ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Clothing Classification and Product Grouping

Status: Draft of future publication

Purpose

1. Our experimental methods to process our web scraped clothing alternative data sources use a supervised machine learning algorithm to classify web-scraped data into consumption segments, then a product grouping method that tracks prices of similar groups of products over time. In this paper we outline some potential challenges and solutions associated with the methods we are exploring.

2. In this paper, we will present results of both the classification and product grouping pipelines and discuss various methods of improvement.

Actions

3. Members of the panel are invited to:

    a. Advise on whether we should prioritise precision over recall for our classification task, through implementing a confidence threshold and using an Fβ score.

    b. Advise on whether our quality adjustment methodology with hedonic regression is acceptable, specifically considering words as explanatory variables and running the regression with hundreds of word dummies.

    c. Advise on the over-homogeneity problem for groups with single products.

    d. Advise on to what extent the low scores for homogeneity of price relatives fail our grouping model.

Background

4. Clothing contributes approximately 5% of the CPI basket in the UK and currently is covered with manually collected data. We obtain web-scraped data from the online shopping websites of main retailers in the clothing sector. We aim to increase product coverage with the high numbers of clothing items collected via web-scraping compared to manual price collection. This helps us to have more representative price data as we collect daily prices and improves granularity of the index since we can cover more varied types of clothing.

5. We obtain web-scraped clothing data since June 2020 from 17 online retailers in the UK which covers around 1000 brands. This makes more than 900,000 unique clothing products in each month, extending our coverage significantly compared to the traditional manual data collection which covers approximately 20,000 products. Working with such a different nature of data at this scale requires extensive data processing and implementing new innovative methods for index calculation.

6. This paper discusses three main pipelines used for incorporating web-scraped clothing data into index calculation. The classification pipeline classifies clothing products into relatively homogeneous categories. The product grouping pipeline aims to address product churn by tracking prices of groupings of similar products over time. Finally, we calculate our experimental indices with web-scraped data using multilateral methods.

7. This paper follows on from previous papers we have taken to the Technical Panel:

    a. In Automated classification of web-scraped clothing data in consumer price statistics (originally taken to APCP-T in April 2020), we described our research into using

supervised machine learning for classifying clothing data – in this paper we now describe a few modifications we may make to the classification procedure.

b. In Dealing with product churn in web scraped clothing data: product grouping methods (PDF), we described our early work within product grouping, which we now expand upon the methods within this paper.

c. In Introducing multilateral index methods into consumer price statistics (originally taken to APCP-T in April 2022), we described our preference for using the GEKS-Törnqvist with a 25-month window and a mean splice on published extension method, which our index methodology prioritises in this paper.

8. Due to the complexity involved in the research and integration of clothing alternative data sources, we remain in an experimental research phase and have not integrated clothing into our implementation work planning.

Section 1: Clothing Classification

9. The first section of this paper will give an overview of the clothing classification pipeline, before outlining some of the main questions we would like the panel's advice on.

10. In the classification pipeline, we have target classes, known as consumption segments, that we classify our raw data to. Consumption segments partition the consumption basket into groups of relatively homogeneous (similar) products. Price changes are measured within each consumption segment, then aggregated through an international classification system known as Classification of individual consumption according to purpose (COICOP). An example of a consumption segment may include "Women's t-shirts", which is in turn aggregated into "Garments for women" and then into "Clothing". Our goal is to ensure that products are assigned to the correct consumption segment.

11. This is a complex task due to the scale and complexity of scraped clothing data. Whilst in other alternative data sources we can use manual classification, the large number of new products entering the clothing market each month means that this would be too strenuous a task for manual labellers. Therefore, we are investigating the use of supervised machine learning to classify our alternative clothing data.

12. Supervised machine learning models require a dataset containing several features (predictor variables) and a set of data that has already been classified to the desired category (labelled data). The model "learns" which predictors are associated with which categories, and then uses the rules it has learned to categorise (classify) new data.

Manual Labelling

13. The first step in the project was to manually classify a sample of products to the clothing consumption segments, creating a "human-labelled" dataset. This dataset is split into a training dataset, from which the model can learn its rules, and a test dataset, on which we can test the performance of the predictions made by the model. We used stratified sampling to ensure that the weights of different consumption segments in the sample dataset were proportional to the number of consumption segments in each age and gender group (boys, girls, infants, men, women). For example, if 35 of our 133 consumption segments were women's, we would have required 26% of web-scraped clothing sample data to be women's. We then stratified retailers with equal weight.

14. Upwards of 30 people within the Office for National Statistics (ONS) Prices Division labelled clothing data using a bespoke in-house application. Labellers labelled the data at multiple levels of granularity, allowing us to compare classification performance at different levels of homogeneity.

15. Despite our best efforts to ensure consistency amongst human labellers, there is an element of subjectivity in clothing classification. For example, a "hooded jacket" can be considered both a "hoodie" and a "jacket". Often, the labeller may find the correct class to be indeterminable. If high levels of inconsistency are occurring, then the machine will not be able to reliably predict how to classify products. To quantify any inconsistencies between our human labellers, we labelled a total of 30,000 products twice. On average, labellers were consistent in 88.8% of cases.

16. The products that split opinion the most strongly were often products which were on the boundary of two possible classes. This suggests that some of the inconsistency is driven by subjectivity in how to place cases that could belong to more than one class rather than labellers making explicit errors. This also means that there is a limit as to how consistent our labels can be, given the subjective nature of some clothing items. This provides something of a benchmark for our automated classification model.

