Modernising the measurement of clothing price indices using web scraped data: classification and product grouping

Liam Greenhough

Office for National Statistics, United Kingdom

9th June 2022

Clothing data and goal

- Goal to introduce web scraped clothing data into consumer price statistics
- Scraping 500,000 unique products per month

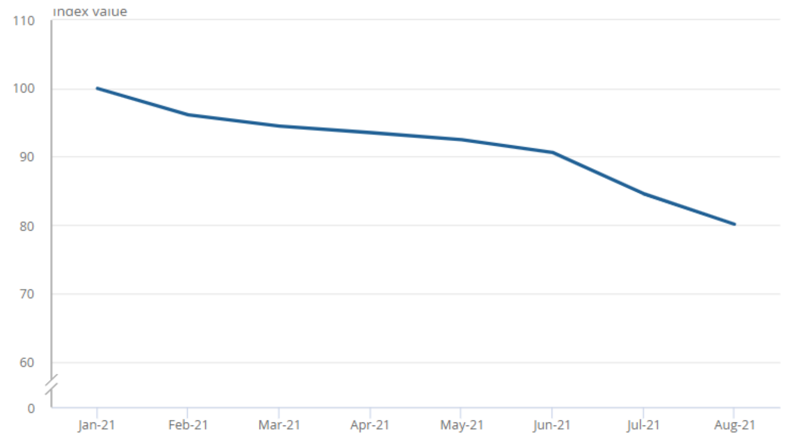
Key problem: churn

Clothing: ~30% monthly churn!

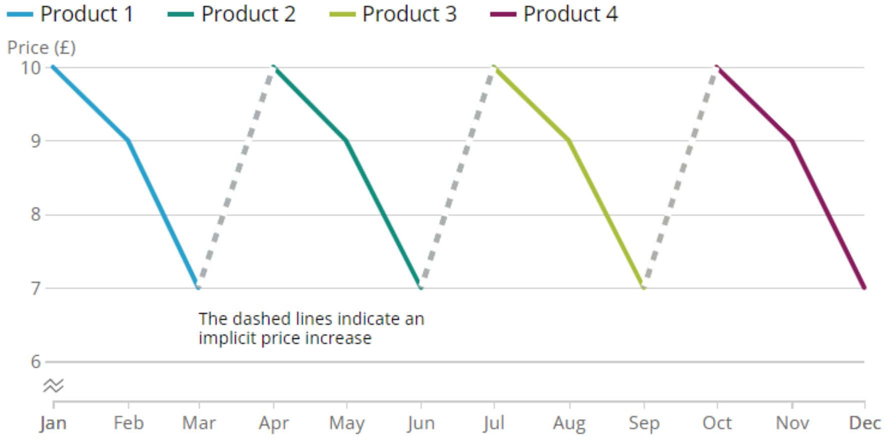
Problems:

- Too many data to classify
- Implicit price increases not captured

Index for women's dress



Leading to rapidly falling indices

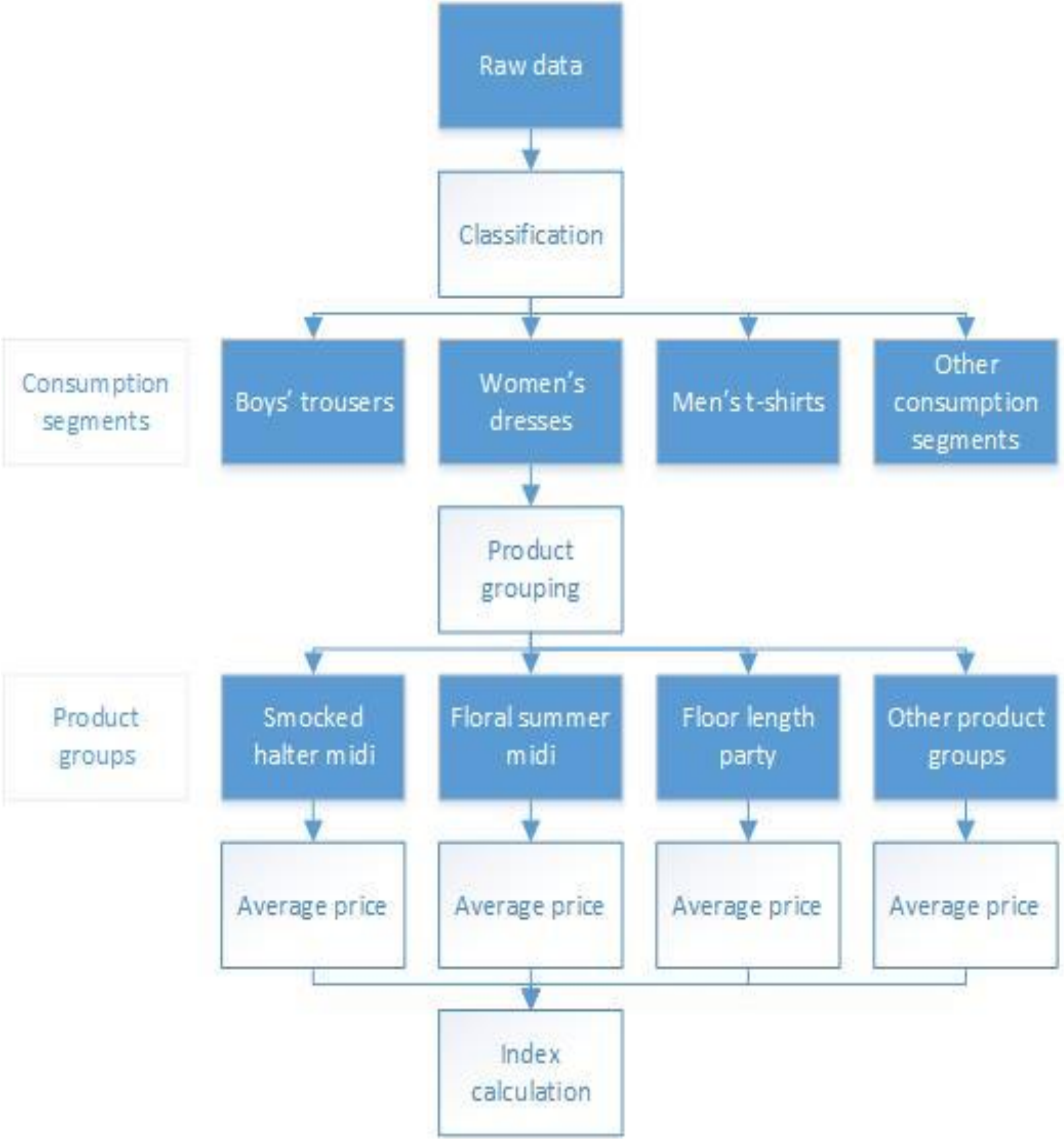


Clothing summary

Perform classification then product grouping:

- Classification: supervised machine learning assigns products to consumption segments which are used as elementary aggregates

- Product grouping: group together similar products within consumption segments, use average prices as inputs into index calculations



Classification

Classification lessons learnt

Topic	Description	Lessons learnt
Labelled datasets	Crowd-sourcing Use of an application	Crowd-sourcing improves quantity; application improves quality!
Feature creation	FastText Text-mined (e.g. regex) age/gender	FastText: similar words = similar vectors Text-mining: for “key” features
Data augmentation	SMOTE	Augments smaller classes so algorithm treats classes with increased importance
Favoured algorithm	XGBoost	Confidence scores; quite fast to fit with GPU; high performing.
Quality	Labelling quality important	[See next slide]

Label consistency

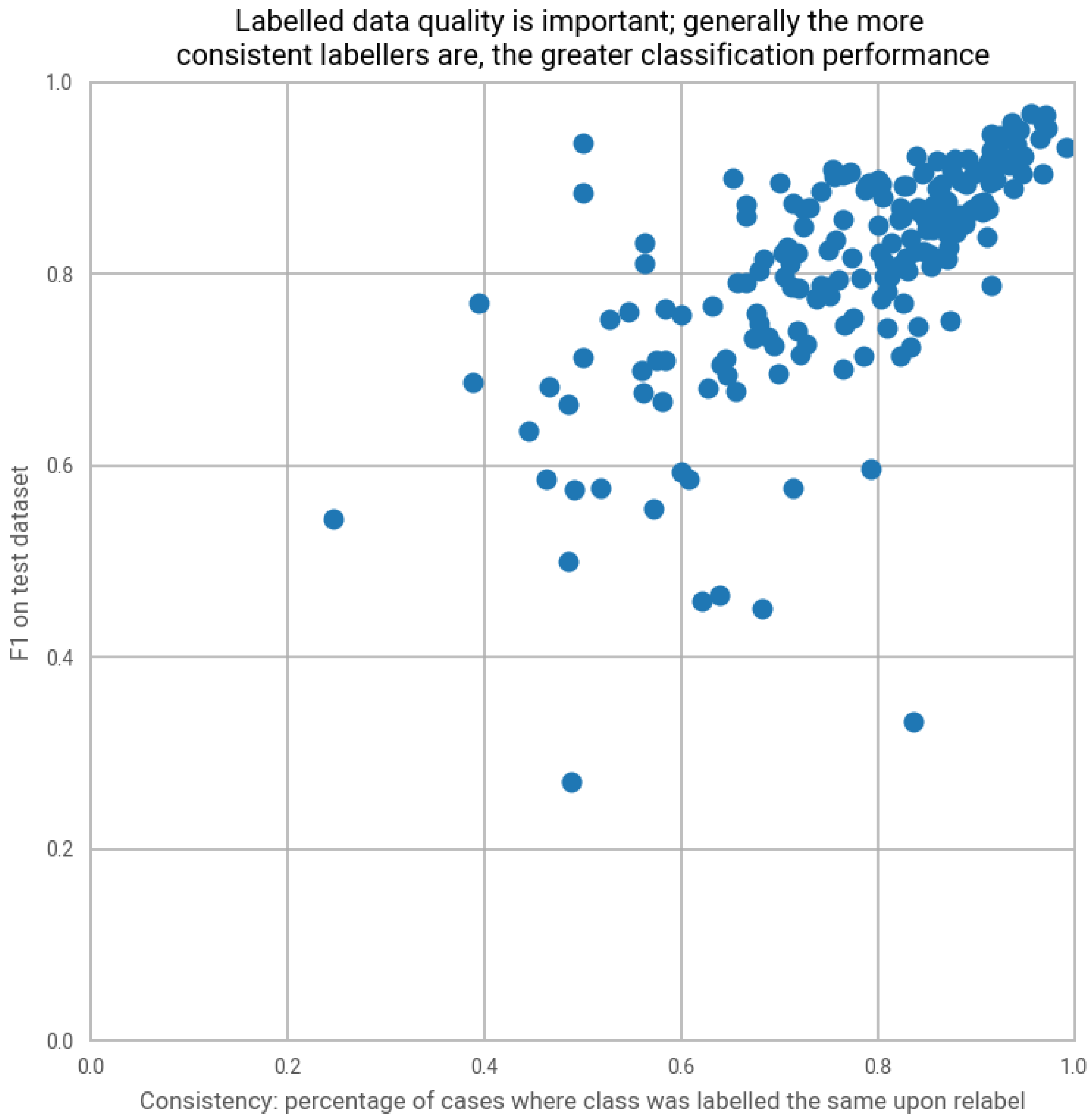
(“sweater”: “sweatshirt” or “jumper”?)

