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

Author: Liam Greenhough

Contents

1. Introduction

New data sources and methods are being introduced into the production of consumer price statistics from 2023. The new data sources have the potential to improve the quality of UK consumer price statistics through increased coverage and more timely data. Our new data sources cover a much wider range of products, and at larger frequencies and quantities, than is possible with any form of manual price collection. However, with these benefits, there are also challenges, and new methods are required to fully utilise these new data sources.

The applicability of these methods depends on the data source, for example, scanner data (point-of-sale transaction data, which tends to have less attribute information available), may be better suited to simpler classification techniques that do not require a lot of additional descriptors to help identify what category a product might belong to.

Clothing web scraped data is one of several data sources that ONS are investigating for introducing into consumer price inflation statistics. Traditionally, our clothing prices are collected from physical outlets, but increased spending in online-only clothing retailers in recent years means our market share coverage has been falling. Web-scraped clothing data provide several advantages to traditionally collected price data, including increased coverage of online-only retailers, but also increased product coverage per retailer and more frequently collected data. In any one month our web-scraped clothing data contain approximately 500,000 unique products. This compares with approximately 20,000 clothing products collected each month using traditional methods.

Clothing is a particularly dynamic market, which makes measuring price inflation challenging. In this paper we explore the factors that make clothing inflation difficult to measure, and the solutions that ONS are exploring using to implement web scraped clothing data into official statistics. Particularly, we explore using supervised machine learning to ensure we can classify at scale, and product grouping, to ensure index methods do not suffer from product churn.

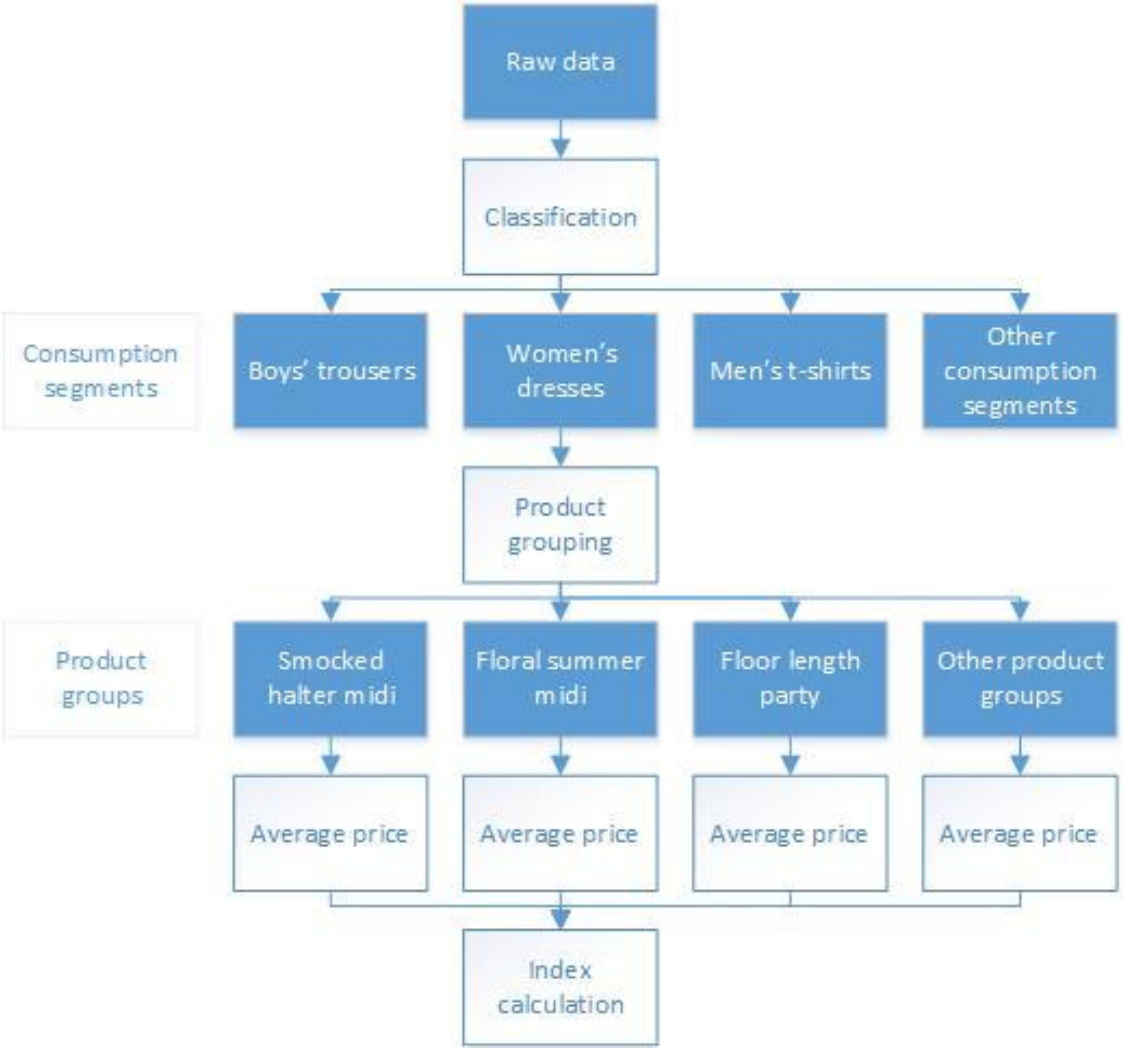
2. Summary of Classification and Product Grouping

In sections 3 and 4 we will discuss classification and product grouping (respectively), in greater detail. To begin with we will summarise the objectives of these two methods.

In Figure 1 we present a process flow for classification and product grouping. In classification, we have target classes (known as consumption segments) that we classify our raw data to. These classes are relatively homogeneous (such as women's dresses) and our goal is to produce inflation statistics for each of these consumption segments. (We will discuss an automated approach for classifying products to consumption segments in section 3.)

However, since clothing data are particularly dynamic, comparisons using index methods can become unrepresentative as more time goes by due to products leaving the dataset. To avoid this, we combine products into very homogeneous product groups, using the average prices within these groups as inputs for index calculations. The goal of product grouping then, is to create groups of products that appear throughout the year (even if individual product lines within the group fall in and out) to reduce product churn, whilst avoiding unit value biases from taking the average price of heterogeneous groups. (We will discuss this further in section 4.)

Figure 1. Products are first assigned a consumption segment during classification and then a product group within product grouping. Product group average prices are then used as an input into index calculations to produce an index for each consumption segment.



3. Clothing Classification

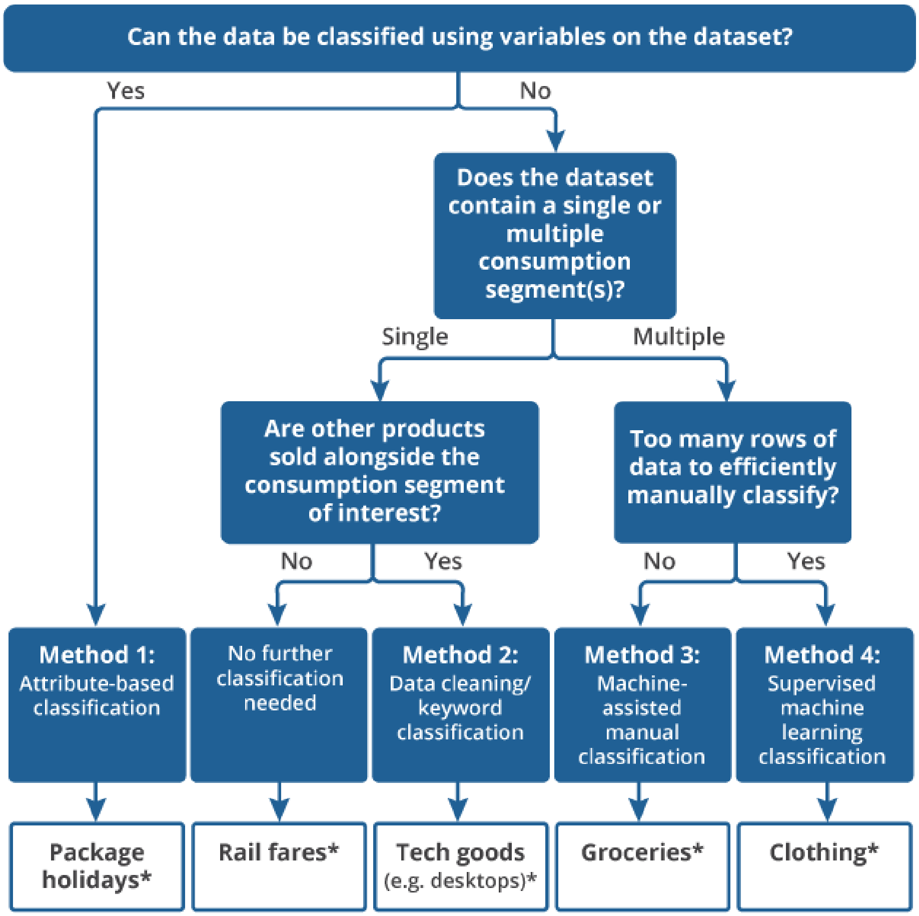
3.1. Introduction to classification

In a previous article, we introduced the concept of "consumption segments", that partition the consumption basket into groups of similar (homogeneous) 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".