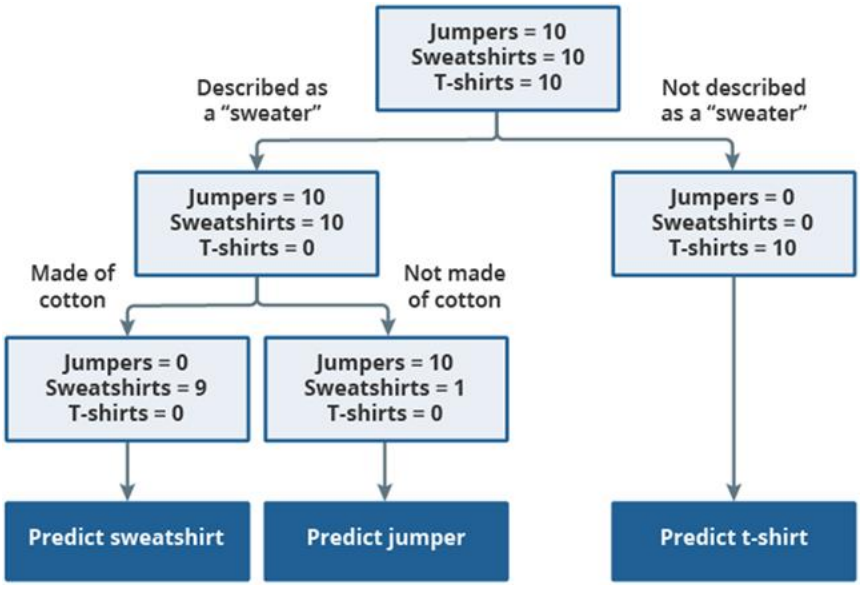
Machine Learning Classification Model

17. Each consumption segment is made of an age and gender group and a clothing type. This includes, for example, "infants' sleepsuits", "men's socks" and "girls' dresses". Since age and gender are important in unpicking the consumption segment to which a product belongs, we use text mining to obtain features (predictor variables) that may indicate gender or age. For example, if a clothing product comes with the size "mg" (medium girls'), this would indicate that the product is for non-infant girls. We also use industry standard word embeddings for our features, including FastText, TF-IDF and bag-of-words. A more thorough view of these word embedding features can be found in the Ottawa Conference paper (PDF, 1.61MB).

18. We are using these features along with our human-labelled data to train and test the performance of machine learning (ML) models. For this paper we demonstrate results based on gradient-boosted trees (specifically XGBoost), as these are currently our highest-performing algorithm.

19. XGBoost uses decision trees as a foundation. A decision tree can conceptually be thought of as a flowchart, as displayed in Figure 1. When training the tree, every product starts in a single node (represented by the top rectangle). Products are continually split into two new nodes using automatically generated binary (yes/no) decisions, the "rules". The goal is to keep splitting the tree until the final nodes mostly represent a single consumption segment.

Figure 1: Decision trees use binary decisions to split the data



20. XGBoost involves training many decision trees sequentially, with each tree trained to improve on the errors made by previous trees. Predictions of the final ensemble model are the weighted sum of the predictions made by all previous tree models.

Measuring Performance of the Model

21. In this section, we will discuss the performance metrics of the model, and whether we should prioritise one over the other. The panel's advice would be appreciated on whether we should implement a confidence threshold, explained below, which would increase precision of the model at the expense of recall.

22. We are using two main metrics to assess the performance of the classifier on our dataset:

    a. Precision: measures the "purity" of a consumption segment. A segment with 90% precision would mean 10% of the elements classified to the segment are from other segments (false positives).

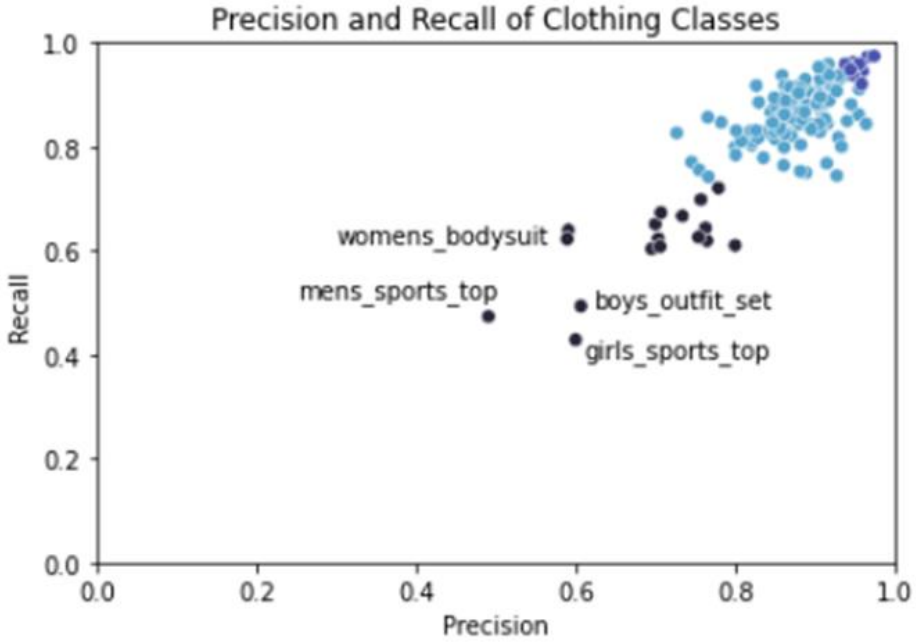
    b. Recall: measures the extent to which all cases from the consumption segment are captured by the classifier. A class with 90% recall would mean 10% of elements that should be part of the segment have been incorrectly classified elsewhere (false negatives).

