Clustering Large datasets into Price indices - CLIP

Matthew Mayhew

Index Numbers Methodology

Office for National Statistics

Overview

01 Web Scraping

02 Overcoming the Product Churn Issue

03 Finding the groups

04 New Data and Forming the Index

05 Results

06 Future Work

01

Web Scraping

Motivation for web scraping

- Consumer Prices Index including Owner Occupied Housing Costs (CPIH) is the most comprehensive measure of inflation in the UK
- Johnson Review published in January 2015, recommended increasing the use of alternative data sources in consumer prices

Web scraping in ONS

- Prices for 33 CPIH items from 3 online retailers

TESCO Sainsbury's Waitrose

- Daily collection (around 8,000 price quotes, compared to 6,800 a month for traditional collection)
- Collects price, product name and discount type
- Ongoing since June 2014

Limitations

- Market coverage

  Large retailers only, permission, regional variation?
- High product churn

  Traditional methods struggle
- Only prices not expenditure

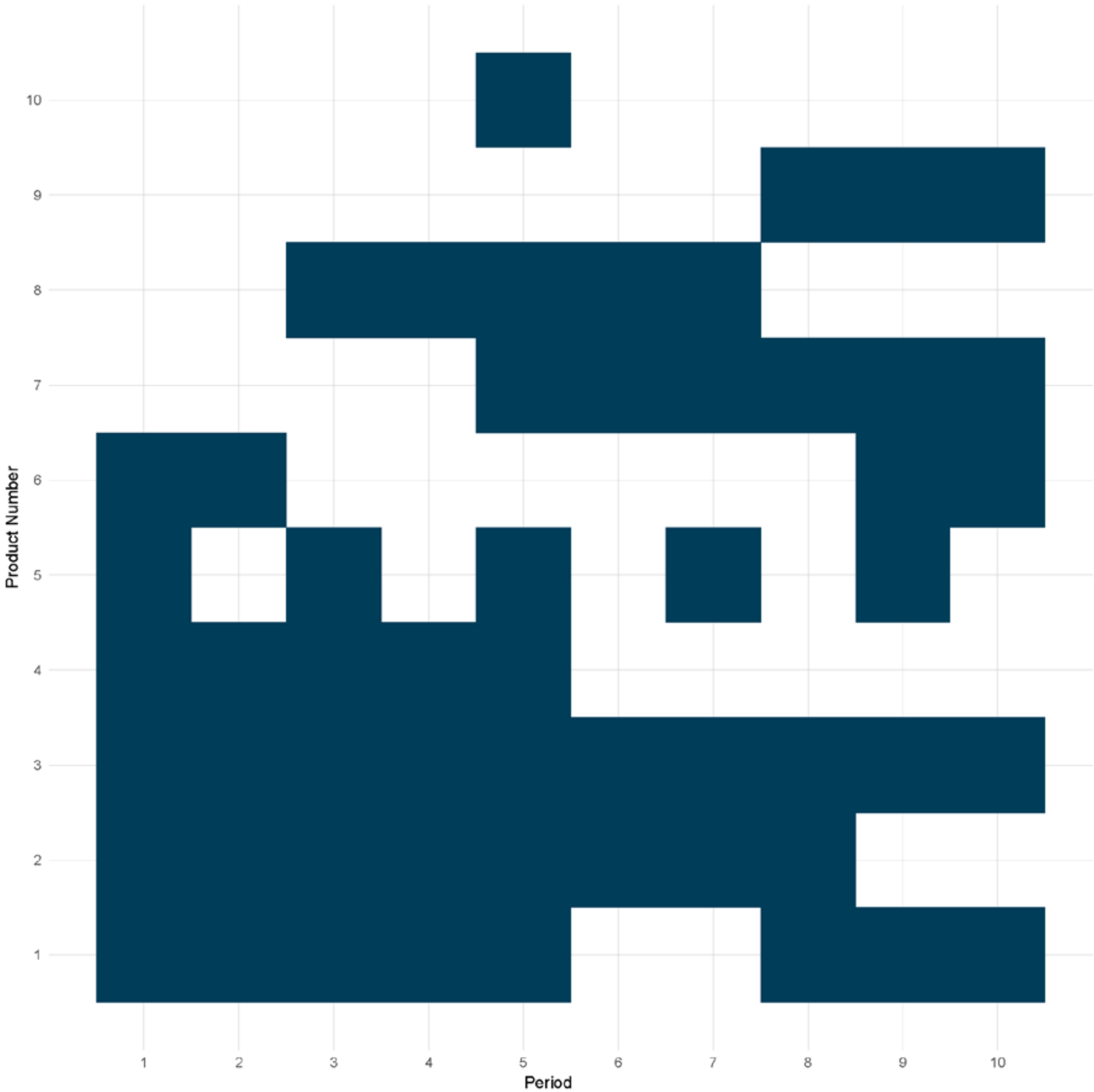
  What do people actually buy?
- Technological difficulties

  Scraper breaks, time and cost

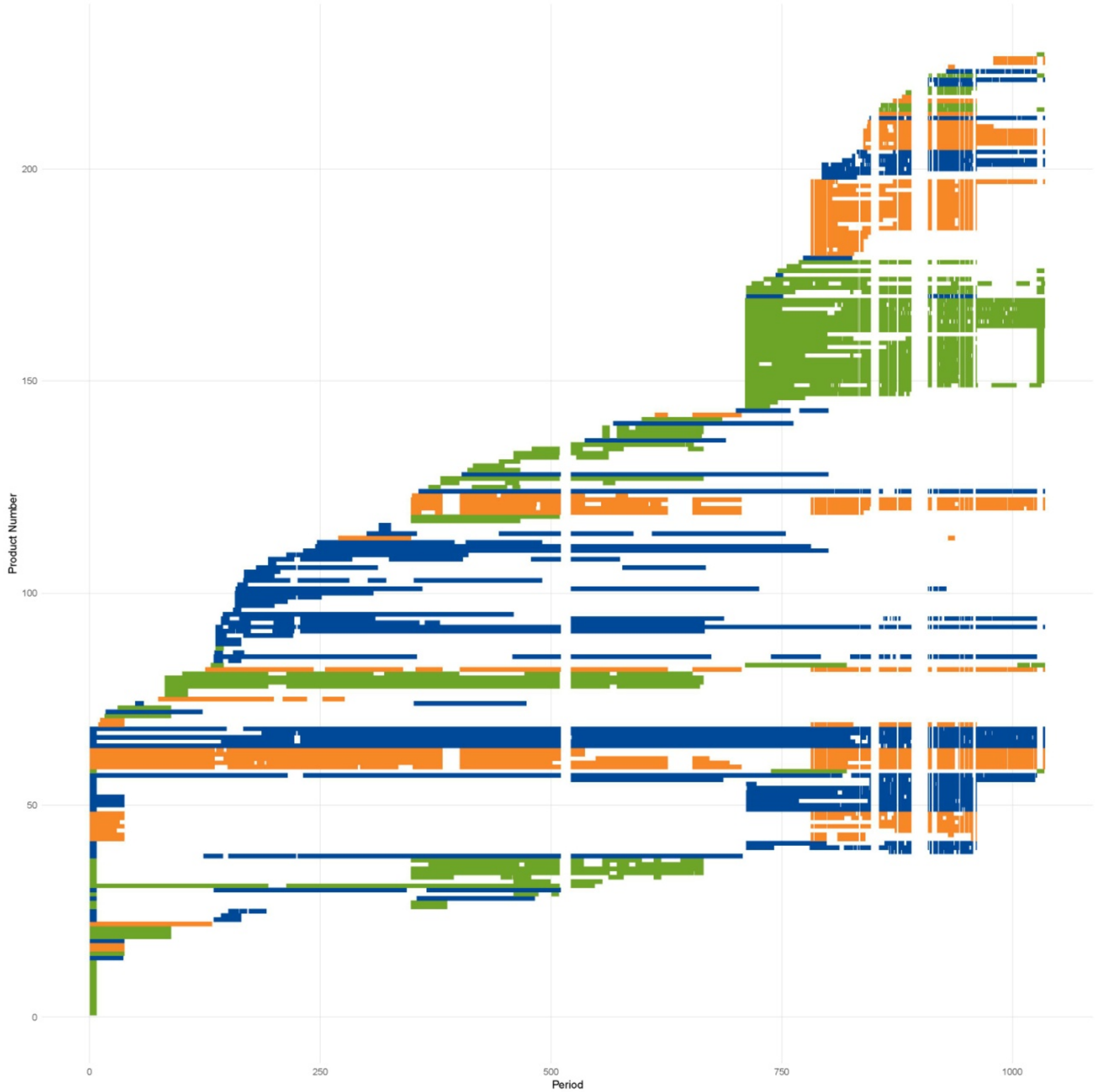
Product Churn

- Product Churn is the process of products leaving and/or entering the sample.
- This can either be:
  - Product goes out of stock, temporally leaves the sample,
  - Product is restocked, and reenters the sample,
  - Product is discontinued and permanently leaves the sample,
  - Product is new to the market
  - Products being rebranded

