Using alternative data sources in consumer prices

Liam Greenhough
Senior Statistical Officer
Prices

21 May 2019

Outline

- What are alternative data sources?
- Plans for alternative data sources
- Pipeline for processing data
- Further research
- Later: experimental results

What are alternative data sources?

Current sources:

- Local collections
- Central collections
- Admin data

Alternative sources:

- Web scraped data
- Scanner data

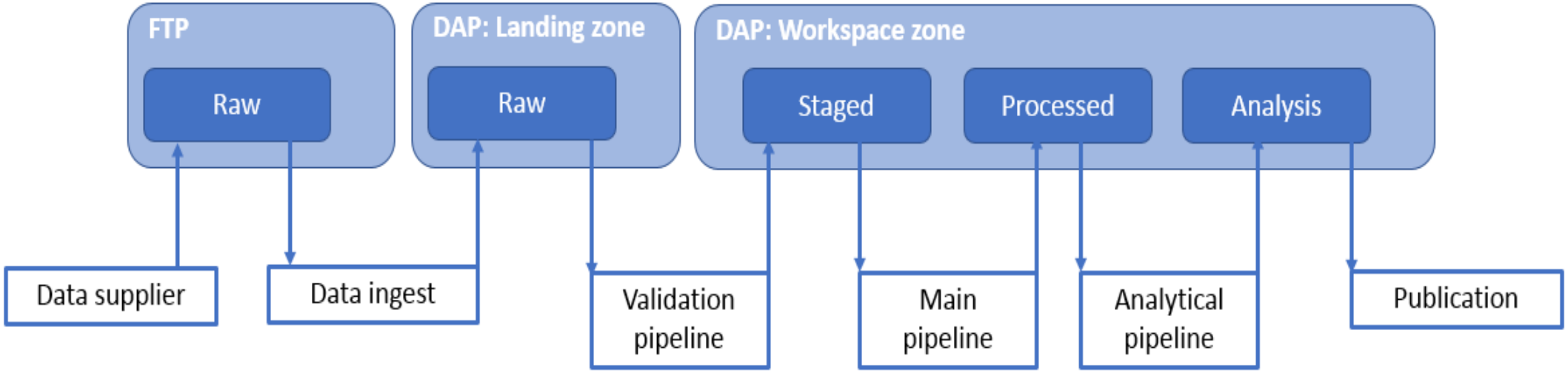
Plans for alternative data sources

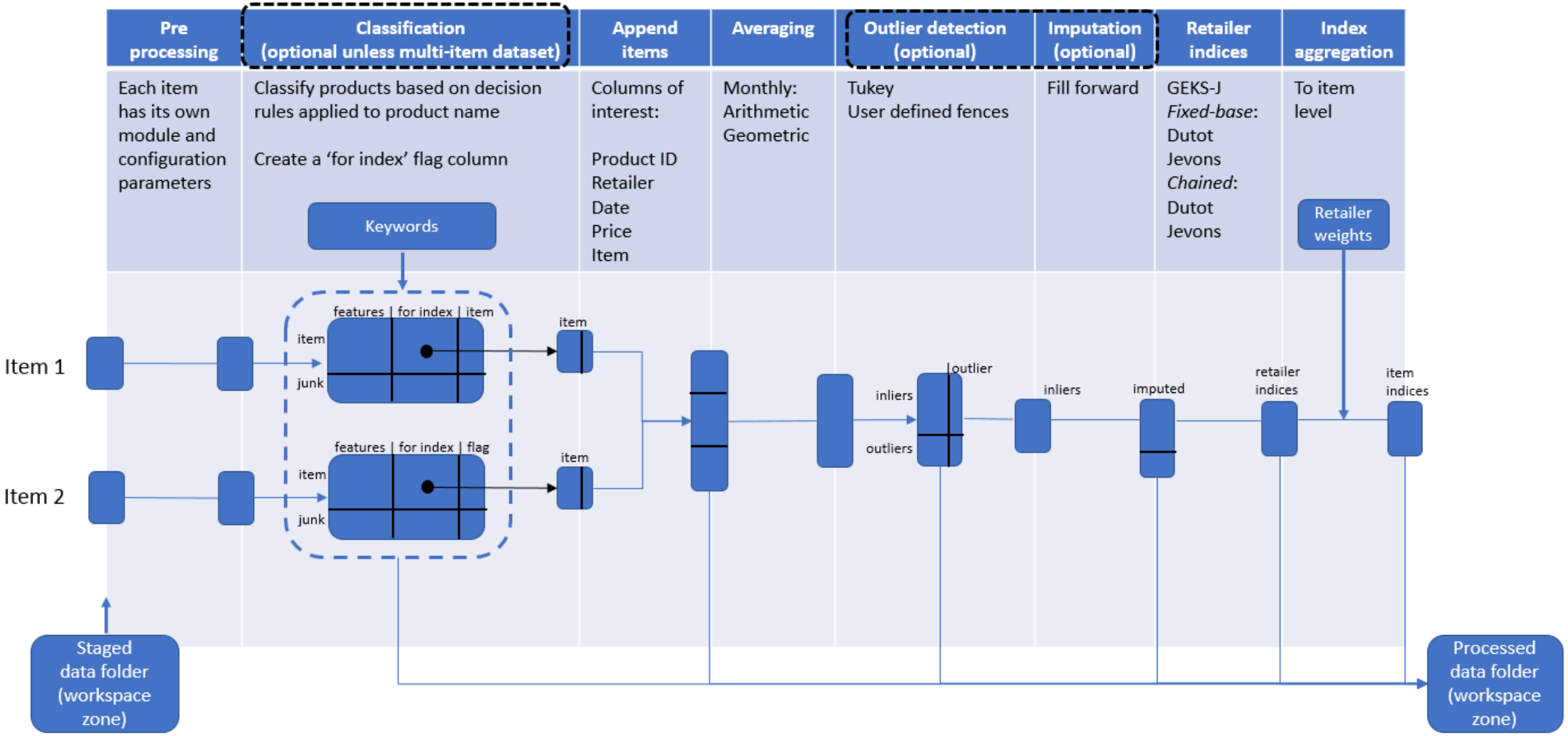
- Incorporate into consumer price statistics by Jan 2023
- Develop new pipeline to process prices data including alternative data sources
- Focus on particular product groups