23. There is a trade-off between precision and recall captured by a third metric, the Fβ score, which is a weighted harmonic mean of precision and recall. The formula for an Fβ score is shown below. This score can be weighted to favour recall (where $\beta > 1$), precision ($\beta < 1$) or give equal importance to the two ($\beta = 1$).

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{(\beta^2 \times precision) + (recall)}$$

24. The precision and recall scores of our 133 consumption segments are displayed in Figure 2.

Figure 2. Precision-recall scatter graph of all classes. Colours indicate whether F1 scores are low (mean – 1SD), high (mean+1SD), or medium (in between).



25. For this task, we may decide that precision should be prioritised over recall. Using the example of women's dresses, this would mean preferring a smaller "true" class of products where everything allocated to that class truly is a women's dress, rather than an "all-

encompassing" class of products where every single women's dress is captured, but the classifier has also erroneously predicted that some women's tops belong to the same class.

26. This prioritisation of precision over recall could be achieved by implementing a "confidence threshold", which sets a threshold for the probability the classifier needs before it allocates a product to a class. For example, setting this threshold at 0.8 would mean that the classifier needs to be more than 80% confident in its prediction before allocating a product to a class. If its confidence is less than 80%, it gives "no prediction". The impact of applying a confidence threshold at different levels to our model is shown in Table 1, below, and an illustration of different classification scenarios is demonstrated in Figure 3.

Table 1. Precision, recall and F1 score of classification models at different levels of confidence threshold.

Threshold	Precision	Recall	F1 Score	F0.33 Score
None	0.86	0.84	0.85	0.86
0.70	0.91	0.69	0.77	0.88
0.75	0.92	0.66	0.75	0.89
0.80	0.92	0.61	0.72	0.88

Figure 3. How confidence thresholds may allow us to prioritise improving precision at the expense of recall

27. In the left-hand diagram, we demonstrate what may happen in our traditional collection, where a comparatively small sample of products for each consumption segment is manually collected. In this example, we group four women's dresses of the same variety into a class called "women's dress". This gives us a precision of 1, because everything in that class is verified manually as truly a women's dress, but we do not capture the full range women's dresses, meaning that recall for this method is low. The middle diagram is an example of what could happen with our machine learning algorithm without a confidence threshold. This gives equal weight to precision and recall, as it groups eight of the women's dresses and one women's top into a class called "women's dress". This gives us an equal level of both precision and recall, but still groups one women's top with the dresses, meaning that precision is lower than we would like. The right-hand diagram shows a grouping which prioritises precision, as could happen in our machine learning algorithm with a confidence threshold. This classifies a wider variety and higher number of women's dresses correctly and does not include any women's tops. We think this is a desirable outcome for our classification task, as it is more in line with our traditional methodology and ensures our indices are not contaminated with products that do not meet the same description.

28. If we decide to use confidence thresholds to target a classifier which prioritises precision over recall, we may then prefer to use an Fβ score weighted towards precision ($\beta < 1$) to assess model performance. For example, in Table 1 where precision is given double the weight of recall (β = 0.33), the moderate increase in precision is seen as more important than the larger decrease in recall, with optimal results occurring with a threshold of 0.75. This is unlike the F1 score, which treats precision and recall equally, and has optimal results without confidence thresholds.

29. The panel's advice would be appreciated on whether expressing a preference for precision, coupled with the use of confidence thresholds and an F-beta metric with $\beta < 1$ are an appropriate addition to the classification methodology.

Combining classes

30. Another way of correcting for issues that we are experiencing in the classification model is to combine classes which the model is confounding, where we may not have the data to create a split. We find that consumption segments that are easy to classify often contain items with an indicative word in their product name. For example, most "jeans" products contain the word "jeans" in their descriptor. Those that the algorithm struggles to classify have words which overlap with other segments. For example, sports clothes such as "men's sports shorts" often have similar descriptors to non-sports clothing like "men's shorts". This mirrors difficulties faced by manual labellers in our consistency experiment (described in the "Manual labelling" section), where distinguishing between such classes was problematic due to subjectivity and a lack of sufficient identifying information. To identify further areas for improvement of the performance of the algorithm, we have produced a confusion matrix which establishes which segments the classifier is struggling to predict accurately by comparing precision and recall for all consumption segments and showing points of contention for each class. We could look at combining classes to create larger classes which have higher performance, but this would mean less homogeneous segments. An example of some of the classes that the classifier struggles to differentiate between is shown in Table 2.

Table 2. Example of classes which have low performance scores, and their points of contention

Class	Point of Contention
Girls' sports top	Girls' top/t-shirt/crop-top
Boys' outfit set	Boys' full tracksuit
Men's sports top	Men's t-shirt
Women's sports top	Women's top/t-shirt/crop-top
Boys' vest	Boys' t-shirt

31. For low-performing classes, we can calculate three metrics to aid decision-making: the weight, the Fβ score, and the change in Fβ score affected by combining the class with its contending class. We have identified around 20 classes which either have very small weights, low Fβ scores, or both. The decision as to how to treat these classes will depend on finding a balance between homogeneity and performance. We could, for example, combine all sports clothes with their non-sporting counterparts (sports tops with other tops, sports shorts with other shorts), raising the Fβ score of the class but resulting in a less homogenous grouping of products. Further exploration will be needed to truly quantify the effect of these changes on the final indices.

32. Note that we must maintain a product grouping model for each consumption segment, so having fewer albeit more heterogeneous consumption segments would make model maintenance less expensive.