The goal of classification is to ensure that products are assigned to the correct consumption segment. This can be a complex task when working with alternative data sources given their size and complexity relative to traditional survey-type data. Several different methods are available for classifying alternative data sources ranging from simple “keyword” classifiers, which link to an item category based on a single keyword in a product name, to more complex methods using supervised machine learning algorithms that identify unseen patterns in the data.

In figure 2, we present a variety of classification methods, and the circumstances in which we use each method, which we published in a previous article.

Figure 2. The classification of new data sources requires a suite of methods


Can the data be classified using variables on the dataset?
Yes
No
Does the dataset contain a single or multiple consumption segment(s)?
Single
Multiple
Are other products sold alongside the consumption segment of interest?
Too many rows of data to efficiently manually classify?
No
Yes
No
Yes
Method 1: Attribute-based classification
No further classification needed
Method 2: Data cleaning/ keyword classification
Method 3: Machine-assisted manual classification
Method 4: Supervised machine learning classification
Package holidays*
Rail fares*
Tech goods (e.g. desktops)*
Groceries*
Clothing*

*Recommendations for these categories are subject to change due to ongoing research. These categories are given as examples and are not an exhaustive list of all categories being explored.

We note that machine-assisted manual classification involves every product classification being manually scrutinised, but the process of finding the right classification is sped up by a machine providing recommendations. For example, within groceries, the machine may provide a set of "top 5" recommendations, avoiding the labeller needing to sift through over 600 labelling options for every product. Here, the goal of the machine is to improve the efficiency of manual labelling.

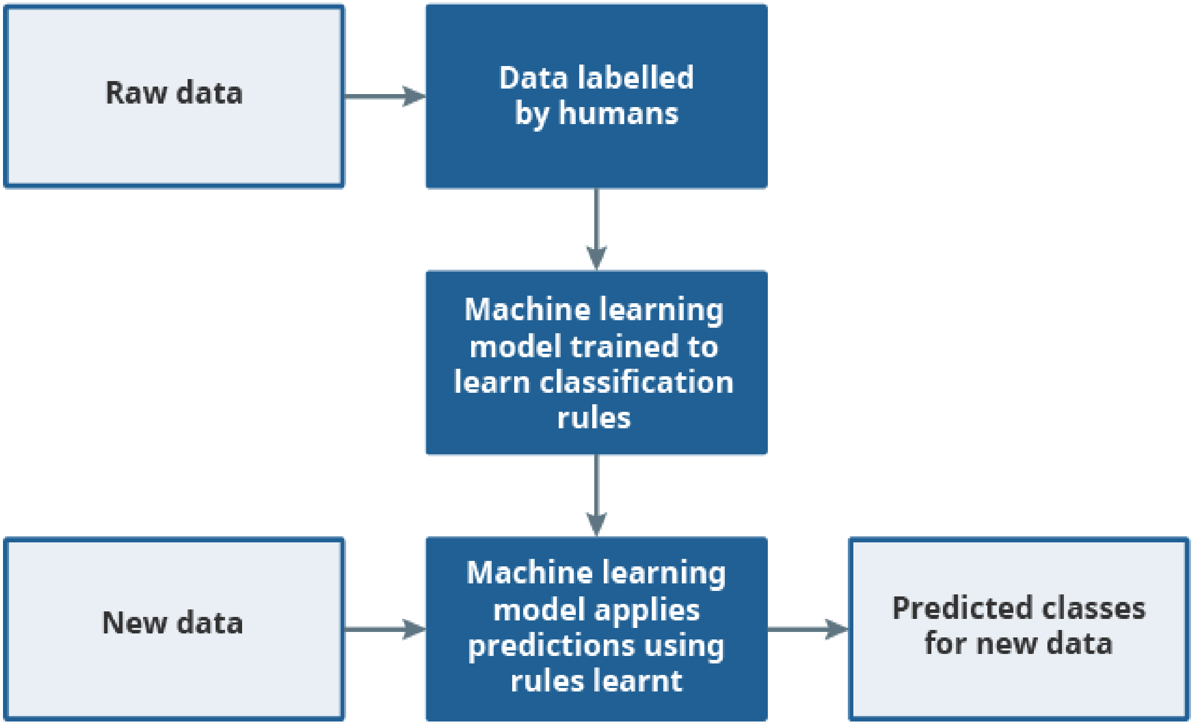
Some of the datasets that we are intending to use in the construction of UK consumer price statistics require so much monthly classification work that manual classification is not possible – even with machine-assistance. These datasets are typically large in product count and/or high in product churn[1]. In these cases, a more automated approach to classification is required to ensure large amounts of data can be classified rapidly. Supervised machine learning provides a potential method for carrying out this classification work.

ONS clothing datasets fall into this category. There are hundreds of thousands of new products each month and it would take thousands of hours of scrutiny to manually check every product. ONS needs a more-automated approach to classification, where the machine predicts rather than advises. In the next section, we discuss how supervised machine learning can be used.

3.2. Supervised machine learning

Supervised machine learning (ML) 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) and are trained to learn rules that associates the two. These rules are then applied to new, previously unseen data to make predictions on the correct classification. This process is shown in Figure 3.

Figure 3: Supervised machine learning uses labelled data to learn rules which are then applied to new unseen datasets


Raw data
Data labelled by humans
Machine learning model trained to learn classification rules
New data
Machine learning model applies predictions using rules learnt
Predicted classes for new data

[1] Product churn is where products enter and leave the market rapidly due to emerging technologies or fashion trends.

To build a supervised ML model, we first 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, from which we can test the performance of the predictions made by the model.

If we were to label a random sample of products, then we would expect the labelled dataset to have many women's shirts and few boys' gilets. The model trained would likely have greater success at identifying women's shirts compared to boys' gilets. To improve the balance between the two (and other types of clothing), we use stratified sampling with weights proportional to the number of consumption segments in that age and gender group (boys', girls', infants', men's, women's). For example, as 22 of our 85 segments are women's, we require 26% of the web-scraped clothing sample data to be women's. This attempts to account for there being a wider variety of women's clothing types sold, relative to boys' clothing types. We then stratify with equal weight the retailers and retailer hierarchy from each age and gender segment to cover all product types and all retailers.

For this analysis we have used a human-labelled dataset containing 1% of six months of web-scraped clothing data, equating to over 54,000 unique products. This was achieved by upwards of 30 people within the Office for National Statistics (ONS) Prices Division labelling clothing data using a bespoke labelling application that we developed in house. Labellers label the data at multiple levels of granularity, allowing us to compare classification performance at different levels of homogeneity. The six currently sampled months are spread across the year to ensure seasonal variations in clothing are accounted for. We have tried to ensure that our human-labelled data are as consistent as possible using the bespoke labelling application, training for human labellers, and a detailed frequently asked questions (FAQ) document.

3.3. Labelling consistency

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". 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, 12 labellers manually classified the same 313 clothing products to consumption segments, without consultation.

Since the placement of products can be subjective, we treat the chosen majority as being the correct consumption segment. For example, if 11 people have labelled a product as a "dress" and only one person has labelled it as a "skirt", then we assume dress to be the correct consumption segment.

Figure 4 shows that, on average, labellers were consistent with the majority consumption segment in 88.8% of cases. Most labellers were closely distributed, showing between 88% and 92% consistency with the majority consumption segment. Consistency with the majority consumption segment was always greater than 81%.

Figure 4: Around 89% of labels are consistent across human labellers

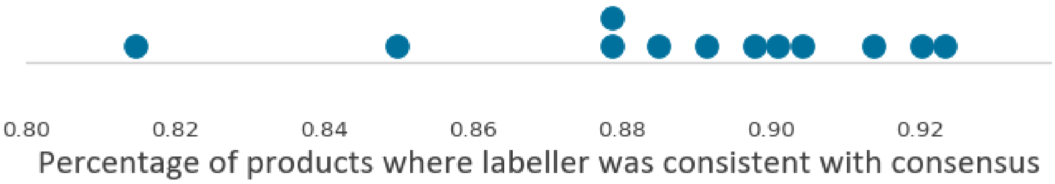
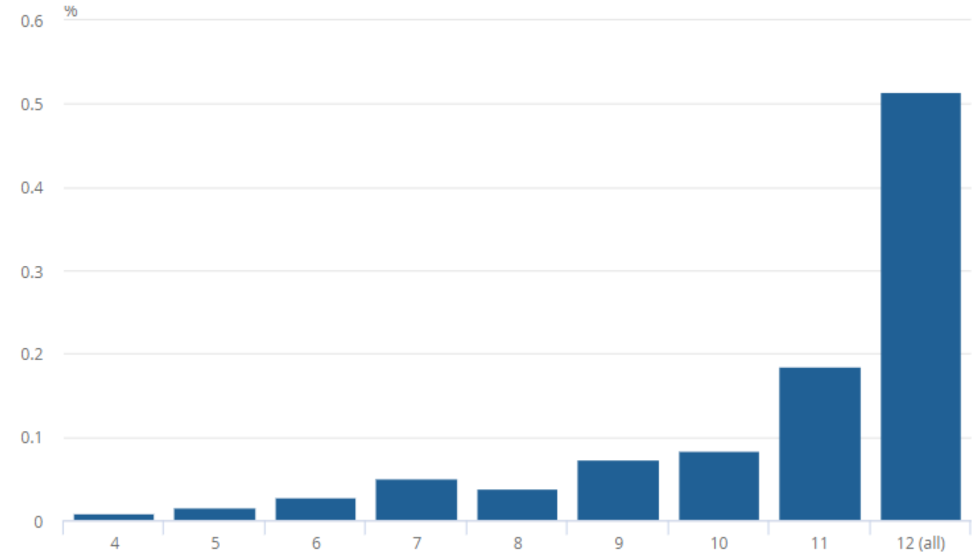


Figure 5 shows for each product, the number of labellers who have chosen the majority class. For more than half of the products, all twelve labellers agreed on how to classify the product, and all but one labeller agreed in a further (approximately) 20% of cases.