Plans for alternative data sources

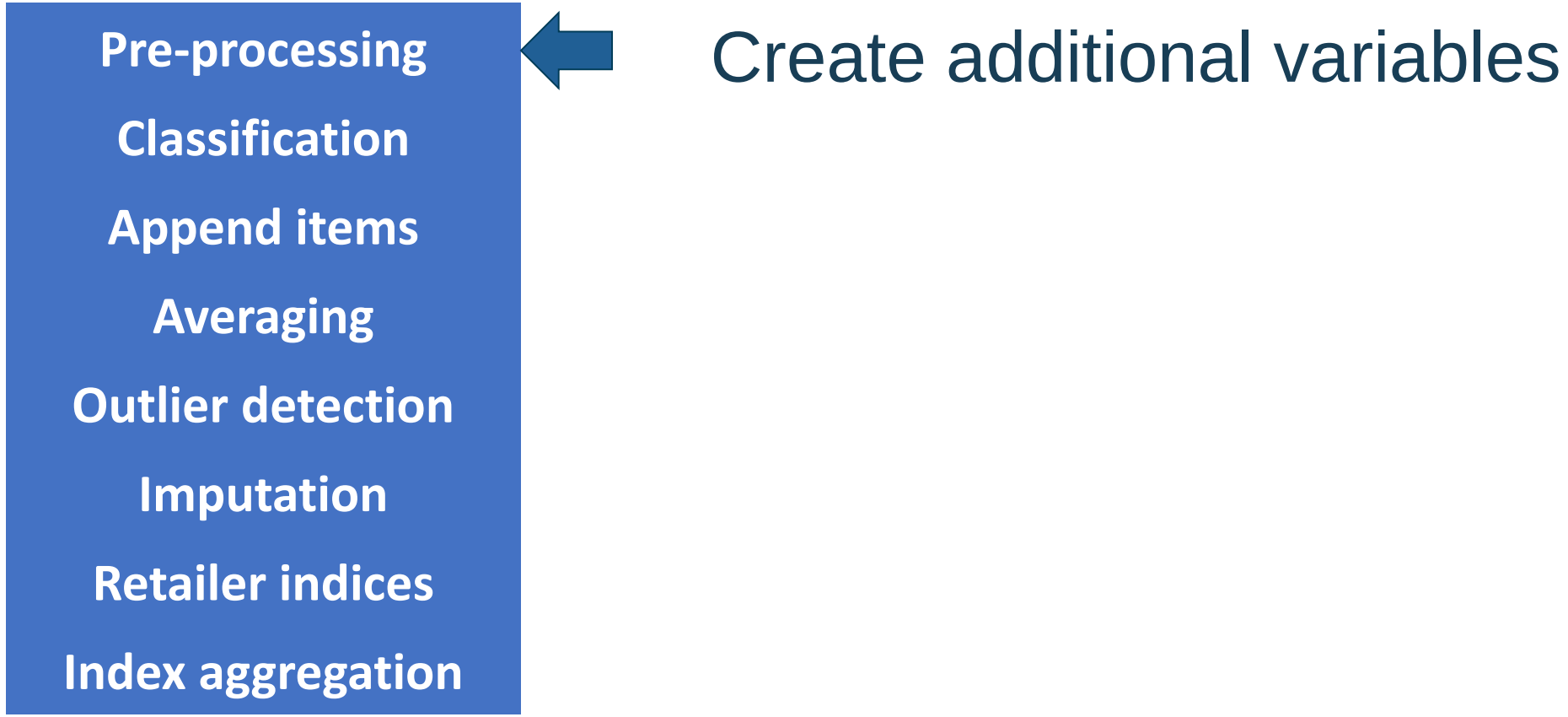
- Regular updates and experimental indices scheduled
- Consultation with advisory panel and interest groups
- Intended formal consultation in 2022
- Impact assessment