Started with smaller experiment: 12 labellers labelling same 313 products. (Findings in paper.)

Expanded experiment to labelling 30,000 products twice. Measured consistency:

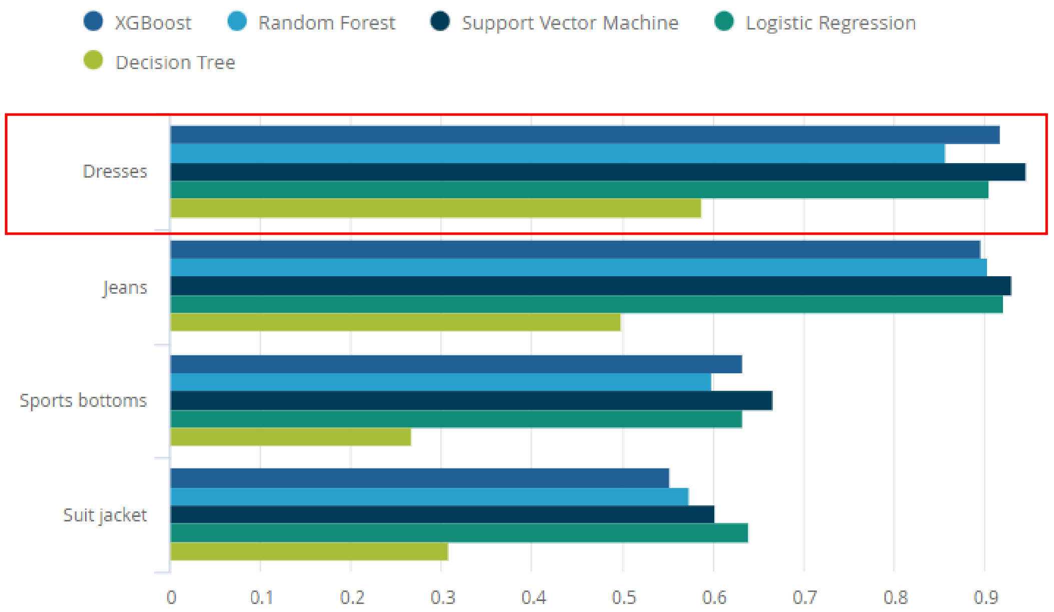
$$Consistency = \frac{Number\ products\ labelled\ same}{Number\ of\ products}$$

