ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Dealing with product churn in web-scraped clothing data: product grouping methods

Status: Work in progress
Expected publication: For publication alongside minutes

Purpose

1. This paper discusses the methods and metrics under consideration for forming homogenous groups of products in web-scraped data, to reduce the effect of product churn. Two different methods of forming groups are presented along with discussion of how to assess the group properties. This paper discusses the options currently under consideration and our plans for further work.

Actions

2. Members of the Panel are invited to:
    a. recommend which elements of product homogeneity are most important for price indices.
    b. consider impact of seasonality on groups and the effect on the choice of base period for comparing groups too.
    c. comment on pros and cons of the two proposed methods of group formation.
    d. comment on the proposed metrics to measure homogeneity in groups.
    e. comment on our plans for future work.

Introduction

1. When following products through time it is natural that some of these products will drop out of the market as they are removed and replaced with new products, known as product churn. Various parts of the market will see this to a lesser or greater extent. In our traditional collection, when a product leaves the market a replacement product is found and any differences in quality between the original and replacement product are adjusted for. With scanner and web-scraped data, due to the sheer volume of data that we are working with, this manual replacement approach becomes unmanageable.

2. In this study we look at clothing products. Here we see a frequent change in products as seasons and fashion trends change several times a year. This makes being able to follow a single clothing product over time from one base period to the next unlikely. With clothing in particular, treating each product as unique means that we do not capture the implicit price increase from new products (of comparable quality) that enter the market replacing products on sale, so the price index can fall dramatically.

3. This problem has been studied by several national statistical institutes and it has been suggested that following groups of homogeneous products through time rather than individual products may alleviate the effect of churn (Chessa, A. 2019, Van Loon, K. 2019, Statistics Canada 2019, ONS 2017). Essentially the product definition is broadened to encapsulate a group or cluster of similar products and the average price of this group or cluster is taken, essentially treating the group as a single product for index construction purposes. This group can then be tracked over time, and products can fall in and out of the group but the overall churn within the elementary aggregate is reduced.

4. In this paper we look at how product grouping could be applied to web-scraped clothing data. This work first examines two different methods by which the groups can be initially formed. It then discusses the properties of these groups and how the grouping strategy could be objectively measured using various metrics. Many of the properties that a human performing the task of grouping clothing into similar products would use are either not available in the web-scraped data or are subjective. This makes grouping products into sensible and helpful groups very hard and the task of assessment uncertain.

5. It is a requirement of price indices that the products that make up the groups are homogenous. This is usually taken to mean that for the consumer all these products could be used or purchased interchangeably. While this might be more obvious for some areas of the market such as groceries, in clothing it is not. Why do people choose one pair of trousers over another? There is often an element of what the product is marketed for such as office wear, sportswear etc. that can indicate use, but unfortunately this information is not uniformly available across all retailers' websites. This limits the methods used here as we want the methods developed to apply universally across all retailers and not be limited to those with high quality attributes. This question of homogeneity is at the heart of our discussion on assessment metrics as we can only attempt to establish if our groups are homogenous enough once we have a clear understanding of homogeneity in this context.

Product grouping and classification

6. As we have discussed in our previous paper (ONS 2020) the web-scraped data will be classified into consumption segments. For clothing data this will be done using a machine learning based method. Product grouping will occur within each consumption segment. For example, the data may be classified to the consumption segment "dresses", but to reduce churn when following individual dresses, groups of homogeneous dresses are formed and the average price within each of these dress groups is used within the index number method chosen.

7. If it was possible for the machine learning classifier to assign data to the product groups, this could be considered. But the volume of human labelled data required to go down to such detail would be vast and not practical to achieve. Instead, we view product grouping as a building up of comparable products into larger homogeneous groupings, such as "long-sleeved v-necked cotton t-shirts" which would group together all such t-shirts regardless of colour. We have also chosen to enforce that groupings will be unique to each retailer as a consumer choosing from one website or store would not necessarily have access to the products of other retailers.

8. As the data will come from the classification provided by the machine learning based classification module of our processing pipeline there will likely be some level of misclassification in the resulting consumption segment. As the classification project continues to improve classification performance, this will become less of a problem.

9. During this research phase of grouping, we have selected to look only at consumption segments with high classification performance, with an f1 score above 0.9. Any products that have been misclassified could potentially all be grouped together to form a single homogeneous product. For instance, all the dressing gowns that make it into the dress

sample could form a group of their own. Whether this is the case in our data has not been tested and the impact on the index of this behaviour has not been investigated.

