Using clustering of rankings to explain brand preferences with personality and socio-demographic variables

Daniel Müllensiefen, Christian Hennig & Hedie Howells

Published online: 15 Jun 2017.

Submit your article to this journal ⬈

Article views: 39

View related articles ⬈

View Crossmark data ⬈

Taylor & Francis
Taylor & Francis Group

Using clustering of rankings to explain brand preferences with personality and socio-demographic variables

Daniel Müllensiefen[a], Christian Hennig [b] and Hedie Howells[a]

[a]Department of Psychology, Goldsmiths, University of London, London, UK; [b]Department of Statistical Science, University College London, London, UK

ABSTRACT

The primary aim of market segmentation is to identify relevant groups of consumers that can be addressed efficiently by marketing or advertising campaigns. This paper addresses the issue whether consumer groups can be identified from background variables that are not brand-related, and how much personality vs. socio-demographic variables contribute to the identification of consumer clusters. This is done by clustering aggregated preferences for 25 brands across 5 different product categories, and by relating socio-demographic and personality variables to the clusters using logistic regression and random forests over a range of different numbers of clusters. Results indicate that some personality variables contribute significantly to the identification of consumer groups in one sample. However, these results were not replicated on a second sample that was more heterogeneous in terms of socio-demographic characteristics and not representative of the brands target audience.

1. Introduction

The primary aim of market segmentation is to identify relevant groups of consumers that can be addressed efficiently by marketing or advertising campaigns. Here we address the issue whether consumer groups can be identified from background variables that are not brand-related, and particularly how much personality vs. socio-demographic variables contribute to the identification of consumer clusters. See Section 2 for an overview of the relevant literature.

We suggest a statistical approach based on clustering to generate consumer profiles and evaluate the association of socio-demographic and psychological variables with clusters of consumers. However, we do not make any assumptions about a 'true' or 'natural' number of consumer clusters, but rather to try to identify 'constructive', pragmatic clusters [34]. In fact, we start from the presumption that the underlying multivariate distribution of brand preferences is continuous and use clustering as a tool to segment this continuous distribution into meaningful groups of individuals who differ in their brand preferences and therefore can be targeted differently with marketing communication. Here we take into consideration that in most practical applications the number of different target groups or

---

market segments is determined by practical constraints dictated largely by logistic complexities and production costs that increase with the number of clusters, that is, target groups that need to be addressed separately. Thus, we follow a profiling approach where we measure the impact of socio-demographic and psychological variables as a function of the number of clusters. For the marketing practitioner this translates to the question: 'For the number of different market segments I can target, how important are psychological vs. demographic variables to describe the different clusters of consumers, and which specific variables are most important?'

We test the approach in studies using two different samples that are similar in size but differ in their socio-demographic characteristics. The first sample is drawn from a population of young urban consumers living in the London area that can be considered demographic target audience for the product categories and brands that we selected in this study based on data from the UK TGI database. The second sample is has a wider spread on all socio-demographic variables and is drawn from a panel representing Scottish adult consumers of all ages and social strata; that is, it is much more diverse than the first sample and cannot be considered to represent as a whole the demographic target audience. Thus, the two studies represent different scenarios: The first study resembles a market research project where prior knowledge about a consumer target group for specific brands and products is used to narrow down the selection of a panel. The second study is closer to academic studies where broader (but at the same time often larger) samples are drawn from existing panels or databases representing the general adult population of consumers.

In both studies we assess consumer preferences for five different brands drawn from each of five different product categories that serve as the primary outcome data. Participants are clustered according to their ranked brand preferences. Cluster solutions from 2 up to 10 consumer clusters are computed. In a second step, each cluster solution is then assessed in terms of how well socio-demographic and personality variables contribute to predicting cluster membership of individual consumers.

Our primary focus is on the contribution of the personality variables when the number of consumer clusters changes. We expect that the personality traits gain importance for describing the differences between clusters of consumers as the number of clusters increases, that is, for finer grained market segmentations. We also expect to observe a stronger contribution of personality variables in Study 1 (target group sample) than in Study 2 (general consumer sample). By way of the same analysis we will also be able to test the assertion that neither demographic nor personality variables are significantly associated with consumer clusters (e.g. [67]). In order to address these issues, we use logistic regression and random forests for explaining the clusters from the explanatory variables.

Section 2 summarises the relevant literature on market segmentation. Section 3 gives some details about the two empirical studies. Section 4 presents that statistical methodology that was applied here. Section 5 summarises the results. Section 6 concludes the paper with a discussion.

2. Relevant literature on market segmentation and personality variables

Market segmentation is based on the assumption that an individual product or brand will not appeal equally to the entire population and that marketing budgets can be used

most effectively when marketing communications are tailored to the characteristics of specific consumer groups who are most likely to purchase a product or engage with a brand [46,60, p. 363] and thus saving considerable resources in comparison to a scattered, shotgun approach [22, p. 512]. In addition, market segmentation can also aim to segment consumers by media consumption or by their value to the brand (e.g. loyal vs. occasional consumers of the brand). But overall the primary aim of market segmentation is the identification of target consumer groups [6, p. 499] and to inform marketers how, when, and where to advertise, as well as indicating who comprises the target market.

While developing new ideas for brand and product communications, advertising creatives often employ simplified notions of prototypical consumers that are the intended audience(s) for specific communications. Similarly, marketing practitioners make use of insights into the behaviour of target consumer groups to maximise the effectiveness of marketing campaigns, for example for media planning and forecasting the effects of campaigns on sales and revenue.

For most practical applications in advertising and consumer marketing, market segmentations are performed on the basis of basic demographic information, most commonly and with little variation over the past decades [52,66], mainly age, gender, household income, and ethnicity. In addition, market segmentation is often based on standard classification schemes for social class (e.g. the 6 occupation-based groups of the NRS scheme as maintained by the Market research Society in the UK) or on neighbourhood area (e.g. the ACORN or Mosaic schemes in the UK). Larger market segmentation schemes, such as the PRIZM NE scheme in the US, combine demographic as well as geographic and neighbourhood with lifestyle data to yield very fine-grained consumer classifications. Information from these standard industry classification schemes is readily available together with product preferences through large survey panels such as the ones provided by MINTEL, the Target Group Index (TGI) or Market Assessment.

In contrast, the term 'psychographics' [15,25] describes the collection of data on variables that reflect consumer personality, attitudes, personal values [40,54], life style [52] and other psychological constructs in order to identify and describe subpopulations of consumers and ultimately to inform marketing processes (see a recent summary account on the use of psychographics in market segmentation in [66]). Early empirical studies testing the psychographic approach only found limited evidence for the predictive value of personality variables impacting on consumer behaviour [9,41,45,68]. Since then psychographic approaches to market segmentation have been repeatedly criticised for being ineffective and only explaining a small amount of variance in consumer choices [59,71]. However, over the last two decades personality psychology has seen great advancements both in terms of psychometric developments, both through increasing the reliability and validity of the measurement of personality traits as well as in terms of the applications of personality measurement ranging from behavioural genetics [61] to core applications in human resources selection and work place contexts [65] and important life outcomes in general [64]. In addition, the Big Five model of personality [11,51] – comprising Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience – has been established as a quasi-standard in the conceptualisation and empirical measurement of personality [39], which enables broad comparisons of research findings across different studies and fields. As a consequence of these recent developments in personality psychology, several academics (e.g. [5,53]) have argued that - with a few exceptions (e.g, [10,56])

– personality variables have received their due credit yet in marketing research and practice and that the consideration of personality traits can play a crucial role in understanding consumer behaviour [21].