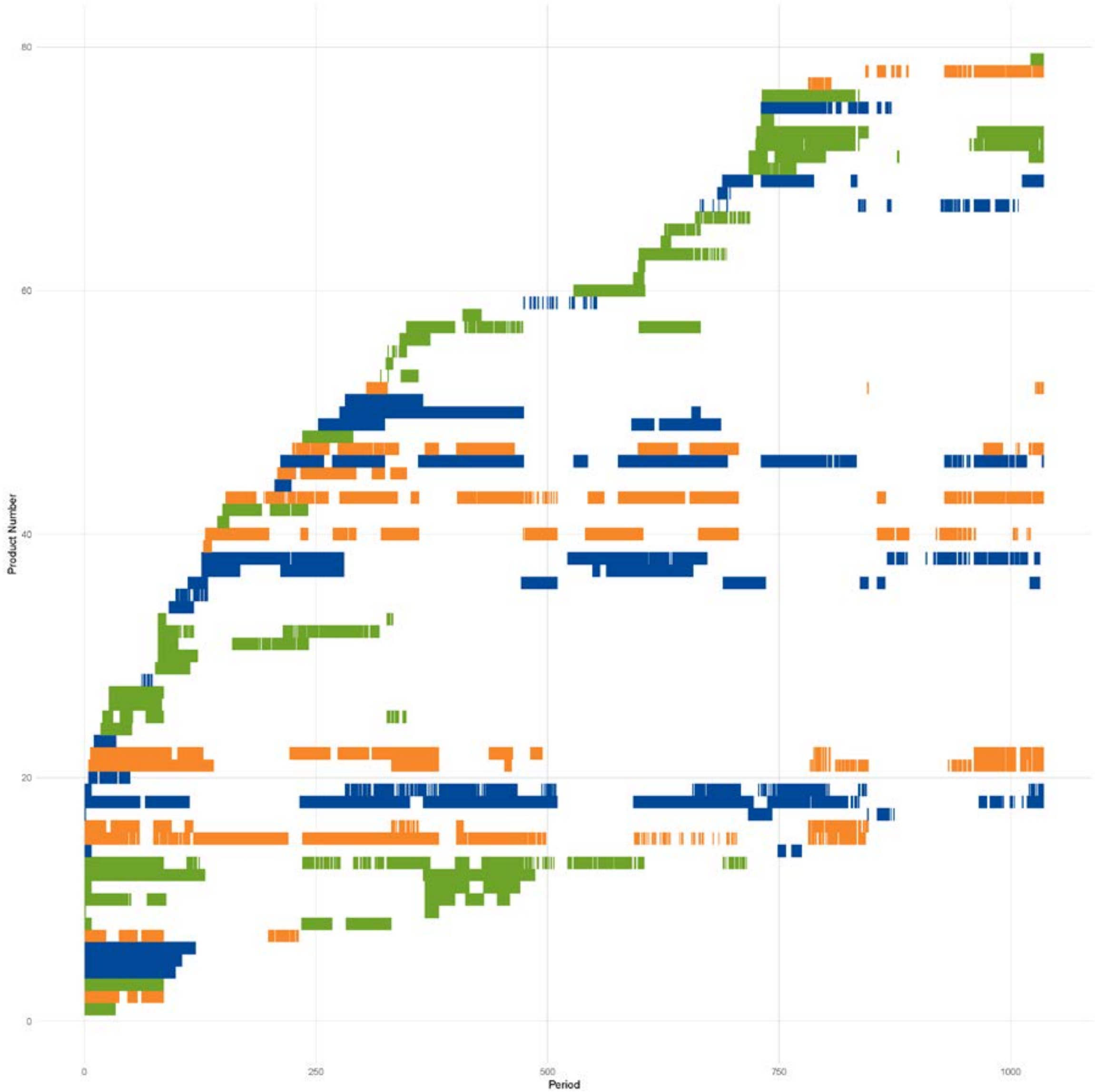
Product Churn – Example



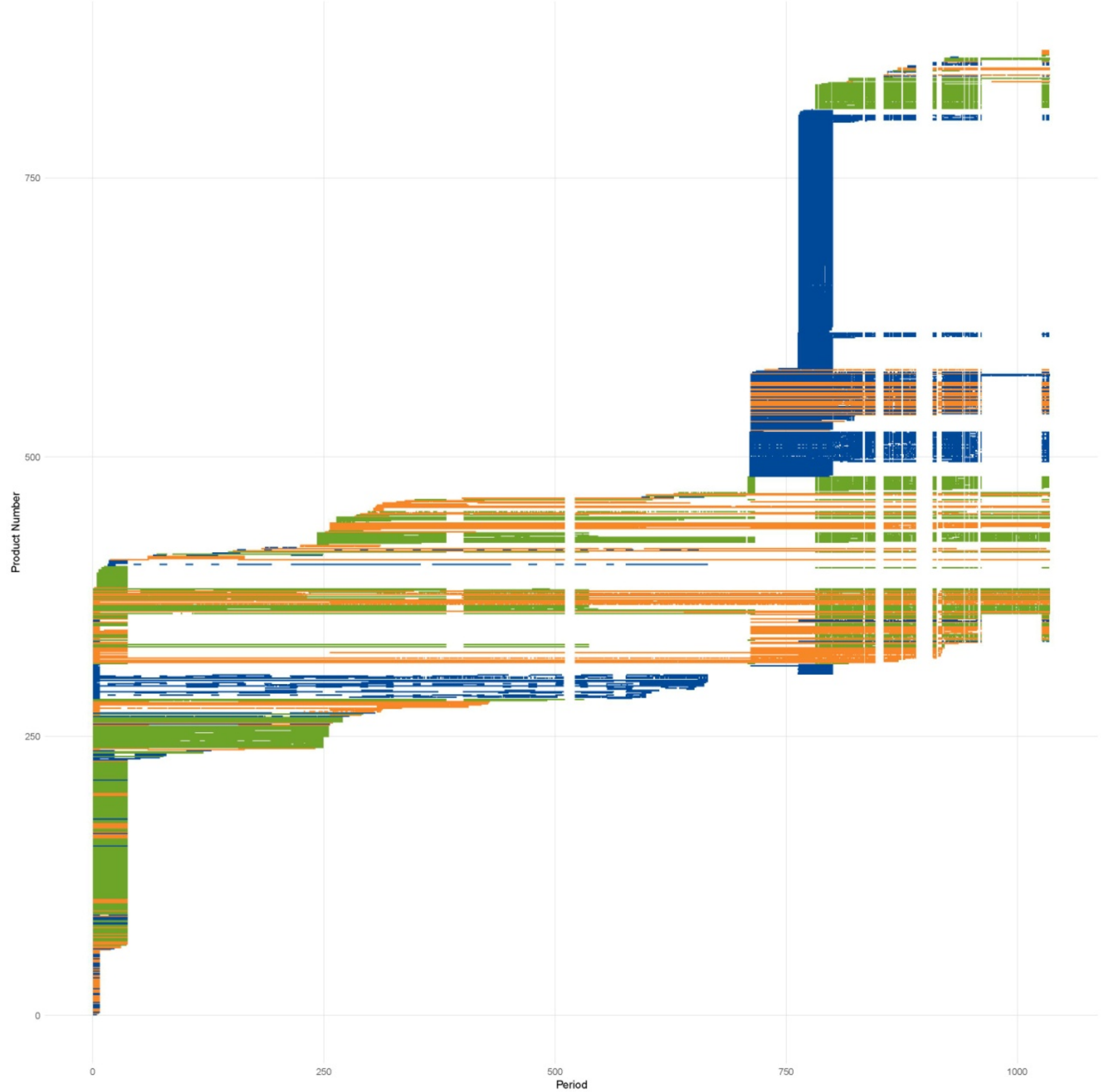
Product Churn - Apples



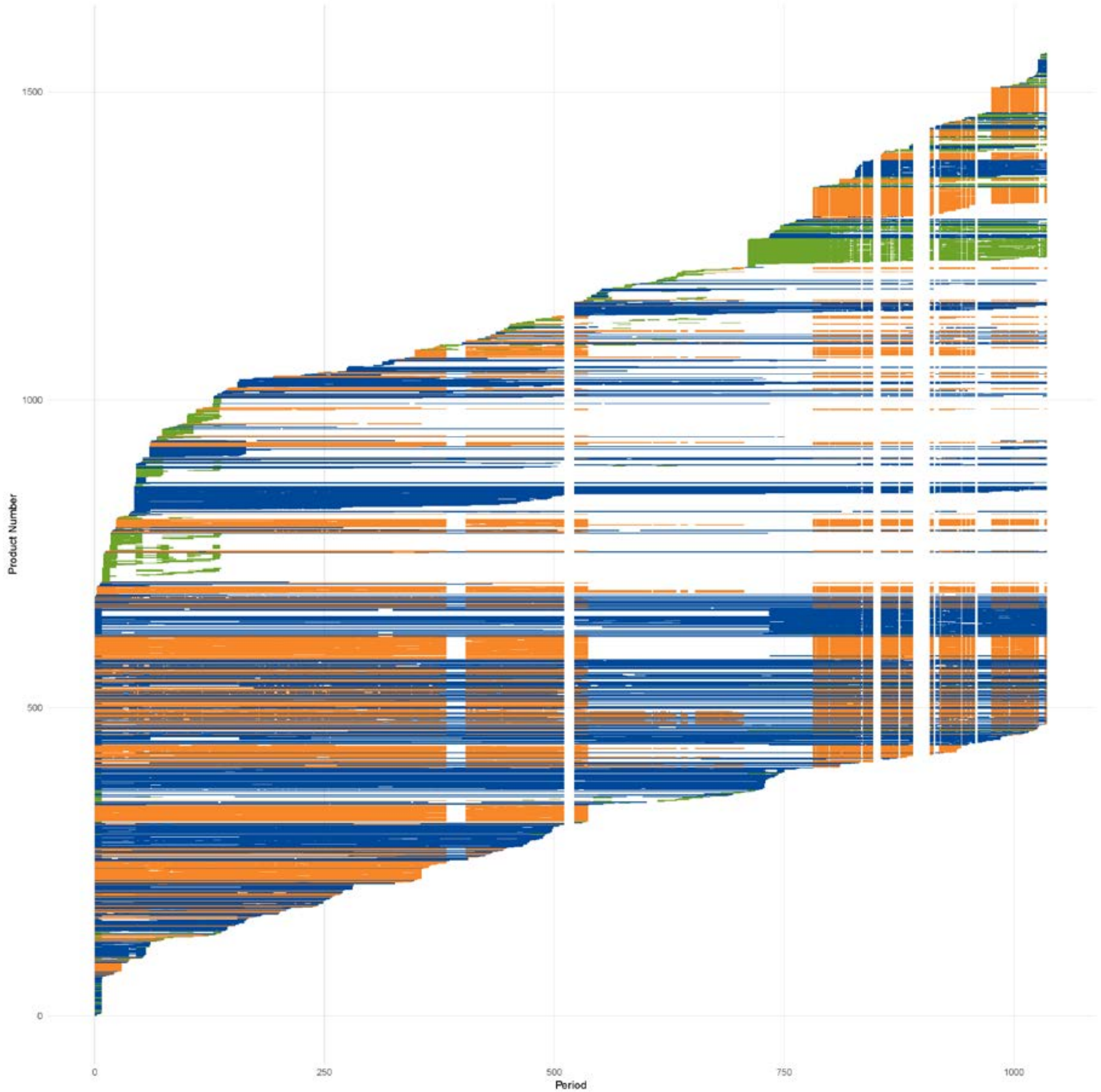
Product Churn - Strawberries



Product Churn - Tea



Product Churn – Red Wine



Overcoming the Product Churn Issue

02

Problems due to Product Churn

- With long datasets there is minimal chance of product being observed in every period, especially and high frequencies
- Causes problems with tradition methods

Possible Solutions

- Impute the missing prices in the appropriate period
  - ITRYGEKS
- Adjust for the change in quality due to the change in products on the market
  - FEWS
- Track groups of products over time
  - CLIP