10. The initial focus for this work has been on women's dresses as this has a high level of classification confidence (f1=0.95), is a large category and has products for different use (formal, casual etc.). Further work will look to expand the methods described here to other consumption segments.

Group formation methods

11. We are currently researching two methods of forming groups. One of these is based on whether keywords are present for a range of product attributes (name, brand etc.) and the second is using unsupervised clustering to form groups with minimal human intervention. Both make use of text to form the groups, but the actual method is distinct.
12. To date we have not performed a direct comparison of the kind of groups formed by these two methods but will in future work. The attribute-based grouping is more directly controlled and has the potential to produce groups that directly follow the properties we desire as there is significant room for human intervention. While unsupervised clustering requires less human interaction, it has more potential to give unintuitive groupings. We discuss the details of these two methods in this section.

Attribute-based grouping

13. The attributes provided directly in the web scraped data are not detailed/standardised enough to create product groups. Therefore, this grouping method requires keywords to be generated. These are searched for in various columns containing the attributes of a product. An example of how this grouping would work is shown in Table 1.

Table 1: An example of attribute-based grouping

Product Name	Material	maxi	Polyester	Cotton	V-neck	Group
Blue v-neck mini dress	Polyester	False	True	False	True	Polyester_v-neck
Floral maxi dress	100% cotton	True	False	True	False	Maxi_cotton
Green floor length maxi dress	95% cotton 5% elastic	True	False	True	False	Maxi_cotton

14. These keywords can be determined in a variety of ways. The first would be to inspect the data and use domain knowledge to generate lists of important keywords for each attribute to group on. This would have the advantage that the keywords would be chosen such that the group properties would be those desired by the humans. For instance, if it was decided that the length of a clothing item was most important the keywords could reflect this. However, this method of identifying keywords is very manual and requires both time and expertise to look through the hundreds of thousands of clothing products that make up each consumption segment and identify the language that the retailers use and which words are most important. This level of manual interaction is impractical given the number of consumption segments being considered for clothing. It will significantly impact the timeliness of the construction of groups when new groupings are required.

15. Instead of relying on a completely manual approach we are also developing a method to automatically identify keywords to use in the grouping procedure. To do this, we use a mixture of natural language processing (NLP) techniques to find the words that occur most commonly for each attribute to be considered.

16. Before counting the most commonly occurring words we apply some data cleaning to remove punctuation and numbers from the strings. We also join commonly occurring phrases such as “short” and “sleeved” into “shortsleeved”. This is applied universally across all columns but in future work we will consider if the numerical values have some valuable information such as in the material column. An attempt is made to standardise the retailer and brand attribute across all data sources removing differences in capital letters and spacing so that they will be grouped together.

17. We also remove stopwords from the text. In NLP stopwords are any word that does not provide additional or helpful information. An example of a stopword would be “*the”* or *“and”* as these occur so frequently in language. We also add custom words to the stopword list that we want to remove from consideration when we are counting the most common occurring words. For example, dress and dresses when we are constructing dress keywords; these do not add information as we already know we are dealing with dresses and that these are likely to be very commonly occurring words. This list of stopwords is currently manually checked before use and adapted as necessary. This could prove to be a time-consuming step for expanding this work to many consumption segments.

18. We are also investigating if stemming will help standardise the words across products and columns when searching attributes. Stemming reduces a word to its root. For instance, if sleeves and sleeved are stemmed they both reduce to sleev and so would be counted as the same word. This works for some products but may not be universally applicable, for instance shorts (the clothing item) would stem to short but is not the same short as in the length of a product. How much of an issue this will be is still to be investigated as we are forming groups on top of classification and so in theory the shorts and short sleeved items will have already been separated.

19. The number of commonly occurring words to be chosen for each attribute column varies. For some columns, such as the retailer, where there are not many unique values the most common words might be all available words. But for columns such as product description where there are many unique words, we must choose the number of words and this can have a significant impact on the groups created. With many words included we get a lot of small groups and taking only a few of the most common words results in larger groups. There is a trade-off here of the homogeneity of a group and the persistence of that group over time. Further work is needed to find the ultimate number of keywords to include for each column and this could vary between consumption segments. Complex items such as dresses may require more groups, and so words, than simpler items such as jeans.