In academic as well as applied contexts market segmentation studies are typically performed either at the level of individual brands or the level of the product category. In these scenarios the research goal is to identify whether or not demographic variables or psychological traits co-vary with the choice of specific brands [33]. However, there has been mixed support for the usefulness of demographic and psychological variables in market segmentation [43,44,58,59], leading Sharp [67, p. 71] to suggest that predicting brand preferences from personal consumer variables is by and large impossible: 'The big discovery (of research on market segmentation) is that customer bases of brands in a category are very similar (. . .), there isn't a vanilla ice-cream buyer and a different type of person who buys strawberry - there are just ice-cream buyers who sometimes buy vanilla and very occasionally buy strawberry'.

In a recent comparative study on data from a large sample, Sandy *et al.* [66] tested the relative contribution of demographic and personality factors in regression models for a large number of individual outcome variables. Outcome variables were each derived from specific statements on individual aspects of media consumption, political and societal views and product choices. In line with Sharps assertion, Sandy *et al.* found only very small effects, even when demographic and personality variables were combined in the same model most of their models explained less than 10% of the variance in the outcome variable. This is in line with the low effect sizes reported by Novak and MacEvoy [59] where the median R^2 value for regression models using demographic variables was 0.04 and median R^2 values for models using List of Values and Values and Life Style psychographic variables were 0.011 and 0.026 respectively. In contrast, Sandy *et al.* [66] found that personality variables from a Big Five inventory contributed about an equal amount to the regression models of their outcome variables of direct and indirect consumer behaviour. In their discussion, Sandy *et al.* [66] attribute the low effect sizes for both types of variables to the fact that their outcome variables only described a narrow aspect of direct and indirect consumer behaviour and they suggest to aggregate indicators of consumer behaviour in order to find larger effects. In addition, they suggest considering brand choices rather than choices for classes of products as primary outcome measures, assuming that personality (as well as demographic) variables may have a larger impact on brands choice.

This study takes up these suggestions from Sandy *et al.* [66] by starting from brand preference rankings within product categories and by using clustering methods to obtain aggregate consumer indicators. Then, using these methodological refinements, we assess (similar to [66]) the absolute and relative importance of personality and demographic variables for segmenting consumers into different groups. In addition, we assess the question of how the relative contribution of personality and demographic variables might change with the number of different consumer groups in a clustering solution. Hence, this study does not assume that the contribution of personality variables is constant across the number of market segments that a created during the segmentation process but might depend on the complexity of the segmentation solution.

It is worth noting that the term psychographics refers to a larger set of measures and scales that aim to capture consumer characteristics and behaviour. Commonly, activities

and interests, commercially relevant attitudes and behaviours as well as values and beliefs along with personality questionnaires are part of the psychographic toolbox in consumer research [70]. However, in terms of practical research applications, psychographic information is usually much more difficult to obtain than demographic information because individuals have to be contacted individually and incentivised to self-report using non-standardised questionnaires. By contrast, demographic information at the household level can be obtained from databases such as the one maintained by PRIZM NE without the need to contact contact consumers directly. Thus, for most practical market segmentation applications, demographic information is already available for most Western countries (at least at the household-level), while psychographic information is generally not available. However, personality constitutes an exception to this rule. The greater availability of personality data compared to other psychograhic measures is helped by the fact that allmost all recent studies use the standard big five model of personality and hence regardless of the actual self-report personality inventory used, data and results can be compared across studies. In addition, recent studies have shown that it is not necessary to obtain personality data from consumers directly through self-report inventories but that personality information can be gathered indirectly from online data such as facebook likes [72], Twitter profiles [62], musical preferences [31] or spending behaviour [50]. Thus, unlike other psychographics measures, personality has become a layer of information that can be obtained from various sources, especially online information, and that has proven to be useful for predicting a broad variety of outcome measures, such as substance use, political attitudes, or purchase satisfaction [50,72]. Hence, this study focuses on the big five personality profile as the only psychograpohic measure to predict aggregate consumer preferences in addition to common demographic information.

Much of the traditional academic research on market segmentation has focused on predicting individual brand choices. However, it has been shown that consumer brand preferences can vary by context and that consumers can have multiple preferences within the same category [16], which can be one cause of the instability of preference models based on single brand choices. In line with the suggestion by Sandy and collaborators [66], a more stable approach may result from aggregating brand preferences into broader consumer profiles across product categories, reflecting associations between multiple brands as opposed to a single preference. Aggregating consumer choices into consumer profiles or clusters often feeds into successful product recommendations [4,63] and prediction tasks. Thus, at least in online retailing, aggregating consumer choices has become a standard approach and been implemented and refined successfully by major online retailers and media services such as Amazon and Netflix. Similar approaches (e.g. [19,55]) have been used very successfully with different types of survey data, for example, from tourism research [17] where the goal is to 'identify groups of tourists who share common characteristics' and target them with a 'tailored marketing mix' [18].

3. The studies

3.1. Methods

For Study 1, data was collected via a survey (paper as well as online) that was distributed to a sample of young adults in London, UK, in 2011–2012.

The aim of Study 2 was to assess to what degree the results from Study 1 hold true with data from a different sample which does not represent the brands demographic target audience. The data for this was collected via an online survey that was distributed to a sample from the ScotPulse panel of adults living in Scotland, in 2012. The ScotPulse panel has around 12,000 members from all age and demographic groups and can be considered representative of the Scottish population. The panel is run commercially by STV.TV Ltd, Glasgow.

3.2. Participants

Study 1 had 343 participants (57.7% women) with a mean age of 23.18 years (SD = 3.48), which were recruited among young adults all currently residing within in the Greater London Area. Two thirds of the participants were from the target age group of up to 24 years while about $\frac{1}{3}$ were between 25 and 34 years old. Participants were mainly living as singles (50%), renting an accommodation (67%), had achieved A-levels (44%) or an undergraduate degree (41%) as their highest level of education, and indicated a diverse range of income brackets for the main earner in their household (almost equal proportions for income brackets from < £ 15,000 to £ 50,000– £ 75,000 per year).

Study 2 had 355 participants (54.1% women), who were recruited from among members of the ScotPulse panel living in urban as well as rural areas of Scotland. The age distribution was markedly different from the sample in Study 1 and included a larger proportion of older participants outside the age range for which we selected the 25 brands. 30.7% of the participants were less than 25 years and 64.8% were between 25 and 34 years old. 42.2% lived in a 2-person household and for 29.0% the yearly household income was between £ 15,000 and £ 30,000. For 27.3% the households main earner was working in a professional or technical profession that requires at least university degree-level qualification. For 18.9% the main earner worked in a non-managerial but non-manual job (e.g. office worker). 45.4% of the participants had achieved an undergraduate degree as their highest level of education. As in Study 1, there were no missing values in the brand choice variables. 8 observations had missing values for one or more socio-demographic variables (one observation in Study 1).

3.3. Materials

Participants took an online survey that was distributed via chain-referral system in February to June 2012. The survey questionnaire was entirely anonymous and participants were not remunerated for their participation. The survey comprised the following measurement instruments.

Personality With a view on possible implementations in practical contexts we chose the Ten Item Personality Inventory (TIPI, [30]), which comprises only 10 brief statements that participants respond to on a 7-point agreement scale and has been shown to have high validity and reliability scores compared to other short-form Big Five personality inventories [24]. Average scores for the five dimensions Extraversion, Agreeableness, Conscientiousness, emotional Stability, and Openness to Experience were derived for each participant.

Socio-demographic variables In line with marketing practice and previous comparative literature (e.g. [58,59,66]) we focused on the most widely used socio-demographic variables that are of relevance with the sample, namely age, gender, education, household main earners income, dwelling type and relationship status.

Brand choice Participants provided preference rankings for each of five brands from each of five different categories. The product categories were chosen to represent largely gender-neutral commodities with a high relevance to the participant target group (young adults living in an urban area). Brands in each category were chosen to be the five most frequently used brands in the age group of the 19–29 year olds according to data from the TGI GB survey for 2011. Categories and chosen brands were:

   Smart phone: iPhone, Nokia, Blackberry, Samsung, Sony Ericsson.

