# Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data*

*Antonio G. Chessa and Robert Griffioen*

## Key Question

Web scraping of online prices is increasingly popular as an alternative data collection method. Statistical institutes may encounter difficulties with the acquisition of scanner data, which is also often a lengthy process. In addition, web scraping enables NSIs to collect prices of entire assortments automatically, an advantage over traditional price collection. However, web scraped data do not provide consumer expenditures, so that essential information for constructing product weights is missing. The question therefore is whether web scraped data can be used for calculating price indices.

## Methodology

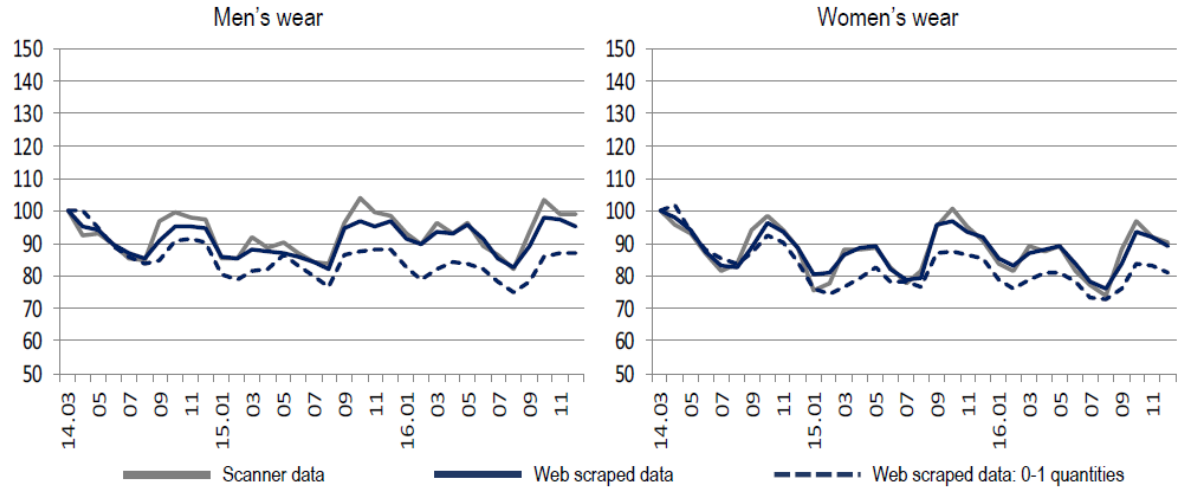
Statistics Netherlands receives scanner data and scrapes online prices and metadata for the same Dutch webshop. Prices and quantities from both data sources are compared for 16 product categories of men's and women's clothing and footwear items, then we use a multilateral method (Geary-Khamis) to calculate price indices with both data sets.

## Main Results

For the product categories compared:

- Scanner data and web scraped prices are most often equal, but the latter higher on average;
- Numbers of sold products and numbers of web scraped product prices show remarkably high correlations over time;
- Although web scraped prices are higher on average, the price indices calculated for the two data sets show small differences for most product categories;
- The differences at COICOP level are only 0.3 percentage point for the year on year indices;
- Replacing the numbers of web scraped prices by binary values (scraped/not scraped) leads to large differences and to downward biases in the web scraped indices (see figure below). This suggests that multiply scraped items should not be deduplicated.

**Comparison of prices from scanner data and prices using numbers *vs* binary value (scraped/not scraped) of web scraped prices**




## Message

The results look promising, but we should keep in mind that they apply only to one retailer. Two questions emerge from the comparisons made: one is about the high correlations found between quantities sold and numbers of web scraped prices over time. The other is about the retailer strategy for organizing his website: does he promote items that sell best? Finally, deriving characteristic patterns from scanner data could further assist in assessing the suitability of web scraped data for price index calculation. This should encourage NSIs to invest in statistical analyses of web scraped and scanner data.

* *Economie et Statistique / Economics and Statistics* N° 509, 2019. https://www.insee.fr/en/statistiques/4203548?sommaire=4203556