Pipeline and methods




Pre processing
Classification (optional unless multi-item dataset)
Append items
Averaging
Outlier detection (optional)
Imputation (optional)
Retailer indices
Index aggregation
Each item has its own module and configuration parameters
Classify products based on decision rules applied to product name
Create a 'for index' flag column
Columns of interest:
Product ID
Retailer
Date
Price
Item
Monthly:
Arithmetic
Geometric
Tukey
User defined fences
Fill forward
GEKS-J
Fixed-base:
Dutot
Jevons
Chained:
Dutot
Jevons
To item level
Keywords
Retailer weights
Item 1
Item 2
features | for index | item
features | for index | flag
item
junk
inliers
outliers
outlier
imputed
retailer indices
item indices
Staged data folder (workspace zone)
Processed data folder (workspace zone)

Main pipeline



Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation

ID	Product name	Retailer	Scrape date	Date	Price
L1	Laptop 1	A	01/04/2019	Apr-19	350
L3	Laptop bag	A	01/04/2019	Apr-19	30
L1	Laptop 1	A	08/04/2019	Apr-19	400
L4	Laptop 3	B	08/04/2019	Apr-19	30000000
L1	Laptop 1	A	01/03/2019	Mar-19	350
L2	Laptop 2	B	01/03/2019	Mar-19	600
L1	Laptop 1	A	08/03/2019	Mar-19	350
L2	Laptop 2	B	08/04/2019	Mar-19	600

Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation

ID	Product name	Retailer	Scrape date	Date	Price	Item
L1	Laptop 1	A	01/04/2019	Apr-19	350	Laptop
~~L3~~	~~Laptop bag~~	~~A~~	~~01/04/2019~~	~~Apr-19~~	~~30~~	~~Junk~~
L1	Laptop 1	A	08/04/2019	Apr-19	400	Laptop
L4	Laptop 3	B	08/04/2019	Apr-19	30000000	Laptop
L1	Laptop 1	A	01/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	01/03/2019	Mar-19	600	Laptop
L1	Laptop 1	A	08/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	08/04/2019	Mar-19	600	Laptop

Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation

ID	Product name	Retailer	Scrape date	Date	Price	Item
L1	Laptop 1	A	01/04/2019	Apr-19	350	Laptop
L1	Laptop 1	A	08/04/2019	Apr-19	400	Laptop
L4	Laptop 3	B	08/04/2019	Apr-19	30000000	Laptop
L1	Laptop 1	A	01/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	01/03/2019	Mar-19	600	Laptop
L1	Laptop 1	A	08/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	08/04/2019	Mar-19	600	Laptop

Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation

ID	Product name	Retailer	Scrape date	Date	Price	Item
L1	Laptop 1	A	01/04/2019	Apr-19	350	Laptop
L1	Laptop 1	A	08/04/2019	Apr-19	400	Laptop
L4	Laptop 3	B	08/04/2019	Apr-19	30000000	Laptop
L1	Laptop 1	A	01/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	01/03/2019	Mar-19	600	Laptop
L1	Laptop 1	A	08/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	08/04/2019	Mar-19	600	Laptop

Other items…

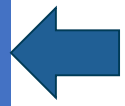
Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation

ID	Product name	Retailer	Scrape date	Date	Price	Item
L1	Laptop 1	A	01/04/2019	Apr-19	350	Laptop
L1	Laptop 1	A	08/04/2019	Apr-19	400	Laptop
L4	Laptop 3	B	08/04/2019	Apr-19	30000000	Laptop
L1	Laptop 1	A	01/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	01/03/2019	Mar-19	600	Laptop
L1	Laptop 1	A	08/03/2019	Mar-19	350	Laptop
L2	Laptop 2	B	08/04/2019	Mar-19	600	Laptop

Main pipeline

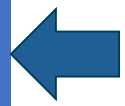
- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation



ID	Product name	Retailer	Date	Price	Item
L1	Laptop 1	A	Apr-19	374.17	Laptop
L4	Laptop 3	B	Apr-19	30000000	Laptop
L1	Laptop 1	A	Mar-19	350	Laptop
L2	Laptop 2	B	Mar-19	600	Laptop

Main pipeline

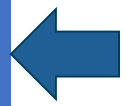
- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation



ID	Product name	Retailer	Date	Price	Item
L1	Laptop 1	A	Apr-19	374.17	Laptop
~~L4~~	~~Laptop 3~~	~~B~~	~~Apr-19~~	~~30000000~~	~~Laptop~~
L1	Laptop 1	A	Mar-19	350	Laptop
L2	Laptop 2	B	Mar-19	600	Laptop

Main pipeline

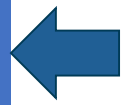
- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation



ID	Product name	Retailer	Date	Price	Item
L1	Laptop 1	A	Apr-19	374.17	Laptop
L1	Laptop 1	A	Mar-19	350	Laptop
L2	Laptop 2	B	Mar-19	600	Laptop

Main pipeline

- Pre-processing
- Classification
- Append items
- Averaging
- Outlier detection
- Imputation
- Retailer indices
- Index aggregation