33. Panel feedback is requested on the topic of what is a desirable level of homogeneity for creating consumption segments within clothing and how this decision could be reached.

Section 2: Product Grouping

34. Once we classify products into consumption segments, we could potentially create separate price indices for each consumption segment and aggregate them to obtain a single clothing price index. However, the extremely dynamic nature of the clothing market, with fast product entry and exit, causes problems for the way our indices are calculated with individual product prices. Prior to calculating indices, we are therefore looking to group homogeneous products together within each consumption segment to use average prices of those groups instead of individual product prices.

35. Clothing products rarely stay in the market for a full year as seasons and fashion trends change several times a year. Therefore, "product churn" is a fundamental problem for the clothing index. Traditional methods of calculating our Consumer Price Index (CPI) requires finding a substitute product when a product leaves the market. This replacement is a manual step in traditional methods; however, these manual processes are not viable with the huge volume of products and prices with web-scraped data. This can cause the index to become unrepresentative due to the rapid product entry and exit in the clothing market.

36. The product grouping pipeline aims to group products which are similar or reasonably substitutable from the consumer perspective. We reduce the effect of the product churn problem by tracking average prices of groups instead of individual products, since product groups are more likely to survive even though some products within group are leaving the market.

37. Product groups should be large enough to control for product churn. In other words, we try to capture a high proportion of products in the index calculation by making broad product groups to survive through a year despite losing some of its constituents. On the other hand, those groups should also be homogeneous in terms of quality and from a consumer perspective so that compositional effects do not bias inflation. These two criteria compete as we need to have finer groups for the sake of homogeneity, but finer groups are likely to fail to survive in time. Therefore, product grouping should have a balance between the homogeneity and survival of groups.

Assessment Measure: MARS Score

38. Chessa (2019) introduced the "Match Adjusted R Squared" (MARS) score to assess the success of groups as a product of match rate and homogeneity:

$$\mathrm{MARS}_t = \mathrm{R}_t \, \mu_t$$

where μ_t is the match rate and R_t is the R-squared measuring in-group price similarity within the current month.

39. **Match Rate:** Match rate is the share of products in a month from groups present in both reference period and the current month. In this paper, we use a 12-month reference period[1] to calculate the match rate for each month in the assessment period to capture the seasonal changes in clothing sector.

[1] The reference period for MARS score is different than the reference period in index calculation. The reference period in MARS score is used to match the surviving groups in time to evaluate the success of our grouping model. For index calculation, we reference our index to a single month.

40. **Homogeneity:** R-squared is a measure of in-group price similarity within the current month. It measures the proportion of explained variance in prices by grouping relative to the total variance in prices without grouping. R-squared formula is

$$R_t^K = \frac{\sum_{k \in K} q_t^k \left(\bar{p}_t^k - \bar{p}_t\right)^2}{\sum_{i \in G_t} q_{i,t} \left(p_{i,t} - \bar{p}_t\right)^2} \quad \Longrightarrow \quad \frac{\text{explained price variance by grouping}}{\text{total price variance in product prices}}$$

where $p_{i,t}$ is the price of product i in month t, $\bar{p}_t$ is the unit value over all items in month t, and $\bar{p}_t^k$ is the unit value for group k in month t.

41. There are two main caveats of this measure. First, the original R-squared within MARS weights these variations with quantities. However, we use an unweighted version as web-scraped data do not include quantity or expenditure information. Secondly, MARS measures homogeneity only with price similarity. We cannot measure homogeneity in terms of purpose or quality as we cannot quantify them with web-scraped data.

42. We also measure the homogeneity of groups in terms of price movements by calculating R-squared with price relatives, which we discuss in paragraph 60.

Rules Based Method

43. We need a set of rules for each consumption segment to assign individual products to product groups. They should reflect the characteristics and quality of the products to form homogeneous groups. Ideally, we can select such rules with a visual inspection of the data and using domain knowledge on the market, as we have done for second-hand cars and rail fares. However, this is not practical for clothing as we have more than hundred consumption segments and each has distinct characteristics. Setting the rules manually for each consumption segment would be very subjective and ambiguous with the lack of well-defined categorical variables other than the textual data. Therefore, we propose an automated approach.

44. We search through the attribute columns[2] in the data to find these rules. As the data scraped from the retailer websites are textual and unstructured, we clean and process attribute columns using NLP techniques to find key words to use as rules forming product groups. We remove punctuation and numbers, remove stop words, and standardise retailer and brand columns to account for different versions of same retailer/brand. We also remove some common words which do not differentiate products within consumption segment and do not provide any quality information, such as "dress" for "women's dresses" consumption segment.

45. Once we have a set of rules; for each product, we flag the words in the rules list if they appear in the attribute columns. The combination of those flag words creates a group identifier for each product and assigns them to a particular group. Table 3 provides a simplified example of this process with three products, two attributes, and two rules from each attribute.

Table 3a. Rules dictionary with two attributes and two rules for each attribute

	Rules Dictionary	
Attributes:	**Product Name**	**Material**
	v-neck	polyester
	maxi	cotton

[2] The six attribute columns are "Retailer, Brand, Product Name, Product Description, Product Style, Material."

