UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneva Switzerland
04 April 2023

Clothing Price Index using Web-Scraped Data

Ahmet Yusuf Aydin, Steven Jones, Laura Christen (Office for National Statistics, United Kingdom of Great Britain and Northern Ireland)
ahmet.aydin@ons.gov.uk

Abstract

Transformation of UK Consumer Price Statistics project aims to incorporate alternative data sources including scanner and web-scraped data into the production of major consumer price statistics. We have developed new methods to process the data and integrate into the production of consumer price index.

Clothing contributes 5% of the Consumer Price Index basket in the UK. We obtain web-scraped data from the main retailers in clothing sector from their online shopping websites, which cover 17 retailers and around 1000 brands. We aim to increase product coverage with the high numbers of clothing items collected via web-scarping compared to manual price collection. We process scraped textual data using NLP and machine learning techniques to build a clothing price index.

To build the index we use three consecutive pipelines: 1. Clothing Classification, 2. Product Grouping, 3. Index Run.

First, we create a classification mapper which maps each item into consumption segments, such as women's dress or men's jeans. We use a gradient-boosted tree machine learning algorithm (XGBoost) to classify web-scraped clothing data. We use human labelled data manually classified by price experts within ONS to train and test the model. Our classifier performs well with an F1 score of 85% on average, while precision is over 90% on some consumption segments.

Secondly, we create a product grouping mapper which maps each item to a product group using a rules-based method. The aim of product grouping is to group similar items and track average prices of each group instead of individual items to increase product match over time. This is crucial for the clothing price index due to the high churn with high product turnover rates and seasonality in the market.

Thirdly, we create a clothing price index using multilateral methods (GEKS) as they allow better use of the dynamic structure of web-scraped data with entering and leaving products. We utilise the classification mapper and product grouping mapper created in the previous stages while building the clothing price index. Classification mapper helps us to build price indices at a granular level of consumption segments and aggregate them to obtain a single index for clothing. Product grouping mapper enables us to overcome the product churn problem and run the index with continuous price data for a high proportion of products.

In conclusion, this project will allow us to modernise UK consumer price statistics by making better use of new data sources and innovative methods.