Paper Title: Maximising the use of Web Scraped Data in the Australian CPI

Australian Bureau of Statistics (ABS)

The ABS uses several modes of collection to obtain prices for the Australian Consumer Price Index (CPI). These include personal visits, online, telephone, and administrative data, including transactions data. More recently, the growth of online retailing, combined with advances in technology and automated scraping software has enabled large scale collection of pricing and product information from the web. The process of extracting data from online retailers through scraping software is referred to as web scraping.

From mid-2017 the ABS has been gradually implementing web scraped prices into the CPI as a replacement for manually collected price data. Approximately 500,000 prices are being collected each week, compared to 1,000 prices with traditional modes of collection. The current approach to compiling price indexes from web scraped data involves using an average price of an item over a given period of time for a sample of items.

While this has enhanced the CPI, it is recognised that more can be done with web scraped data. This paper describes research findings focused on the elementary aggregation of web scraped data. The ABS has undertaken research on the conceptual and practical challenges by addressing three important considerations:

1) options for defining products as homogeneous sets or ‘clusters’ of items within each elementary aggregate;
2) weighting approaches to aggregate products together to construct elementary aggregate indexes; and
3) the assessment of the most appropriate index aggregation method.

The paper concludes by outlining a roadmap for the ABS to implement greater use of web scaped data in Australian CPI.