How the ABS conducts web scraping?

• Significant amount of product churn – can be mitigated by defining items based on common characteristics (i.e. item group).
• Churn can be characterised as new, ending, new and ending, and intermittent.
• Price behavior of continuing products is different to price behavior of new/ending items (life cycle effects).

Considerations for using web scraped data

1. Product definition

• Web scraped item IDs are granular – has the potential to overstate churn and lead to downward drift.
• Characteristics (e.g. brand, dress type) are extracted from text string to construct ‘item group’ definition.

2. Assigning weights to products

• Web scraped data does not have expenditure information – natural choice would be to use an unweighted price index formula.
• Some authors have experimented with approximating expenditure weights using assumptions on quantities purchased.
• Estimated weighted price indexes can be used to compare against unweighted index formula.

3. Bilateral/multilateral indexes

• Multilateral indexes increasingly being used to solve ‘chain drift’ problem with scanner data.
• Multilateral indexes can be applied to web scraped data – they have an advantage of using all product matches across a window of data.
• Limited length of data available allows for a 13 month window.

Comparison of price index methods

1. Product definition

What do we observe?

• Chained indexes exhibit downward drift – susceptible to life cycle effects.
• Unweighted indexes report lower inflation compared to estimated weighted indexes.
• Multilateral indexes look most promising – broadly comparable with current published indexes.

Conclusion and way forward

• Expand investigations to more retailers – interested if empirical results generalise to other retailers and studies.
• Interested in defining item definitions in automated manner (e.g. CLIP method proposed by ONS).
• Elementary aggregation is one component of project to maximise the use of web scraped data – other components include upper level aggregation and analytical/decomposition tools.