Office for National Statistics

Research indices using web scraped data

Tanya Flower

UN CPI Experts Group Meeting, Geneva, 2nd - 4th May 2016

Outline

1. Introduction & context
2. Data wrangling
   - Classification, Cleaning, Imputation
3. Data limitations
4. Results
   - CLIP, GEKS
5. Summary & Next Steps
6. Links

ONS alternative data sources for prices

- Scanner data
- Bulk extraction of web scraped data
- Replace/supplement central collection using import.io software
- Hedonic adjustment using import.io
- Admin data

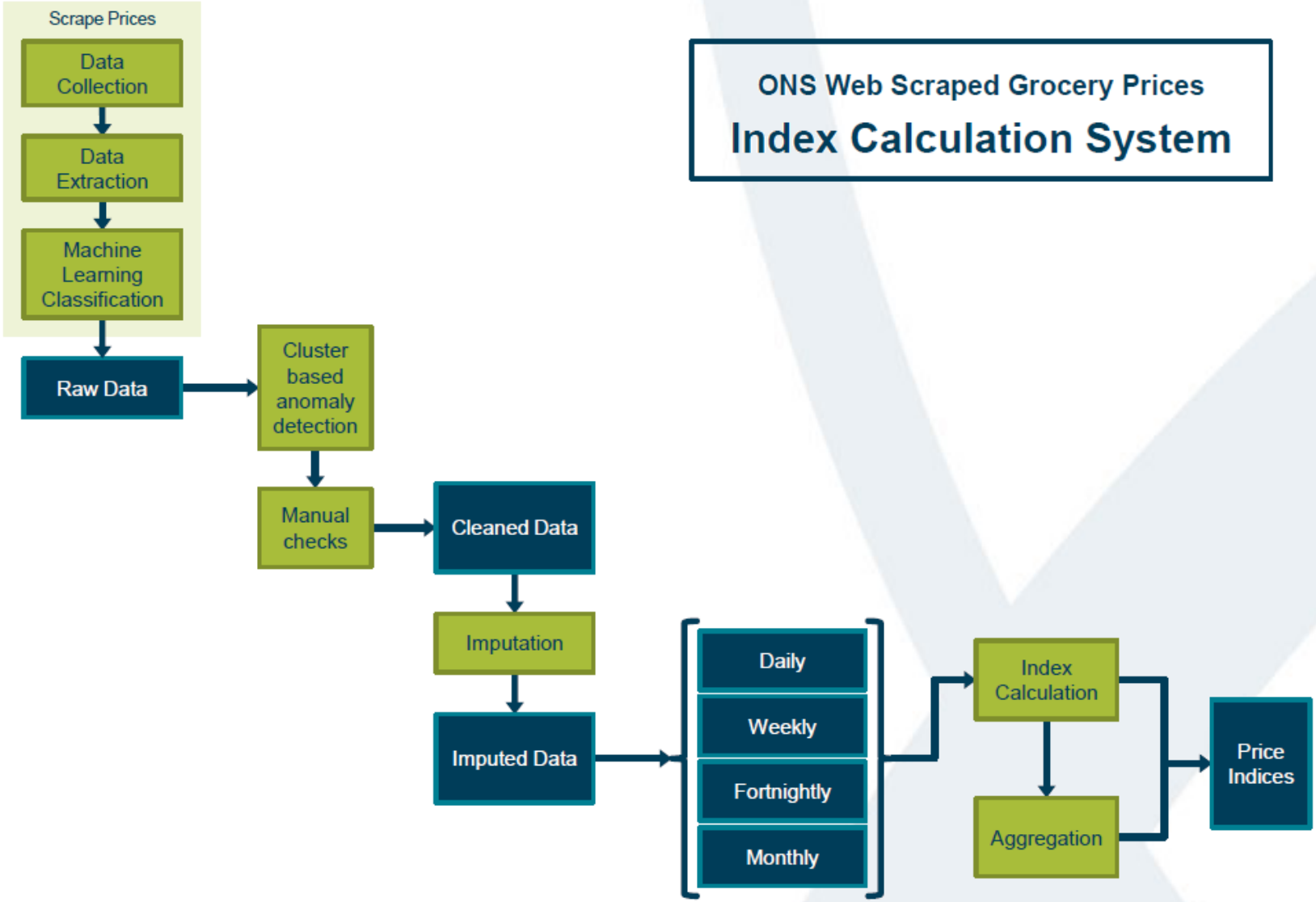
Web scraping

- Prices for 33 CPI items from 3 online retailers

Waitrose Sainsbury's

- Daily collection (around 6,500 price quotes)
- Collects price, product name and discount type
- Period = Jun 2014 to Feb 2016
- Next ONS summary article released on 23rd May 2016