Why track groups not products?

- Consumers have preferences.
- Preferences might be product specific, i.e.
  - Product A $\prec$ Product B
- Preferences might be characteristic specific instead
  - Characteristic 1 $\prec$ Characteristic 2

Why track groups not products?

- Therefore there might be a group of products who’s have the consumer’s preferred characteristics.
- The consumer would be indifferent to those products with their preferred characteristics
- This group is what is tracked over time

03

Finding the groups

How to find these groups?

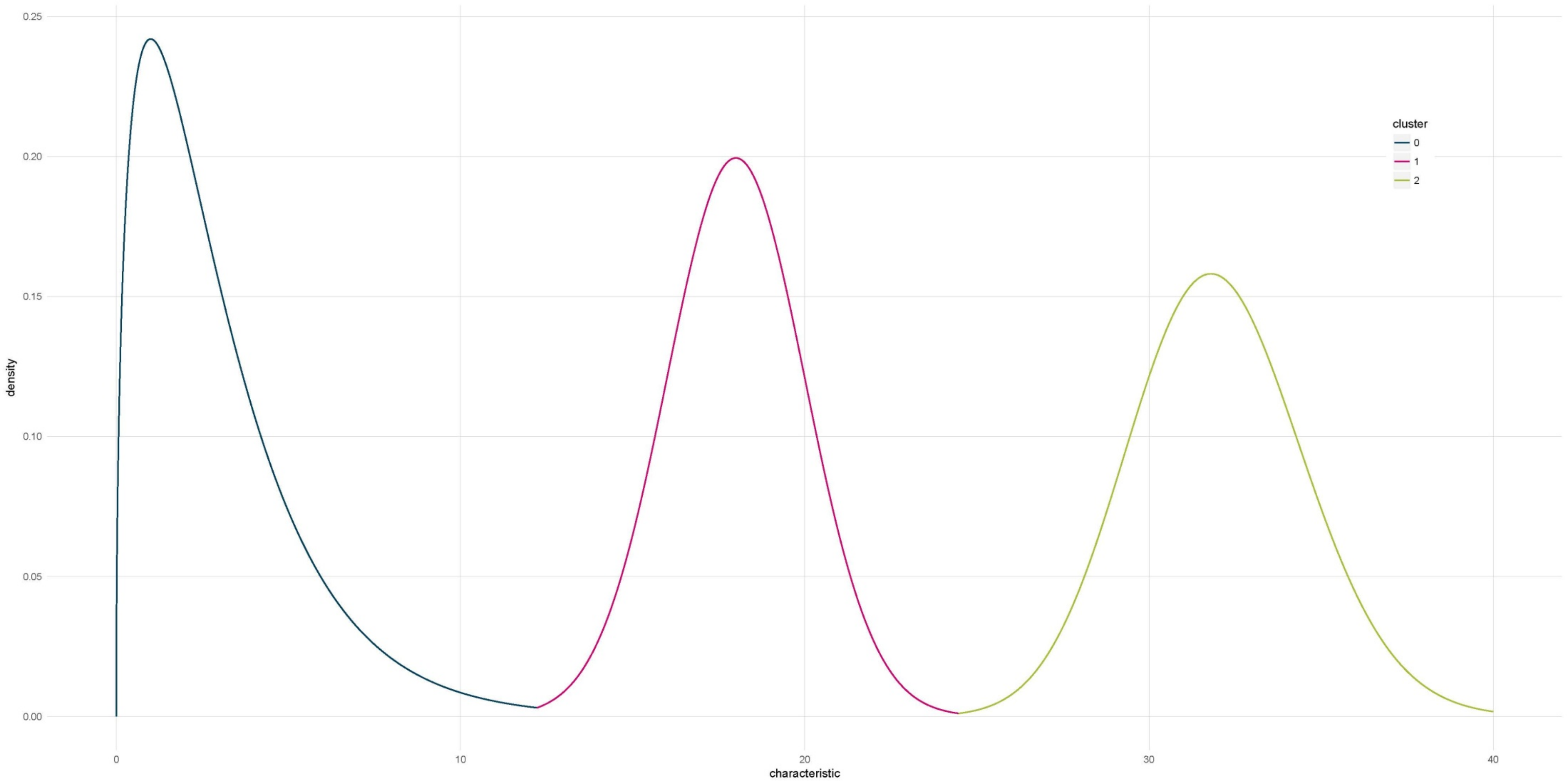
- Usually the preferences would be determined by finding utility functions and maximising under a budget constraint.