20. Once the keywords have been identified the columns are searched for these words and their presence identified as in Table 1 above. When the final groups are identified we concatenate the keywords found in that row into a group identifier. The keywords are sorted into alphabetical order before concatenation so that ‘green_blue_red’ and another ‘blue_green_red’ would still be classed as the same group.

21. Once groupings have been set up, we use the MARS (Match adjusted R squared) scoring method to measure the difference between different groupings in both the price homogeneity and the continuity of groups over time (Chessa, A. 2019). See section on group properties below for further discussion of assessment metrics. This work is at an early stage and we do not have firm results, but we are forming groups with similar scores between the human manufactured list and the automated keywords. This is encouraging but more work is needed to further refine the method and look to create an optimization procedure to further automate the finding of the keywords. This optimization procedure will use an iterative process and a metric to find the keyword combination that optimizes the metric. We have not yet explored different optimization methods and so do not have a preferred method.

Unsupervised clustering

22. An alternative way to form clusters into groups is to use unsupervised clustering. This method looks to find structure in the data and form it into groups. As we do not have a fixed criteria or classification structure for the groups the problem is more suited to an unsupervised approach. Although we do not have a fixed classification, we do need the groups to have specific properties related to product homogeneity. So, we need to bear in mind that the structure the algorithm finds may not always suit the problem we have of grouping similar products together. We must assess how well the formed clusters work for our use case.

23. Some of our earlier work looked at automated grouping using clustering through the CLIP method (ONS 2016, ONS 2017). The work here deviates from the method of the CLIP although there are some similarities. The differences between the CLIP and this work will be discussed at the end of this section.

24. There are many different unsupervised clustering algorithms each with different properties and suited to different data and problems. As we are not restricting the number of groups we are forming, we rule out any cluster formation methods, such as k means, that need this knowledge a priori. We also rule out density-based methods, such as DBSCAN and OPTICS[1], as these do not assign every point to a cluster. Points in regions of extremely low density will all go into a single unclustered group which is not the behaviour we want for clustering products.

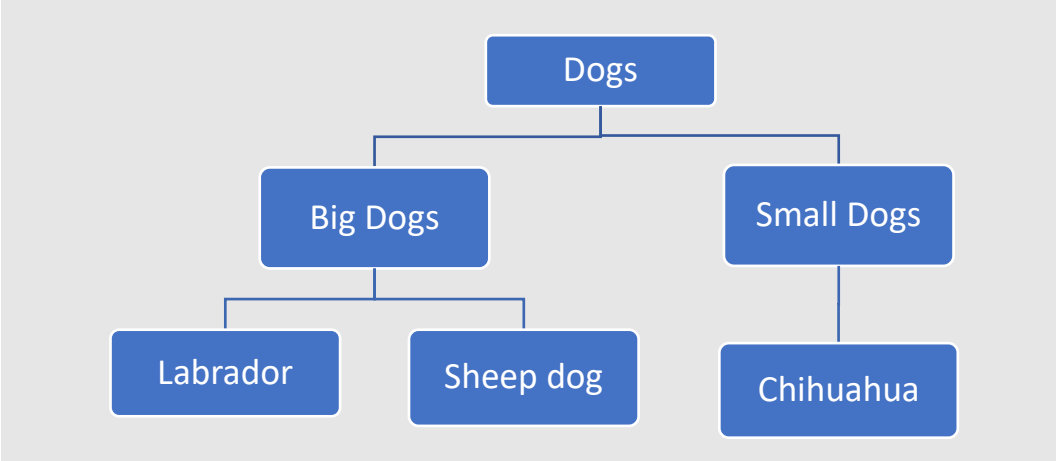
25. We have focused our work here on hierarchical clustering as all the data are assigned to a cluster (even if that cluster only has 1 member) and can be calculated in a reasonable compute time. In work to date we have made use of the SciPy ward linkage algorithm and Euclidean distances to form the clusters (Pauli Virtanen, et. al. 2020, Ward 1963). This method works from the bottom up joining together similar products into larger and larger groupings until all points are in a single cluster.