Figure 5: There was a broad consensus on how to classify most products



The products that split opinion the most strongly were unsurprisingly 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. Example products include:

- "fleece zip through jacket" (fleece=5 / jacket=5 / waterproof=2)
- "sporty hooded jacket" (hoodie=2 / sports jacket=5 / jacket=5)
- "[sports brand] zip top" (sports top=4 / sports jacket=4 / other=3 / t-shirt=1)

Further inconsistencies were seen in products that were ambiguous with regards to their pack contents. For example, a product was described as a "two-piece jogger set". How labellers classified this product was dependent on what the labeller perceived "two-piece" to mean: a two-pack of jogging bottoms; or a full tracksuit, including a top and a pair of jogging bottoms.

As well as these ambiguities, some explicit labelling errors were also observed. For example, one product was described as "floral print shorts (30 waist)" with no further supplementary information. Some labellers did not pick out that a "30 waist" is typically a men's size in British clothing standards and labelled the product as unisex.

We can use this information to refine our guidance to our human labellers. But while consistency of labelling can be improved through further training and detailed guidance, there remains a limit as to how consistent our labels can be, given the subjective nature of some clothing items as we have seen.

This "ceiling effect" in consistency therefore also provides something of a benchmark for our automated classification models. We are unlikely to accurately classify every product through either manual or automated classification, but we can strive for our automated classification to classify products to the same level of consistency as our human labellers have demonstrated.

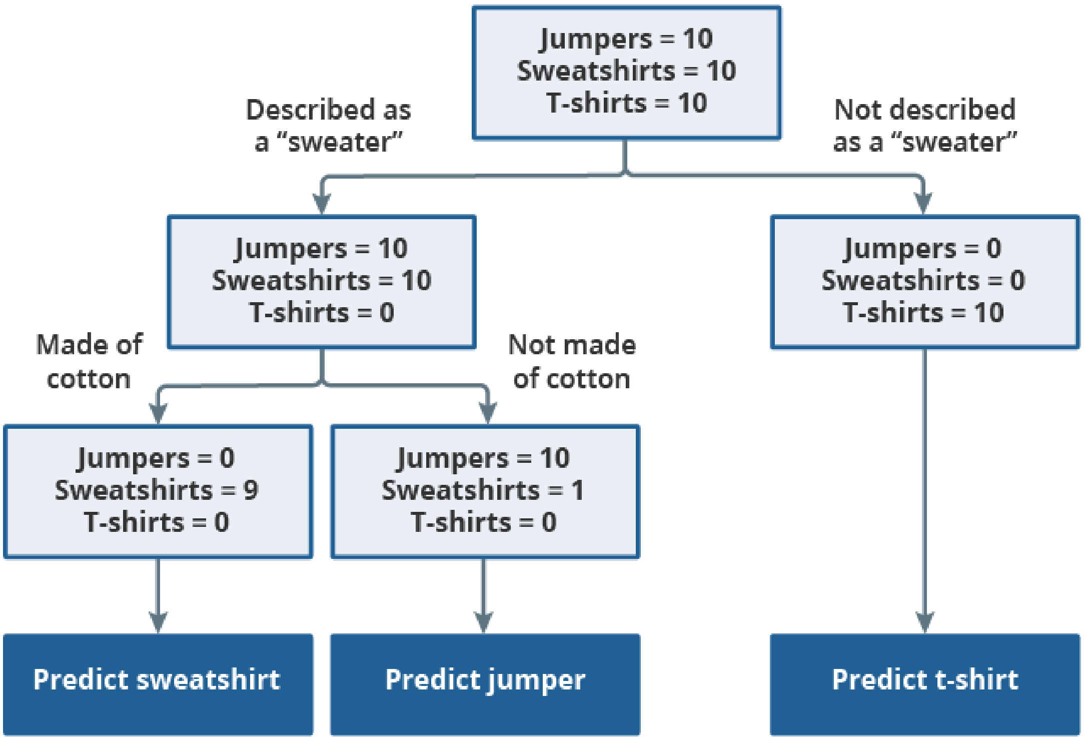
3.4. Machine learning classification models

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 various 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).

We are using these features along with our human-labelled data to train and test the performance of machine learning (ML) models. There are many ML algorithms that have been developed to perform classification tasks. Not all algorithms are suitable for all datasets or tasks, and different algorithms have different levels of complexity and transparency. For this article we primarily demonstrate results based on gradient-boosted trees (specifically XGBoost), as these are currently our highest-performing algorithm.

XGBoost uses decision trees as a foundation. A decision tree can conceptually be thought of as a flowchart, as displayed in Figure 6. 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. These rules are then applied to new, previously unseen, data to make predictions as to the correct product classification.

Figure 6: Decision trees use binary decisions to split the data



Decision trees are among the most basic classification algorithms and are relatively poor performers for complex classification tasks. However, they can be trained very quickly so are often used in ensemble models, such as gradient-boosted trees. An ensemble model involves training numerous classifiers and then using them in conjunction to provide a final decision on classification.

Gradient-boosted trees (such as the XGBoost algorithm used in our analyses) involve training many decision trees sequentially, with each tree trained to improve on the errors made by previous trees. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by all previous tree models.

3.5. Measuring the performance of our machine learning classification models

As we have seen, even with well-trained machine learning (ML) classification models it is unlikely that the predictions will be accurate 100% of the time given the ambiguities in clothing data and imperfect human-labelled datasets. To be able to include these classifiers in our official UK consumer price statistics we must have appropriate metrics to measure and monitor classifier performance over time. We also need to provide a performance threshold that a classifier will need to exceed to be used and understand the impact of classification inaccuracies on our consumer price statistics.

There are numerous metrics to quantify classifier performance on a dataset, each with different properties. In earlier work, we published an article for the Technical Advisory Panel on Consumer Prices (PDF, 1.2MB), discussing the appropriateness of various classification metrics. For this article, we focus on two lower-level metrics:

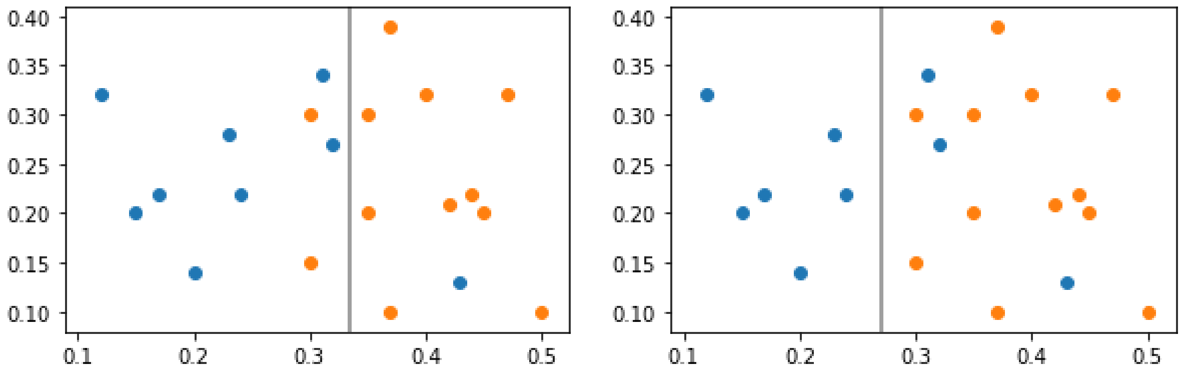
- Precision: measures the purity of a consumption segment. A segment with 90% precision would mean 10% of the elements classified to the segment are from other segments (false positives).
- Recall: measures the extent to which all cases from the consumption segment are captured by the classifier. A class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

There is a trade-off between precision and recall. To capture the trade-off between precision and recall, we report a third metric:

- F1: the harmonic mean of precision and recall. In this article, precision and recall are equally weighted. In future we may choose to weight precision as more important than recall, or vice versa, producing a more general F-score.

Figure 7 shows the trade-off between precision and recall. Two consumption segments are split by two different classifiers. Classifier 1 (left) splits the data to give greater recall for the blue class (it captures most of the blue cases, but is impure in that it also captures a couple of orange cases), whereas Classifier 2 (right) provides greater precision (it captures only blue cases, but a few blue cases are also lost to the orange class).