   Chocolate snack: Dairy Milk, Galaxy Milk, Kit Kat, Maltesers, Creme egg.

   Clothing retailer: Topshop/Topman, River Island, H&M, Primark, George at ASDA.

   Coffee shop: Starbucks, Caffe Nero, Costa Coffee, Department stores own coffee shop, Local coffee shop.

   TV show: X-Factor, Live at the Apollo, The Simpsons, Shameless, Harry Hills TV Burp.

In Study 2, participants took an online survey that was distributed randomly to the members of the ScotPulse panel. The survey questionnaire was entirely anonymous and participants were remunerated with a typical panel credit or their participation. Similar to the questionnaire instruments used with the London sample in Study 1, the survey in Study 2 used the TIPI as a personality inventory and questions on the socio-demographic variables age, gender, household main earners income, the work position of the households main earner (similar to the ESeC classification of socio-economic status), the number of children living in the shared home and highest level of education achieved. Participants ranked the same 25 brands that were used in Study 1 according to their preferences.

4. Statistical methodology

4.1. Cluster analysis and distances: methodology

Market segmentation is done by clustering the brand choices. The resulting clusters are then used as response variable to be explained by the socio-demographic and personality variables. We are particularly interested in whether there is evidence that the personality variables have an impact on the brand choice clusters. The present section explains what was done. Section 4.2 presents a thorough discussion of the choices involved in this approach.

The brand choices for a participant $i = 1, \ldots, n$ can be represented as a 25-dimensional vector $\mathbf{r}_i = (r_{ijk})_{j=1,\ldots,5,k=1,\ldots,5}$, where r_{ijk} is the rank given to brand k in category j by participant i, and for fixed i,j: $\{r_{ijk} : k = 1, \ldots, 5\} = \{1, 2, 3, 4, 5\}$, where 1 denotes the first preference. We use 'partitioning around medoids' (PAM, [42]), which is a distance-based method for cluster analysis; for a given number of clusters G, PAM looks for the G centroids in the data set such that the sum of the distances of all objects to their closest centroids is minimised. Rather than postulating that there is a single true number of clusters and trying to estimate it, we use all PAM solutions for $G = 2, \ldots, 10$ in order to monitor how

the contribution of the socio-demographic and personality variables looks like over all the values for G (we tentatively looked at $G > 10$ but this did not bring any additional insight).

A key issue for the definition of the distance between brand choices is that usually regarding brand preferences the higher ranks can be assumed to be more important than the medium and lower ranks; many consumers have favourite brands but would not differentiate much between several brands for which their preference is not so high, see also [8]. For this reason we introduce a score function s for the ranks, with $s(1) = 1, s(2) = 5$, $s(3) = 7, s(4) = 8, s(5) = 9$. We then define the distance between two rankings as the 'Footrule'– (L_1-)distance between scores:

$$d(\mathbf{r}_{i_1}, \mathbf{r}_{i_2}) = \sum_{j=1}^{5} \sum_{k=1}^{5} |s(r_{i_1jk}) - s(r_{i_2jk})|. \tag{1}$$

We also ran analyses separately for each category, for which the definition of d only requires a single sum over the 5 brands.

In order to assess the degree of 'clustering' in the data (on which our methodology does not rely, although it may be interesting in its own right), we show 2-dimensional plots produced by classical multidimensional scaling (MDS, [49]) as implemented in the R-function 'cmdscale', and we compute two cluster validation indexes, namely the Average Silhouette Width (ASW) and the normalised version of Hubert's Γ based on the Pearson correlation ($P\Gamma$, [32,42]). The former assesses the quality of a clustering based on the contrast between within-cluster homogeneity and between-cluster separation with high values if there are clear gaps between the clusters. The latter formalises to what extent the distance is represented by the 'clustering-induced distance', which is 0 between two observations in the same cluster and 1 between two objects in different clusters. This is particularly relevant here because the clusterings are used, even in absence of a true clear clustering of the data, to summarise the information in the brand preferences encoded by the distances.

4.2. Cluster analysis and distances: discussion

The main aim of cluster analysis here is pragmatic. The information in the rankings is very complex, and the standard methods used in Section 4.3 could not have been applied for explaining the full information in the rankings directly. Given the moderate sample sizes, some information reduction is required in order to apply any method for exploring how the explanatory variables affect the rankings. Forming clusters is our approach to summarise the ranking information. Occasionally, cluster analysis is used for reducing more complex information for use in explanatory methods (e.g. [3]), but the specific requirements of such clusterings are rarely discussed. Hennig [34] calls such a pragmatic clustering task 'constructive', meaning that the aim is not to find true underlying 'real' clusters, but rather to organise the data in a suitable way for the requirements of the specific application.

In Section 4.3, the clusters are treated as a response for multinomial logit regression (MLR) and random forests, and we are interested in whether explanatory variables have an impact on the brand rankings that are encoded through the clusters. Therefore it is important here that the clusters are homogeneous, that is, that the within-cluster dissimilarity of rankings is low (otherwise it would be problematic to interpret the impact of the explanatory variables in terms of the rankings). We also want to represent the similarity

structure well by the clustering, which is measured by $P\Gamma$. Another desirable feature of the clusters is that their sizes should be fairly uniform, so that the observations can be used in a balanced manner for predicting all clusters in Section 4.3.

Separation of the clusters is less important (actually the MDS plots and also not shown higher dimensional MDS information make us believe that there are no strongly separated clusters in the data), and it is not desirable here to rely on the model assumptions required for postulating and fitting a probability model for the ranks. This also means that we do not use any definition of a 'true' number of clusters G, for which reason we do not attempt to estimate G. Instead, we track results over various values for G.

Because there is also no particular reason why a hierarchy of clusters should be imposed, we decided to use the non-hierarchical PAM method that focuses on homogeneity of clusters rather than separation. Because of the triangle inequality, no within-cluster distance can be bigger in PAM than the sum of the two biggest distances of objects to the cluster centroid. For comparison, we ran Average Linkage and Complete Linkage hierarchical clustering, cutting the dendrograms at $G = 2, \ldots, 10$ clusters. This led to clearly worse values in all cases (both studies, all values of G) regarding the average within cluster distance and the uniformity of cluster sizes, and in almost all cases regarding ASW and $P\Gamma$.

The Footrule distance defined in Equation (1) is an L_1-distance on the vector of scores, and as such it fulfils the standard properties of a metric (identity, symmetry and the triangle inequality). We follow the philosophy of Hennig and Liao [35], Hennig [34] here, according to which distance design is about formalisation of what counts as 'similar' or 'distant' in the given application, which cannot be estimated from the data alone but will always involve user input. The main feature that we wanted to achieve with the choice of the distance and the specific scores is to emphasise the first rank compared with the differences between the others, as we believe that consumers often have a 'favourite brand', which often will have an impact on their buying behaviour, whereas we rather expect, at least for the lower ranks, that customers will normally construct a ranking only when asked, and that this ranking will have far weaker implications for their behaviour. We quantified this in a subjective manner, but there is no objective alternative, because there is no information in the given data about this (one could imagine data from certain questionnaire questions or buying behaviour that gives some information about how much more important the first or higher ranks are to a typical consumer, but such data was not available to us). Note that the Footrule distance with the scores s defined above is not rank-invariant (invariant against applying the same permutation to both rankings, [48]), because this is in direct conflict to emphasising the top ranks. It is label-invariant, that is, invariant against permutation of the labels (listed brands) as rank differences are just summed up over brands.

The 25 score distances are aggregated in the Footrule distance in an L_1-manner, which seems appropriate to us because it gives the rank difference for every brand the same weight, whereas the Euclidean distance uses squares, which upweights brands with larger rank differences.