Table 3b. Rules-based product grouping with rules dictionary provided in Table 3a

	Product Name	Material	Group identifier
Product 1	**v-neck** dress	**polyester**	v-neck_polyester
Product 2	floral **maxi** dress	100% **cotton**	maxi_cotton
Product 3	white **maxi** dress	**cotton** elastic	maxi_cotton

Finding the Rules: Most Frequent Words Dictionary

46. Frequency of words in the attribute columns is a good starting point to find the key words that identify similar products. We extract the most common words dictionary by searching for a fixed number of most frequent words in each attribute column. A naïve approach would be to use all words in this dictionary as rules. In that case, having large number of frequent words from each column would create very small groups which are not surviving in time. If we extract a smaller number of frequent words, on the other hand, we miss some key words with less frequency but having greater importance to distinguish products. Therefore, we implement quality adjustment and optimisation steps over this dictionary to obtain a list of rules which maximises the MARS score for each consumption segment and improves grouping performance.

Quality Adjustment

47. We first combine all words in the most frequent words dictionary except retailer names to get a single bag of words as candidate rules. Then, we run a hedonic regression with dummies of each word to quantify the impact of each key word on the price of a product. The regression equation is $P_i = \sum_{j=1}^{N} \beta_j D_{ij} + \varepsilon_i$, where P_i is the price of product i, N is the number of words in the combined most frequent words dictionary, and D_{ij} is a dummy for word j being contained in product i.

48. We drop the words with statistically insignificant coefficient and re-rank the remaining words according to their contribution to price. This gives us a quality adjusted rules list. We always keep all retailer names as rules to have separated groups for retailers. We do not allow groups to mix up across retailers as we create initial indices for each retailer separately.

49. We ask for feedback from the panel whether our quality adjustment methodology with hedonic regression is acceptable, specifically considering word dummies as explanatory variables and running the regression with hundreds of word dummies. We hesitate over the use of too many explanatory variables due to the potential risk for overfitting.

Optimisation

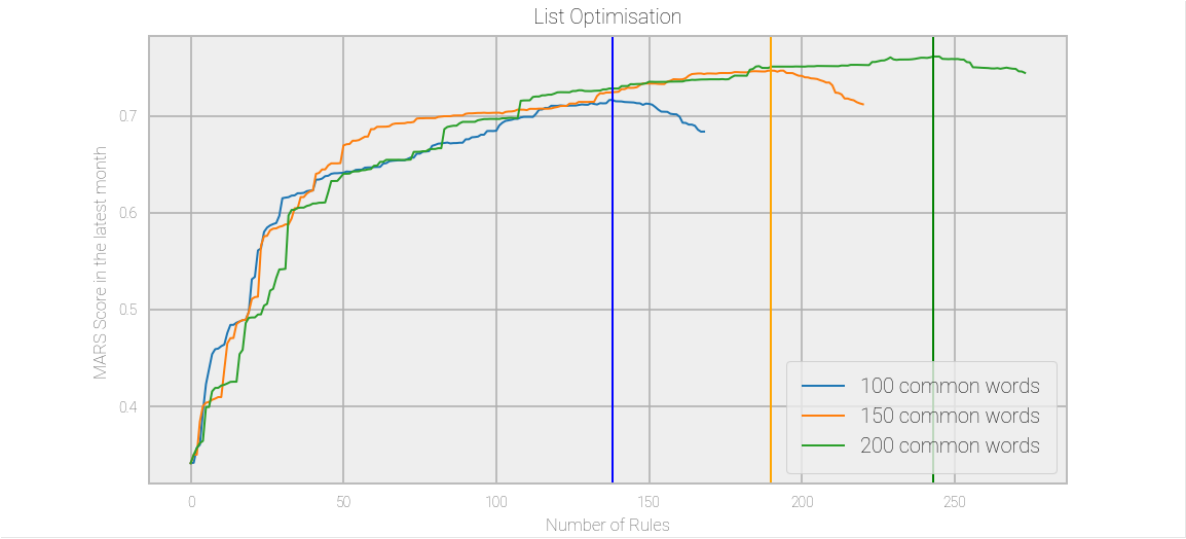
50. The number of rules used to create product groups has a direct impact on the MARS score. Increasing the number of rules would narrow down the groups to be more homogeneous, but also decrease the match rate simultaneously. Therefore, we should find the optimum balance between homogeneity and match rate which maximises the MARS score. We can find the optimal number of rules with an optimisation algorithm.

51. The optimisation algorithm starts by grouping all products into a single group for each retailer and adds one rule in each iteration from the quality adjusted rules list. We run grouping in each iteration and calculate the MARS score in the latest month of the assessment period. Iteration stops when there is no improvement in MARS score for at least 30 iterations. The grouping with the highest MARS score would give us the optimum number and set of rules.

52. Figure 4 shows the results of grouping optimisation with “women’s dress” consumption segment. Each line represents optimisation with quality adjusted rules lists obtained from dictionaries with different number of most frequent words from each column. Increasing the

number of common words results in higher peak points; however, it does not make a significant difference after some point.

Figure 4. Optimisation of grouping with "women's dress" consumption segment



Retailer by Retailer Optimisation

53. We run this optimisation algorithm for each retailer individually to allow for retailer specific optimum number and set of rules. We run the algorithm with different number of most frequent words[3] and pick the best performing model for each retailer. Table 4 shows the results of this optimisation for each retailer in comparison to their corresponding MARS scores with optimisation as a whole instead of retailer-by-retailer. The optimum MARS scores and optimum number of rules are quite different as the wording and data quality change across retailers.