- Utility functions can’t be calculated with web scraped data – lacking quantity information

Groups by clustering

- Groups are instead found by clustering the products
- Clusters are found using the *Mean Shift* algorithm
- Mean Shift was used as no a priori choices about cluster shapes and number of clusters

Forming Clusters



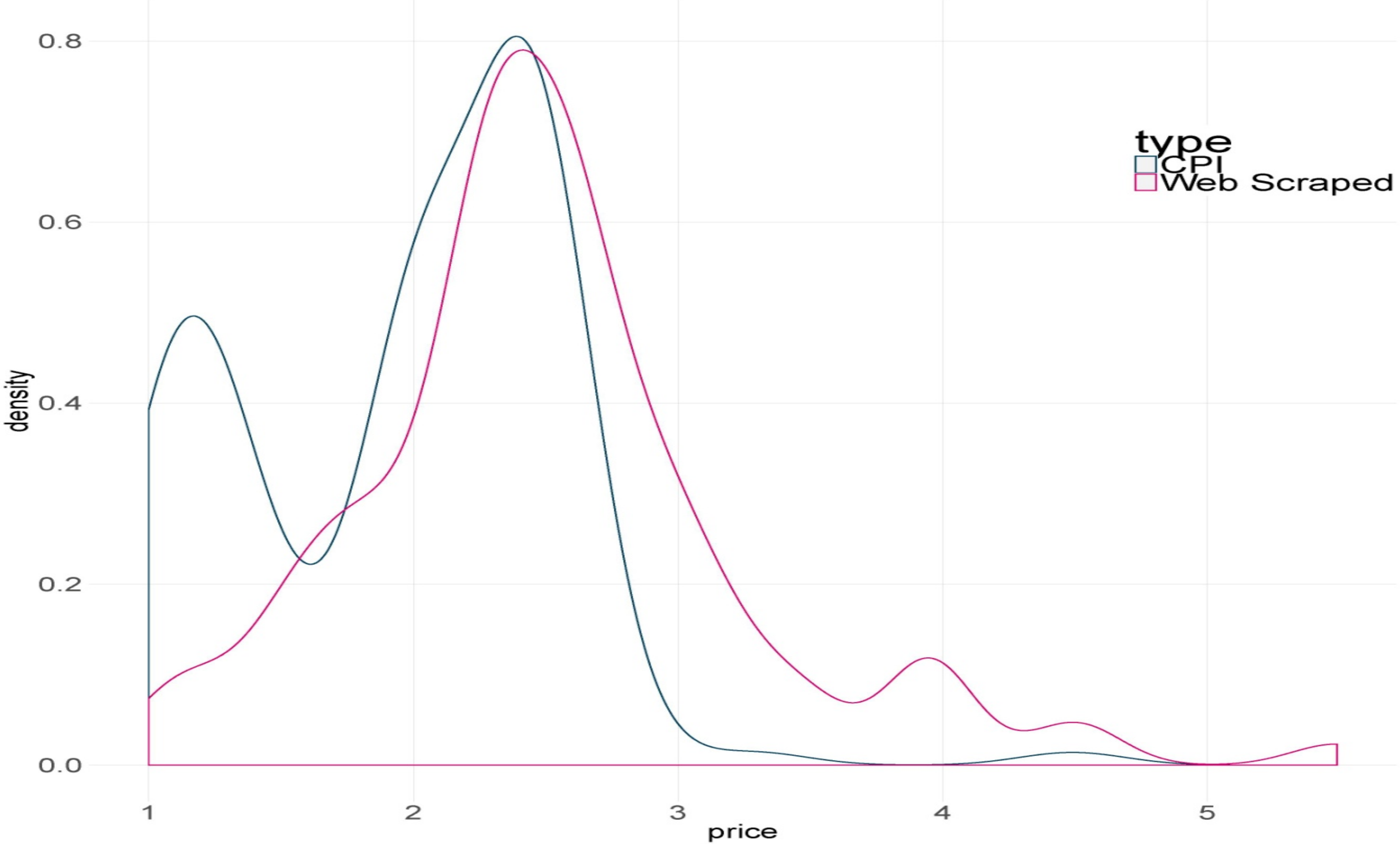
Characteristics used to form clusters

- Product Name
- Store
- Offer
- Price

Clustering - Tea

Clustering - Tea

Price Distributions



Clustering - Tea

New Data and Forming the Index

04

What to do with new data?

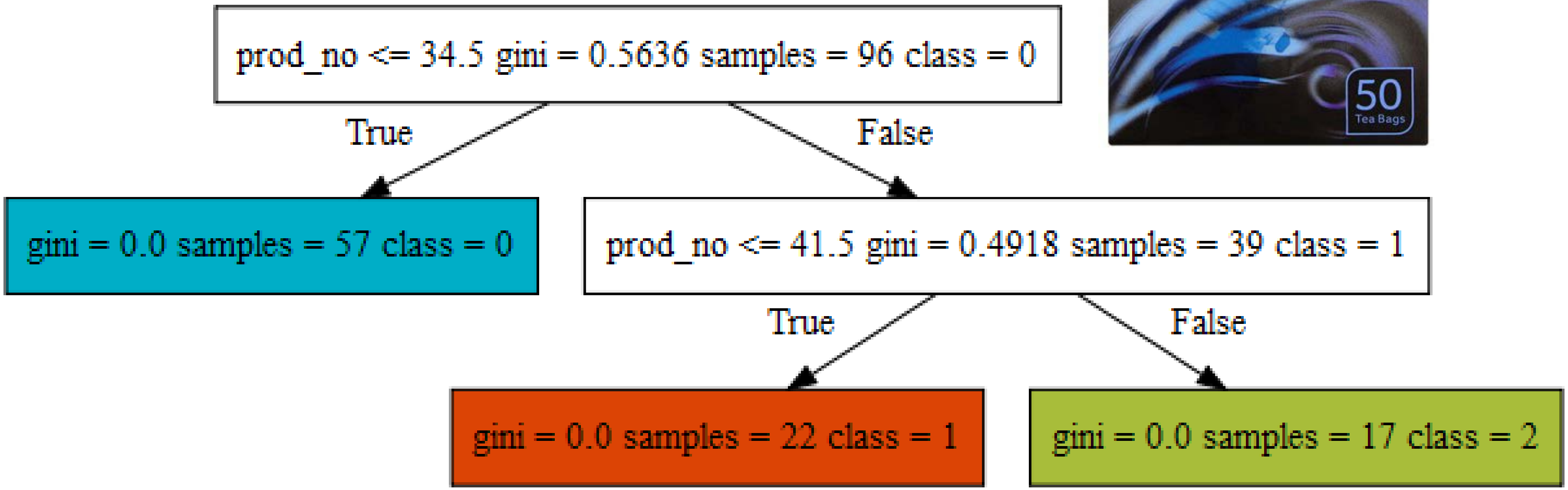
- Solution 1: Recluster the data
  - Problem completely new clusters will be found
- Solution 2: Assign Data to Clusters
  - This is done using a decision tree

Assigning Data

- The decision tree finds the underlying rules that make up the cluster.
- Price is removed as a characteristic when finding the rules.
- In subsequent months when new data is collect the products are the classified using this tree
- The product mix in each cluster will vary but the cluster itself is the same