There are a number of alternatives in the literature. Brentari *et al.* [8] base a distance on a generalisation of Spearman's rank correlation and the directly related Spearman's distance, which also gives higher weight to the better ranks, but in less pronounced and less direct manner (according to their definition, rank 3 is still more distant from rank 5 than rank 1 from rank 2); see also [14] for a comparison of various versions of Spearman's correlation. Ignoring the issue of giving higher weights to better ranks, there are some distances

for ranking data other than Spearman's. Marden [48] lists for example Kendall's τ- and the Footrule distance and compares the distances in certain ways, most of which are not very relevant to clustering. We are not aware of strong arguments regarding the choice between Kendall's τ-, Footrule and Spearman's ρ-distance relevant in our setup and stick to the Footrule distance because of the straightforward implementation of our non-standard scoring and because L_1-aggregation seems to be most appropriate over different brands.

There is also some work on model-based clustering of rank data, see, for example, [12,13,27–29,38,57]. Murphy and Martin [57] is distance-based and may yield good within-cluster homogeneity. As already stated, we are not convinced that assuming that there is an underlying 'true' partition to be estimated is helpful in the given situation. Most of these methods do not readily generalise to non-standard scorings and multivariate rankings, although the Rankcluster package in R [38] analyses multivariate rankings. We applied this to our data but the resulting clusterings had extremely imbalanced cluster sizes, which in our situation is not desirable. Note that all comparisons between candidate methods referred to in this section were run ignoring the explanatory variables, in order to avoid biasing the results in Section 4.3.

We acknowledge that the model-based approach comes with its own advantages, which include an account of the choice process in some models for rank data [23,47] and a generic possibility to quantify uncertainty (although this is based on model assumptions that we are not willing to make); so there is certainly some potential for future work in this direction for similar applications; however, we think that the approach taken by us is superior for producing homogeneous clusters (regarding within-cluster distances and cluster sizes), which is a key issue for our pragmatic use of clustering for information summary.

4.3. Explaining the clusters from socio-demographic and personality variables

Once the clustering has been obtained, we use two different approaches to explain the clusters (which can be seen as simplifying proxies for the full ranking information) from the socio-demographic and personality variables.

The first one is the MLR model [36]. Let $y_i \in \{1, \ldots, G\}$ be the cluster to which the brand choices of participant i belong, and Y_i the corresponding random variable. Let $\mathbf{x}_i = (x_{ij})$, $j = 1, \ldots, p$ be the vector of the values of explanatory variables for participant i. The MLR assumes that $Y_1, \ldots, Y_i$ are independent and

$$\log\left[\frac{P(Y_i = g)}{P(Y_i = 1)}\right] = \beta_{g0} + \beta_{g1}x_{i1} + \cdots + \beta_{gp}x_{ip}$$

for $g = 2, \ldots, p$, with cluster 1 as reference category. We estimate this model using the function 'multinom' in the R-package 'nnet' [69]. Assume that the variables $j = 1, \ldots, q$ are the socio-demographic variables and variables $j = q + 1, \ldots, p$ are the personality variables. We are then particularly interested in the deviance test of the null hypothesis $\beta_{gj} = 0$ for all $g = 2, \ldots, G$, $j = q + 1, \ldots, p$, meaning that none of the personality variable contributes significantly to explaining the brand choice clusters in the presence of the socio-demographic variables. The individual tests for all variables j separately, testing $\beta_{gj} = 0$ for all $g = 2, \ldots, G$, are also considered, that is, the null hypotheses that there is no contribution of variable j beyond what can be explained by the other variables. Analyses here are conditional on the clusterings, which have been derived independently of the

explanatory variables. No formal model selection has been applied, rather the presented model comprising all variables of interest without interactions is the biggest one that seems reasonable with the given numbers of observations. Therefore the standard theory of the MLR applies.

Because we do not fix the number of clusters, this leads to a large number of tests being performed. Statistical hypothesis tests and p-values have been criticised because in many studies including the present one the computation of multiple p-values and the potential of even more testing because of pre-processing decisions that could have been made differently (such as our scoring of ranks) make it too easy to find individual significant p-values (e.g. [26]). We use p-values here in an exploratory manner for visualisation, highlighting particularly consistent (or inconsistent) significances over all or many numbers of clusters, without interpreting individual p-values according to the classical theory, which would implicitly assume that each of them was the only one we computed.

As a second approach we fitted random forests [7] as implemented in the R-package 'randomForest') to the classification problem of predicting $Y_1, \ldots, Y_G$ from $\mathbf{x}_1, \ldots, \mathbf{x}_p$. The random forest is a powerful data mining technique and in contrast to the multinomial logistic regression model, random forests allow to model nonlinear relationships and higher-order interactions between predictor variables. On the other hand, a lack of distribution theory does not allow to compute standard significance tests.

Random forests classify observations by majority vote based on many classification trees computed from non-parametric bootstrap samples. The influence of specific predictors in the presence of the other predictor (explanatory) variables as quantified by computing the mean decrease of classification accuracy for models that do not include a specific predictor variable. This and the overall classification error are computed as out-of-bag (OOB) error, that is, for each observation only those trees are used for which the observation was not involved in growing the tree.

5. Results

We present the results using Equation (1) for all five brand categories combined in detail; detailed results for the individual categories can be obtained from the authors.

Figure 1 shows MDS visualisations of the distances. The data from Study 1 do not show strong clustering, but two slightly denser areas can be observed, and on the left side some participants with rather atypical preferences are scattered. The data from Study 2 lack any visible clustering structure, although one can obviously still partition them into clusters with as low as possible within-cluster distances. A marketing practitioner could be interested in the specific clusterings, but we do not focus on them here.

Figure 2 shows that the values of the ASW and PΓ are rather low for both data sets (both of these are between −1 and 1 and values should be substantially higher than 0 in order to indicate clear clusters) but slightly better for Study 1. The plot for Study 1 can be seen as some weak indication in favour of $G = 7$, but is still quite ambiguous when it comes to the optimal number of clusters; in Study 1, in agreement with the MDS, no specific G looks convincing.

The five different categories of brands are rather heterogeneous; correlation coefficients between the vectors of distances from Equation (1) applied to individual categories are mostly between 0 and 0.1 with only three correlations (all in Study 1) between 0.1 and
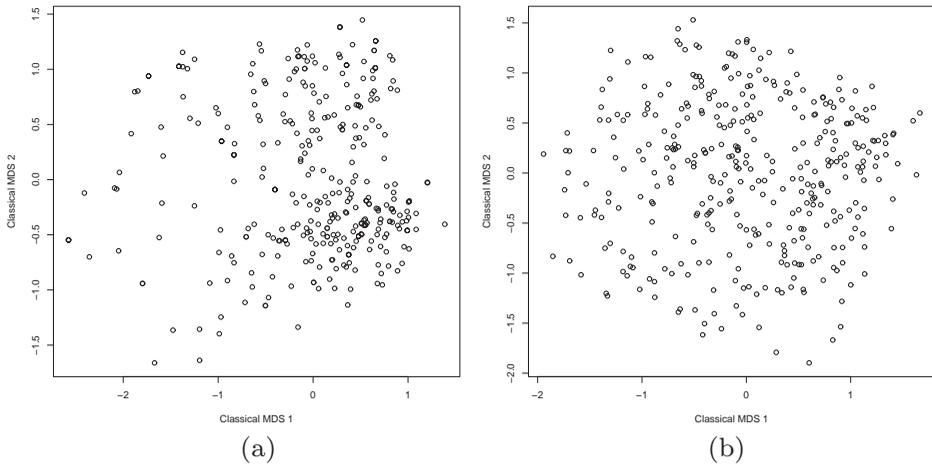
Figure 1. Two-dimensional classical MDS for brand preferences. (a) Study 1 (London) and (b) Study 2 (Scotland).
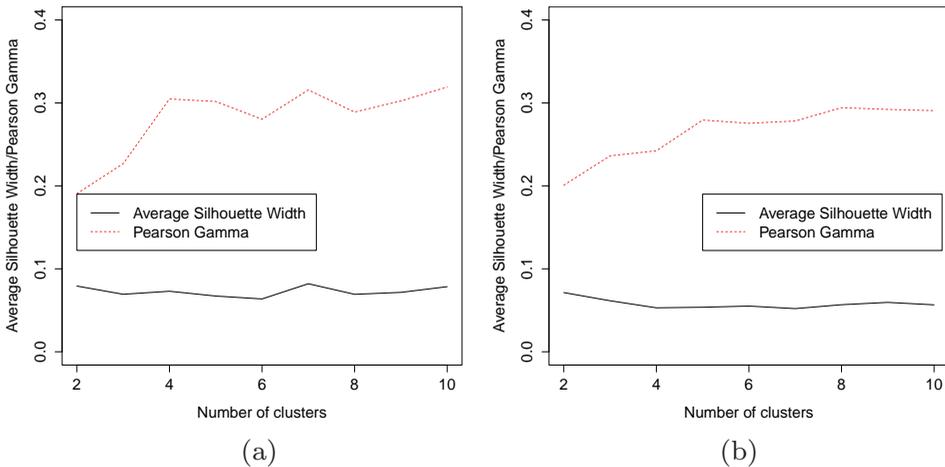
Figure 2. ASW and PΓ for PAM with 2–10 clusters. (a) Study 1 (London) and (b) Study 2 (Scotland).