Figure 7: Classifier 1 gives the blue class better recall whereas Classifier 2 gives the blue class better precision



By taking an average value (with equal weighting) for the precision, recall and F1 across all consumption segments, we can report classifier performance at an aggregated level. This is known as macro-averaging; other methods of averaging are also available and discussed in our Technical Advisory Panel on Consumer Prices article, but are not covered in this article.

In Table 1 we report on a small number of our 85 current consumption segments, alongside the macro-averaged figure across all 85 consumption segments. We see that there are some very high-performing consumption segments, but the macro-averaged F1 score is weighed down by some lower-performing consumption segments. Note that, for the results presented in this article, we have trained our models on three months of labelled data and tested our models on the remaining three.

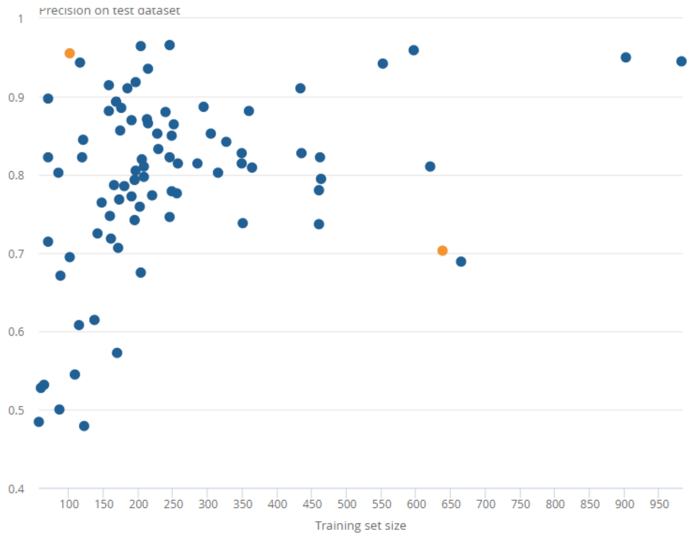
Table 1: Precision, recall and F1 scores for five classes are given, along with macro-averaged scores

	Precision	**Recall**	**F1**
Women's dresses	0.95	0.89	0.92
Women's jeans	0.96	0.84	0.9
Women's skirt/shorts	0.88	0.8	0.84
Women's sports bottoms	0.71	0.57	0.63
Women's suit jacket	0.53	0.58	0.55
Macro-averaged	**0.791**	**0.757**	**0.772**

The table gives an intuitive sense as to where some of the classification challenges lie. For example, the classification of jeans is high performing, likely because the classifier can often determine jeans simply through the word “jeans”. By contrast, sports bottoms use similar terminology to non-sports bottoms (such as “bottoms”, “shorts” and “leggings”), which may lead to confusion in the classifier, and worse performance.

Another challenge is the relative size of different consumption segments. There are many more dresses in our training dataset than suit jackets and the predictive performance likely reflects this. Figure 8 shows the relationship between the training set size and precision scores on the test dataset.

Figure 8: Classes with more training data generally classify better



As can be expected, segments with more training data generally perform better and we are expecting performance to continue to improve for several of our segments as we label more data. However, some segments perform relatively well on low volumes of training data and some segments perform relatively badly on high volumes of training data. For example:

- Infants’ outfit sets (represented by the orange dot on the right hand side of Figure 7) have high volumes of data and relatively poor precision; this likely reflects the variety within the class and the high overlap with other classes, for example, an outfit set that contains a t-shirt and jogging bottoms may cause the classifier to incorrectly identify the product as a t-shirt or jogging bottoms.
- Girls’ swimwear (represented by the orange dot on the left hand side of Figure 7) has low volumes of data and relatively high precision; this likely reflects that simple words such as “swimwear” are enough to identify such a broad category of swimwear because there is such a low overlap between swimwear and other classes.

The latter exception demonstrates the trade-off between the requirement for homogeneous consumption segments and classifier performance. Breaking down the swimwear class into swimsuits and two-piece swim sets will increase homogeneity but results in consumption segments with smaller amounts of training data, and it may require the classifier to learn more complex rules to distinguish between different swimwear types.

A confusion matrix can be used to compare the machine’s predictions against human labels. We can use confusion matrices to understand the consumption segments that the classifier is struggling to distinguish between. A small portion of our confusion matrix is available in Table 2. As might be expected, the ML classifier sometimes confuses Women’s suit jackets with Women’s outerwear jackets.

Table 2: A confusion matrix can show which classes the machine struggles to distinguish between

	Classifier prediction			
Class given by labeller	**Girls' coat or jacket**	**Women's coat or jacket**	**Women's suit jacket**	**Other categories**
Girls' coat or jacket	197	1	0	37
Women's coat or jacket	5	167	14	44
Women's suit jacket	0	9	29	12
Other categories	29	31	12	21098

In an example from our wider confusion matrix, the classifier also often makes mistakes between male underwear and swimwear, likely since “trunks” can appear in both. Note that the classifier is generally quite good at distinguishing between age groups; there is little evidence of misclassification between girls’ coats and jackets and women’s coats and jackets. This may be because of the feature that targets age.

When exploring the consistency of human labelling, we saw that labellers are approximately 89% consistent in their labelling, setting a benchmark for what a classifier may be able to achieve. In this section we have seen that, currently, our best performing classifier achieves a macro-averaged precision score of around 79%.

In Table 3 we compare macro-averaged performance of various classification algorithms. We note that XGBoost and Support Vector Machine appear to perform best. It is worth noting that, aside from classification performance, there are other factors that go into how suitable a classification method is. For example, although Support Vector Machines are competitive in classification performance in this table, they are also (often) comparatively time-consuming to train and make predictions with which may make them less practical for use in production. There are other factors to consider such as interpretability, whether the method can provide a confidence score alongside its prediction and the resource needed to maintain the system. Our final assessment of the most appropriate method will also have to consider these factors. For most of the paper we have used XGBoost as our primary algorithm since it is competitive in both performance and training times.

Table 3: We have observed XGBoost and Support Vector Machines to perform particularly well in complex multiclassification tasks

Classifier	Macro F1	Macro precision	Macro recall
XGBoost	0.772	0.791	0.757
Support Vector Machine	0.799	0.783	0.826
Logistic Regression	0.759	0.722	0.818
Random Forests	0.726	0.679	0.802
Decision Tree	0.42	0.389	0.487

3.6. Classifier performance: what is good enough?

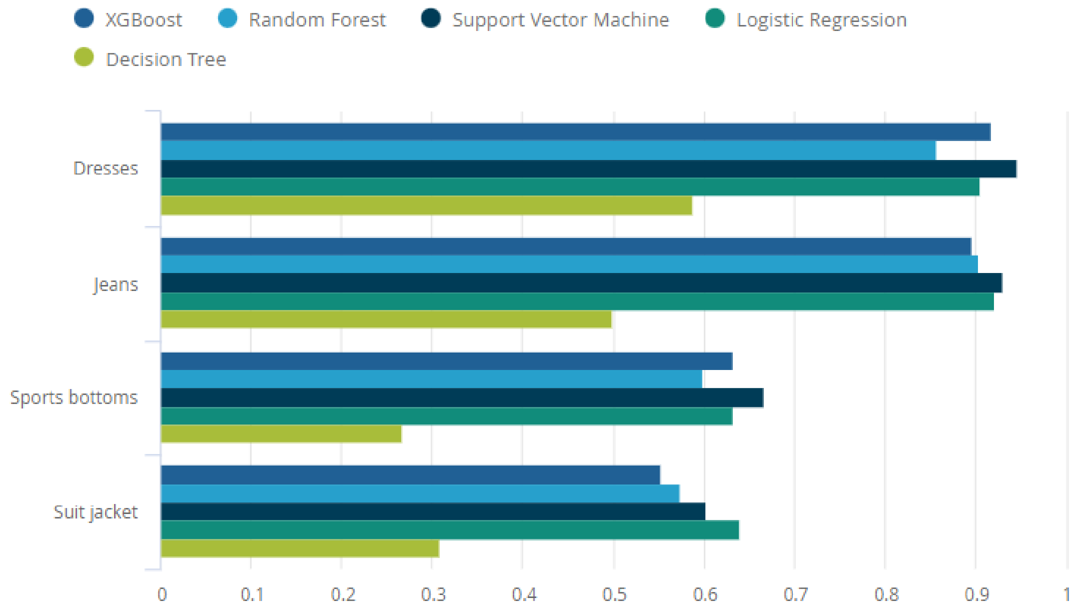
It is not realistic to expect that a machine learning (ML) model would be correct in every single instance. The question then is: when is a machine learning classifier perceived to be good enough for use in the production of consumer price statistics? The goal should be to produce a classification model, such that the indices formed on its predictions are close to the indices that would be produced if all classifications were correct. That is, classification performance is good enough to produce unbiased indices.

At present we do not have a long series of consecutive months of labelled data to compare indices based on our model predications against. Instead, we consider the closeness of the indices for four different Women’s clothing consumption segments (dresses and jeans are chosen as high-performing segments, sports bottoms and suit jackets are chosen as low-performing segments).