Table 4. Retailer by retailer optimisation results for "women's dresses" consumption segment

Retailer	MARS Score with Overall Optimisation	MARS Score with Retailer Optimisation	Optimum Rule Number
Retailer A	0.70	0.73	194
Retailer B	0.52	0.59	133
Retailer C	0.62	0.78	335
Retailer D	0.69	0.77	218
Retailer E	0.61	0.67	83
Retailer F	0.67	0.73	274
Retailer G	0.45	0.69	59
Retailer H	0.50	0.60	90
Retailer I	0.61	0.69	34
Retailer J	0.51	0.79	55
Retailer K	0.45	0.66	35
Retailer L	0.16	0.61	40
Retailer M	0.37	0.42	147
Retailer N	0.43	0.69	74
Retailer O	0.27	0.79	27
OVERALL	0.79	0.80	

[3] From 100 to 350 with 50 increments

54. This optimisation method is our preferred method from different alternatives we considered for four main reasons. First, we reach the maximum overall MARS score with this method. Secondly, it results in significantly higher MARS scores for each retailer individually. Thirdly, it allows different model configurations for each retailer. Lastly, it considers different wording structures across retailer websites.

Group Homogeneity

55. The number of rules used for each product to create a group identifier ranges between 1 and 15 rules, while retailer name is a default rule for each product. Therefore, some groups contain no other keyword than the retailer name or just a few key words, which makes them quite heterogeneous. Table 5 lists a few examples of relatively heterogeneous and homogeneous product groups for the "women's dresses" consumption segment.

Table 5. Sample group identifiers[4] for "women's dresses" consumption segment

Group is	Group Identifiers	Group Size
Relatively heterogeneous	retailerA	93530
	retailerE_polyester	17692
	retailerD_brandX	9408
	retailerA_brandY	982
Relatively homogeneous	retailerE_keyword1_midi_polyester_keyword2	24
	retailerG_keyword3_keyword4_keyword5	11
	retailerA_brandZ_keyword6_linen_keyword7_keyword8	10
	retailerB_brandW_shortsleeved_keyword9	5

56. The groups with 1 or 2 rules contain 52% of the products. There are two main reasons behind this. First, some products have missing or scarce data within attribute columns; hence, they do not contain any key words from the rules list. Secondly, the optimisation algorithm allows for big groups to keep product match rate high enough. Therefore, we consider imposing a minimum number of rules requirement for creating a group to avoid very heterogeneous groups like "retailerA" or "retailerE_polyester".

Minimum Rules Requirement

57. We set a requirement of minimum 3 rules to create a group identifier. This requirement returns null group identification for the products containing fewer than 3 key words in their attribute columns from the rules list. Incorporating this requirement to the optimisation algorithm described above returns an improved rules list as now it takes the null products into account when calculating the MARS score in each iteration. Finally, the product grouping with new optimum rules list returns around 14% of the products with null group identifications. The remaining products previously with 1 or 2 rules have now identified with a greater number of rules due to the improved rules list after optimisation with minimum rules requirement.

58. Table 6a shows MARS scores in the latest month for grouping with and without the minimum rule requirement. We observe a significant decrease in product match rate when we enforce the minimum rule requirement both due to the null products with low data quality and more refined groups with improved rules list. However, we avoid very large and non-homogeneous groups with this requirement. Table 6b shows the summary statistics of the groups for "women's dress" consumption segment after running the grouping with and

[4] Retailers, brands, and some keywords are encoded to avoid possible identification of retailers.

without the minimum rule requirement. The large non-homogeneous groups on the upper end of group size distribution are eliminated and we end up with a significantly higher number of rules.

Table 6a. MARS Scores in the latest month

	Without min rule requirement	With min rule requirement of 3	
		Excluding Null Products	Including Null Products
R-squared	0.87	0.94	0.83
Product Match	0.92	0.67	0.58
MARS Score	0.80	0.63	0.48

Table 6b. Summary statistics of the groups for 13 months (2021-06 - 2022-06)

	Without min rule requirement	With min rule requirement of 3
Number of Products	1,310,372	1,310,372
Number of Null Products	0	185,141 (14%)
Number of Groups	19,982	64,359
Group Size Statistics		
Mean	65	17
Standard Deviation	976	82
Minimum	1	1
Median	10	8
Maximum	93,530	7,747

Over-homogeneity

59. On the lower end of the group size distribution, small groups are another concern since they are more likely to drop out due to churn. For all month-group pairs within 2021-06 and 2022-06, 65% of them contain only one product within a month[5], which means they are not grouped together with any other product in the same month. This brings the product match rate down significantly. We are seeking advice from the technical panel to tackle this problem.

Relative Price R-squared

60. The aim of the grouping is to group similar products together. The R-squared within MARS scores measure the extent to which grouping explains the price variance across products. We also expect the product prices within a group to move together if the groups are homogeneous. Therefore, we calculate a different version of R-squared measuring how well the grouping explains the variance in price relatives across products. We first calculate the price ratio of each product relative the previous month if it exists. Then, we plug those price relatives into the R-squared function. Monthly price relatives R-squared for "women's dresses" ranges between 37% and 47%, which are quite low compared to price levels R-squared. We ask for advise on to what extent the low scores for homogeneity of price relatives fail our grouping model.

Section 3: Index Analyses

61. In this section, we present indices for three consumption segments (women's dresses, men's jeans and men's socks) following product grouping.

[5] This ratio is 48% when grouped without minimum rules requirement.

62. Note that as shown in the match rate analysis in the product grouping section, product groups are not necessarily persistent: new groups can form, and previous groups can disband. In Introducing multilateral index methods into consumer price statistics we expressed a preference for using multilateral index methods, specifically the GEKS-Törnqvist (with mean splice on published extension method and a 25 month window), for handling dynamic products when we want to maximise our use of data.