0.2. All of these are correlated at around 0.4–0.5 with the distance based on all categories combined, which therefore seems to be a well balanced compromise. The adjusted Rand indexes [37] between the clusterings based on individual categories and combined categories are mostly around 0.1 (between −0.017 and 0.299), indicating a rather moderate similarity.

Figure 3 shows the log10-transformed p-values for the inclusion of all five personality variables as one block into the logistic regression model in addition to the 6 socio-demographic variables across all nine cluster solutions. For Study 1, it shows that the contribution of the five personality variables is significant for all models at a significance level of 0.05 for all categories combined. Their weakest contribution is found for $G = 2$; for $G \geq 4$ and particularly $G \geq 8$, the personality variables are strongly significant. Finer
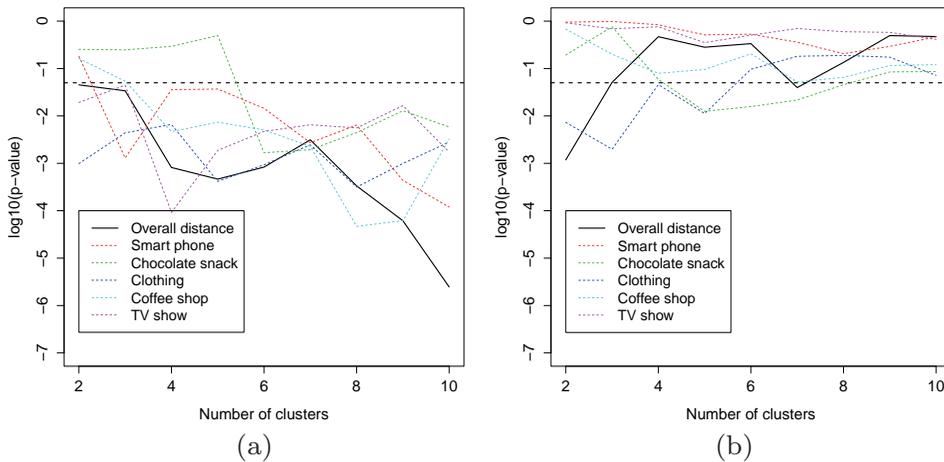
Figure 3. log10-transformed *p*-values for testing the H_0 that all of the coefficients for the personality variables are zero from multinomial logistic regression with clustering with 2–10 clusters as response (the solid line corresponds to the clustering with all categories combined). (a) Study 1 (London) and (b) Study 2 (Scotland).

segmentation solutions show a clearer impact of the personality variables. The results for individual categories are in line with this.

For Study 2, Figure 3 shows that the personality variables taken together as a block, for all categories combined, only have a significant impact on predicting cluster membership in the multinomial logistic regression models for the 2-cluster solution. Given the number of tests, this is a very weak indication regarding the impact of the personality variables. Results for individual categories show some weakly significant impact for clothing and chocolate snacks.

Looking at the individual variables, Figure 4 shows a breakdown of the association between each individual socio-demographic and personality variable and cluster membership for all categories combined. Gender and Openness to Experience are the only variables with significant associations across all cluster solutions in Study 1. In addition, Agreeableness has a significant influence for the more finer-grained solution from 7 to 10 clusters and Dwelling Type is also significant for the solutions with 7 to 9 clusters. The only other variable that shows a significant influence for at least some solutions is the Highest Education Level achieved.

In Study 2, Gender shows a strongly significant influence over all clustering solutions. Homeshare, Education and WorkPosition are significant for some numbers of clusters. Out of the personality variables, Conscientiousness is significant for $G = 2$ and $G = 3$ though not for larger G. The corresponding results for the individual categories are summarised in Table 1.

The variable-wise results from the random forest model (Figure 5) in Study 1 largely confirm the findings from the logistic regression models regarding the importance of Gender and Openness to Experience. In addition, the trend for Agreeableness to gain importance for finer grained cluster solutions is also visible in the random forest models. However, in contrast to the results from the logistic regression models, Age plays an important role in the random forest models across all clusterings and Extraversion and
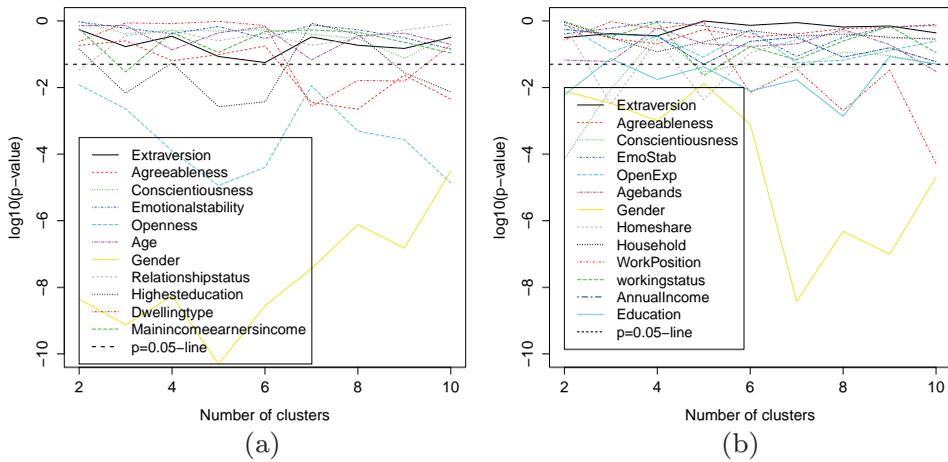
Figure 4. log10-transformed *p*-values for testing the H_0 that the coefficients for every individual variable are zero from multinomial logistic regression with clustering with 2-10 clusters as response (the solid line corresponds to the clustering with all categories combined). (a) Study 1 (London) and (b) Study 2 (Scotland).

Table 1. Overview of results (based on interpretation of the graphical displays).

	MLR evidence for personality variables	MLR most influential variables	Random forest most influential variables
Study 1 London			
Categories combined	Strong	Gender, Open, Educ., Dwelling, Agreeableness	Gender, Age, Open, Extraversion, Agreeableness
Smart phone	Strong	Gender, Emotion, Conscientiousness	Agree., Age
Chocolate snack	Weak	Open, Relationship, Gender	Age, Income
Clothing	Strong	Open, Education, Relationship, Income	Age
Coffee shop	Strong	Extraversion Agreeableness	Income, Age Extraversion
TV show	Strong	Gender, Emotion, Income, Open	Emotion, Age, Extrav., Gender
Study 2 Scotland			
Categories combined	Weak	Age, Education, WorkP., Homes. Conscientiousness	N/A[a]
Smart phone	None	none	N/A[a]
Chocolate snack	Weak	Open, Gender	N/A[a]
Clothing	Weak	Open, Age	N/A[a]
Coffee shop	None	none	N/A[a]
TV show	None	none	N/A[a]

[a] Random forest not clearly better than baseline.

