ONS experience with alternative data sources

Tanya Flower and Helen Sands
ONS

07 October 2020

Timelines for ADS implementation

2020	2021	2022	2023
Research + Development	Application + Development	Engagement + Parallel Run	Launch

Priority item categories

Groceries
CPIH weight: 11.4%

Used Cars
CPIH weight: 1.4%

Chart Items
CPIH weight: <1%

Clothing
CPIH weight: 4.1%

Air Fares
CPIH weight: <1%

Rail Fares
CPIH weight: <1%

Tech Goods
CPIH weight: <1%

Package Holidays
CPIH weight: 3.4%

Data acquisition

Web scraped

- in house
- third-party supplier
- direct from websites via API/ S3 bucket

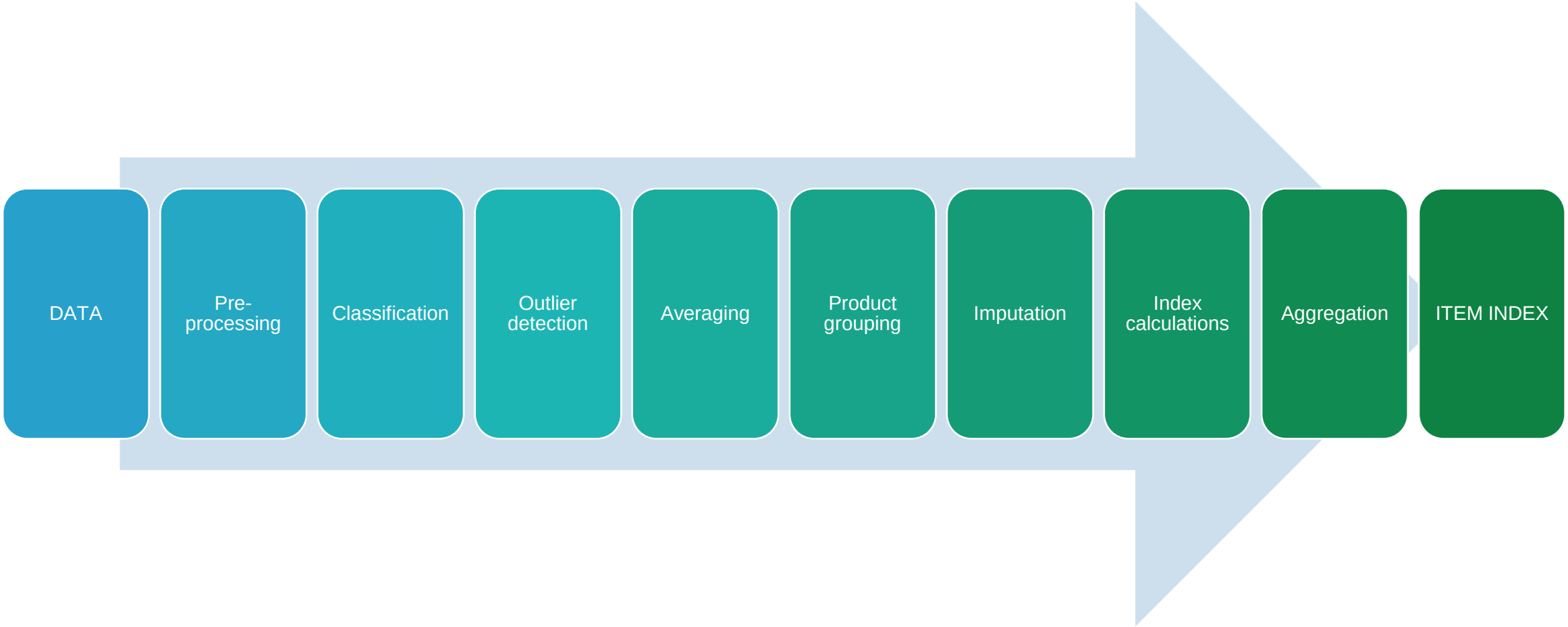
Scanner data

- groceries
- transaction data for railfares

Other

- GS1 product information

Pipeline


DATA
Pre-processing
Classification
Outlier detection
Averaging
Product grouping
Imputation
Index calculations
Aggregation
ITEM INDEX

