process of procuring web scraped prices.

3. ONS research to date on using alternative data sources in Consumer Prices

Since 2014, we have made significant progress in a number of areas including the development of in-house scraping capability and associated data processing issues such as data storage, classification, cleaning and imputation. We have also made important contributions to the price index methodology literature, including the development of a new index: clustering large datasets into price indices (CLIP). In October 2015, ONS received additional funding for this work from Eurostat under the grant agreement "Web scraping as a source for HICP"

Our work over this grant period has focused on three areas of the CPIH basket of goods and services: grocery items, items that are currently collected centrally by our price collectors, and clothing items.

Our research into grocery items has enabled us to explore methods of collecting web scraped prices in-house, including the development and maintenance of custom-built scrapers in Python. This has led to wider benefits for ONS in general, in particular an increase in knowledge and experience that has contributed to the success of other Big Data projects such as web scraping job vacancies. We have also made a number of improvements to the way that web scraped data is cleaned and classified, which can be applied to other alternative data sources.

The second strand on centrally collected items assesses the feasibility of using "off the shelf" point and click web scraping tools to facilitate the collection of these prices. The project so far has given us valuable experience in navigating the use of web scraped data. In particular, the need to scrape a wider range of retailers' compared with the grocery items collection has required us to develop new guidance around the legal and ethical aspects of web scraping. A feasibility report at the end of the project will summarise findings and provide recommendations on how web scraping can be used in our central collection.

The treatment of clothing prices in consumer price statistics is a topic of interest for many NSIs and the use of alternative data sources can potentially reduce a number of measurement challenges. It has also allowed us to explore web scraped data provided by a third-party, with external cleaning and processing already applied to the data.

Finally, there is still much discussion internationally over the best way to calculate price indices from these data. Our methodology review has shown that different methods are suitable for different areas of the CPIH basket. Practical issues should also be taken into account before deciding on which index should be used.

4. Future work plan

To achieve this vision there are four key areas of further work that will be required with:

1. Obtaining robust sources of alternative data

Assessing all areas of the basket for viable alternatives, launching the procurement of web scraped data, negotiation with large retailers and ensuring correct storage and transfer of the data between ONS and third parties

2. Methods research

   Further consideration is required for example on suitable index methodology, dealing with the lack of expenditure weights for web scraped data and assessing the inclusion of higher frequency data.

3. Assessing the impact on consumer price statistics

   Piloting the use of price indices including alternative data sources and understanding the overall impact on headline statistics

4. Development of systems to support the inclusion of new data sources in consumer prices statistics

   Identification of system requirements to take on and process web scraped and scanner data as well as including them in the aggregation of price statistics