Decision Tree

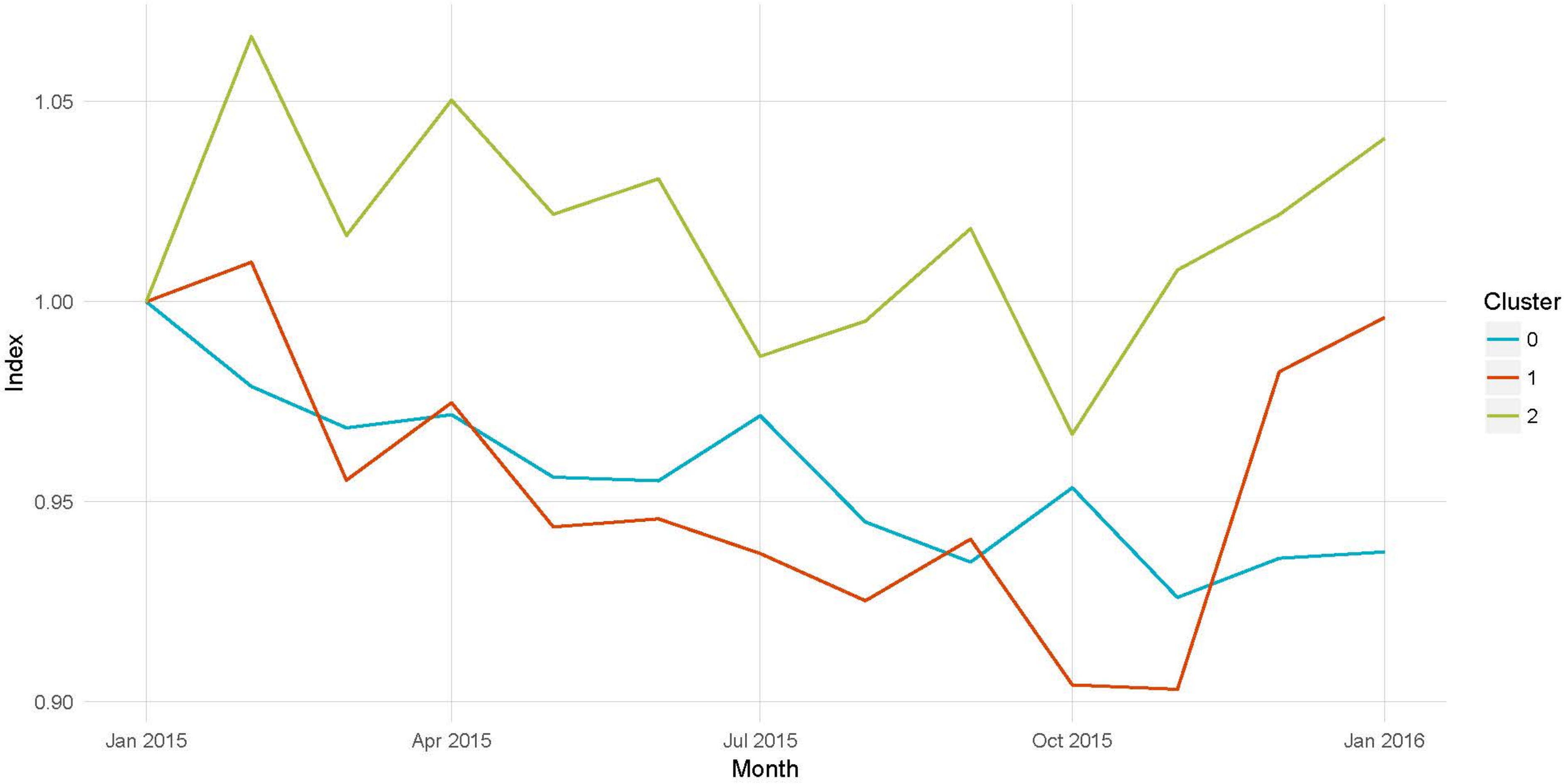
Characteristics:
Product Number = 37
Store = Tesco
Offer = NA



Forming the Index

- The price for a specific cluster is calculated as the geometric mean of the products in that cluster.
- The price for that cluster is then compared to the price for that cluster in the base month.

Price Relatives Per Cluster



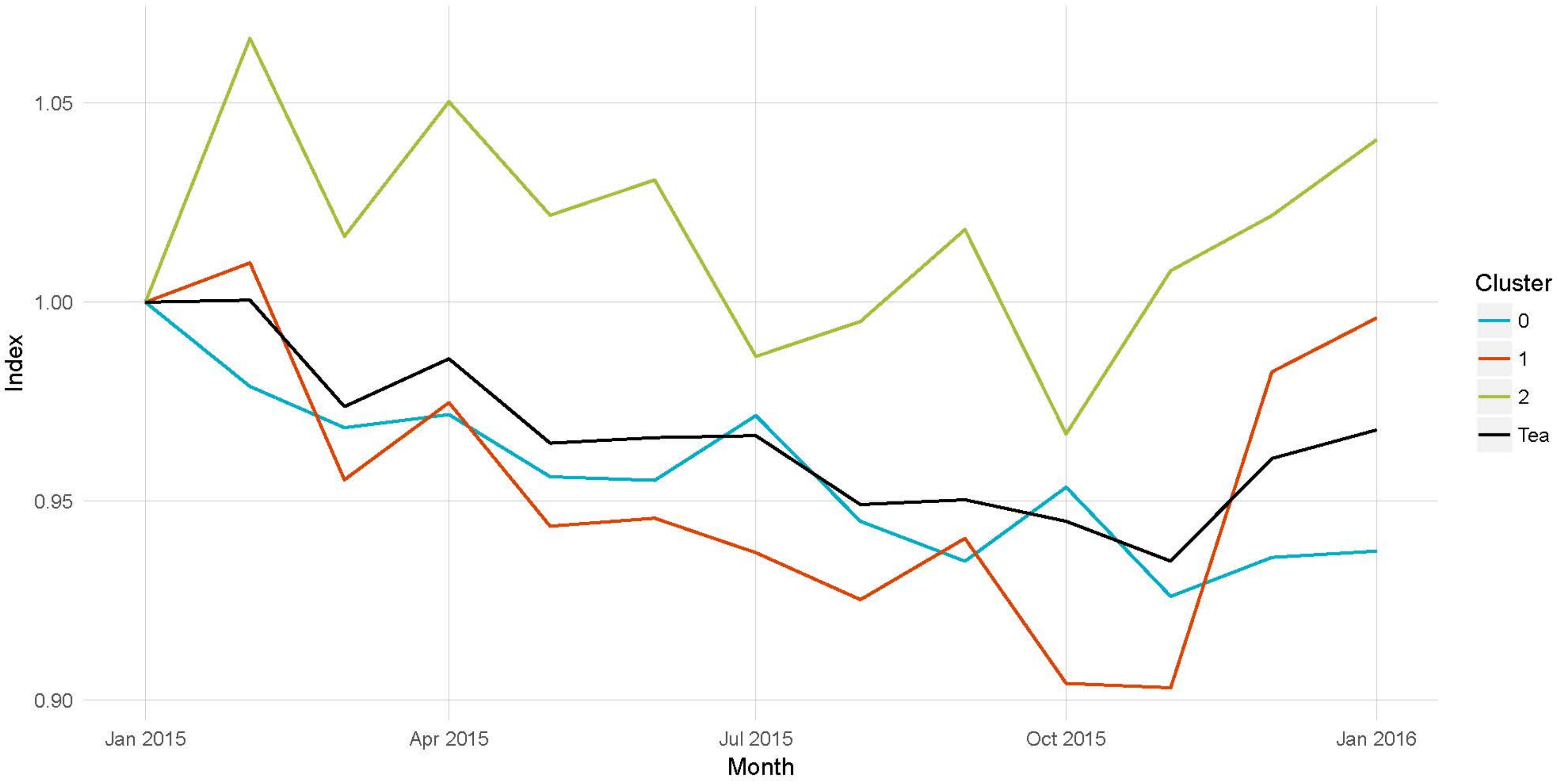
Aggregating over cluster

- The Price relatives are then aggregated over clusters to form the item index.
- These are weighted together with the following weights:

$$w_i = \frac{|C_i^0|}{\sum_k |C_k^0|}$$

- So for this Tea Data w_0=0.61, w_1=0.22 and w_2=0.17

Tea CLIP

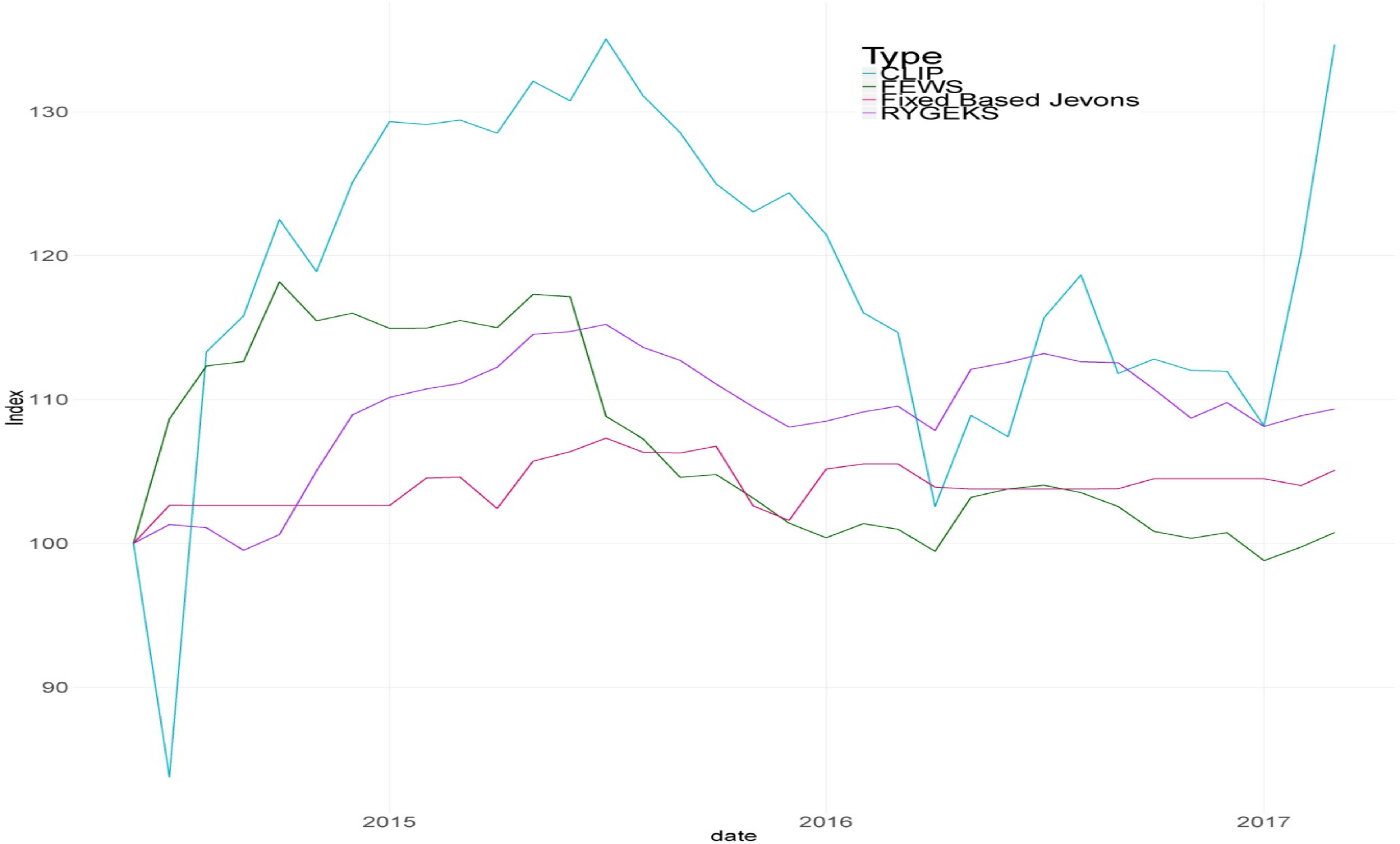


1.05
1.00
0.95
0.90
Index
Jan 2015
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Month
Cluster
0
1
2
Tea