ID	Product name	Retailer	Date	Price	Item
L1	Laptop 1	A	Apr-19	374.17	Laptop
L2	Laptop 2	B	Apr-19	600	Laptop
L1	Laptop 1	A	Mar-19	350	Laptop
L2	Laptop 2	B	Mar-19	600	Laptop

Main pipeline

Elementary indices created at retailer-level

Expenditure weights used to aggregate to item-level

Configurability

Module	On or off?	Method
Classification	On	Rules-based classifiers
Averaging	Always on	Geometric
Outlier Detection	On	User defined fences
Imputation	On	Pull forward previous value
Index method	Always on	Fixed base Jevons

Further research

- Classification
- Further development of pipeline
- Aggregation of data sources
- Impact of returns and discounts

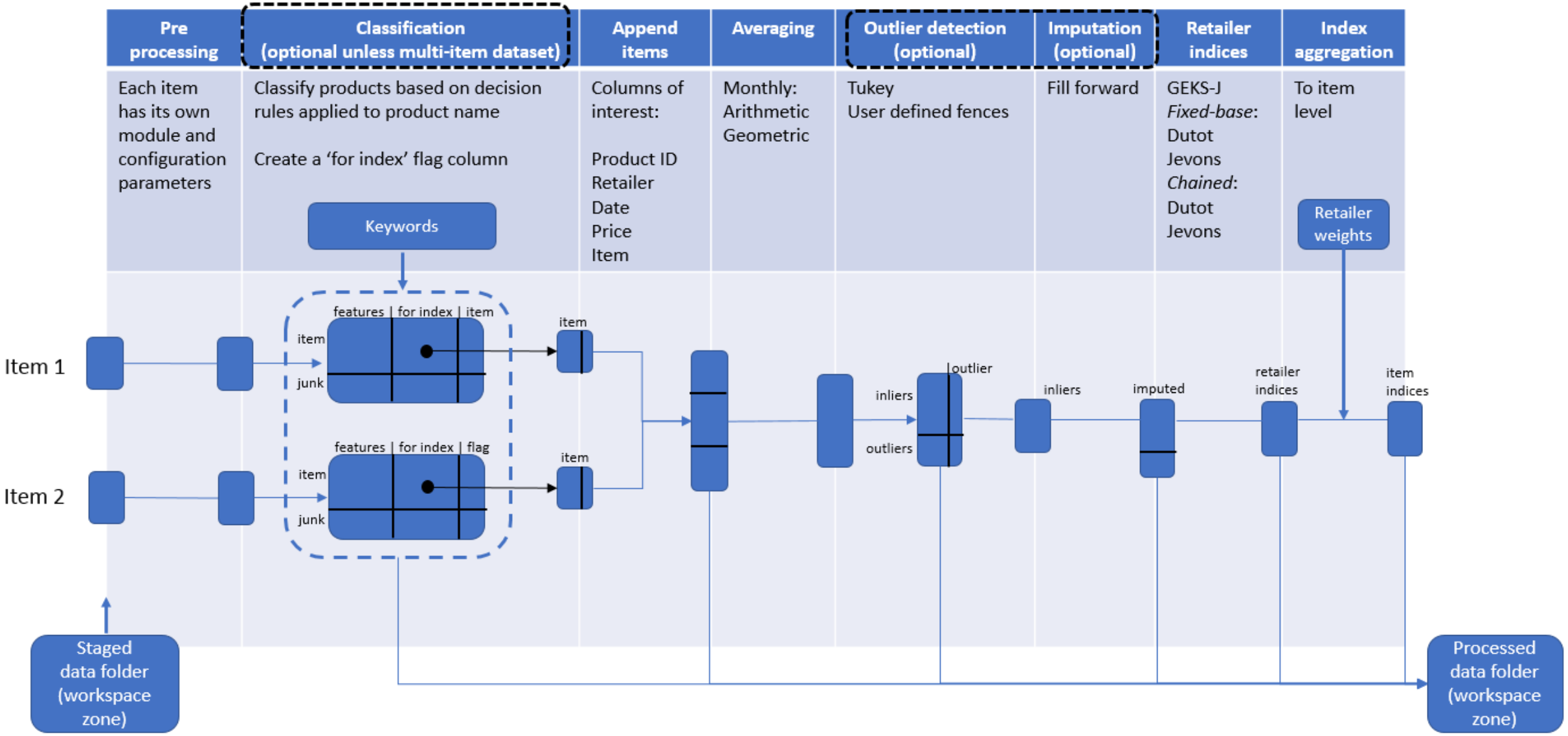
Quick wins…

- Can we replace existing data sources whilst using existing systems and methods?

- Can we use data elsewhere in the basket?