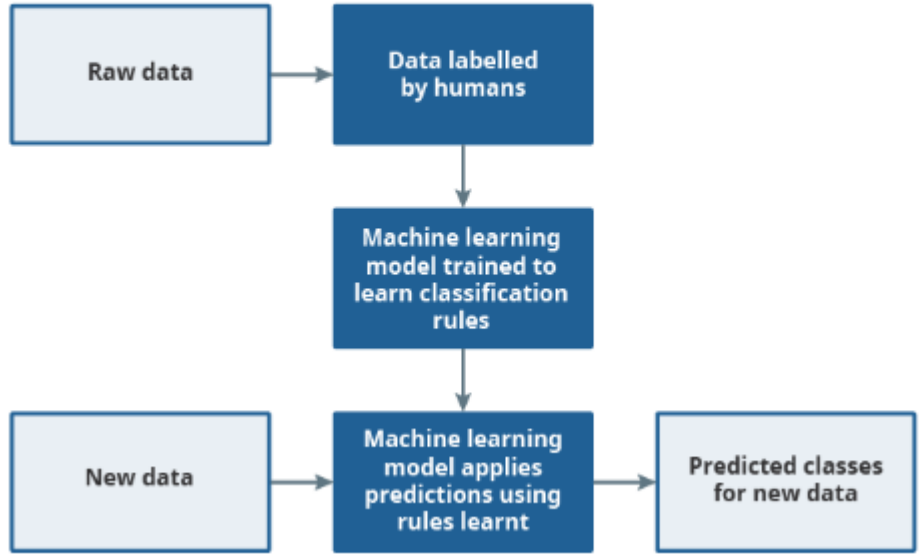
Research… classification

Clothing data is 0.5-0.75m unique products per month, and they quickly refresh. Can't manually classify all of them!

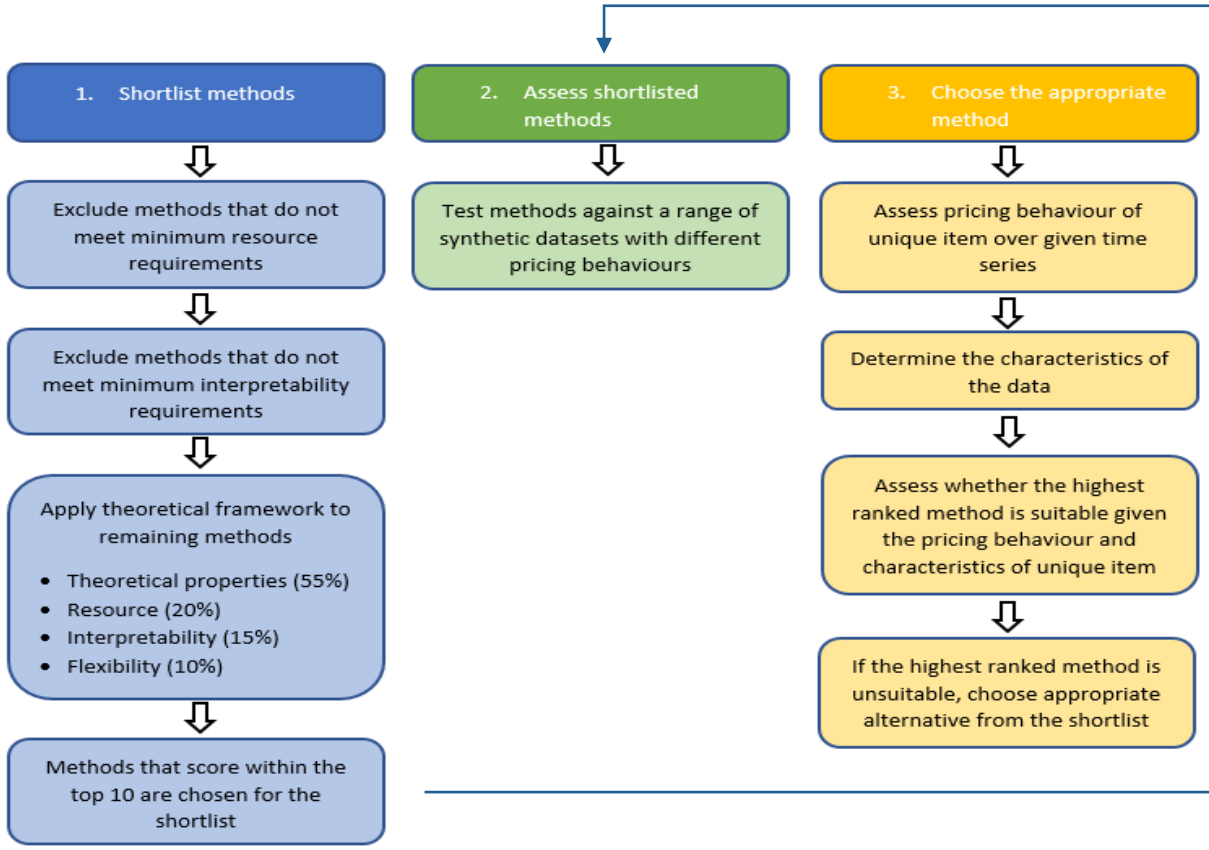
Goal: label a sample of the data, fit a model, apply to rest of the data.