63. Note that we lack weights in a true sense since web scraped clothing data does not have any explicit weights. However, note that the number of products contained within a group may provide a better estimation of a group's genuine weight, than to simply use no weights. For example, a group with 100 products is likely (but not guaranteed) to outsell a group with five products. We therefore create two models:

    c. Model 1: GEKS-Törnqvist (where each group is weighted in accordance with its size)

    d. Model 2: GEKS-Jevons (where each group is given the same weight)

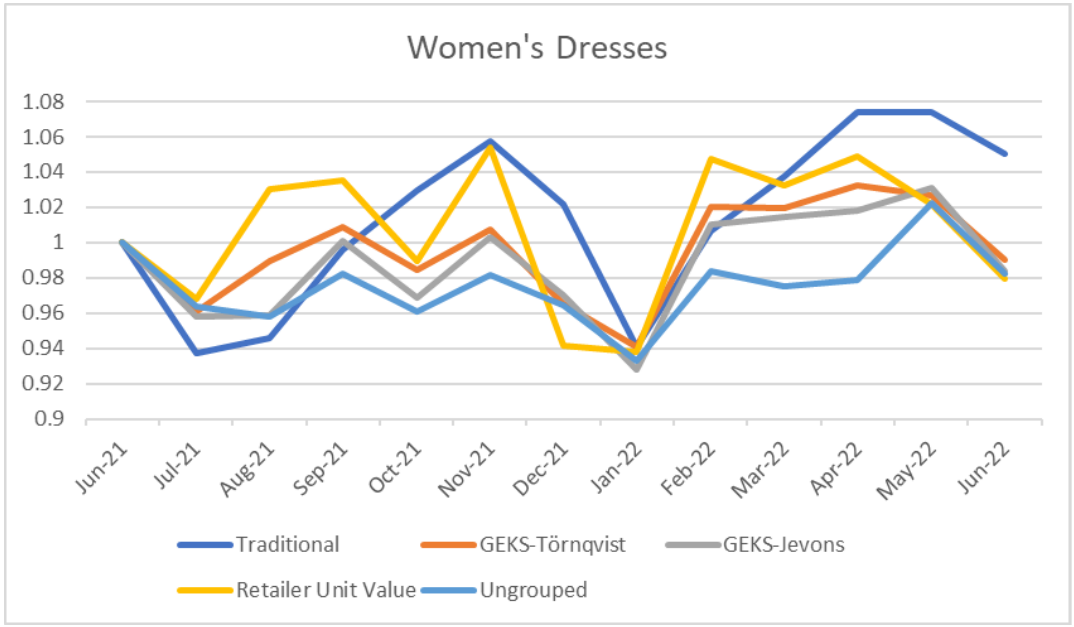
64. Since our product grouping model attempts to balance homogeneity and match rate, it is natural to compare our methods to two "benchmarks" based on extreme scenarios of whether we prioritised homogeneity or match rate, to see whether our methods are behaving characteristically like one of these methods:

    e. Model 3: The unit value index is the average price within the consumption segment (which may contain substantial unit value bias)

    f. Model 4: The ungrouped index uses a GEKS-Jevons over ungrouped prices (which may suffer from a lack-of-matches problem due to the extreme churn observed in clothing)

65. The results are shown in Figures 5 (women's dresses), Figure 6 (men's jeans), and Figure 7 (men's socks) in comparison to the traditional indices[6].

Figure 5. Experimental indices for "women's dresses" in comparison to traditional index



[6] Traditional indices are item level indices in CPIH using monthly manual price collection data. "Women's Dresses", "Men's Jeans", and "Men's Socks" are three of the over 700 items in CPIH basket.

Figure 6. Experimental indices for "men's jeans" in comparison to traditional index