26. An example of the hierarchical structure is given in Figure 1. The clusters can then be taken at any of the given levels of the tree. In the example given if we cut the clusters at the lowest level, we get very homogenous clusters, akin to taking lots of keywords in the attribute

[1] Density-based methods are being considered for outlier detection precisely because they do not group all data into clusters, see APCP-T presentation on outlier detection for more details.

method. Whereas if we cut high, we get a very broad cluster and so loose the homogeneity. The level at which to take the cut is a parameter that needs to be chosen.

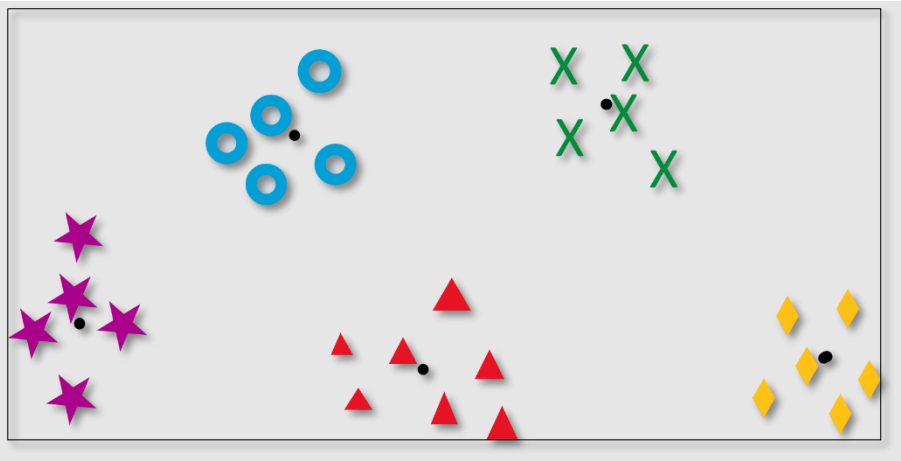
Figure 1: Example of hierarchical clustering



27. Using this form of clustering has a drawback that it is not easy to combine categorical and numeric datatypes. For this reason, our work to date has only used words converted to numerical vectors and not any of the extra categorical information that was used in the attribute-based approach. This is a significant omission currently for the clustering work and future work will look at methods to combine data types. This could be achieved using a different distance metric such as Gower distance (Gower 1971) which is able to combine both data types. However, it is significantly slower to compute than numerical data alone. We will also investigate if any other ways to combine numerical and categorical data exist.

28. Once clusters have been formed, we then need a method to assign new products to one of the clusters or decide it belongs to none and must be considered a new cluster. To do this we calculate the centroids of each cluster and then assign new products to clusters based on how close they are to the nearest centroid, see Figure 2 below where each coloured group is a cluster, and the black dots are the approximate centroids. Currently the products are assigned to a new cluster if it is closer than average intra-centroid distance. Products further than this away will be either dropped or added as new clusters, which option to take has not been investigated to date. The choice of the average intra-centroid distance is currently arbitrary, and it will be necessary to investigate the effect of tuning this parameter on the clusters produced.

Figure 2: Example of 5 coloured clusters and associated centroids (in black)



29. After new products have been added to clusters, we can either leave the centroids the same as that calculated with the original cluster members or we can update it taking account of any new members. Both approaches have their merits and their detractions and the effect of doing either will be investigated in future work.

30. As with attribute-base grouping there are parameters that control how the groups are formed such as the threshold for which level to take the hierarchical clusters at and the distance from a centroid to assign products to existing clusters. Optimization algorithms using a metric, like MARS, can help in setting these parameters as discussed for attribute-based grouping.

31. It will also be important to manually inspect the groups produced in clustering as these are more likely to not represent the properties that we want, such as all the cotton maxi dresses in a group, than with the attribute grouping. This is due to the algorithm finding unexpected clusters in the data that are valid but do not fit with the properties we would expect the groups to possess. For instance, all the red dresses could be grouped together, this would result in a perfectly valid group of red dresses but within that you may well have long, short, formal, and casual dresses. This is not desired for the price index due to the heterogeneity of the product group and so this grouping, while valid, is not correct. Therefore, some level of manual checking would be needed. The amount of such checking will depend on whether we can develop a metric to measure homogeneity and how far we can rely on it.

32. As previously mentioned, our earlier work on product grouping used the CLIP method. This method also made use of unsupervised clustering followed by assigning clusters to groups. As the datasets have grown the clustering method used in the CLIP, mean-shift, has been found to be slow to compute and so we have adopted the hierarchical clustering approach. We would like in future work to do a rigorous assessment of the difference of the groups using both mean-shift and hierarchical clustering.