05

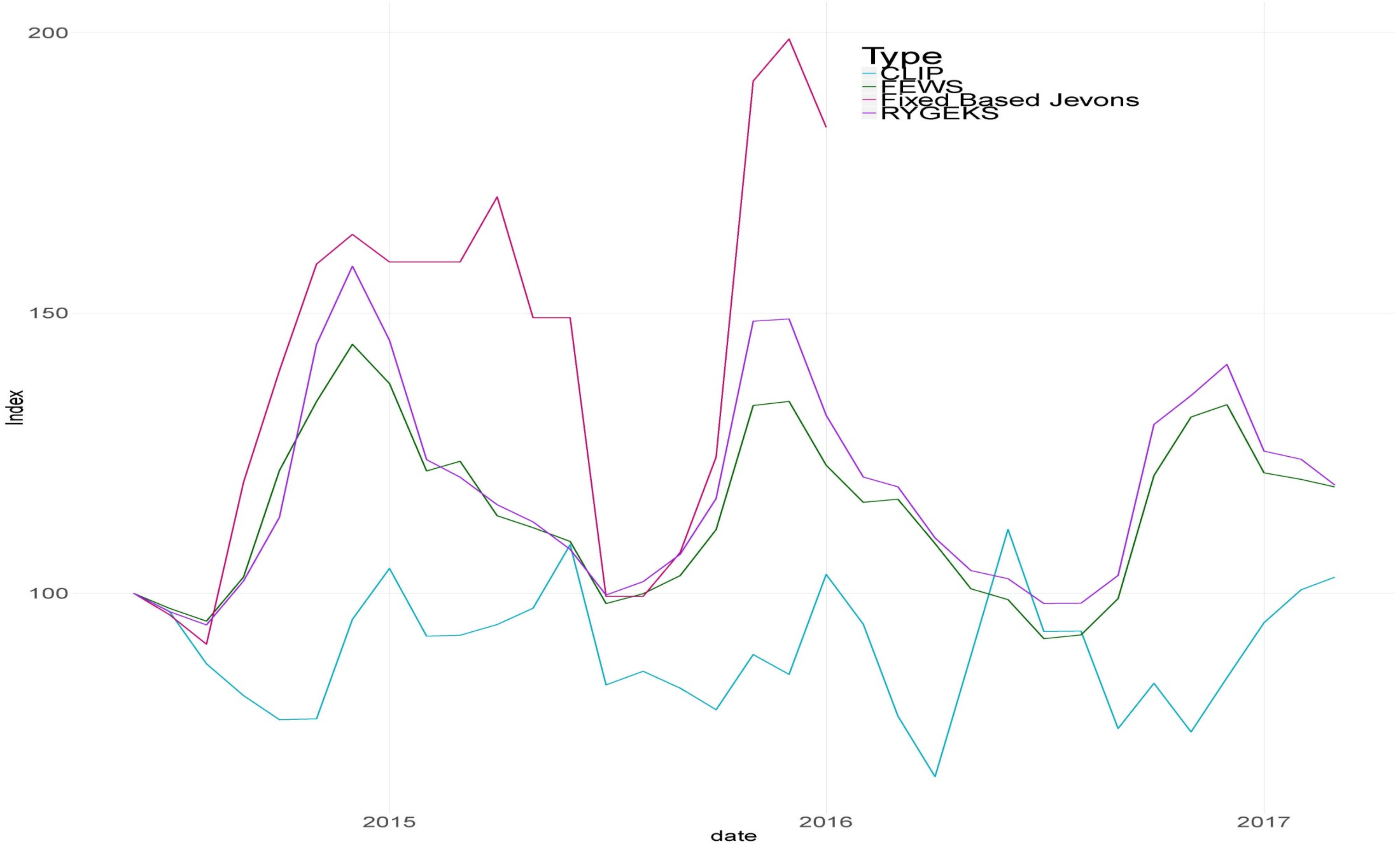
Results

Apples

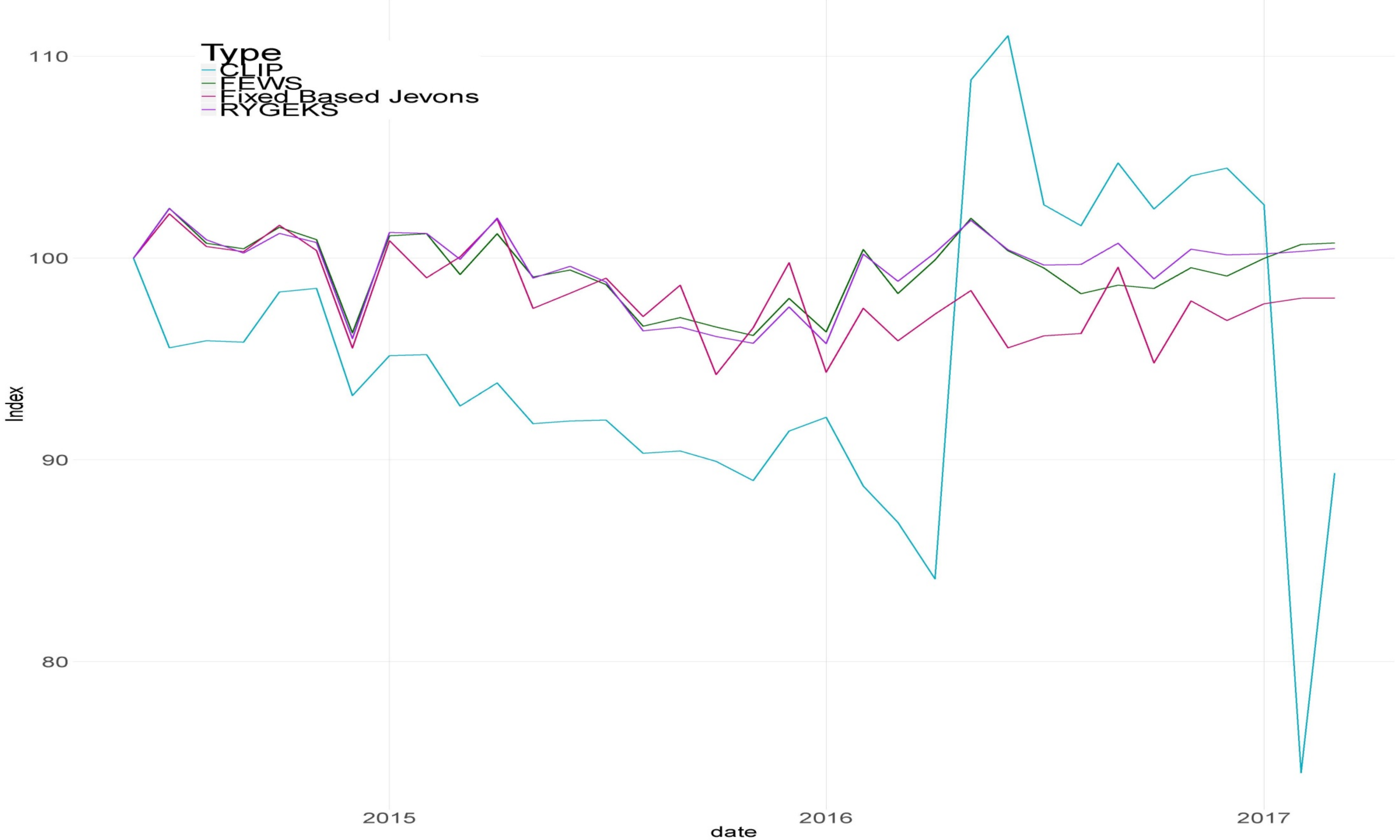


Type
CLIP
FEWS
Fixed Based Jevons
RYGEKS
Index
130
120
110
100
90
2015
2016
2017
date

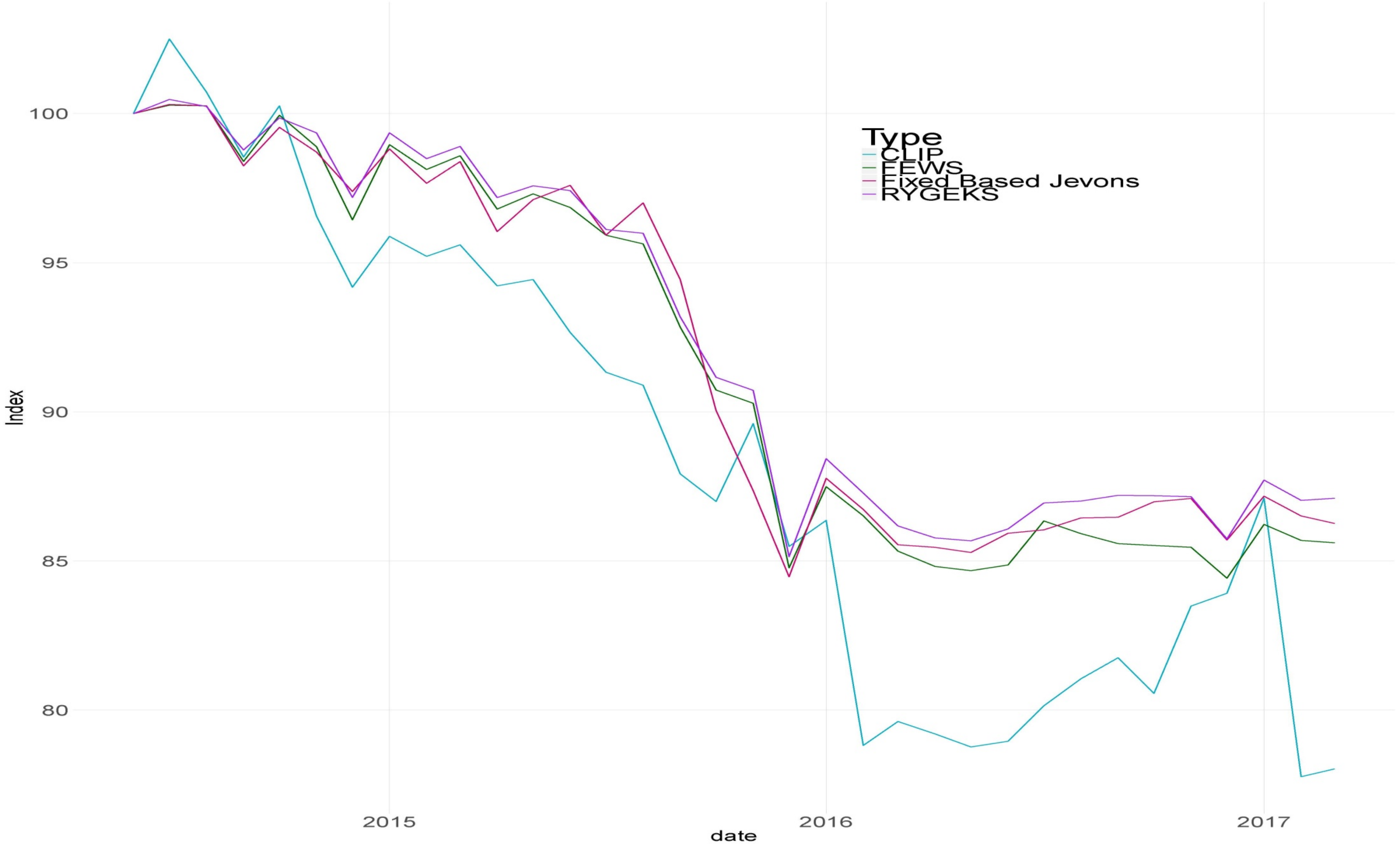
Strawberries



Tea


Type
CLIP
FEWS
Fixed Based Jevons
RYGEKS
110
100
90
80
Index
2015
2016
2017
date

Red Wine


Type
CLIP
FEWS
Fixed Based Jevons
RYGEKS
100
95
90
85
80
Index
2015
2016
2017
date

Future Work

06

Assessing against approach to Index Numbers

- Assessed against the Test/Axiomatic approach only fails the identity, time reversal and Price Bounce tests (Note: FEWS does as well)
- To do:
  - Economic Approach
  - Statistical Approach