These indices are produced based on classification predictions made by five different ML algorithms. These are XGBoost, Random Forest, Support Vector Machine, Logistic Regression and Decision Trees. We compare indices using the rolling year GEKS-Jevons index method as this is currently our top shortlisted index number method for unweighted data.

As seen in Table 1, dresses and jeans classify relatively well using XGBoost, whereas sports bottoms and suit jackets classify relatively poorly. In Figure 9, we show the F1 score for these items using a range of automated classification algorithms.

Figure 9: High F1 scores are shown for dresses and jeans; low scores for sports bottoms and suit jackets



Note that decision trees struggle with complex multiclassification tasks and performs much worse than the other four algorithms for each of these consumption segments. The four other classifiers perform relatively similarly, with higher scores for dresses and jeans and lower scores for sports bottoms and suit jackets.

The price indices presented in this section are experimental and should not be taken as official estimates of market behaviour. All graphs exhibit a downwards trend which is not necessarily true. In our product grouping work (section 4) we will cover the factors that lead to these downward biases. Instead of focusing on specific numbers, we consider the closeness of the indices presented.

Figures 10a and 10b shows price indices for our two high-performing classes: dresses (10a) and jeans (10b). Unsurprisingly, indices using products classified by decision trees behave differently to indices produced using products that have been classified by higher performing classifiers. This is likely because of the previously mentioned low classification performance of decision trees resulting from their overly simplistic design. Reassuringly, the four other classifiers result in relatively close indices for dresses and jeans, suggesting that they are likely classifying the same products.

Figure 10a: When classification performance is high (as for dresses), indices are tight and co-move closely, whereas decision trees are poor-performing classifiers and diverge

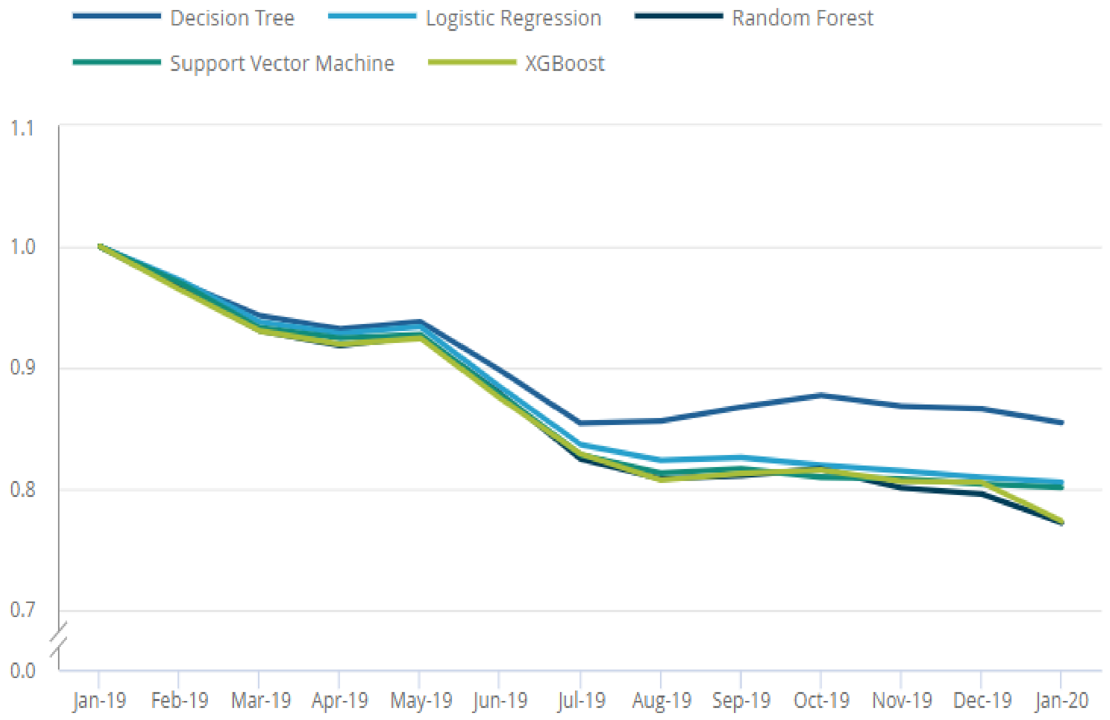


Figure 10b: The classifier performs well at detecting jeans and the resulting indices are close

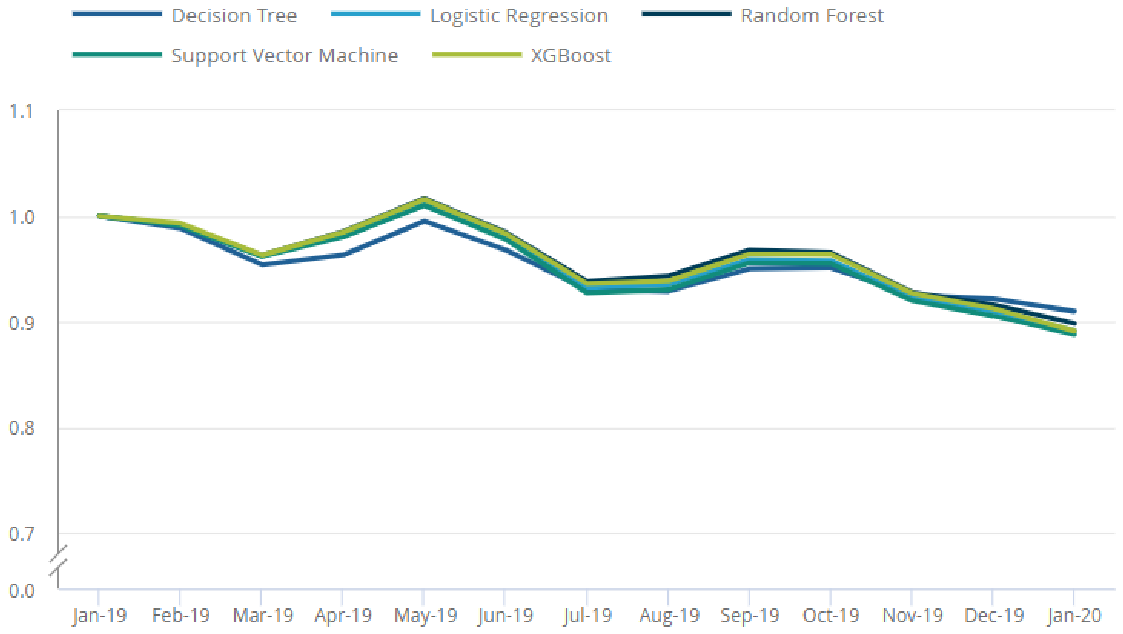


Figure 11a and 11b show indices for two lower-performing classes: sports bottoms (11a) and suit jackets (11b). The differences between the indices are now more pronounced. This is particularly the case for suit jackets where misclassification combined with small sample counts leads to volatility in the indices. Despite being one of our lowest-performing classes, the indices for sports bottoms co-move reasonably well showing that the indices can be quite resilient to classification error.

Figure 11a: When classification performance is lower, the difference between the indices are more pronounced

