Abstract – alternative data sources pipeline

Alternative data sources such as web scraped and point of sale scanner price datasets are becoming more commonly available, providing large sources of price data from which measures of consumer inflation can be calculated. The Office for National Statistics (ONS) has been carrying out research into these data sources since 2014. ONS has recently acquired a robust source of web scraped data from a third-party supplier and are continuing to pursue scanner data.

While previous research has focused on individual aspects of processing these data (for example, the choice of index method), ONS has started an innovative new stage of research that sketches out a proposed end to end pipeline. In practice, this means that we need a system that takes the raw input data, processes it, and outputs low level price indices which are required as inputs into a final production platform.

These new data sources impact choices made at every stage of this pipeline, for example the size of data means that the current manual validation and classification approaches will not be suitable in future. The choice of index method also affects these different stages, for example if a multilateral or chained price index method is chosen, the imputation stage is less important than for a fixed basket price index method. For each module, we have looked at the different methods that could be used, and how they may be affected by the different data sources and final index methods chosen.

This paper will cover the modules required to create low level price indices from big datasets, before showing some experimental price indices calculated by ONS using this prototype pipeline. The work will also include what the impact is of changing some of the underlying assumptions/methods behind earlier stages of the pipeline on the final headline index.