Lots to consider though…

Figure 2: Supervised machine learning uses labelled data to learn rules which are then applied to new unseen datasets



Research… index framework



Rank	Method
1	Quality adjusted Geary Khamis (QU-GK) using a Fixed Base Monthly Expanding window (FMBE)
2	GEKS-Törnqvist using a Movement Splice
3	GEKS-Fisher using a Movement Splice
4	**GEKS-Jevons using a Movement Splice**
5	GEKS-Törnqvist using a Geometric Mean Splice
6	GEKS-Fisher using a Geometric Mean Splice
7	GEKS-Törnqvist using a Window Splice
8	GEKS-Fisher using a Window Splice
9	GEKS-Jevons using a Geometric Mean Splice
10	GEKS-Jevons using a Window Splice

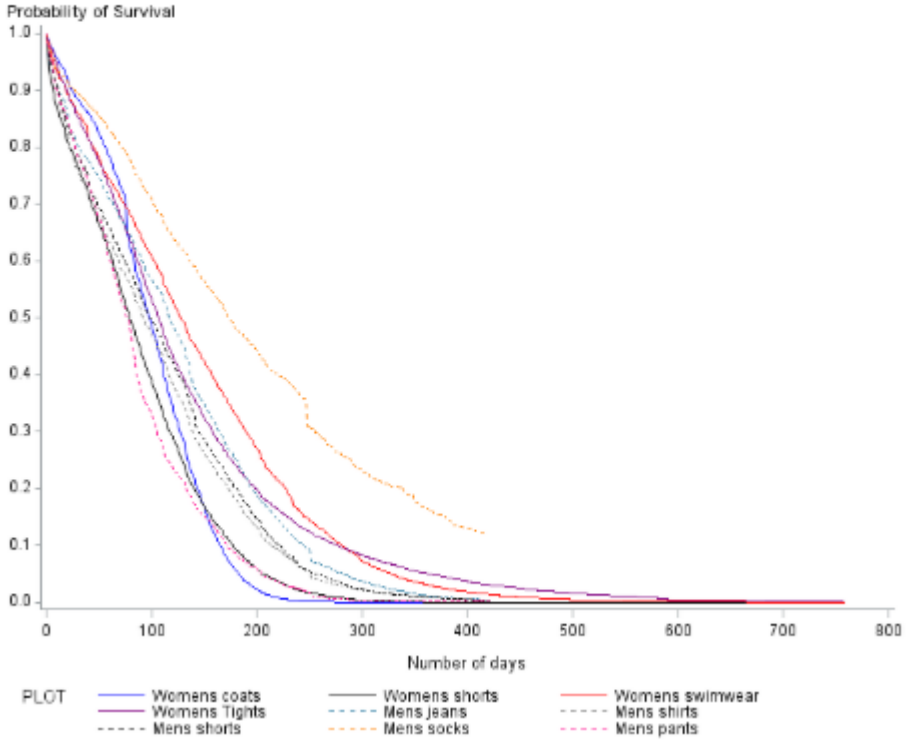
Research… product grouping (1)

Index methods require matched product prices over time. But within a few months, few clothing products remain (see right) so few matches.

Remaining products unrepresentative and does not capture ‘implicit price increase’ from new products replacing clothing on sales i.e. index likely to fall rapidly.