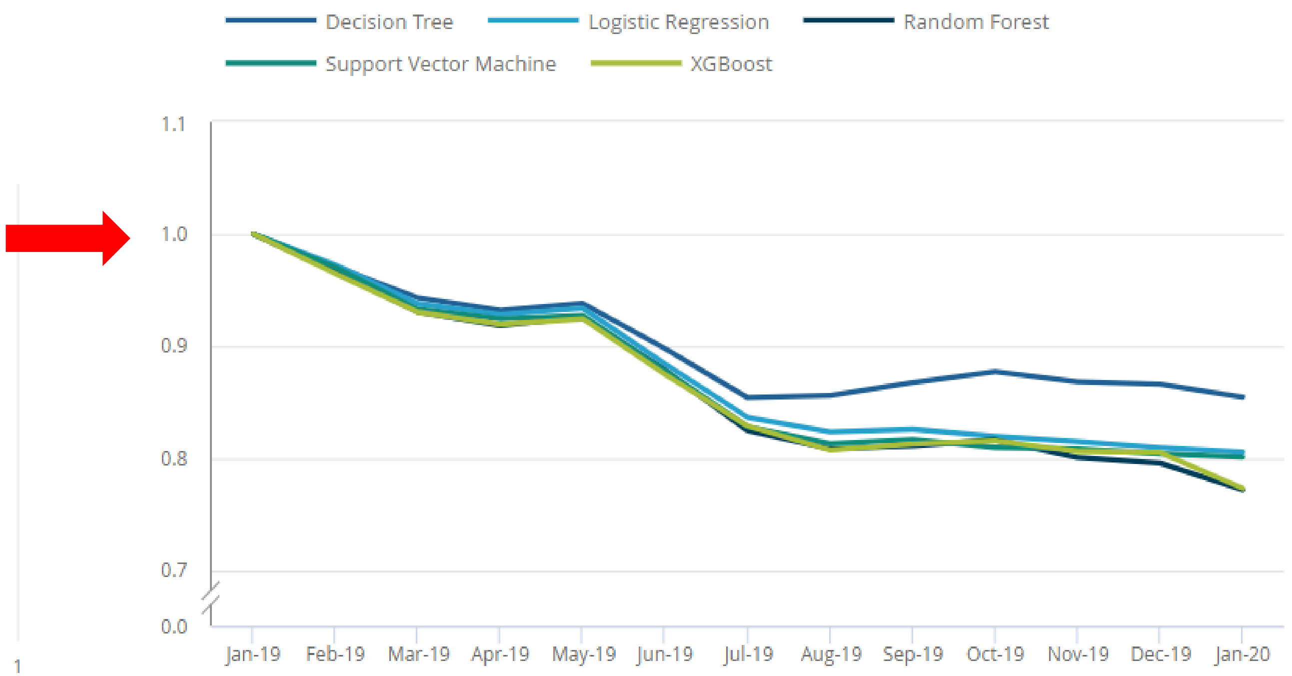
Strong relationship between consistency and performance! Machine only as good as the data it is trained on!