Index Methods

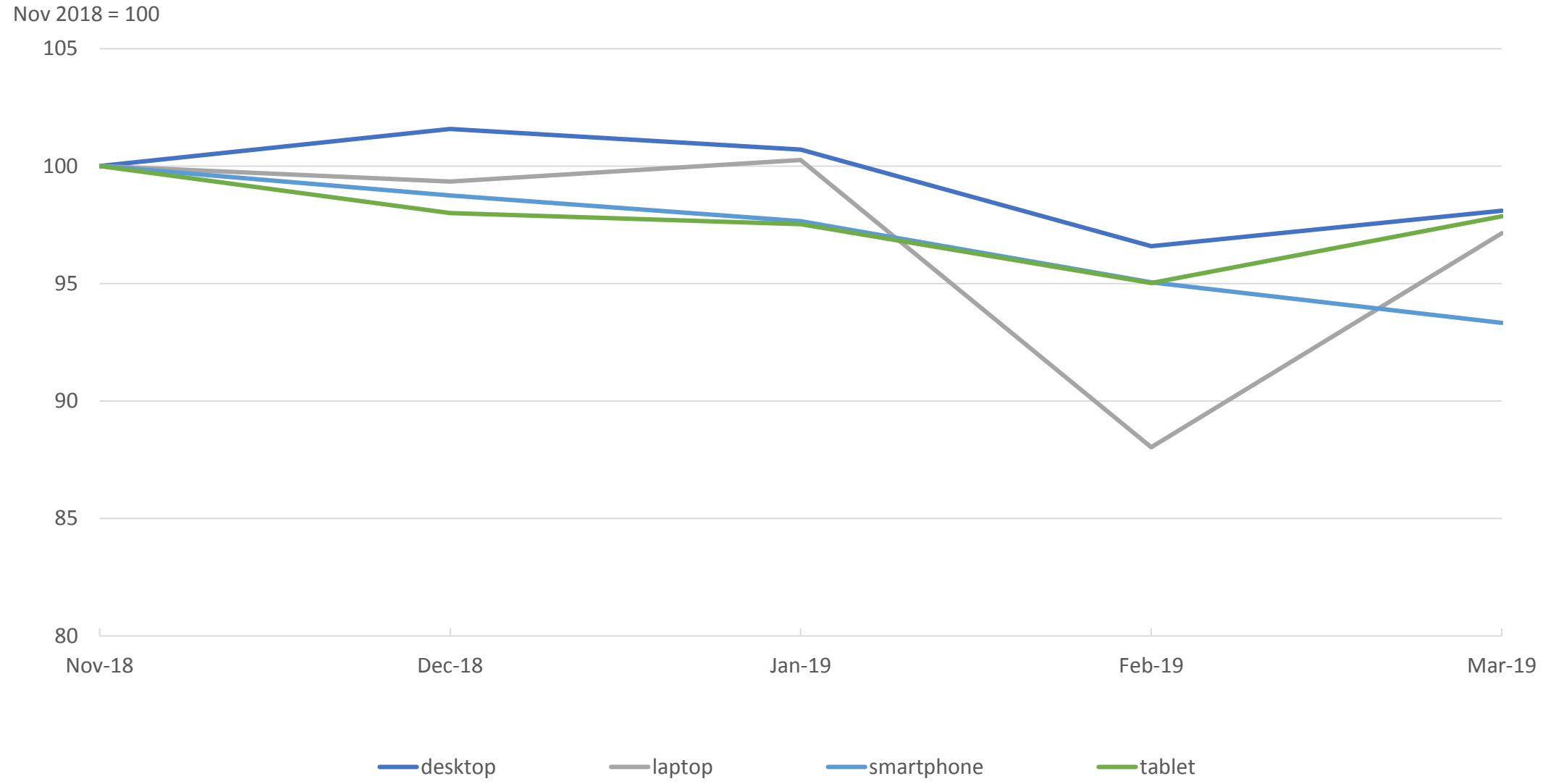
- Framework for assessing quality of consumer price indices produced using alternative data sources
- Product level expenditure weights for web scraped data
- Product definition


Pre processing
Classification (optional unless multi-item dataset)
Append items
Averaging
Outlier detection (optional)
Imputation (optional)
Retailer indices
Index aggregation
Each item has its own module and configuration parameters
Classify products based on decision rules applied to product name
Create a 'for index' flag column
Columns of interest:
Product ID
Retailer
Date
Price
Item
Monthly:
Arithmetic
Geometric
Tukey
User defined fences
Fill forward
GEKS-J
Fixed-base:
Dutot
Jevons
Chained:
Dutot
Jevons
To item level
Keywords
Retailer weights
Item 1
Item 2
features | for index | item
features | for index | flag
item
junk
inliers
outliers
outlier
imputed
retailer indices
item indices
Staged data folder (workspace zone)
Processed data folder (workspace zone)