Figure 6: Survival plot for clothing items

UK, men’s clothing: August 2014 to October 2015; women’s clothing: September 2013 to October 2015



Research… product grouping (2)

	Homogeneous product	Month 1	Month 2	Month 3
Product 1	High-quality HP	30		
Product 2	High-quality HP		29	
Product 3	High-quality HP			30
Product 4	High-quality HP	28		
Product 5	Low-quality HP	20		
Product 6	Low-quality HP		20	
Product 7	Low-quality HP			20.5
Other products	Other HPs			

Homogeneous product	Month 1	Month 2	Month 3
High-quality HP	29	29	30
Low-quality HP	20	20	20.5

Other research streams

- Relaunch linking
- Outlier detection
- Imputation
- Returns and discounts
- Weighting and aggregation
- Pack size standardisation
- Expenditure proxies

ADS and Covid-19

- COVID-19 meant user demand for more timely, granular, measures of inflation
- Online price changes for high-demand products (23 items; such as toilet roll, medication, long-life food)
- Data collected using in-house web scrapers, and processed using an adapted version of our pipeline
- Also have now developed wider food & drink basket weekly indicator

Food & drink indices

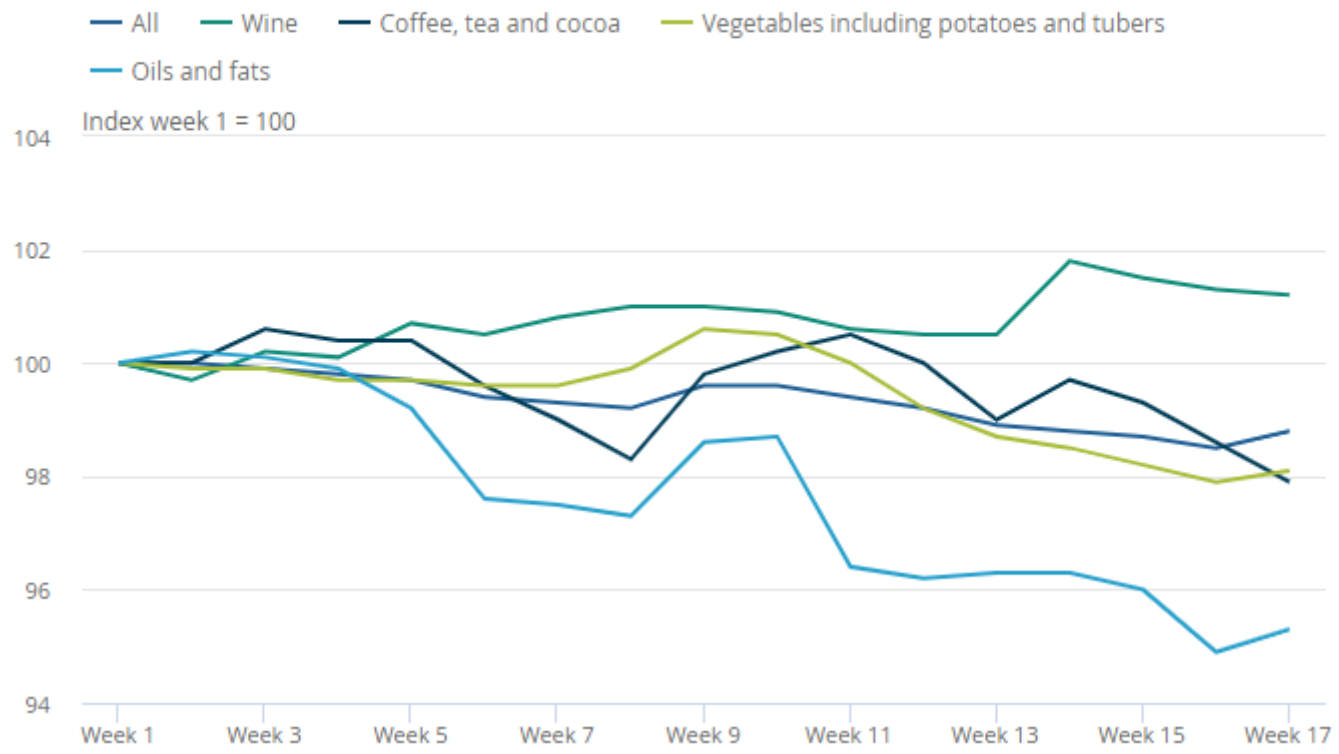
7 supermarket chains

138 food & drink products, 14 higher level categories

25,000 unique products

Daily collection = ~ 175,000 prices processed per week

Online price change of selected food and drink products 1 June to 27 September: index week 1 (1 to 7 June) = 100, UK



Expenditure data

- Regular data feed from 1 retailer used in internal briefing on consumer expenditure, looking to expand to include further regular feeds shortly