Dresses (high F1) indices

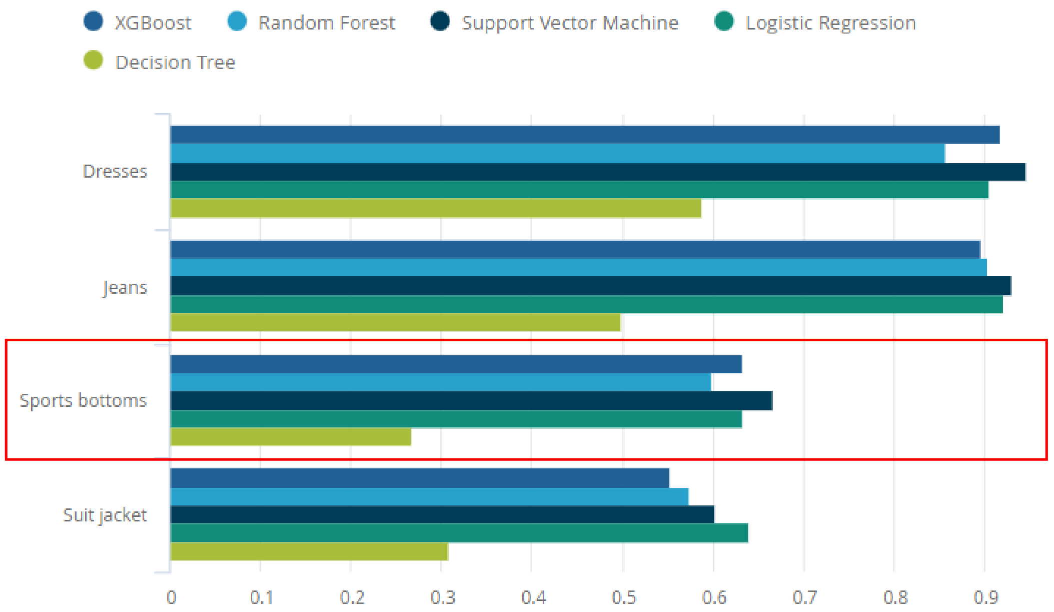
Classification F1 scores

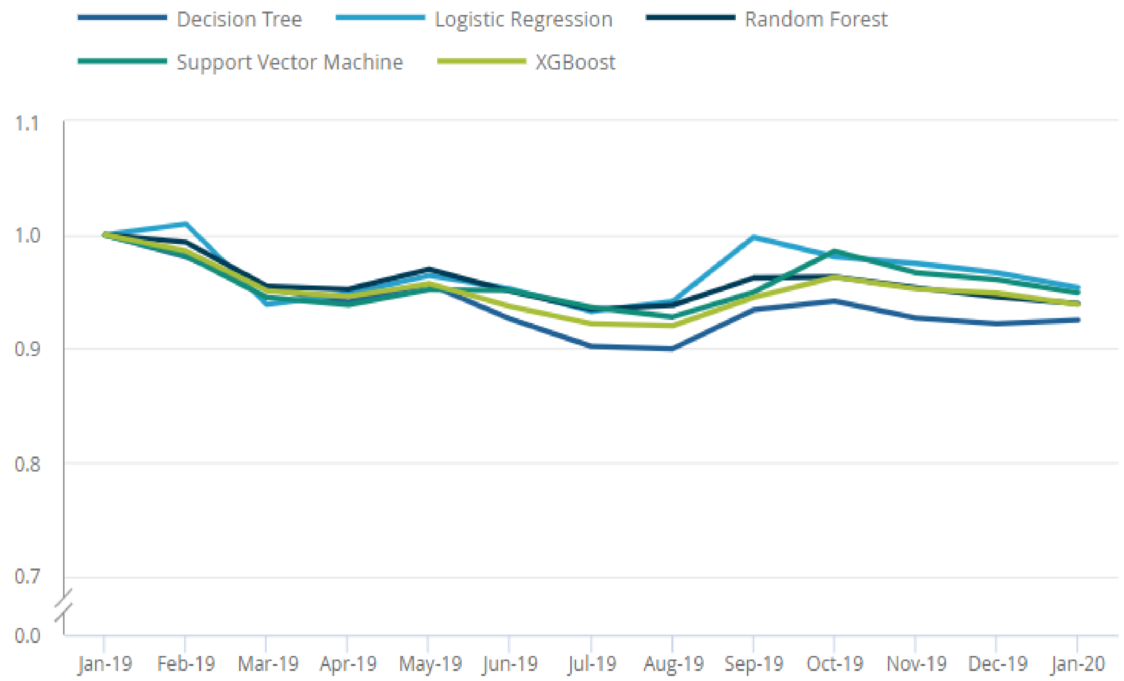




Sports bottoms (low F1) indices

Classification F1 scores





Product Grouping

Problem

Due to rapid product churn, can only use single product match in index

Product	Price, Jan	Price, Aug
Floral winter dress 1	39	
Floral winter dress 2	38	
Floral winter dress 3	44	
Floral summer dress 1	25	20
Floral summer dress 2		45
Party midi dress 1	100	
Party midi dress 2		90

Grouping – extreme examples

"Every product in single group":

- Group homogeneity: low.
- Match rate: 1.

Group	Product	Price, Jan	Price, Aug	Price change
1	Floral winter dress 1	39		
1	Floral winter dress 2	38		
1	Floral winter dress 3	44		
1	Floral summer dress 1	25	20	
1	Floral summer dress 2		45	
1	Party midi dress 1	100		
1	Party midi dress 2		90	
1	All dresses group	49.2	51.6	1.05

Note:

- Group homogeneity: in-group variance of prices.
- Match rate: propensity for inputs into indices to be available in both months

Grouping – extreme examples

"Every product is own group":

- Group homogeneity: 1.
- Match rate: low.

Group	Product	Price, Jan	Price, Aug	Price change
1	Floral winter dress 1	39		
2	Floral winter dress 2	38		
3	Floral winter dress 3	44		
4	Floral summer dress 1	25	20	
5	Floral summer dress 2		45	
6	Party midi dress 1	100		
7	Party midi dress 2		90	
1	Floral winter dress 1	39		
2	Floral winter dress 2	38		
3	Floral winter dress 3	44		
4	Floral summer dress 1	25	20	0.8
5	Floral summer dress 2		45	
6	Party midi dress 1	100		
7	Party midi dress 2		90	

Note:

- Group homogeneity: in-group variance of prices.
- Match rate: propensity for inputs into indices to be available in both months

Product grouping

"Product groups":

- Group homogeneity: medium-high.
- Match rate: medium-high.