Configuration for results

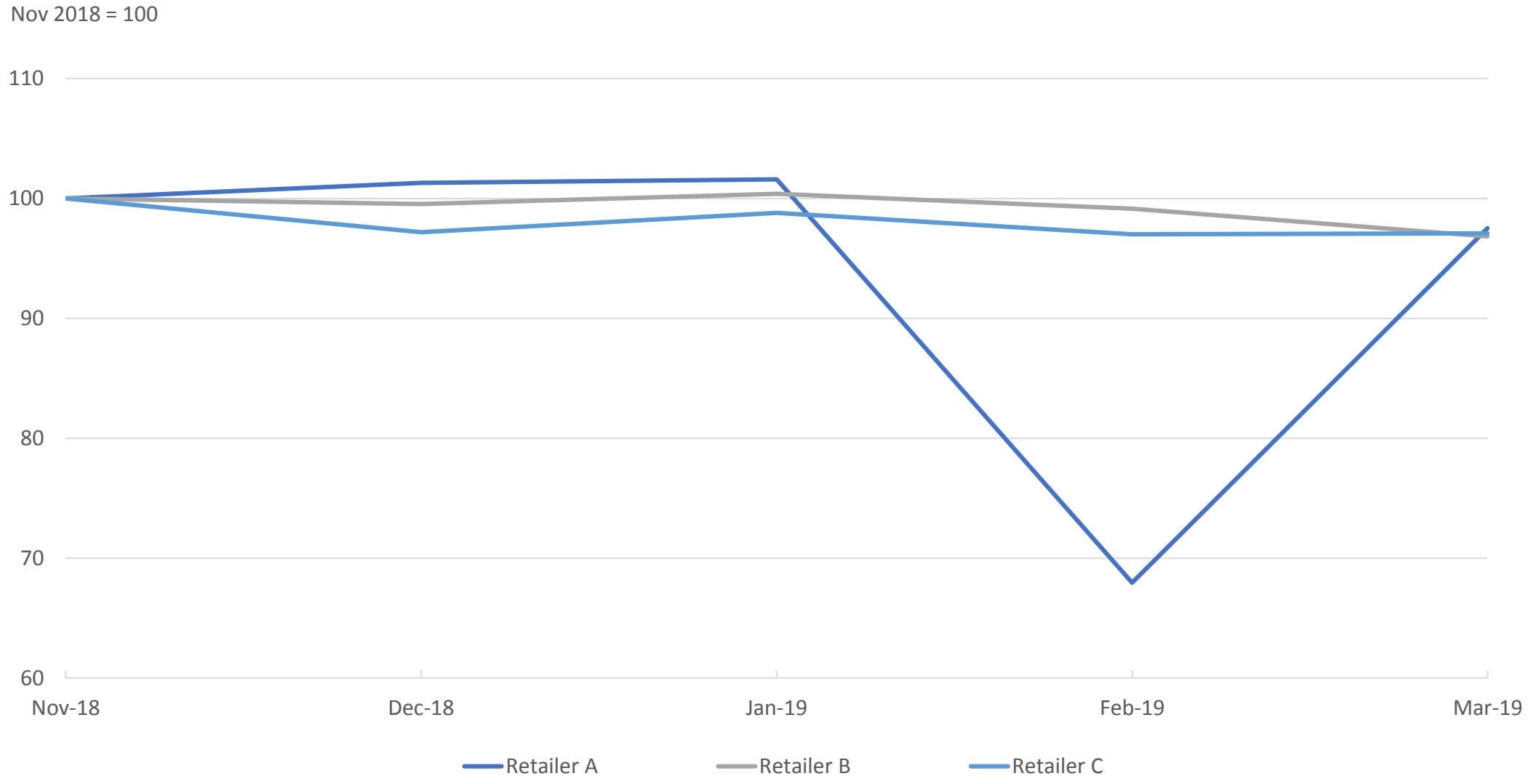
Module	On or off?	Method
Classification	On	Rules-based classifiers
Averaging	Always on	Geometric
Outlier Detection	On	User defined fences
Imputation	Off	N/A
Index method	Always on	Fixed base Jevons