33. The most notable deviation of this work from the earlier CLIP is in how new products are assigned to groups. The CLIP did this using a decision tree to predict the group that a product should belong to. Decision trees are a method of supervised machine learning that uses features to predict a quantity. They have the disadvantage that they very easily overfit to the training data and so may generalise very poorly to new data that may have new and

different features. The training of the decision tree would introduce many extra parameters to tune and manually check when constructing it. For these reasons we have chosen to switch to the centroid based approach that we have presented here.

Group properties

34. Once we have formed groups, we will need to assess their properties such as whether they are sufficiently homogeneous and whether they persist over time (effectively reducing churn). These group properties are not always easy to define and objectively measure. In the next sections we will first discuss the properties that we currently measure in the groups we form and some of the limitations of these measures. We will then pose some new questions and suggestions of other ways to measure group properties. These are untested and further research will be required to determine which best fits the needs of this project.

Properties currently measured

35. To date the work to assess the quality of the groups has focused on looking at simple group properties such as group sizes, maximum and minimum size, price variance in groups and whether groups persist overtime.

36. The size and range of sizes of groups give us some immediate indication as to whether the grouping has produced useful groups. We find quite often that most groups have a similar size but in both the clustering and attribute-based approach we find that a small number of groups contain very large numbers of products. These very large groups usually occur when there is either little or no information about products in the group. This is often because the information is not available in the data. But we also find situations where the same product description has been scraped for many items in the dataset. For example, a description like "everyday skater dresses and party-ready bodycon styles" will cover a very wide range of products and can result in large groups. These very large groups of over 100 products are likely not to be sufficiently homogenous for our use case.

37. We have, in the main, adopted the MARS metric developed at CBS as a metric of group quality (Chessa, A. 2019). This metric is made up of two halves that balance each other. One half is the homogeneity score (R) which is the proportion of explained variance in group prices compared to the total variance of product prices. The other half is the product continuity (μ), whether a product is present in both a base period and the current month under consideration. The two halves the then multiplied together to give:

$$M_t^K = R_t^K \, \mu_t^K,$$

where K is a given grouping of the data in month t. These two properties are opposites as higher homogeneity within a group will typically lead to lower continuity. A balance between the properties is therefore required. In the original formalisation of MARS, the products are weighted by expenditure weights. However, as we have been working with web-scraped data we do not have expenditure weighting information and instead have treated all products to be equally weighted. If in future a suitable expenditure proxy is identified (as discussed in panel paper Approximating Sales Quantities for Web Scraped UK Grocery Data) we could use these in our metric.

38. The product continuity half of the MARS measure relates directly to product churn, which we are seeking to decrease with the grouping of products. How to determine the base period with which to do the comparison is complex and there are pros and cons to different selection choices. One choice would be to choose only one month, such as January, to compare all subsequent months to. This however locks the comparison to the type of products that are present in January only. Whereas some seasonal products could be valid group members but not seen in January.

39. If we instead expand the base period to cover groups present in a full previous year of data we would hope to remove or reduce the effect of seasonality, summer shorts would be expected to be present every summer. However, that group, which is present in the base period, may not be in the data for large portions of the year. In this sense the churn due to seasonality would persist and would not be captured in the continuity score. This is because the score looks at if a group in the current month is present in base and not if a base month group is still present in the current month.

40. The question of if we wish to eliminate this element of seasonal churn is not settled. These seasonal groups would come and go but could be considered the same group as the base period just not present all year. We also suspect that this problem of seasonal group persistence in web-scraped data maybe reduced due to websites tending to still show summer products in winter and vice versa. However, we have not tested this theory and are currently working on plotting a full churn history for both attribute and clustering groups. This does lead naturally to the question of whether consumers are purchasing such out of season products and this issue with groups disappearing for long periods may be more present in scanner data where only sold products are reflected.

41. Another element of seasonality is the impact of sale price on the R squared homogeneity score. If price is being used as an indicator of quality, we propose that we will want to use the non-sale price as this is the normal price in the market. However, when end of season sales occur, the sale price might represent the change from one type of product to another, and this could represent a change in quality. We intend to use the non-sale price, where available, in assessing the group homogeneity but will look at what effect using the sale price has on R squared scores.