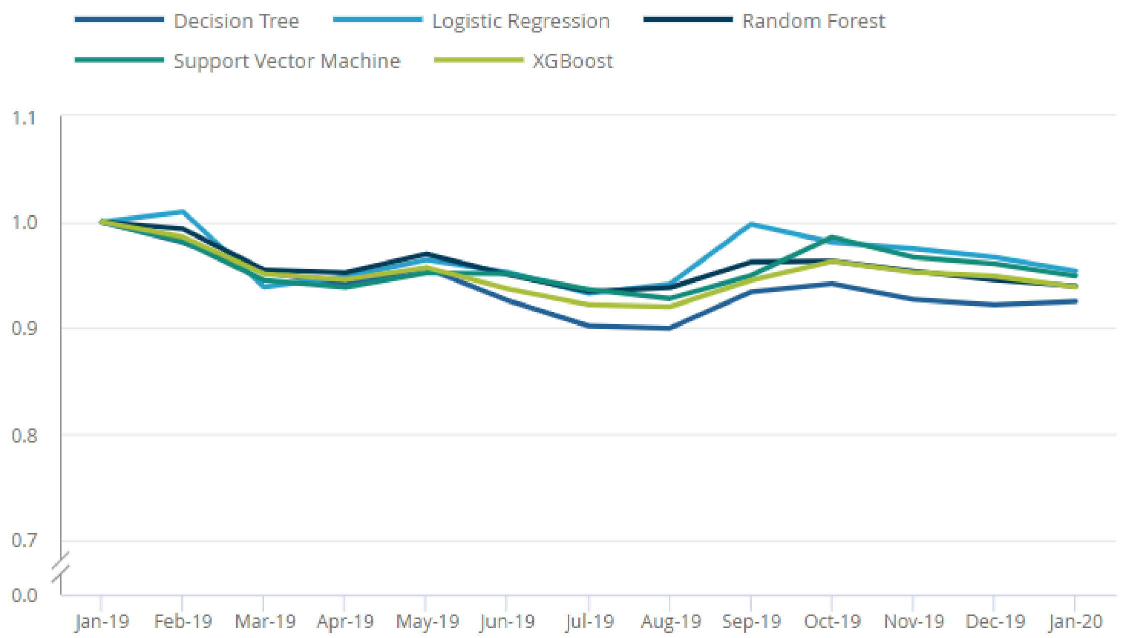
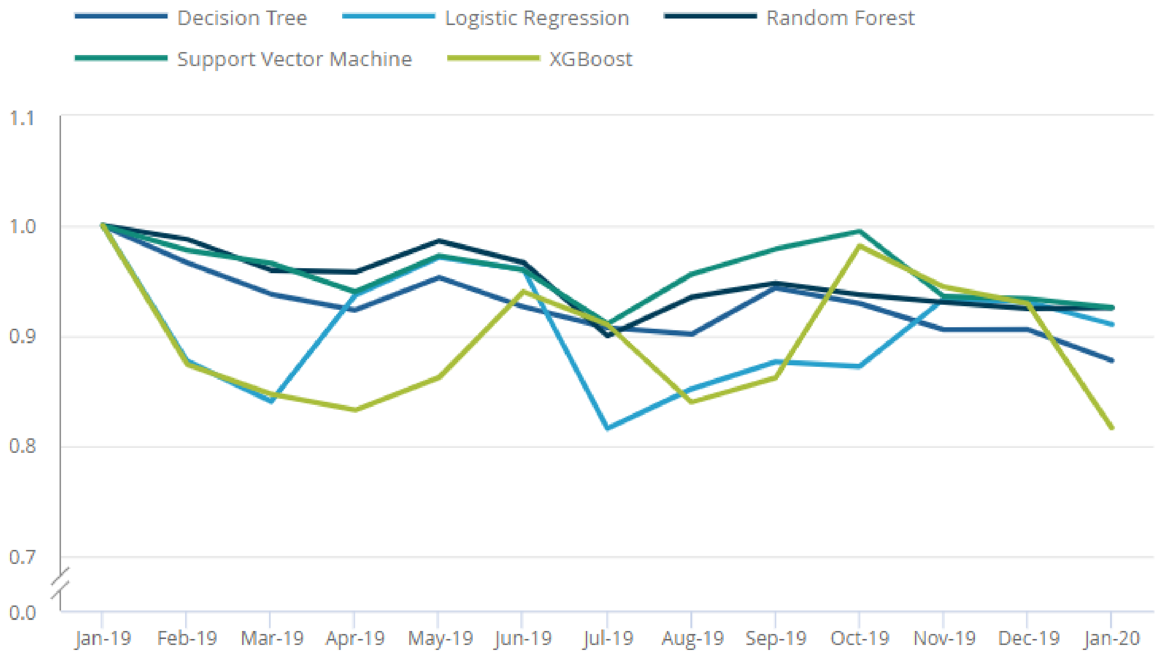


Figure 11b: Poor classification performance paired with low sample counts can result in divergent indices, as is the case for women's suit jackets



Figures 10 and 11 show that higher classification performance generally reduces the variability in index values, showing the need to classify products well. However, there is a possibility that the four high-performing classifiers in Figure 9 are making the same errors, and these errors are causing a bias in the index. For example, it may be that all four classifiers (XGBoost, Support Vector Machines, Random Forests and Logistic Regression) make similar errors and fail to capture denim dresses (perhaps misclassifying them as jeans). If there is a rise or fall in the manufacturing costs of denim, the four indices presented would not be affected but the "true index" would.

4. Product Grouping

4.1. Challenges associated with measuring clothing inflation

The nature of clothing markets results in measurement challenges primarily due to the pace that products enter and exit the market. These challenges have also been experienced historically (PDF, 61KB) but, with the proposed move to include web-scraped prices, we must develop new methods to cope with the larger scale and complexity of these data.

Inflation is traditionally measured by tracking prices of individual products and aggregating them through time. Clothing products rarely exist for more than a few months; they experience high "product churn". In our web-scraped clothing data, around 150,000 of the 500,000 (approximate) products leave and enter the market each month. Table 4 shows how, under a simple example of these conditions, price indices can be unrepresentative as they only use products that exist in both periods.

Table 4: Due to product churn, the price index for August can only make use of only one product

Product	Price, Jan	Price, Aug	Price change (Aug/Jan)
Floral winter dress 1	18		Cannot form
Floral winter dress 2	18		Cannot form
Floral winter dress 3	24	18	0.75
Floral summer dress 1	60		Cannot form
Floral summer dress 2		45	Cannot form
Floral summer dress 3		45	Cannot form
Party midi dress 1	100		Cannot form
Party midi dress 2		90	Cannot form
		Price index	**0.75**

Furthermore, clothing products typically enter the market at a high price and leave at a low price, often on a clearance sale. This creates an implicit price increase when products are replaced which is not normally captured by index methods (as shown in Figure 12). In our web-scraped data, prices of individual products fall on average by 3% between two consecutive months.

Figure 12: Index methods may only capture falls in prices of individual product lines (shaded lines) and not price resetting when product lines are replaced (dotted lines)

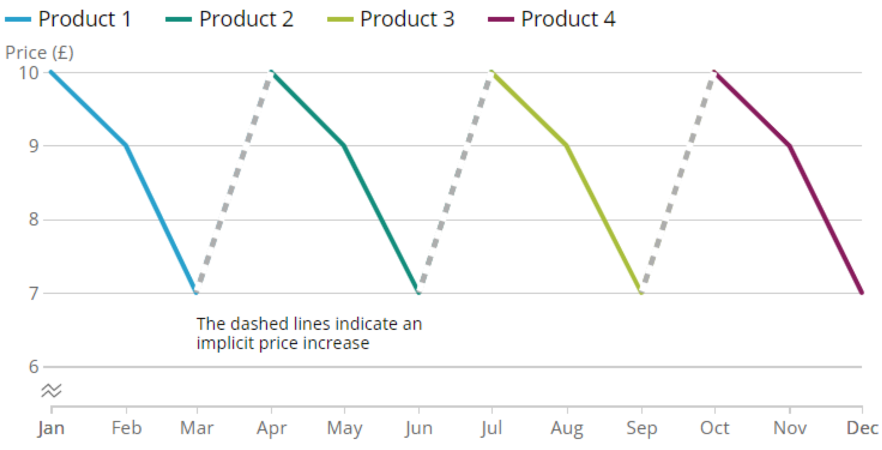
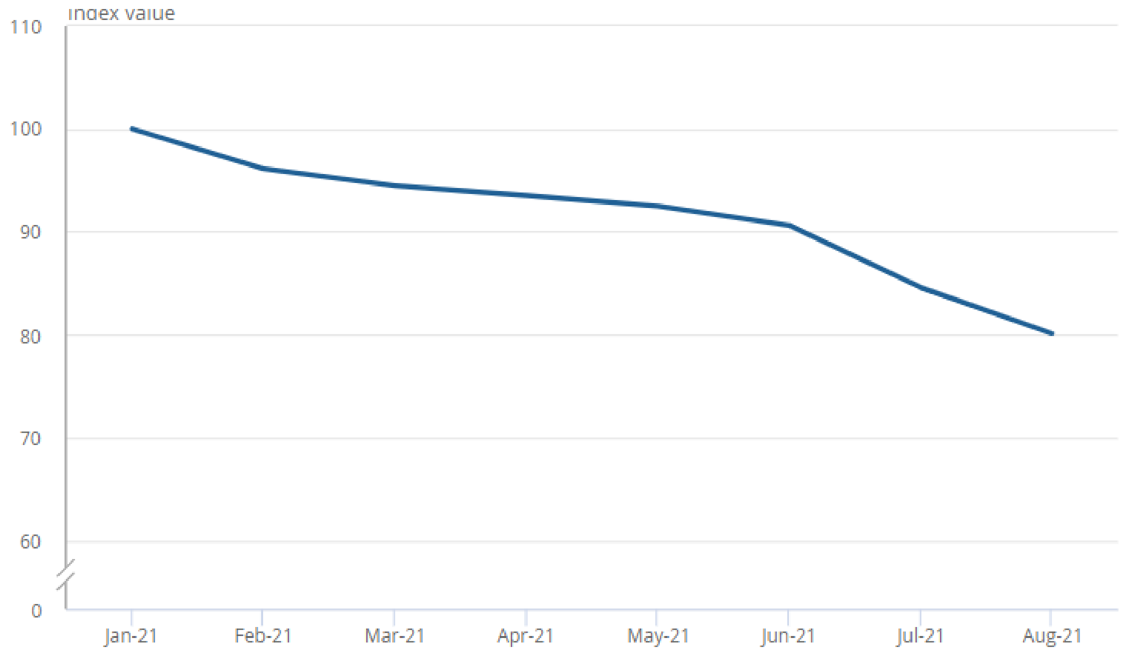


Figure 13 shows a GEKS-Jevons index (our currently preferred unweighted method) on web-scraped women's dresses data, without applying methods to account for these measurement challenges. Within just seven months, prices fall 20%: the index captures the declining price of individual products, but not implicit price resetting that occurs when new equivalent product lines are released.

Figure 13: An unadjusted price index for women's dresses using web-scraped data falls by 20% in seven months



Traditionally, we account for clothing measurement challenges by manually linking products leaving the market with suitable replacement products. However, it is unfeasible to scale this manual replacement process for the 150,000 (on average) leaving products every month when using web-scraped data. Instead, we explore using a method we refer to as product grouping.