Fixed base Jevons indices for desktops, laptops, tablets and smartphones



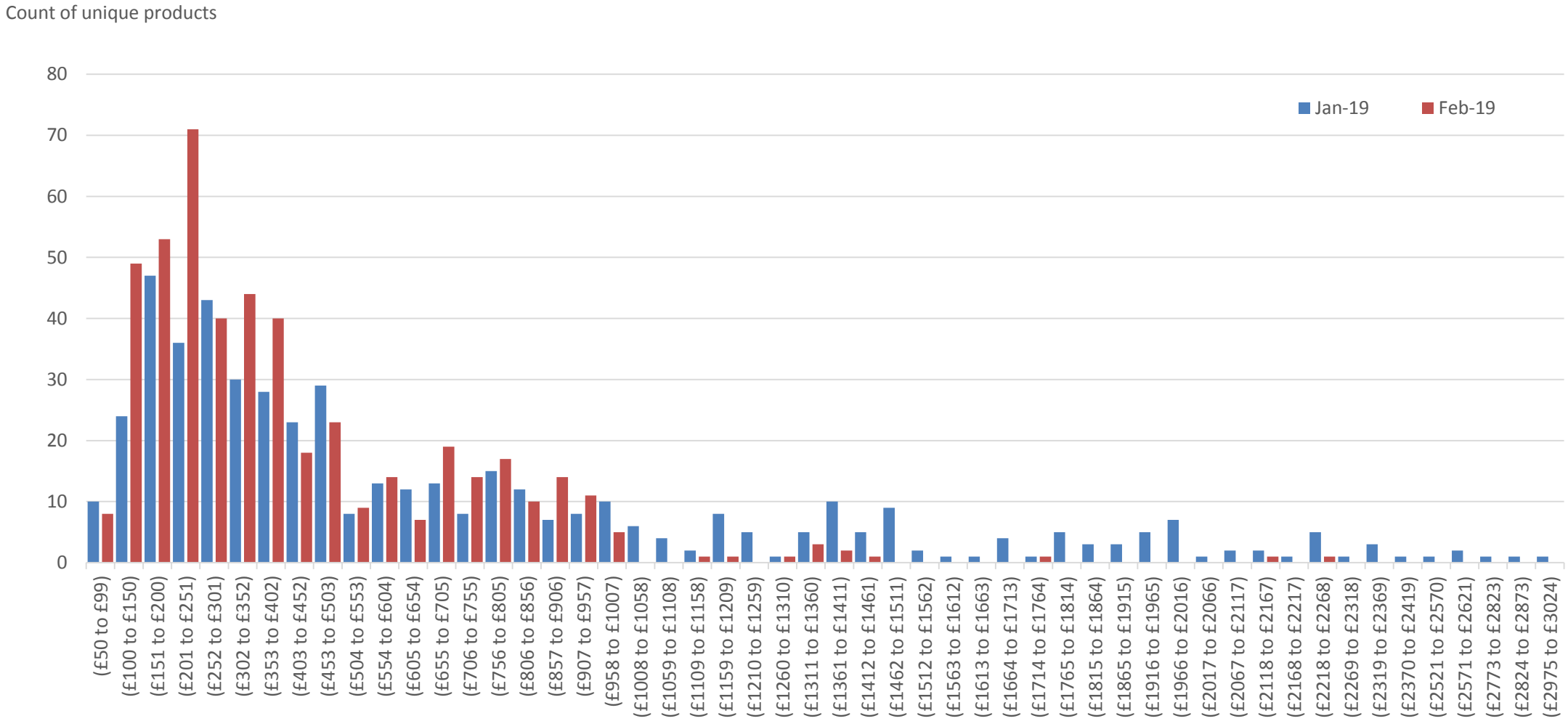
Office for National Statistics

Fixed base Jevons indices for laptops by retailer

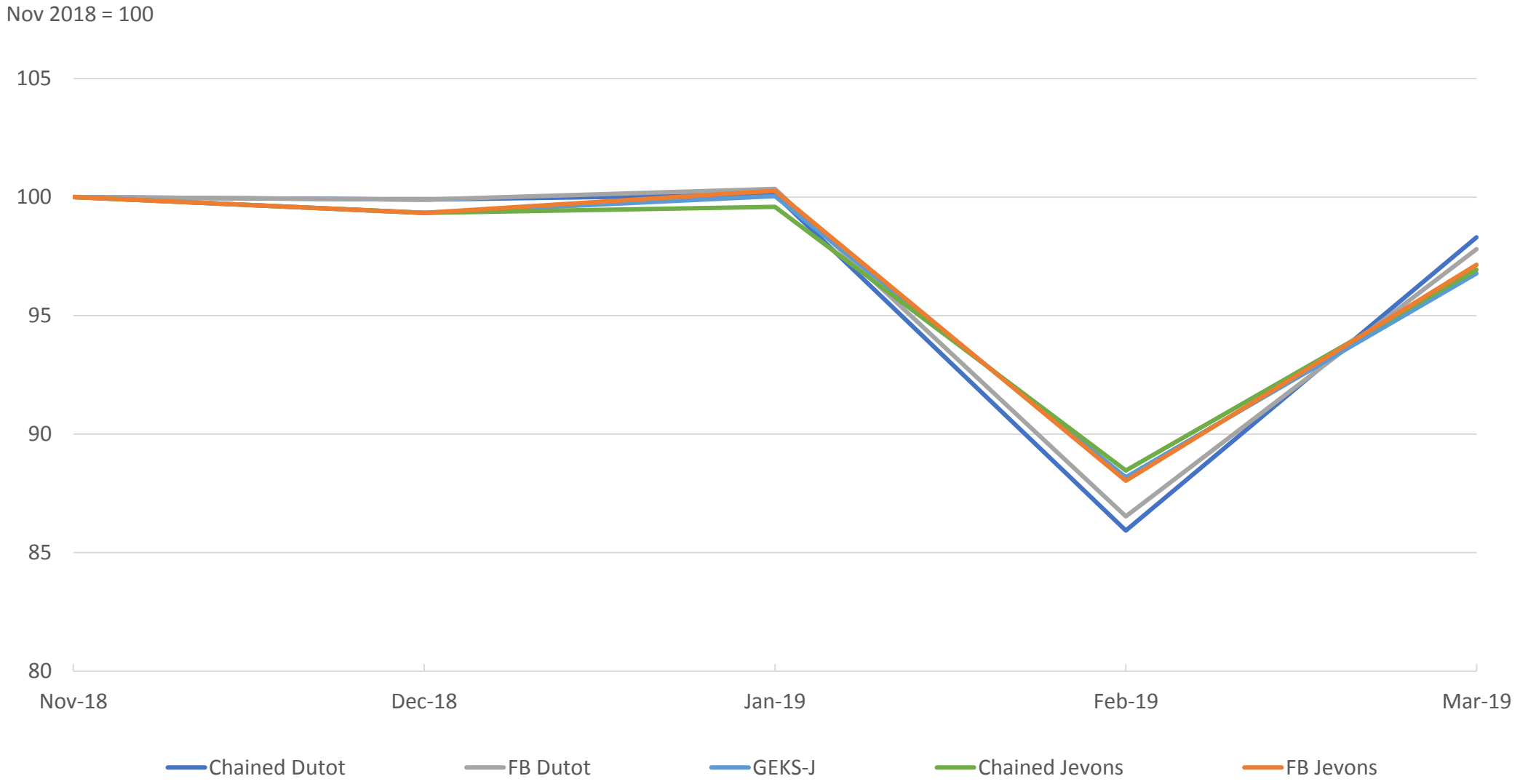


Office for National Statistics