Measuring non numerical group properties

42. The MARS metric focuses on the price of items as the measure of homogeneity, but this is not the only, or perhaps not the most important element, of product homogeneity. There are many other aspects of a product that could be considered, such as the material, brand, shape and target use of the product. These are much harder to measure as they are not easily numerically quantified (which many metrics require) and can be very subjective.

43. We are at the start of exploring other measures for homogeneity and present here some ideas which are not yet tested and would require further development and testing before use. How they could and should be aggregated across different groups and features to give a single metric has not yet been considered.

44. One way that we could consider assessing the groups would be to construct a similarity matrix as we have done in the clustering and use some standard clustering assessment

metric coefficients or indices either singularly or in conjunction. Constructing the similarity matrix is not a simple task and there are several ways to be considered. Currently we are looking at:

a. Using each feature of interest in turn. We can construct numerical vectorisations of the text in each column and obtain a metric for each feature. This would be simple for creation of the vectors but would treat each feature as an independent indicator of homogeneity which may not be the case in reality.
b. Construct a metric using several columns and features into a single similarity matrix. This would potentially be possible using the Gower distance (Gower 1971), or other similar distance metric, to combine categorical and numerical (vectorised) columns. However, we would need to do feature engineering to obtain categorical columns. One example of a feature we might create is the main material of a product. This could be a categorical variable but is not currently present exclusively in the data. Producing such features maybe something that we wish to do for the group formation methods anyway. Once the Gower distance matrix has been constructed it can be used to calculate standard clustering metrics.

45. Although we have stated that there are standard clustering metrics available in python packages such as scikit-learn (Pedregosa, 2011) we have not discussed some of the limitations. In general, these metrics assume some form of spherical symmetry and well-defined group boundaries in the feature space in which they are being considered. These assumptions may not hold for groups formed with either construction method. Even with this drawback we think these methods have potential as we do not feel price alone is sufficient for assessing the clusters. The three metrics under consideration are the Silhouette Coefficient (Rousseeuw 1987), Clainski-Harabasz Index (Calinski & Harabasz 1974), and Davies-Bouldin Index (Davies & Bouldin 1979), these are presented in more detail in Annex A.

46. We are also considering a simple counting of the feature entries that are the same compared to the total number of group members. An example of how such a system could work is given in table 2 for colour, note that this example is illustrative and not representative of the types of groups that we might expect.

Table 2: Illustrative example of categorical groups feature values

Group 1	Group 2	Group 3
Green	Red	Black
Blue	Red	Brown
Green	Red	Grey
Green	Red	Black
3/4	4/4	2/4

47. We can then sum the fractions of similarity for each group and divide by the number we would expect if all groups only had members that were the same. In this example we would get 2.25/3= 0.75. The higher the value the more similar the categorical value are. As with the other metrics we still need to consider how this metric could be combined with others to give one score that could be optimized.

48. While all these metrics could help with automatically measuring homogeneity, they will all have pros and cons. We need to do further research in how they behave with our web-scraped data. We will also need to manually review the groups to see if those groupings that the metrics find to be homogenous are also considered so by human judgment.

Conclusions and next steps

49. In this work we have developed two methodologies for forming products into homogeneous groups. These methods tackle the problem very differently one using keywords and the other using unsupervised clustering. There are pros and cons associated with each in terms of the manual intervention, control over what is grouped and the scalability to large data. This last point we have not yet explored or discussed here.

50. Forming the groups is only half of the challenge, we also require that they have the properties needed for the formation of price indices. These properties are not easy to define objectively and the concept of "homogenous of use" may depend on who is asked the question.

51. The numerical concept of price homogeneity can be measured relatively easily for a given grouping of products using the MARS formalism. However other aspects of homogeneity are not so easily captured into an indication of group quality. We have proposed some, yet untested, ways to tackle assessing other aspects of group quality as well as ways to weight these metrics together. All of these will require some human oversight to see if clusters that the metrics indicate are homogenous are intuitive to humans.

52. There are still some key questions that are unresolved and will impact the future direction of the work. First and most important of these is around what aspects of homogeneity are most important in the context of price indices. Do the more intangible attributes hold more importance for the groups? This will feed into the development of alternative metrics for assessing groupings. The second is around how the concept of seasonality interacts with the groups and product churn and the sale price. The last major question is whether it is possible to produce metrics that can measure non-price aspects of homogeneity and so automate the grouping process to a degree.