Main Earners Income also rank among the more important variables for predicting cluster membership but without any clear trend in variable importance across cluster solutions. The fact that the importance of Age, Extraversion and Main Earners Income was not visible in the results given in Figure 4 is due to correlations among some of the predictor variables and the different handling of the contribution of correlated predictors in the multinomial
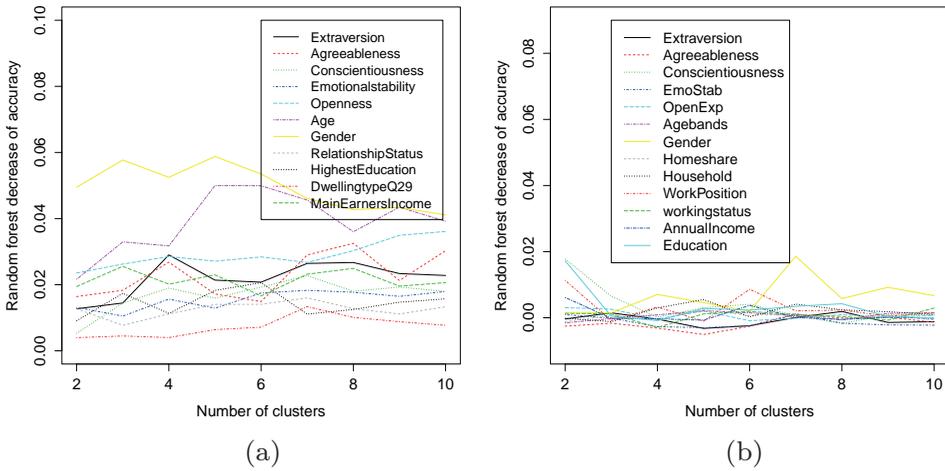
Figure 5. Random forest decrease of accuracy for all variables. (a) Study 1 (London) and (b) Study 2 (Scotland).

logistic regression compared to the random forest model. Specifically, Extraversion shows a considerable but not unusual [2] correlation of $r = 0.42$ with Openness and in the current sample, Age was correlated with Gender ($r = -0.38$) and Main earners income ($r = -0.13$). While the specific contribution of each variable in the logistic regression model is assessed in addition to all other variables being present in the regression model, the random forest spreads the contribution more equally across correlated variables. From the present data it is not possible to decide whether the true contribution to cluster membership does arise from Age or Gender or both. Therefore is seems sensible to consider both variables as potential contributors in line with the variable importance score from the random forest models.
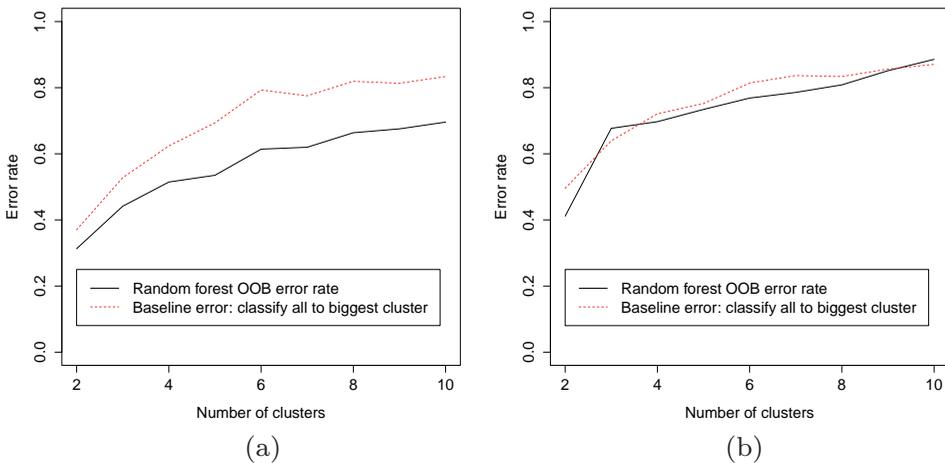
Figure 6. Comparison of random forest OOB-error with error from the trivial rule to classify every observation into the largest cluster. (a) Study 1 (London) and (b) Study 2 (Scotland).

Figure 6 depicts the OOB classification error of the random forest model across the nine cluster solutions and in comparison with the baseline error. We define baseline error as the error resulting from the assignment of all participants to the largest cluster. The random forest model consistently does a bit better but not much better than the baseline error in Study 1 (between 8% for the 2-cluster solution and 18% for the 6-cluster solution).

In Study 2 the random forest by and large hardly improves on the baseline error, which indicates that it does not provide specific evidence for the influence of any of the explanatory variables. Accordingly, Figure 5 does not show clear differences between variables.

6. Discussion

Our empirical results derived from the data of a sample of young adults living in a metropolitan area (Study 1) show that personality variables become more important for market segmentation as the number of target segments increases. Thus, personality variables play an important role in finer grained market segmentations which supports earlier evidence along the same lines [10,21,56]. However, the relationship between the systematic increase in the number of consumer groups and the importance gained by the big five personality variables becomes very apparent through the clustering empirical approach used here. The variable importance scores of the random forest model clearly indicate the primary importance of the socio-demographic variables, Age and Gender, but personality variables Openness, Extraversion and Agreeableness play a considerable role as well, at least segmentations with 6 or more target groups.

Hence, the data from this study does not agree with Sharps notion that brand preferences are not at all linked to consumer traits. However, our methodological procedure differed from previous studies (e.g. the widely cited study by Evans [20]) in as much as we aggregated brand preferences across five product categories to cluster consumers (as opposed to predicting preferences for a single brand). Thus, this study used a different dependent variable compared to the earlier studies cited by Sharp [67]. Moreover, for this study we selected a specific sample of young adults living in an urban area and matched product categories and brands to it that are most relevant and most familiar to this narrow demographic group. Thus, in marketing terms the sample represents a main target audience for the selected brands.

Study 2 uses the same brands and survey set-up but presents them to a sample of adults from a different geographical area in Britain with much more heterogeneous demographic characteristics. The demographically much broader sample in Study 2 does not yield such positive results. This sample was not only much more heterogeneous in its socio-demographic characteristics but also less representative of the chosen brands target audience which might have resulted in many more random preference ratings.

So overall there are mixed results regarding whether personality variables can have predictive power for distinguishing between groups of consumers, which can probably be explained by the different characteristics of the samples regarding homogeneity and representation of the target audience of the brands, although this is somewhat speculative given that there are only two samples. In any case the results from the two samples in the present study already indicate that the association between personality and demographic variables

on one hand and consumer cluster of brand preferences on the other hand is not universal but does depend on the sample of consumers.

From the present data we could not replicate the result reported by Sandy *et al.* [66] that personality variables are generally as important for market segmentation as demographic variables. But at least for the data in Study 1 our results demonstrate that personality variables can have a considerable impact on market segmentation solutions in combination with socio-demographic variables. However, it is worth remembering that in the present study we used a 10-item personality short form to measure traits on five personality dimensions while [66] used four dimensions of the much longer NEO- PI-R inventory. While the pseudo R^2-values reported by Sandy *et al.* [66] are not directly comparable to the classification rates and variable importance scores obtained from our random forest model, it appears nonetheless that the measures derived from the 10-item short form had a similar relative impact for the finer-grained segmentation solutions in Study 1.