4.2. Product grouping

Web-scraped data are first classified to clothing consumption segments, such as women's dresses, as discussed in section 3. Within consumption segments, groups are formed of similar products, such as floral summer dresses. We then track the average price of each group instead of each individual product. We have explored similar concepts previously in Clustering large datasets into price indices. Countries such as the Netherlands (PDF, 860KB) and Belgium (PDF, 436KB) have also explored similar applications.

Table 5 shows a stylised example of how we could group a sample of dresses, based on their product names, to reduce the effect of product churn. Note that every product is impacting the average prices, and these average prices are used in the index calculation, so every product can influence the index. This is in contrast to table 4, where only one product was used.

Table 5: Grouping products and taking their average prices allows us to use products even when matches cannot form over months

Product	Price, Jan	Price, Aug	
Floral winter dress 1	18		
Floral winter dress 2	18		
Floral winter dress 3	24	18	
Floral summer dress 1	60		
Floral summer dress 2		45	
Floral summer dress 3		45	
Party midi dress 1	100		
Party midi dress 2		90	**Price change (Aug/Jan)**
Average price: Floral winter dresses	20	18	0.9
Average price: Floral summer dresses	60	45	0.75
Average price: Party midi dresses	100	90	0.9
		Price index	**0.85**

Since price changes are measured within product groups, that are more likely to be available throughout the year than individual products, product churn is reduced. Furthermore, since products entering and leaving the market influence the group average price, the implicit price resetting as new products enter the market is also captured.

However, using product grouping introduces a new measurement challenge. Consider table 6, where instead of winter and summer dresses, all floral dresses are combined into a single group. This combined group shows a price increase, however, when considered independently, floral summer and winter dresses both show price reductions (as shown in table 5). The floral group rises in average price because the composition shifts from winter to summer dresses, that are more expensive in this example. This grouping is too broad; products not similar in price and purpose are grouped together causing the index to change because of composition rather than price. In other words, the group is not “homogeneous” enough. We will describe this in the next section.

Table 6: By being too broad, the average price in floral dresses shifts due to compositional effects, demonstrating the need for homogeneity

Product	Price, Jan	Price, Aug	
Floral winter dress 1	18		
Floral winter dress 2	18		
Floral winter dress 3	24	18	
Floral summer dress 1	60		
Floral summer dress 2		45	
Floral summer dress 3		45	
Party midi dress 1	100		
Party midi dress 2		90	**Price change (Aug/Jan)**
Average price: Floral dresses	30	36	1.2
Average price: Party midi dresses	100	90	0.9
		Price index	**1.05**

4.3. Assessing product grouping

The objective of product grouping is to create groups that are large enough to control for product churn but of a similar enough quality (known as "homogeneous") that compositional effects do not bias inflation. To measure these competing goals, Chessa (2019; PDF, 860KB) introduced "Match Adjusted R Squared" (MARS). MARS is the product of two components, both in the range [0, 1]:

$$MARS_{t0,t} = (match\ rate)_{t0,t} \times R_t^2$$

Where:

- Match rate measures product churn as the proportion of matching products or groups from the base to current month[2]
- R-squared measures in-group price similarity within the current month

MARS accounts for homogeneity within groups based on price similarity, but this is only one measure of homogeneity. Guidance from the International Monetary Fund recommends homogeneity should also account for purpose. For example, a group containing a £20 t-shirt and a £20 pair of shoes are homogeneous in price, but not purpose. We are exploring measuring purpose homogeneity through human evaluation of product similarity within groups, but this is ongoing research and is not presented in this article.

4.4. Attribute-based product grouping

Attribute-based product grouping creates groups of products that share characteristics. For clothing data, we primarily form attributes through text matching (for example, whether products contain the word “cotton”). A simplified example of this approach is provided in Table 7.

Table 7: Groups are formed based on text matching on “important” words found in the product name and material columns

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

We need grouping models for every consumption segment, such as women's dresses, boys' t-shirts and men's jeans. While models can be created from user-defined words, an automated approach is preferred for scalability.

For each clothing type, we generate groups using the most-commonly occurring words. Note that common words are likely to express quality-defining characteristics of products, for example, style, material and colour. To improve the grouping, common non-quality defining text are removed, including punctuation, numbers and stop words (such as "and", "but", "to"). We choose how many of the top words to use based on performance, currently assessed through MARS.

[2] Chessa (2019) uses expenditure to weight the match rate. Since we lack this information in web scraped data, we use an unweighted variant.

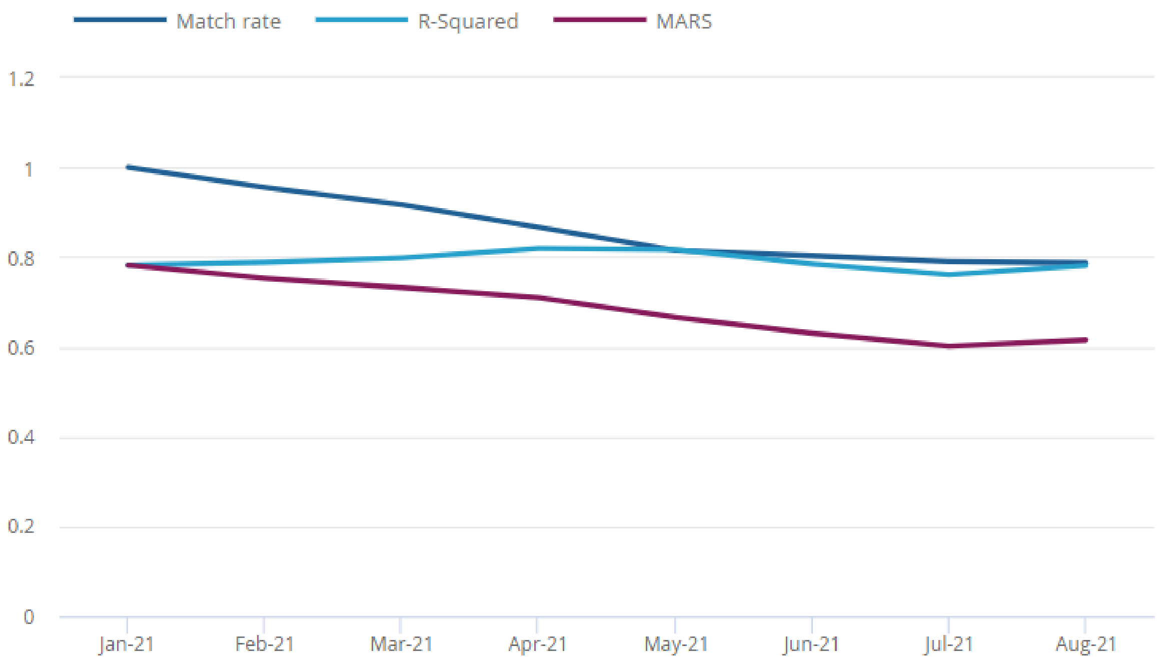
This method is relatively simple to deploy, scalable and we have found it has some desirable advantages relative to more advanced clustering methods that we have tried. More detail on this approach and our clustering approach can be found in Dealing with product churn in web-scraped clothing data: product grouping methods (PDF, 416KB).

4.5. Results

We fitted the model using web-scraped data from June to December 2020, which determines the most common words and how many to use (with a longer data supply, we will use a minimum of a year to capture seasonal variation). We then applied the model to data from January 2021 to August 2021 to create product groups, measure "Match Adjusted R Squared" (MARS) and produce price indices.

In Figure 14, MARS scores are shown for women’s dresses after applying product grouping. The method scores reasonably high on both the match rate and R-squared components. However, these results are experimental as we are continuing to refine our methods.

Figure 14: MARS scores for women’s dress groups January 2021 to August 2021



In Figures 15 and 16 we compare our R-squared and match rate against two benchmarks: tracking individual dresses and tracking a single group of all dresses. A perfect R-squared can be achieved by tracking individual dresses (Figure 4), but this results in a low match rate (Figure 5). Conversely, a perfect match rate can be achieved through grouping all products into a single group (Figure 5) but results in low price similarity within this single group (Figure 4). Our product grouping for dresses is shown to balance these two desired features of products used to construct a price index.