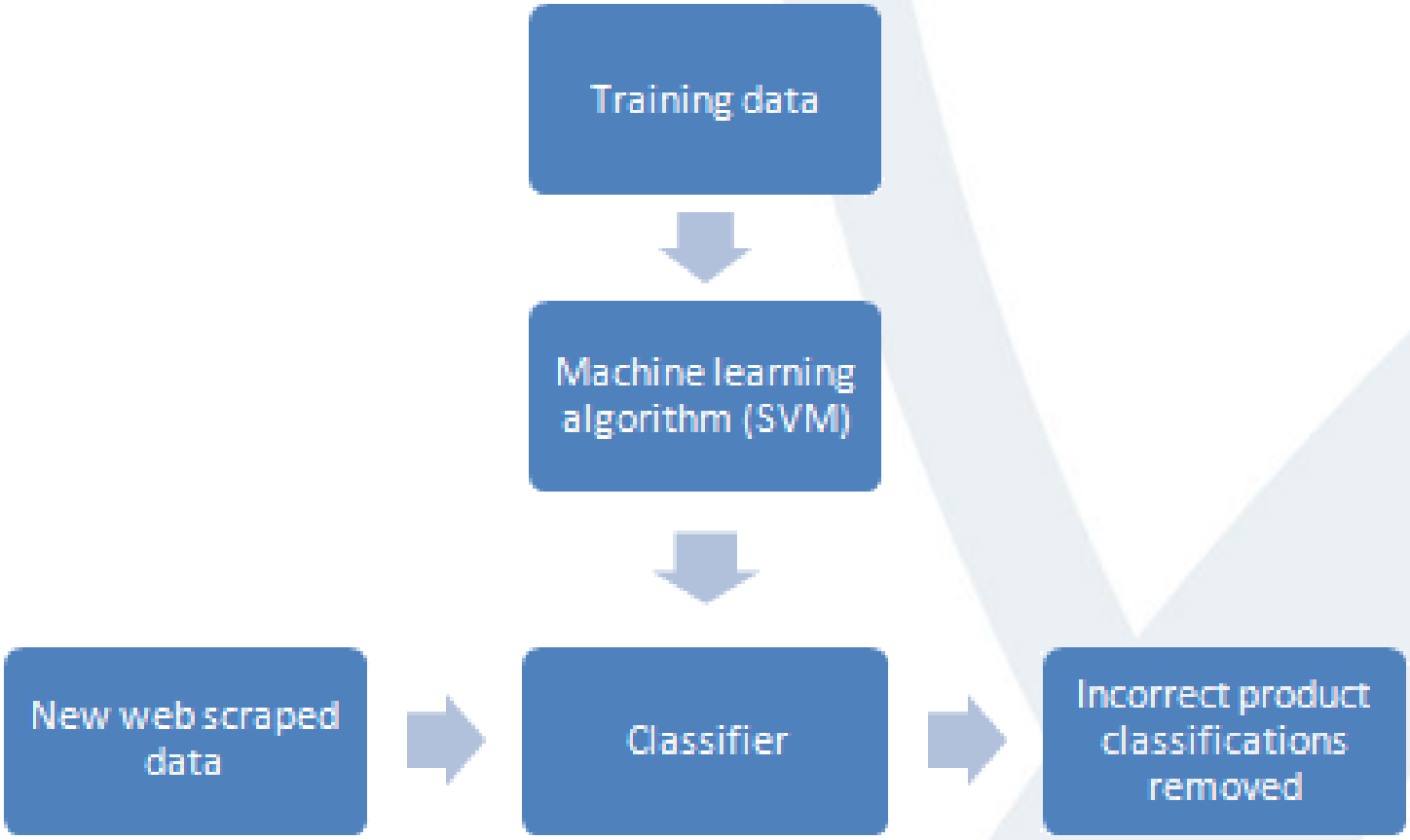
Web scraping – current process



Classification

- Supervised machine learning
- Training data used to teach the Support Vector Machine (SVM)
- Associates words/tokens with correct item classification
- Average accuracy of 85% (F1 score)

Classification



Cleaning

- Cluster based anomaly detection
- Uses Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Identifies clusters that follow a different distribution to correctly classified data
- Use results of previous classification exercise to validate results & ensure new products are retained