Group	Product	Price, Jan	Price, Aug	Price change
1	Floral winter dress 1	39		
1	Floral winter dress 2	38		
1	Floral winter dress 3	44		
2	Floral summer dress 1	25	20	
2	Floral summer dress 2		45	
3	Party midi dress 1	100		
3	Party midi dress 2		90	
1	Floral winter dresses	40.3		
2	Floral summer dresses	25	32.5	1.3
3	Party midi dresses	100	90	0.9

Note:

- Group homogeneity: in-group variance of prices.
- Match rate: propensity for inputs into indices to be available in both months

Assessment: MARS (Chessa)

$$MARS = (match\,rate) \times R^2$$

Where:

- $(match\,rate) \in [0,1]$ measures match rate
- $R^2 \in [0,1]$ measures in-group homogeneity

Goal:

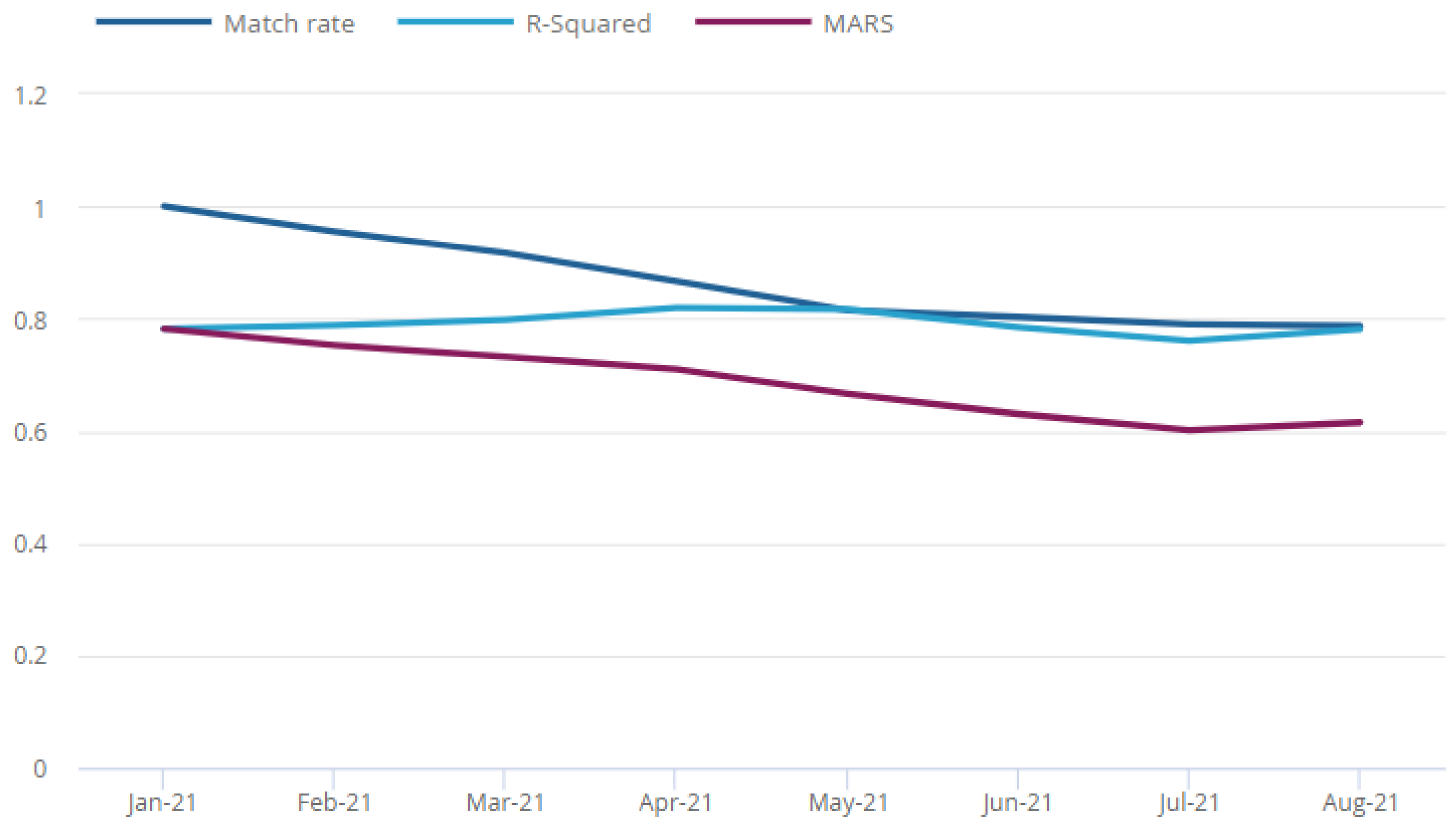
- Produce groups with high MARS, balancing homogeneity and match rate