Figure 15: Attribute-based product grouping creates dress groups with greater price similarity than tracking a single group of all dresses (R-squared measures similarity of prices within groups)

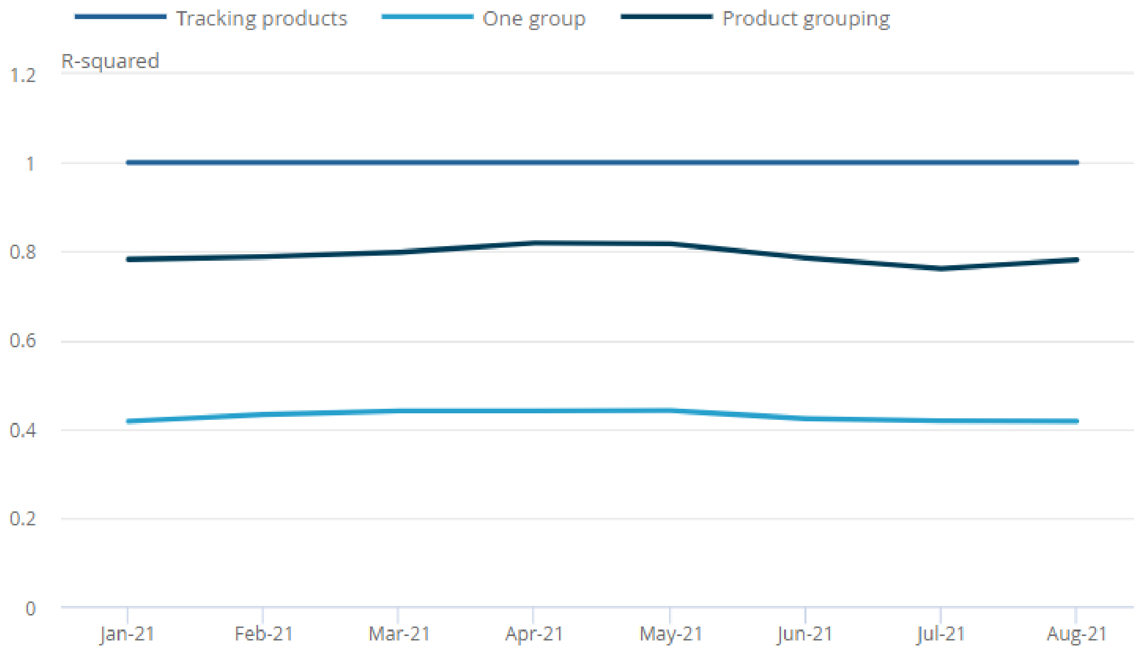


Figure 16: Attribute-based product grouping creates dress groups with a higher match rate than tracking individual dresses (match rate measures how long groups remain on the market)

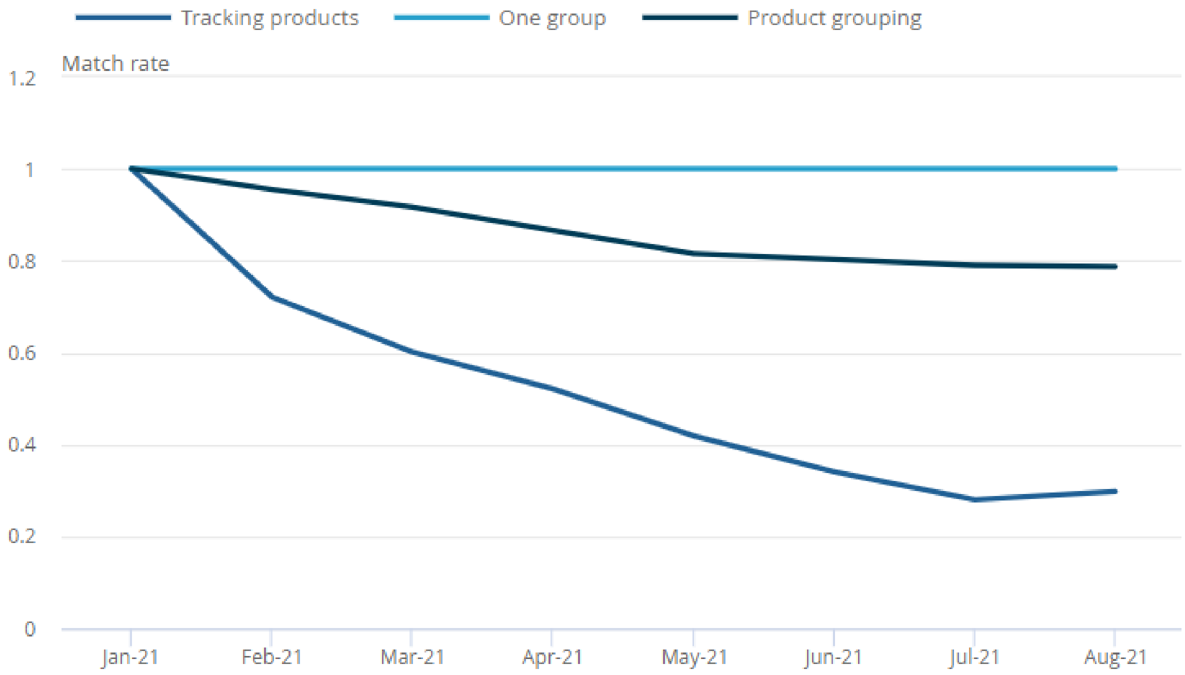
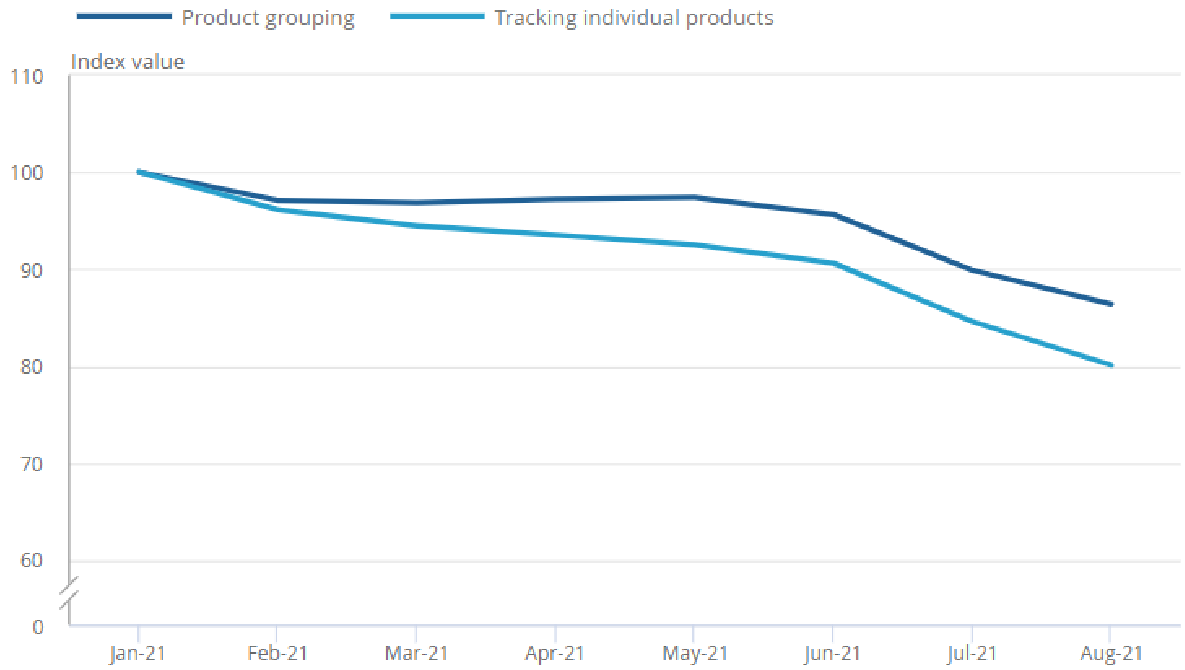


Figure 17 compares the price index before using product grouping (as shown in Figure 13) with the index after using product grouping. The method appears to capture the same seasonal patterns as not applying grouping but mitigates some of the fall in the index.

Figure 17: Product grouping appears to capture the same seasonal patterns as tracking individual products but mitigates some of the fall in the index



Note that although product grouping promisingly appears to mitigate the fall in the index, it is still unclear whether women's dresses falling in price by approximately 13% between January and August 2021 is realistic.

In our future work we will look to:

- use more data to both fit our model (to ensure variation in seasonal terminology is captured) and extend the index data time series to understand the longer-term impact of product grouping on indices
- productionise the system to ensure the method and these analyses can be scaled and applied to other clothing consumption segments
- make further refinements to the method with the goal of improving "Match Adjusted R Squared" (MARS); in particular refining which words are chosen to form groups
- explore measuring purpose homogeneity to determine whether the groups formed make an intuitive sense to consumers
- consider how we can better account for consumption patterns within the index, by applying retailer market shares and weighting individual products, or product groups, by estimates of consumer expenditure

5. Summary

In this paper, we have explored how to solve two issues. Firstly, how to classify web scraped clothing data at the scale we need to; and secondly, how to avoid an index that drops rapidly due to extreme product churn. Our proposed methods involve using supervised machine learning to resolve the former, and product grouping to resolve the latter.

Within supervised machine learning classification, we have seen mixed success. The method is scalable and so we can classify the entire dataset in a way that would not be achievable using more manual methods. However, the performance of the classifier varies by consumption segment. In general, the ability of the classifier to classify to the correct classification is dependent on the labeller's ability to distinguish between different clothing types. Where labellers struggle to distinguish between different clothing types, such as determining whether an item of clothing is sportswear, the machine also struggles to make predictions.

Within product grouping, our methods are promising for at least mitigating the (unrealistic) fall in the index that is observed when producing indices from individual product lines. However, it is unclear without a longer time series whether our current proposal mitigates or fully solves this problem. We are also looking towards ways of improving our methods to improve the MARS scores that underpin the quality of our product grouping method.