Imputation

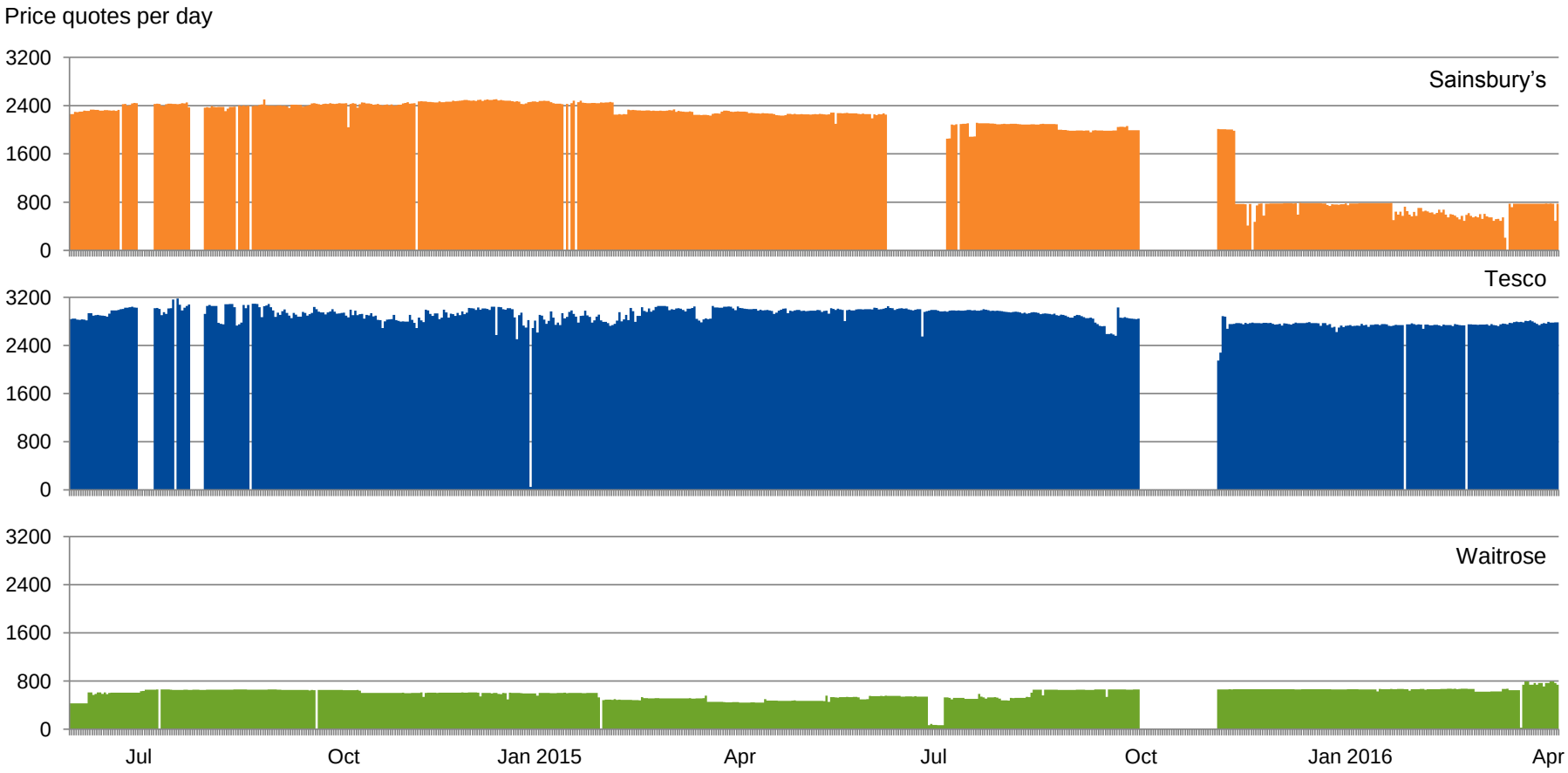
- Imputation can be used when product is temporarily unavailable (out of stock) or there has been a scraper break

- Rules applied:
  - For missing prices carry forward previous price
  - Prices will be carried forward for a maximum of three days
  - If a scraper break is identified, then a price may be carried forward for seven days

Data limitations

- Currently only collect from three stores with an online presence
- High product churn – traditional methods of calculating price indices (matching products) reduce sample size
- All prices are scraped regardless of expenditure & therefore are treated equally
- Reliance on websites not blocking scrapers
- Technological difficulties (scraper breaks)

Data limitations



RESULTS

CLIP index

Results to be published on 23rd May when article link will be updated

GEKS index

Results to be published on 23rd May when article link will be updated

Summary

- Benefits to using web scraped data in consumer price statistics
- Contributed to our understanding of working with large datasets (data wrangling steps)
- CLIP can be applied to large datasets with high product churn

Next steps

- Continue streamlining our data wrangling steps
- Develop a strategy for extreme price changes
- Research further methods for high frequency data
- Expand the web scrapers to collect additional CPI items
- Further analytical work (online vs offline prices, discounts etc)

Thank you for your attention..

Links

- Research indices using web scraped price data: https://www.ons.gov.uk/releases/researchindicesusingwebscrapedpricedatamay2016update