Price distribution for laptops, Retailer A, January and February 2019

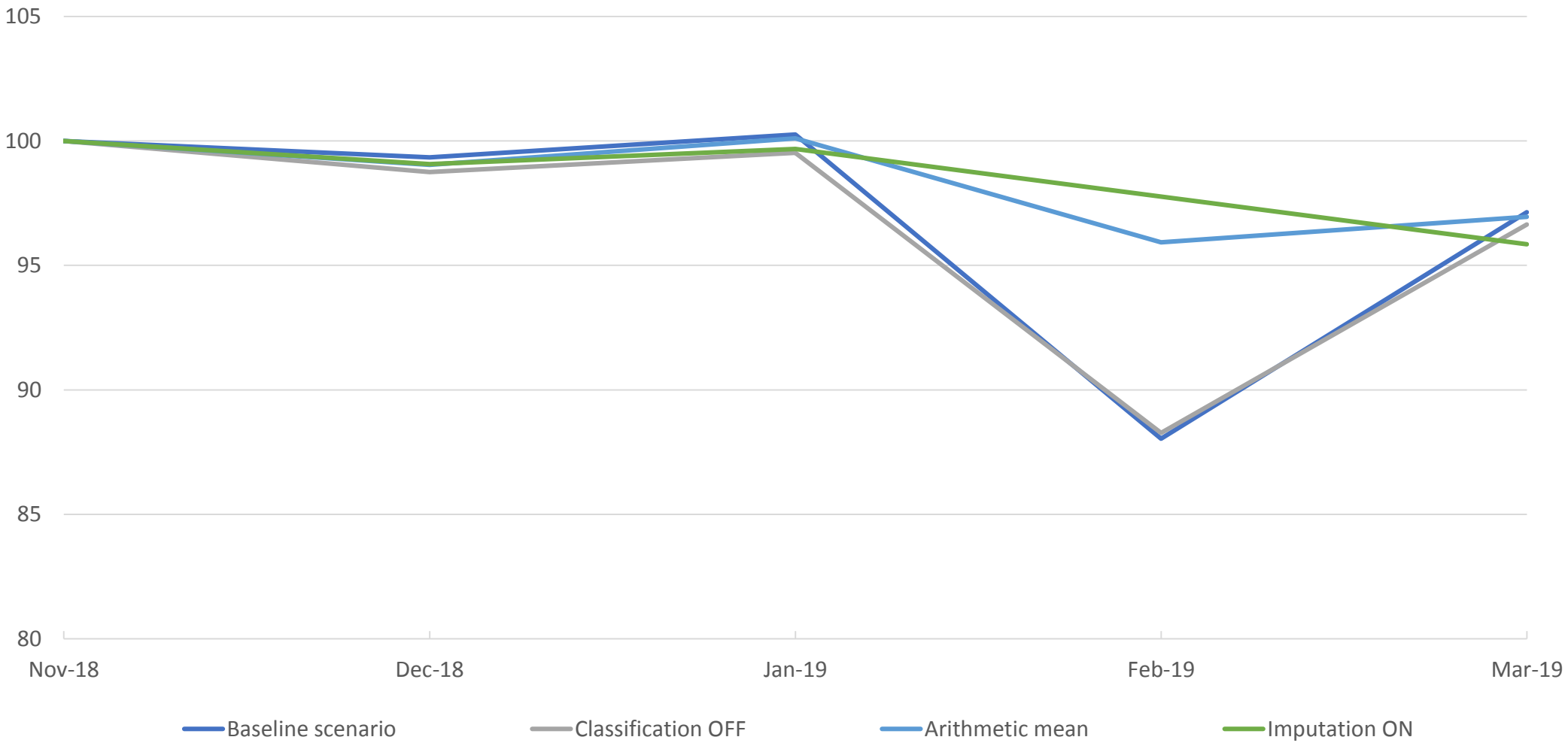


Price indices for laptops, different index methods



Fixed base Jevons indices for laptops, alternative scenarios

Nov 2018 = 100



Office for National Statistics