In Study 1 the classification model performed clearly better than the baseline error rate, refuting the notion that market segmentation based on consumer background variables is a largely impossible undertaking [67]. However, one has to bear in mind that we used a different dependent variable compared to the earlier studies that [67] and others have referred to as evidence for the ineffectiveness of market segmentation. Hence, it is possible that Sharps interpretation and the results from our clustering approach are ultimately compatible if the aggregate of several brand preferences represents more stable measurement than single brand preferences. Additionally, the use of the random forests as a modern data mining technique with a strong predictive power might also have contributed to the clear segmentation results of this study.

In sum, the results of this study have shown that is possible to segment consumers according to their brand preferences into clusters. While there seemed to be no optimal or natural number of clusters, any division of the sample in Study 1 into clusters derived from the PAM clustering procedure had a significant association with socio-demographic and personality variables, with personality variables being more important in segmentation solutions with a higher number of consumer groups. This means for marketing practitioners that the decision on the number of target segments of a market can be based on practical considerations and costs. The results from Study 1 show that market segmentations with a classification accuracy of up to 70% can be achieved with as few as 11 variables derived from 16 question items. Given that several variables (e.g. Emotional Stability/Neuroticism, Conscientiousness, Dwelling Type) did not contribute much to the prediction accuracy of the random forest models, it seems possible to reduce the length of the consumer questionnaire even further without losing a large amount of predictive power.

In addition, the segmentation approach we propose here has the conceptual advantage that it generates data-driven model solutions that are very close to practical notions of brand personality [1] and profiles of prototypical consumers which are often defined by associations between habits, preferences and demographic profiles and which often seem helpful to guide the development of creative ideas and planning of marketing campaigns. Thus, this clustering approach to market segmentation which builds on personality and socio-demographic variables not only has comparatively good predictive power but can also blend in conceptually with current practice in advertising and marketing.

On the statistical side, a key feature of this paper is that our use of clustering does not rely on the existence of a true underlying clustering or a true number of clusters. Clustering

is used to summarise and simplify the brand preferences, and analyses run over a range of numbers of clusters rather than a single one. Because the clusterings are constructed so that large within-cluster distances are avoided, clusters can still be interpreted by marketing practitioners, while their 'constructive' rather than 'real' nature is acknowledged. Another observation is that both the multinomial logistic regression and the random forest deliver valuable and largely complementary information about the impact of the various variables on the clusterings.

The methodology can be used by a practitioner to use and interpret specific clusters. The estimated coefficients of the multinomial logistic regression can be interpreted regarding the specific contribution of the variables regarding specific clusters, although this is beyond the scope of this paper.

A very elegant option for future studies as suggested by an anonymous reviewer would be to combine clustering and the fit of explanatory models for the clusters into a single model. However, this beyond the scope of the present paper work and standard tests of the MLR could no longer be used within such an approach.

Acknowledgements

Disclosure statement

Funding

ORCID

Christian Hennig 📛 http://orcid.org/0000-0003-1550-5637

References

[1] J.L. Aaker, *Dimensions of brand personality*, J. Mark. Res. 34 (1997), pp. 347–356.
[2] A. Aluja, O. Garcia, and L.F. Garcia, *A comparative study of Zuckerman's three structural models for personality through the neo-pi-r, zkpq-iii-r, epq-rs and goldberg's 50- bipolar adjectives*, Pers. Individ. Differ. 33 (2002), pp. 713–725.
[3] M. Ammann, R. Frey, and M. Verhofen, *Do newspaper articles predict aggregate stock returns?* J. Behav. Finance 15 (2014), pp. 195–213.
[4] A. Ansari, S. Essegaier, and R. Kohli, *Internet recommendation systems*, J. Mark. Res. 37 (2000), pp. 363–375.
[5] H. Baumgartner, *Toward a personology of the consumer*, J. Consum. Res. 29 (2002), pp. 286–292.
[6] C. Blamires, *Segmentation*, in *The Handbook of International Market Research Techniques*, R. Birn and P.N. Hague, eds., Kogan Page, London, 2000, pp. 497–518.
[7] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.

[8] E. Brentari, L. Dancelli, and M. Manisera, *Clustering ranking data in market segmentation: A case study on the Italian McDonald's customers preferences*, J. Appl. Stat. 43 (2016), pp. 1959–1976.

[9] R.P. Brody and S.M. Cunningham, *Personality variables and the consumer decision process*, J. Mark. Res. 5 (1968), pp. 50–57.

[10] S. Choo and P.L. Mokhtarian, *What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice*, Transp. Res. A, Policy Pract. 38 (2004), pp. 201–222.

[11] P.T. Costa and R.R. McCrae, *Neo PI-R Professional Manual*, Psychological Assessment Resources, Odessa, FL, 1992.

[12] M.A. Croon, *Latent class models for the analysis of rankings*, in *New Developments in Psychological Choice Modeling*, G. De Soete, H. Feger, and K. Klauer, eds., Elsevier, North Holland, 1989, pp. 99–121.

[13] M.A. Croon and R. Luijkx, *Latent structure models for ranking data*, in *Probability Models and Statistical Analyses for Ranking Data*, M.A. Fligner and J.S. Verducci, eds., Springer, New York, 1993, pp. 53–74.

[14] L. Dancelli, M. Manisera, and M. Vezzoli, *On Pinto Da Costa & Soares and other weighted rank correlation measures deriving from the Spearman's ρ*, Adv. Appl. Stat. 36 (2013), pp. 83–104.

[15] E.H. Demby, *Psychographics revisited: The birth of a technique*, Mark. Res. 6 (1994), p. 26.

[16] W.S. DeSarbo, S. Atalay, D. LeBaron, and S. Blanchard, *Estimating multiple ideal points from context dependent survey data*, J. Consum. Res. 35 (2008), pp. 142–153.

[17] S. Dolnicar, *A review of data-driven market segmentation in tourism*, J. Travel Tour. Mark. 12 (2002), pp. 1–22.

[18] S. Dolnicar, S. Kaiser, K. Lazarevski, and F. Leisch, *Biclustering: Overcoming data dimensionality problems in market segmentation*, J. Travel Res. 51 (2012), pp. 41–19.

[19] S. Dolnicar and F. Leisch, *Evaluation of structure and reproducibility of cluster solutions using the bootstrap*, Mark. Lett. 21 (2010), pp. 83–101.

[20] F.B. Evans, *Psychological and objective factors in the prediction of brand choice Ford versus Chevrolet*, J. Bus. 32 (1959), pp. 340–369.

[21] E. Ferguson, J.J. Heckman, and P. Corr, *Personality and economics: Overview and proposed framework*, Pers. Individ. Differ. 51 (2011), pp. 201–209.

[22] R. Fitzgerald, *Products, firms and consumption: Cadbury and the development of marketing, 1900–1939*, Bus. Hist. 47 (2005), pp. 511–531.

[23] M.A. Fligner and J.S. Verducci, *Multistage ranking models*, J. Amer. Statist. Assoc. 83 (1988), pp. 892–901.

[24] A. Furnham, *Relationship among four big five measures of different length*, Psychol. Rep. 102 (2008), pp. 312–316.

[25] A. Furnham and B. Gunter, *Consumer Profiles: An Introduction to Psychographics*, Routledge, London, 1992.

[26] A. Gelman and E. Loken, *The statistical crisis in science*, Amer. Sci. 102 (2014), pp. 460.

[27] I.C. Gormley and T.B. Murphy, *Analysis of irish third-level college applications data*, J. Roy. Statist. Soc. Ser. A 169 (2006), pp. 361–379.

[28] I.C. Gormley and T.B. Murphy, *A mixture of experts model for rank data with applications in election studies*, Ann. Appl. Stat. 2 (2008), pp. 1452–1477.

[29] I.C. Gormley and T.B. Murphy, *A grade of membership model for rank data*, Bayesian Anal. 4 (2009), pp. 265–295.