Our grouping method

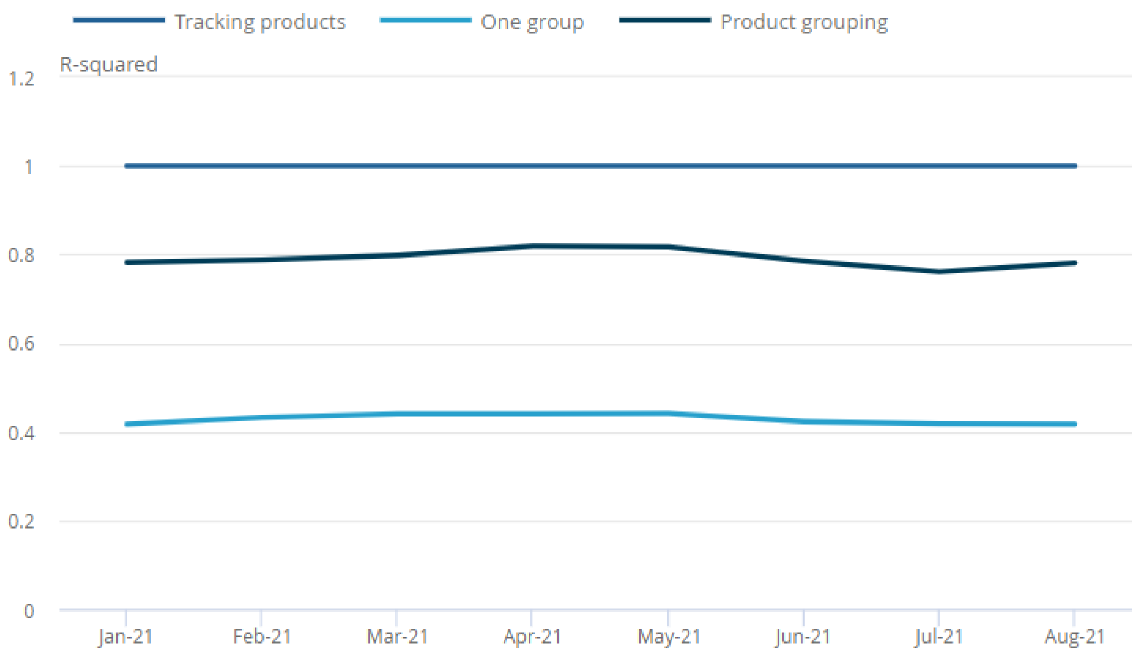
1. Remove non-quality defining stopwords/punctuation
2. Rank words in chosen columns by commonality
3. Select top X words (X chosen to maximise MARS)
4. Groups are a combination of these words:

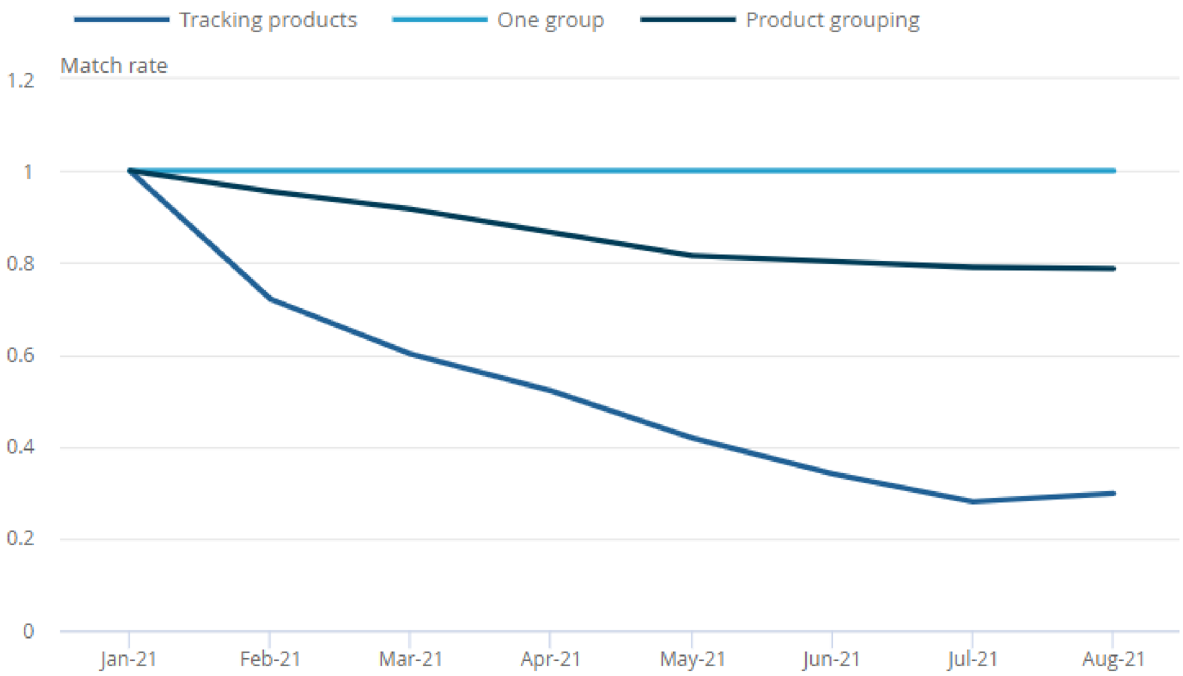
Product name	Material	Group
v-neck dress	polyester	polyester_v-neck
floral maxi dress	100% cotton	maxi_cotton
floor length maxi dress	cotton, elastic	maxi_cotton

