Food Quality and Preference 33 (2014) 27–36

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Classification trees in consumer studies for combining both product attributes and consumer preferences with additional consumer characteristics

Rosaria Romano [a,*], Cristina Davino [b], Tormod Næs [c,d]

[a] University of Calabria, Department of Economics, Statistics and Finance, 87036 Arcavacata di Rende, Cosenza, Italy
[b] University of Macerata, Department of Political Sciences, Communications and International Relations, Piazza Oberdan 2, 62100 Macerata, Italy
[c] Nofima Mat AS, Osloveien 1, NO-1430 Ås, Norway
[d] University of Copenhagen, Department of Food Science, Rolighedsvej 30, DK-1958 Frederiksberg, Denmark

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form 7 October 2013
Accepted 17 November 2013
Available online 23 November 2013

Keywords:
Classification trees
Acceptance pattern
Questionnaire data
Validation

ABSTRACT

The main objective of this paper is to describe and discuss the use of classification trees in consumer studies. Focus will be given to the use of the method in relating segments of consumers, based on their acceptance pattern, to additional consumer characteristics, including attitudes, habits and demographics variables. Advantages of the method in handling typical issues from consumer studies will be discussed. Primary interest will be given to the validation of the results, which will also be compared with results from alternative methods widely used in consumer studies. The approach will then be illustrated by using data from a conjoint study of apple juice.

1. Introduction

In experimental consumer studies, one of the primary aims is to obtain information about consumer preference or purchase intent for a number of products. One of the most used methodologies in this field is *conjoint analysis* (Green & Srinivasan, 1978; Gustafsson, Hermann, & Huber, 2003), which studies the effect of a number of product characteristics on consumer acceptance. In conjoint studies, product information is organized into a number of factors, each combination giving rise to a trial product to be presented to consumers. Consumers then express their degree of liking (or another hedonic characteristic) for each combination, or alternatively their ranking of the products or their choice (Louviere, 1988; Louviere, Hensher, & Swait, 2000). The data are generally analyzed using Analysis of Variance (ANOVA) (Næs, Brockhoff, & Tomic, 2010; Searle, 1971) or via rank order logistic modeling (Train, 1986; McCullagh & Nedler, 1989).

Due to the intrinsic heterogeneity in consumer acceptance patterns, it is extremely important to investigate not only the drivers of liking at the population level, but also to explore individual differences among consumers (Gustafsson et al., 2003; Næs et al., 2010). In addition, it is very important for the purpose of planning appropriate marketing strategies to relate the individual differences in acceptance pattern to other consumer characteristics, including attitudes, habits and demographics. (Benton, Greenfield, & Morgan, 1998; Wedel & Kamakura, 1998) or to other external information such as sensory data (McEwan, 1996; Schlich & McEwan, 1992; Vigneau & Qannari, 2002). To achieve this, many approaches have been proposed. One possible strategy is to segment acceptance values using some type of cluster analysis and then relate the obtained segments to the additional consumer variables by tabulation; another option might include regression analysis (Næs, Kubberod, & Silvertsen, 2001) or discriminant analysis (Ripley, 1996). Other important possibilities are based on ANOVA with the incorporation of effects for additional consumer characteristics (Næs et al., 2010), combinations of ANOVA modeling and multivariate analysis (Endrizzi, Menichelli, Johansen, Olsen, & Næs, 2011) and combining all datasets into one single multivariate analysis using the L-PLSR method (Martens et al., 2005; Vinzi, Guinot, & Squillacciotti, 2007). An alternative procedure allowing for a simultaneous analysis of the three blocks of information (product hedonic scores, product sensory descriptors and consumer attributes) has been recently proposed by Vigneau, Charles, & Chen (2014). The different blocks of information may also be related to each other in different ways by using some type of structural equations modeling (Guinot, Latreille, & Tenenhaus, 2001; Olsen, Menichelli, Sørheim, & Næs, 2012; Tenenhaus, Pagès, Ambroisine, & Guinot, 2005).

* Corresponding author. Address: University of Calabria, Department of Economics, Statistics and Finance, Cubo 1/C, 87036 Arcavacata di Rende, Cosenza, Italy. Tel.: +39 0984 492448; fax: +39 0984 492421.
E-mail addresses: rosaria.romano@unical.it (R. Romano), cdavino@unimc.it (C. Davino), tormod.naes@nofima.no (T. Næs).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.11.006

The purpose of this paper is to present an alternative approach based on the use of *classification trees* (Breiman, Friedman, Olshen, & Stone, 1984; Hastie, Tibshirani, & Friedman, 2009) for combining product attributes and consumer preferences with additional consumer characteristics. The focus will be on explaining consumer segments in terms of questionnaire data in a situation where the segments are based on the acceptance pattern of the consumer groups (a posteriori segmentation, see Næs et al., 2001).

This study will discuss and illustrate the main advantages and possible challenges of this approach, which is not often used in consumer science. In particular, we will discuss variable selection, multicollinearity among predictors, missing values, joint use of predictors at different scales and interpretation of results. Furthermore, particular attention will be given to the validation of the results. For comparison, we will also consider *Partial Least Squares Regression* (PLSR; Martens & Næs, 1983; Wold, 1995), which is another natural and often used choice for this type of study. The method will be illustrated using data from a conjoint study of apple juice (Olsen et al., 2011).

