Research into using alternative data sources in the production of consumer price indices, ONS

**Keywords:** product churn, CPI, web scraping, scanner data, CLIP

# Introduction

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 could potentially be calculated. The 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.

Given this progress with acquiring alternative data sources, ONS has started a new stage of research to sketch out a proposed end to end pipeline, comprised of individual modules required to process the data, for example ‘classification’. For each module, we have looked at the different methods that could be used, and how they may differ for the different data sources. In practice, this means that we need a pipeline that takes the raw input data, processes it, and outputs item level indices which are required as inputs into a final production platform.

One of the major obstacles with this pipeline is the product churn (the volume of products entering and leaving the sample). Methods to define suitable clusters of homogeneous products are seen as a way of solving this problem however they remain an open question in the international research at the moment.

This presentation will touch on the modules required to create item level indices from big datasets, before focusing on the clustering large datasets into price indices (CLIP) approach developed by ONS as a way of solving the issues associated with high product churn.

# Modules needed for building a pipeline for alternative data sources

There are a number of stages required for processing alternative data sources before they can be used to calculate price indices. These include -

Pre‐processing is the first stage of the pipeline and is needed to prepare the raw data for the following stages. It includes cleaning, feature engineering, and highlighting excessive missing rates and imbalanced variables. Each item and data source is going to have its own functions for pre‐processing depending on its own unique characteristics. For example, a function that corrects the format for the attribute column “RAM size” is applicable only for laptops.

Classification is about ensuring that we have the right products in each individual dataset to produce an index for a specific item of the basket i.e. the aim of this stage is to create a specific dataset for each defined ONS item. To achieve that we need to make a decision about whether each product/row in the initial dataset should be included or excluded from the individual item dataset. There are a number of methods that can be used to classify the data, ranging from the simplest (mapping the existing retailers’ classification to an ONS item definition) to the most complicated (using supervised machine learning techniques such as a support vector machine).

Imputation is a procedure for entering a value for a specific data item where the response is missing. Rather than leaving missing items blank, imputation allows us to reduce non‐ response bias; manage systematic bias and preserve variance; and preserve relationships between variables. Missing prices can be a problem particularly for price indices constructed from alternative data sources, which may have higher rate of product churn than in the current collection. However, the choice of whether to impute or not is highly interdependent with the index methodology chosen and should not be taken in isolation.

The final module is index number methods, which takes the cleaned and classified data from previous modules in the pipeline. It uses price and expenditure data (where available) to calculate a price index at the elementary aggregate level. Methods include traditional fixed basket indices such as Jevons or Laspeyres, compared with multilateral methods such as GEKS which make better use of the large quantity of data now available.

# Methods for dealing with product churn

In the traditional methodology, an individual product is chosen in the base month by the price collector and then the price of that same product is collected over time. If the product is no longer available, the price collector identifies a comparable product with the same quality characteristics (if possible). If a comparable replacement is not available, a non‐comparable replacement is chosen with a suitable quality adjustment applied to the price in the base period.

## Dealing with relaunches in alternative data sources

With alternative data sources, it may not be possible to identify a suitable comparable replacement if a product goes out of stock due to the sheer volume of data collected. Generally, a product is defined and linked at the GTIN level (i.e. the individual barcode). However, defining products at this level can cause problems where there are lots of “relaunches”. This is where existing products are re‐introduced into the stores with a new GTIN, but there are no changes in the quality of this product.

For example, such GTIN changes are usually the result of changes to the packaging of items, which may be reshaped in order to fit a new product line, while quality characteristics such as brand, package volume and composition/ingredients often stay the same. If GTINs are treated as unique products, the price differences between old and new items will then be fully ascribed to quality changes. However, if these new GTINs are of comparable quality to old GTINs, we are then missing this price change. The same problem occurs if new models are introduced which are of comparable quality to existing models in the dataset. Defining products at GTIN level can therefore affect the trend of price indices calculated using matched methods (i.e. use price relatives as input, for example GEKS or traditional bilateral methods). Research has shown that these indices show a strong downward trend when applied to consumer goods with high rates of churn and with prices that decrease after their introduction to the market.

## Clustering large datasets into price indices (CLIP)

For market segments that experience high rates of product churn, a different approach may be to define a broader concept of product that combines different GTINs of homogenous quality into the same unit. Index methods could then be applied to the average unit value and expenditure of these “clusters”, rather than the individual GTIN level.

The CLIP is applied at the item level, which is the lowest level of classification used to calculate the UK CPI e.g. "Breakfast cereal 1, sugar/chocolate coated" (Figure 1). Within each of these items, the available products are clustered together into similar groups using the information that has been scraped from the website (price, product name, shop, discount marker). The CLIP is calculated by measuring the price change over time between the average price of these clusters. To maintain a fixed basket, the clusters are formed and set for the base month (January, to maintain consistency with the UK CPI), and then the same clusters are formed for each time period over the year (in this case, monthly). The product make-up within each cluster can vary over time, as products move in and out of the market. This therefore allows for high product churn in the data. The clusters are weighted according to the number of products within each cluster.

# Results

There are a number of different approaches that have been developed to calculate price movements for products collected via alternative data sources, each with its own merits and drawbacks. However, there is no defined way of testing which most accurately measures changes in price over time. Figure 1 and Figure 2 compare the CLIP using web scraped data to the published CPI/CPIH. This will give an indication of the impact of using web scraped data with a CLIP approach.



Figure . Comparison of the CLIP and a special aggregate of the published CPI item indices for food and non-alcoholic beverages, Index June 2014 = 100

Similar downward trends are shown for both the published CPI and the CLIP price indices (Figure 1). CPI has seen largely negative contributions from grocery prices over the period since February 2015. While we may not expect the CLIP to behave in the same way as the CPI due to its different methodology and source data, supermarkets have been engaged in a price war since the beginning of 2015 and have therefore reduced prices accordingly to attract consumers. The CLIP also provides evidence for this behaviour.



Figure 2. Comparison of CPIH aggregate and indices based on web scraped data for all clothing

The CPIH all clothing aggregate index remains relatively stable over the period, with a slight upward trend (Figure 2). This is also the case for the men’s and women’s clothing aggregates. The CLIP matches the CPIH trend, but also appears to be more volatile. However, this may be a better reflection of the seasonality displayed in the fashion industry. This is distinct from another method that was tested - a Chained Bilateral Jevons, which decreases over the period investigated.

# Conclusions

ONS are making progress to include alternative data sources within their consumer price statistics. Innovative methods such as the CLIP have been developed to get the most information from the substantially large data sets that are available from scanner and web scraping sources. Promising preliminary results show that these methods have potential to be included within consumer price production in the future and will be further tested within the prototype processing pipeline.

# References

1. Proposed pipeline for processing alternative data sources, Advisory Panel on Consumer Prices (2018)
2. Research indices using web scraped price data: clustering large datasets into price indices (CLIP) (2016)
3. Research indices using web scraped price data: clothing data (2017)