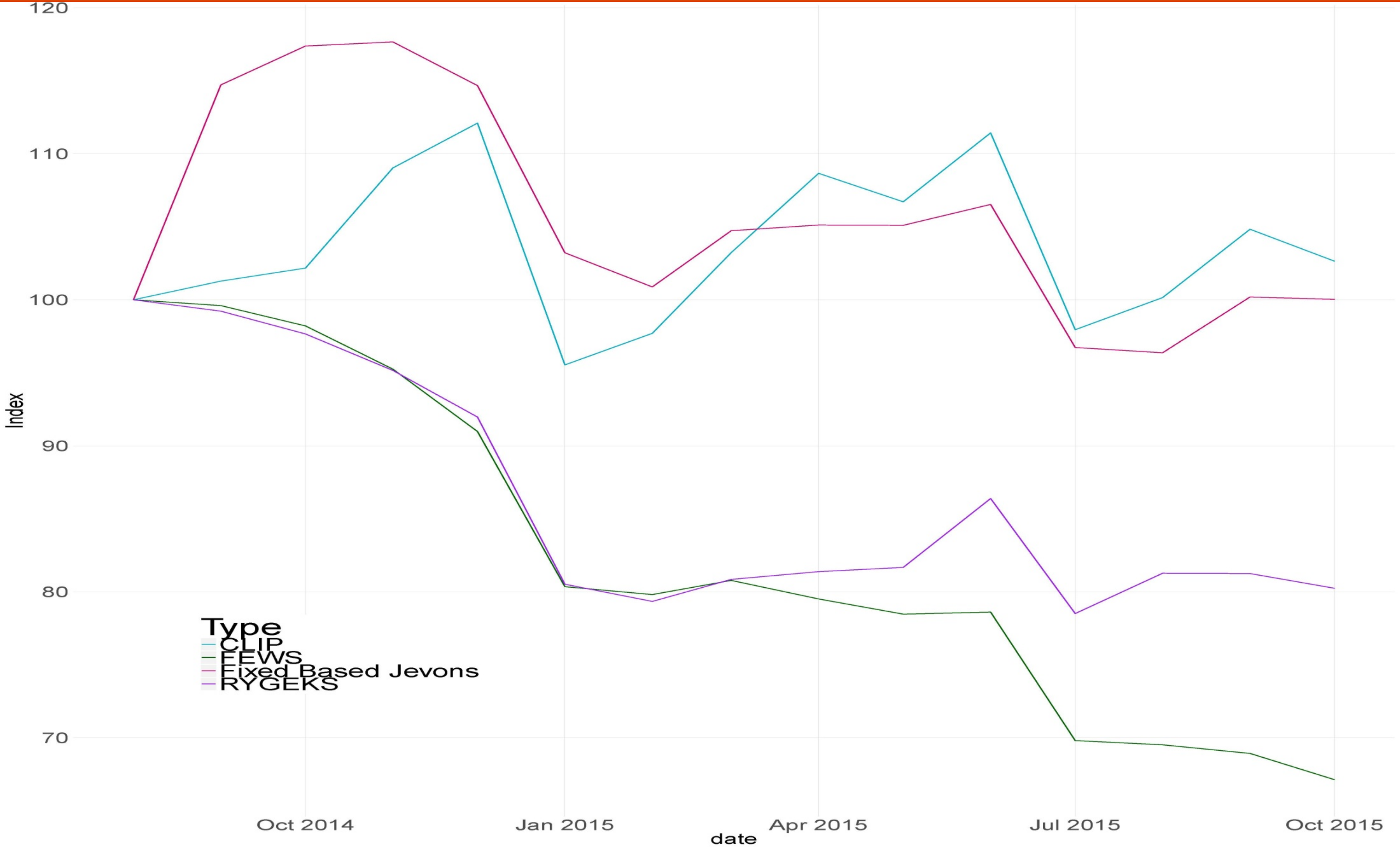
Test Assumptions about Substitution

- Do consumers substitute within clusters?
- Do consumers substitute between clusters?

Clothing and other forms

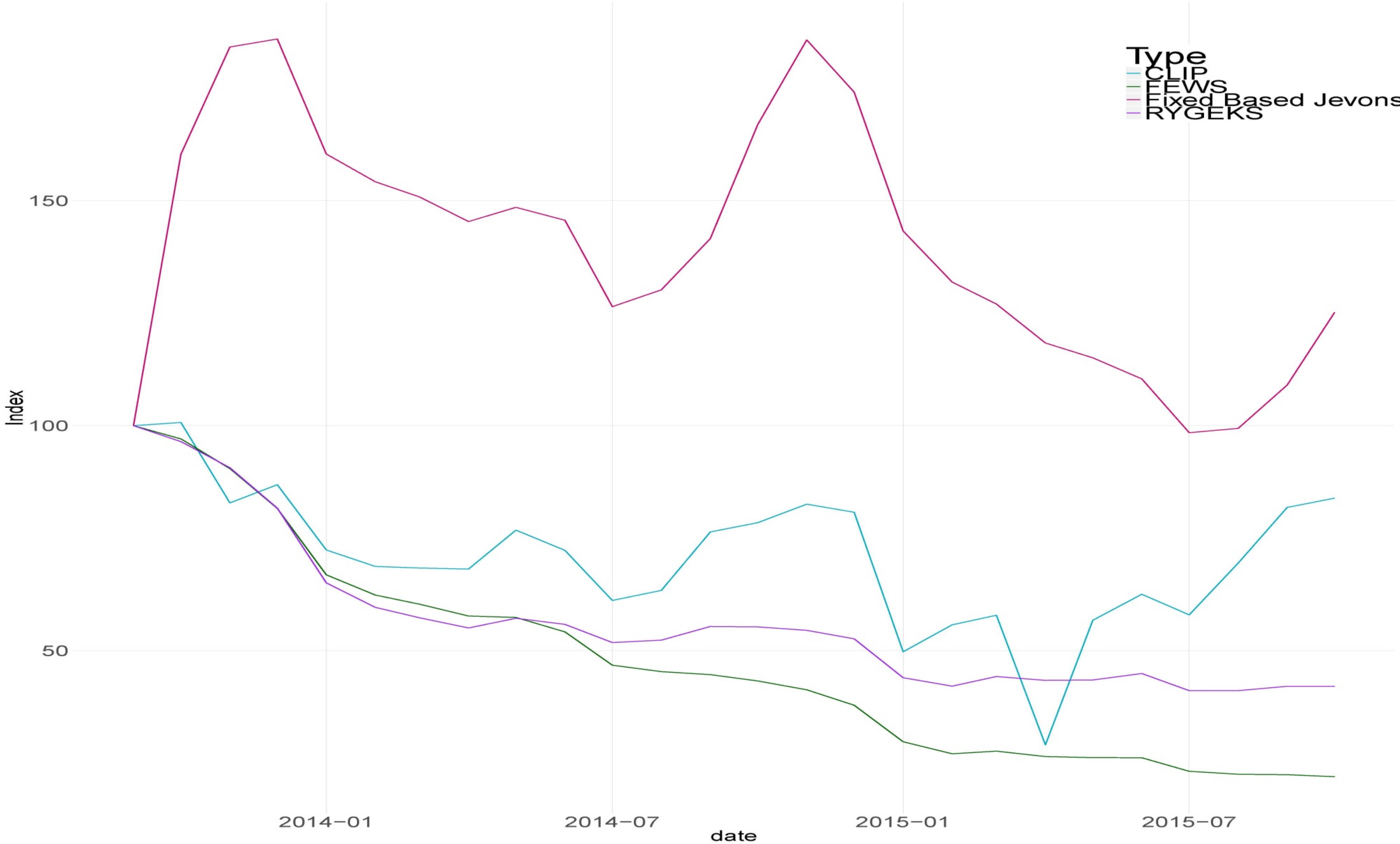
- CLIP might be more suited to Clothing Items
  - ONS is to release research into this
- Testing a geometrically aggregated CLIP as well as other variants of the index

Men’s Jeans



120
110
100
90
80
70
Index
Type
CLIP
FEWS
Fixed Based Jevons
RYGEKS
Oct 2014
Jan 2015
Apr 2015
Jul 2015
Oct 2015
date

Women’s coats



More Information

- More information on the CLIP along with more results can be found on the Office For National Statistics website.
  - https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip

Questions?

- Contact Details
- Matthew.mayhew@ons.gov.uk
- methodology@ons.gov.uk
- For CPIH enquiries please contact
- CPI@ons.gov.uk