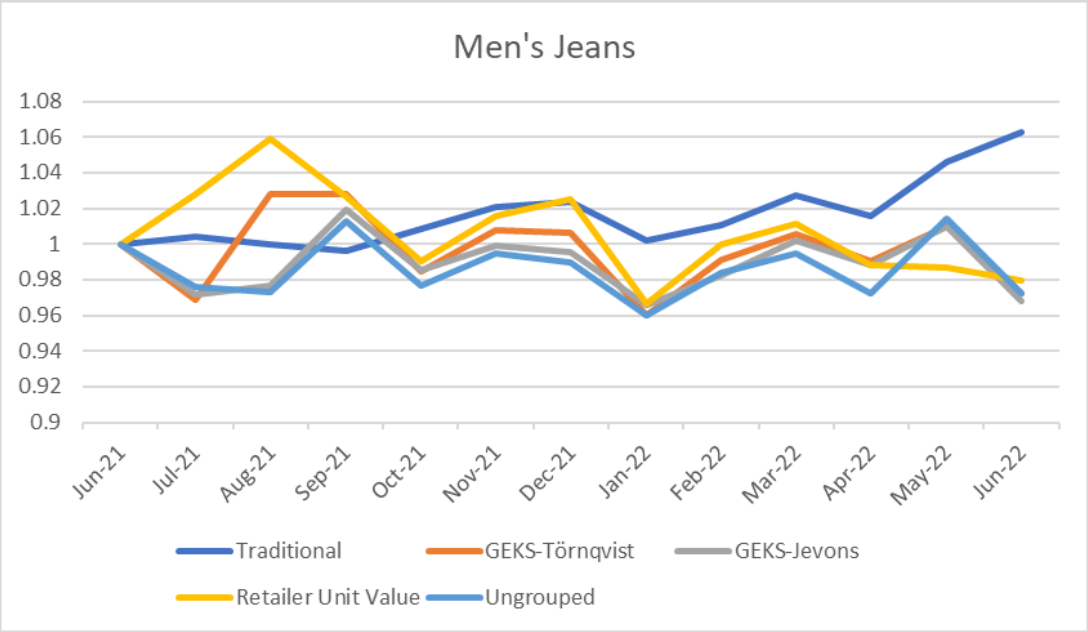
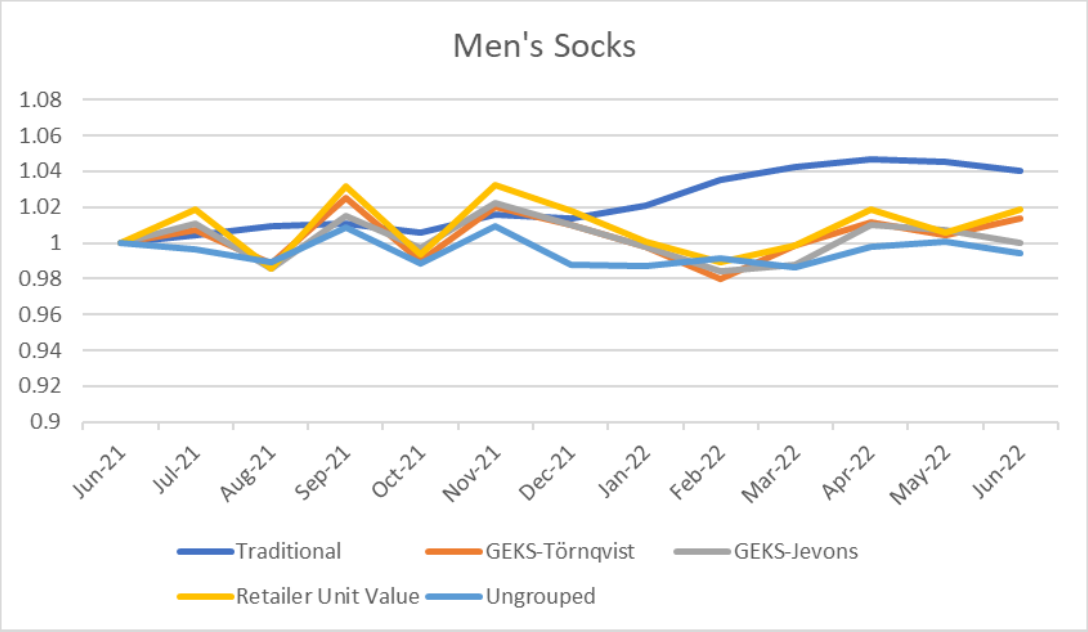


Figure 7. Experimental indices for "men's socks" in comparison to traditional index



66. GEKS-Jevons and GEKS-Törnqvist indices mostly lie between ungrouped index and unit value index for all three consumption segments. Unit value indices are in general higher than the other experimental indices, which could be due to unit value bias. On the other hand, ungrouped indices are generally lower than the others as they are likely downward biased due to the clearance prices on individual products coupled with an extreme churn problem in the clothing markets. GEKS-Jevons and GEKS-Törnqvist indices address these two opposite biases to some extent with product grouping. The GEKS-Törnqvist index is slightly higher

than GEKS-Jevons as it puts more weight on large and potentially less homogeneous groups which may be more prone to unit value bias.

67. Experimental indices follow a mostly similar trend with traditional indices, although there are some significant differences in levels. The experimental indices would capture a wider range of products for each consumption segment while the traditional indices are containing prices of much narrowly defined products. The retailer coverage is also different as web-scraped data capture online-only retailers as well while missing retailers without online shopping website. These could create difference across traditional and experimental indices with web-scraped data.

Section 4: Future Work

68. Following the advice from the technical panel, we will continue to work on improving product grouping performance. Setting a minimum rules requirement helped us to deal with the extremely heterogeneous groups at the upper end of group size distribution. We will continue our research to deal with the vastness of groups with single product at the lower end of the distribution as they are more likely to drop out due to churn.

69. We may explore amendments to the MARS metric to prioritise homogeneity over match rate, on the condition that the match rate is above a given threshold, since multilateral index methods can cope with some degree of failed matches.

70. Regarding the concerns about our quality adjustment method with hedonic regression, we can consider to research on other alternatives to adjust our rules dictionary to go beyond the naïve frequency-based approach.

71. There are alternative approaches we may yet explore for clothing. For example, we may be able to use regression-based decision trees to create product groups, giving us finer control over the minimum and maximum size of the groups. Alternatively, we may consider a more-fundamental switch to a more traditional "static framework" which uses a (larger) fixed basket with automated replacement techniques.

72. As discussed previously, we remain in an experimental research phase for clothing and have not planned an implementation date at present. However, clothing scores highly in our prioritisation framework (covered in APCP-S(23)06) so remains of interest.

Laura Christen, Ahmet Aydin, and Liam Greenhough
Methodology Division and Prices Division, ONS
October 2023

Annex A: References

Chessa, Antonio G. (2019) MARS: A method for defining products and linking barcodes of item relaunches - 16th Ottawa group Meeting

Eurostat Classification of individual consumption by purpose (COICOP)

Martindale, Hazel, Edward Rowland, Tanya Flower (2019) Semi-supervised machine learning with word embedding for classification - 16th Ottawa group Meeting

ONS (2020) Automated classification of web-scraped clothing data in consumer price statistics - Office for National Statistics

ONS (2022) Introducing multilateral index methods into consumer price statistics