MARS scores for women's dresses

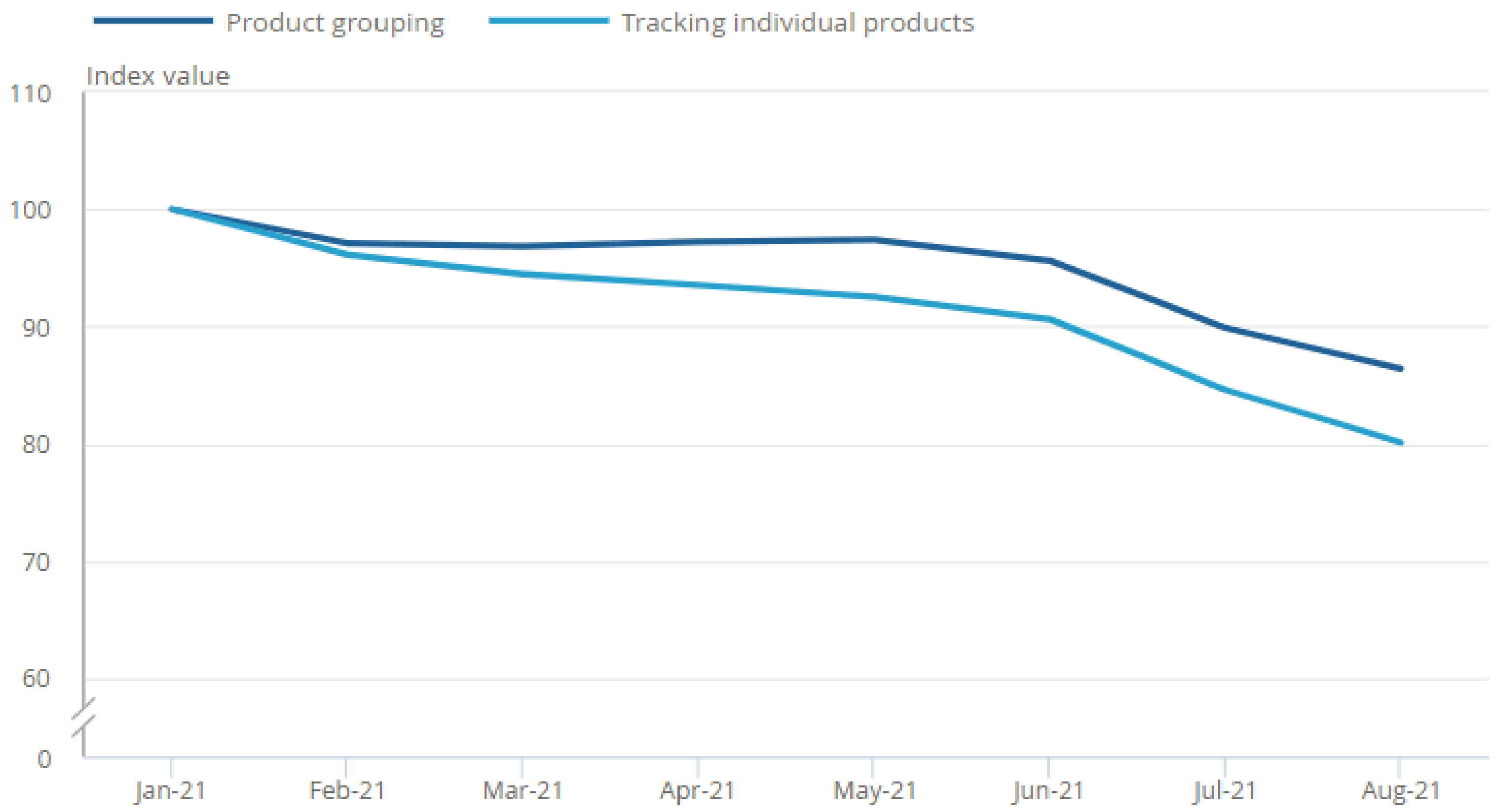


R-squared (left); match rate (right)





How index is affected



Any questions?

Future work:

Classification
Productionise and efficiency gains
Improve labelling consistency!
Choose suitable number of consumption segments
Explore precision/recall trade-off
Extend time series of analysis
Other pre-trained word vector models

Product Grouping
Productionise and efficiency gains
Extend time series of analysis
Explore product group sizes as weights (GEKS-T)
Improve algorithm word choices
Other measures of homogeneity beyond MARS
Generalise across clothing categories