DEUTSCHE BUNDESBANK
EUROSYSTEM

15th Meeting of the Ottawa Group
10 – 12 May 2017

Session 8: Issues with new data sources

Research indices using web scraped data: clustering large datasets into price indices (CLIP)

Elizabeth Metcalfe, Office for National Statistics
Tanya Flower, Office for National Statistics
Thomas Lewis, Office for National Statistics
Matthew Mayhew, Office for National Statistics
Edward Rowland, Office for National Statistics

Alternative data sources such as web scraped and point of sale scanner price data sets are becoming more commonly available, providing large sources of price data from which measures of consumer inflation could potentially be calculated. However, utilising these data without a continuous time-series available for each product is a challenge that a number of National Statistics Institutes (NSIs) are currently facing.

This article puts forward an alternative approach to aggregating large data sets into price indices: Clustering Large datasets Into Price indices (CLIP). The CLIP uses all the data available by creating groups (or clusters) of similar products and monitoring the price change of these groups over time. Unsupervised and supervised machine learning techniques are used to form these product clusters.

The article ends by applying the CLIP to grocery data that has been web scraped from online retailers by the Office for National Statistics (ONS) between June 2014 and July 2016. The experimental price indices presented in the previous ONS web scraping articles are also updated to July 2016 and compared to the CLIP. Charts for each of the web scraped items and aggregate indices are presented in the “Data” section of this release.