The paper is organized as follows. In the next section, the theoretical aspects of the statistical methods used in the paper are presented. Thereafter, the data from a conjoint study of apple juice are described in Section 3. The empirical results are presented and discussed in Section 4, and some conclusions are given in Section 5.

2. Theory

2.1. Segmentation in conjoint analysis

As discussed in the introduction, a number of approaches to segmentation in conjoint analysis have been put forward. In this paper, we will concentrate on the method proposed in Endrizzi et al. (2011), which was founded on interpretation-based visual inspection of Principal Component Analysis (PCA) plots. However, we emphasize that the classification tree method can be used for any type of segmentation approach. The method in Endrizzi et al. (2011) for analyzing individual differences is based on residuals from an ANOVA model containing all possible product effects and interactions (a saturated model) plus the consumer main effect. In this type of modeling, all information about individual interactions with product factors is collected only in the residuals. The residuals from the saturated model give rise to double-centered data that are useful for both visualizing and interpreting individual differences, as well as for providing a basis for visual (or a statistical) interpretation based segmentation. The segments used for this study will be based on visual inspection of the first principal component (i.e. the dimension with the largest variability) of the residual data, but other segmentations could also be envisioned.

When relating segments to consumer characteristics, many technical problems may arise. For example, a huge number of variables are generally included in the questionnaire, and they may be on completely different measurement scales. Additionally, there may be strong multi-collinearities among them, and there may be several missing values. A statistical approach often used for solving these problems is the PLSR (Martens & Næs, 1983). This approach has been used in Endrizzi et al. (2011) and will be discussed in the following sections. The aim of the present paper is, however, to present an alternative approach based on the *classification trees*, which has certain advantages for this type of study.

2.2. Classification and regression trees

Classification and regression trees (CART) are statistical methods introduced by Breiman et al. (1984). *Classification trees* aim to determine the membership of objects/units to a number of preselected classes while *regression trees* aim to build a model for a dependent quantitative variable. The two approaches can allow for quantitative and categorical explanatory variables simultaneously. In both cases, the CART tree is a binary recursive partitioning procedure that tries to identify variables and split points of the explanatory variables that predict the response in the best possible way.

The procedure starts at the *root node* (the entire dataset). Then, the data are split into two so-called "children", which are in turn split into so-called "grandchildren", etc. The procedure stops when no further splits of the data are possible due to lack of data, unless a stopping rule is defined. The last children in the tree model, defining the *terminal nodes*, identify units belonging to the same class (classification tree) or representing the same value of the quantitative dependent variable (regression tree). The CART algorithm includes automatic missing values handling by the use of so-called "surrogate variables" (Breiman et al., 1984). A surrogate split is a variable whose pattern in the dataset with respect to the dependent variable is similar to the primary split in the sense that it can predict almost the same partition of the parent node into the two child nodes. It can also be used for ranking of the importance of the explanatory variables, and it is invariant to the relative scaling of the input variables, which may be important if the variables are measured on different scales.

For the rest of the paper, focus will be on the *classification tree*. In the following section, we introduce this area. Some possible advantages of the method will also be discussed.

2.2.1. Classification tree procedure

Let X be the matrix of the M independent variables observed on N units. It will be assumed that each of the N units belongs to one of J pre-specified classes. For each step of CART (i.e. for each node in the tree), the method attempts to find the most important variable X_i ($i = 1,\ldots,M$) and the most important splitting point s for accounting for the variability in the dependent variable, here defined by the J different groups. By splitting point for a continuous variable X_i, we mean the value s for which units with $X_i > s$ and $X_i \leqslant s$ represent two subsets ("children") of the dataset found to be optimal for the prediction ability at that point. Several splitting points for a variable are possible, but it is generally recommended to make only a binary split at each node. For binary input variables, there is obviously only one splitting point, but for a categorical variable with l levels, there are $2^{l-1}-1$ possibilities. To know what one means by an optimal split, a *splitting criterion* is needed. The criterion used for the classification tree procedure is described in Appendix A.

The final nodes of the tree are named terminal nodes, and one of the J groups is assigned to each of them by choosing the group most frequently represented in the node. To assess the quality of the classification procedure, the *percentage of correctly classified* (or its complement to 100, *misclassification rate* or *cost*) can be computed as an average of the percentage of correct classifications for each terminal node (number of units belonging to the group assigned to the node over the number of units in the node) weighted by the cardinality of each node.

Since node splitting depends on the homogeneity of the dependent variable inside the child nodes, the presence of outliers does not affect the results. This because usually the outliers correspond to only a few units in a dataset.

The CART method also provides a measure of the variable's importance, which is based on the sum of the improvements in all nodes in which the attribute appears as a splitter. More specifically, if the parent node 1 is split into child nodes 2 and 3, the importance of the split variable improves by $(r_1 - r_2 - r_3)/n$, where r_i are the risks and n is the total number of nodes in the tree. The risk is defined as $r_i = p_i * e_i$, where p_i is the node probability (proportion of units in the node) and e_i is the node impurity (proportion of misclassified units in the node).