[30] S.D. Gosling, P.J. Rentfrow, and W.B. Swann, *A very brief measure of the big-five personality domains*, J. Res. Pers. 37 (2003), pp. 504–528.

[31] D.M. Greenberg, M. Kosinski, D.J. Stillwell, B.L. Monteiro, D.J. Levitin, and P.J. Rentfrow, *The song is you*, Soc. Psychol. Pers. Sci. 7 (2016), pp. 597–605. Available at http://dx.doi.org/10.1177/1948550616641473.

[32] M. Halkidi, M. Vazirgiannis, and C. Hennig, *Method-independent indices for cluster validation and estimating the number of clusters*, in *Handbook of Cluster Analysis*, C. Hennig, M. Meila, F. Murtagh, and R. Rocci, eds., Chap. 26, Chapman & Hall/CRC, Boca Raton FL, 2015, pp. 595–618.

[33] K. Hammond, A.S. Ehrenberg, and G.J. Goodhardt, *Market segmentation for competitive brands*, Eur. J. Mark. 30 (1996), pp. 39–49.

[34] C. Hennig, *Clustering strategy and method selection*, in *Handbook of Cluster Analysis*, C. Hennig, M. Meila, F. Murtagh, and R. Rocci, eds., Chap. 31, Chapman & Hall/CRC, Boca Raton FL, 2015, pp. 703–730.

[35] C. Hennig and T.F. Liao, *How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification*, J. Roy. Statist. Soc. Ser. C 62 (2013), pp. 309–369.

[36] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., John Wiley and Sons, Inc., New York, 2000.

[37] L. Hubert and P. Arabie, *Comparing partitions*, J. Classif. 2 (1985), pp. 193–218.

[38] J. Jacques and C. Biernacki, *Model-based clustering for multivariate partial ranking data*, J. Statist. Plann. Inference. 149 (2014), pp. 201–217.

[39] O.P. John, L.P. Naumann, and C.J. Soto, *Paradigm shift to the integrative big five trait taxonomy*, in *Handbook of Personality: Theory and Research*, O.P. John, R.W. Robins, and L.A. Pervin, eds., Vol. 3, Guildford Press, New York, 2008, pp. 114–158.

[40] L.R. Kahle, *Social Values and Social Change: Adaptation to Life in America*, Praeger Publishers, Westport, CT, 1983.

[41] L.R. Kahle, S.E. Beatty, and P. Homer, *Alternative measurement approaches to consumer values: The list of values (lov) and values and life style (vals)*, J. Consum. Res. 13 (1986), pp. 405–409.

[42] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data*, Wiley, Chichester, 1990.

[43] R. Kennedy and A. Ehrenberg, *There is no brand segmentation*, Mark. Res. 13 (2001), p. 4.

[44] R. Kennedy and A.S. Ehrenberg, *Competing retailers generally have the same sorts of shoppers*, J. Mark. Commun. 7 (2001), pp. 1–8.

[45] D.R. Lehmann, *Television show preference: Application of a choice model*, J. Mark. Res. 8 (1971), pp. 47–55.

[46] G.L. Lilien and A. Rangaswamy, *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning*, 2nd ed., Prentice Hall, Engelwood Cliffs, 2002.

[47] R.D. Luce, *Individual Choice behavior: A Theoretical Analysis*, Wiley, New York, 1959.

[48] J. Marden, *Analyzing and Modeling Rank Data*, Taylor & Francis, Oxford, 1996.

[49] K. Mardia, *Some properties of classical multidimensional scaling*, Comm. Statist. Theory Methods 7 (1978), pp. 1233–1241.

[50] S.C. Matz, J.J. Gladstone, and D. Stillwell, *Money buys happiness when spending fits our personality*, Psychol. Sci. 27 (2016), pp. 715–725. Available at http://dx.doi.org/10.1177/0956797616635200, PMID: 27056977.

[51] R.R. McCrae and P.T. Costa, *A five-factor theory of personality*, in *Handbook of Personality: Theory and Research*, L.A. Pervin and O.P. John, eds., Guilford Press, New York, 2003, pp. 139–153.

[52] R.D. Michman, *Lifestyle Market Segmentation*, Praeger Publishers, Westport, CT, 1991.

[53] G. Miller, *Spent: Sex, evolution, and Consumer Behavior*, Penguin, London, 2009.

[54] A. Mitchell, *The Nine American Lifestyles: Who We are and Where We're Going*, Viking, London, 1983.

[55] E.A. Mooi and M. Sarstedt, *A Concise Guide to Market Research: The Process, Data, and Methods using IBM SPSS Statistics*, Springer, New York, 2011.

[56] R.C. Mulyanegara, Y. Tsarenko, and A. Anderson, *The big five and brand personality: Investigating the impact of consumer personality on preferences towards particular brand personality*, J. Brand Manag. 16 (2009), pp. 234–247.

[57] T.B. Murphy and D. Martin, *Mixtures of distance-based models for ranking data*, Comput. Statist. Data Anal. 41 (2003), pp. 645–655.

[58] M.B. Naseri and G. Elliott, *Market segmentation revisited: The predictive utility of demographics*, in *ANZMAC 2006 Abstracts & Programme: Brisbane, Queensland 4–6 December 2006: Advancing Theory, Maintaining Relevance*, E.M.I. Larkin, ed., Brisbane Queensland, Brisbane, 2006, 1–8.

[59] T.P. Novak and B. MacEvoy, *On comparing alternative segmentation schemes: The list of values (lov) and values and life styles (vals)*, J. Consum. Res. 17 (1990), pp. 105–109.

[60] J.P. Peter and J.C. Olson, *Consumer Behavior and Marketing Strategy*, McGraw-Hill, London, 2010.

[61] R. Plomin and A. Caspi, *Behavioral genetics and personality*, in *Handbook of Personality: Theory and Research*, O.P. John, R.W. Robins, and L.A. Pervin, eds., 2nd ed., Guilford Press, New York, 1999, pp. 251–276.

[62] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, *Our Twitter profiles, our selves: Predicting personality with Twitter*, in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, IEEE, 2011.

[63] P. Resnick and H.R. Varian, *Introduction to special section on recommender systems*, Commun. ACM. 40 (1997), pp. 56–58.

[64] B.W. Roberts, N.R. Kuncel, R. Shiner, A. Caspi, and L.R. Goldberg, *The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes*, Perspect. Psychol. Sci. 2 (2007), pp. 313–345.

[65] J.F. Salgado, *The five factor model of personality and job performance in the european community*, J. Appl. Psychol. 82 (1997), pp. 30–43.

[66] C.J. Sandy, S.D. Gosling, and J. Durant, *Predicting consumer behavior and media preferences: The comparative validity of personality traits and demographic variables*, Psychol. Mark. 30 (2013), pp. 937–949.

[67] B. Sharp, *How Brands Grow: What Marketers Don't Know*, Oxford University Press, Oxford, 2010.

[68] F.J. Thumin, *Consumer behavior as related to personality, intelligence, age and education*, Psychol. Rep. 23 (1968), pp. 1185–1186.

[69] W.N. Venables and B.D. Ripley, *Modern Applied Statistics with S*, 4th ed., Springer, New York, 2002.

[70] A. Weinstein, *Handbook of Market Segmentation: Strategic Targeting for Business and Technology Firms*, Haworth series in Segmented, Targeted, and Customized Marketing, Haworth Press, 2004. Available at https://books.google.de/books?id=IAwKAQAAMAAJ.

[71] W.D. Wells, *Psychographics: A critical review*, J. Mark. Res. 12 (1975), pp. 196–213.

[72] W. Youyou, M. Kosinski, and D. Stillwell, *Computer-based personality judgments are more accurate than those made by humans*, Proc. Natl. Acad. Sci. 112 (2015), pp. 1036–1040. Available at http://www.pnas.org/content/112/4/1036.abstract.