53. The next steps proposed for this work are as follows:

    a. Further investigation of the alternative metrics of group assessment to work in tandem with human assessment. The level of human review required will need to be considered as this could be a time-consuming occupation.
    b. Make use of optimization techniques such as hill climbing to use metrics to optimize the grouping performance.
    c. We have yet to fully assess how much churn has been reduced by these groups and, as this is the problem that grouping is trying to solve, we need to fully examine this. This will allow us to explore the elements of churn that we can address with grouping.
    d. As the groupings will significantly affect the resulting price index this must be examined. What is the impact of different grouping strategies, reference periods and introduction of new groups through the year? This will form a large part of future work.

Hazel Martindale and Matt Eddolls
Methodology Division and Prices Division, ONS
April 2021

List of Annexes

Annex A – Clustering evaluation metrics

1. There are three clustering evaluation metrics that we are considering in this work as part of developing a measure of group homogeneity based on product attributes. The first of these is the Silhouette coefficient (Rousseeuw 1987). This measure looks at how similar any individual member of a cluster is to the whole cluster compared with other clusters. This is done using distance measures in whatever feature space the clustering has been performed. The inter-cluster distance is the mean distance of a point to all other points in the cluster and the intra-cluster distance is the mean distance between the point with all members of the next nearest cluster. The silhouette value for an individual point is given by:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

where $a(i)$ is the inter-cluster distance for point i and $b(i)$ is the intra-cluster. The silhouette coefficient for the total samples to be clustered is the mean $s(i)$ value for all samples. The silhouette coefficient is bounded to lie between -1 and 1 with a higher the value the more dense and well separated the clusters are. Values around 0 show the clusters are overlapping. This metric is time consuming to compute in large datasets as the ratio must be calculated for every point.

2. The second measure we consider is the Calinski-Harabasz index or variance ratio criterion (Calinski & Harabasz 1974). This measure is the ratio of the intra-cluster dispersion and the inter-cluster dispersion. Here dispersion is defined as the sum of distances squared.

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_e - k}{k - 1},$$

where $tr(B_k)$ is the between group dispersion and $tr(W_k)$ is within cluster dispersion, n_e the size of the data and k the number of clusters. The higher the value of this metric the better separated and more dese are the clusters. This metric is fast to compute.

3. The final metric under consideration is the Davies-Bouldin Index (Davies & Bouldin 1979). This measures the ratio between the clusters scatter and the clusters separation. The cluster scatter is measured as the average distance between each point in a cluster and its centroid and the separation is the distance between cluster centroids. This is defined for cluster pair i and j is given as:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}},$$

where s_i is the scatter of cluster i, s_j the scatter of cluster j and d_{ij} is the distance between the two cluster centroids. The final total metric is then given by:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

where k is the number of clusters. This index is opposite to the others as values closer to 0 show a better partitioning of the data. This metric is simpler to compute than the silhouette metric but still complex.

4. All these metrics measure different aspects of cluster properties and may play a different role when assessing clusters or groups for homogeneity. Which is best suited for our use case is not clear from the metric properties alone.

Annex B – References

Caliński, T. & Harabasz J. (1974) A Dendrite Method for Cluster Analysis, Communications in Statistics-theory and Methods 3: 1-27

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

Davies, David L., Bouldin, Donald W. (1979) A cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 2224-227

Gower (1971) A general coefficient of similarity and some of its properties. Biometrics 27 857–874.

ONS (2016) Research indices using web-scraped price data: clustering large datasets into price indices (CLIP)

ONS (2017) Research indices using web-scraped price data: clothing data

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

ONS (2021) Approximating Sales Quantities for Web Scraped UK Grocery Data, APCP-T panel paper

Pauli Virtanen, et al. and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.

Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

Peter J. Rousseeuw (1987) Silhouettes: A Graphical Aid to the interpretation and Validation of Cluster Analysis, Computational and Applied Mathematics 20: 53-65

Statistics Canada (2019): Research on Using Web Scraped Data for Clothing Price Index

Van Loon, K. (2019) Redefining what products are in the context of scanner data and web scraping, experiences from Belgium. 16th Ottawa group Meeting

Ward, J. H., Jr. (1963) Hierarchical grouping to optimize an objective Function, Journal of the Americal Statistical Association 58 236-244

Wes McKinney. (2010) Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56