Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Knowledge extraction from a large on-line survey: a case study for a higher education corporate marketing

K. Fernández-Aguirre [a] , M. I. Landaluce-Calvo [b] , A. Martín-Arroyuelos [a] & J. I. Modroño-Herrán [a]

[a] Departamento de Economía Aplicada III, University of the Basque Country (UPV/EHU), Bilbao, Spain

[b] Departamento de Economía Aplicada, University of Burgos (UBU), Burgos, Spain

Available online: 20 Apr 2011

Journal of Applied Statistics
2011, iFirst article

Taylor & Francis
Taylor & Francis Group

Knowledge extraction from a large on-line survey: a case study for a higher education corporate marketing

K. Fernández-Aguirre[a]*, M.I. Landaluce-Calvo[b], A. Martín-Arroyuelos[a] and J.I. Modroño-Herrán[a]

[a]*Departamento de Economía Aplicada III, University of the Basque Country (UPV/EHU), Bilbao, Spain;* [b]*Departamento de Economía Aplicada, University of Burgos (UBU), Burgos, Spain*

(*Received 28 April 2010; final version received 2 February 2011*)

For a higher education public institution, young in relative terms, featuring local competition with another private and both long-established and reputed one, it is of great importance to become a reference university institution to be better known and felt with identification in the society it belongs to and ultimately to reach a good position within the European Higher Education Area. These considerations have made the university governors setting up the objective of achieving an adequate management of the university institutional brand focused on its logo and on image promotion, leading to the establishment of a university shop as it is considered a highly adequate instrument for such promotion. In this context, an on-line survey is launched on three different kinds of members of the institution, resulting in a large data sample. Different kinds of variables are analysed through appropriate exploratory multivariate techniques (symmetrical methods) and regression-related techniques (non-symmetrical methods). An advocacy for such combination is given as a conclusion. The application of statistical techniques of data and text mining provides us with empirical insights about the institution members' perceptions and helps us to extract some facts valuable to establish policies that would improve the corporate identity and the success of the corporate shop.

Keywords: data and text mining; knowledge extraction; clustering; principal component analysis; correspondence analysis; PLS path modelling; logit models

1. Introduction

Before marketing was first established as an articulate discipline, practices targeted to customers' needs were already common inside business. That is, the target of such activities was precisely the customer. Thus, a feasible definition of the emerging marketing area would be that it is a set of principles, methodologies and techniques addressed to the simultaneous search of customers' need satisfaction and profit generation.

*Corresponding author. Email: karmele.fernandez@ehu.es

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664763.2011.567248
http://www.informaworld.com

After 1950, different related concepts have been introduced, such as corporate identity, corporate branding, corporate image, corporate reputation and corporate communications. All have evolved independently, but it seems adequate to attach all of them to the wider concept of what is known as corporate marketing [3] or corporate-level marketing [4].

Among several developing lines of corporate marketing, we focus on the concept of corporate identity, with a high relevance on the institutional priorities. Thus, in [52], it is pointed out that corporate identity incorporates the three basic concepts of communication, symbolism and behaviour of institutional members, on which, adequate actions will need to be taken to achieve an adequate corporate reputation [11].

Given that any institution members' behaviour will have a clear effect on the identity and the corporate image [25], any organization should consider, as one of its main goals, to manage employees' identification with its values and objectives. In this sense, some factors such as internal communication, corporate culture, institutional prestige or pride about institutional membership are very important. With this target in mind, the specific institution should develop techniques of internal marketing leading to an improvement in the degree of identification of its members with itself. This would then translate into an improvement in its own external image.

In the present work, we focus on a particular kind of institution: the university. In this type of institutions, all techniques establishing as a goal the improvement of corporate identity are very useful. This is particularly true when considering current competitiveness related to student interchange programmes, interest for achieving quality certifications and, particularly in the European Union, the creation of the European Higher Education Area (EHEA) with all issues related to the ongoing curricular reform. Universities, more than ever, must establish marketing strategies to successfully face peer competition, one of the consequences of this new reality. These strategies will have to focus on the differential characteristics of each institution, which must be effectively communicated to all relevant agents. As pointed out in [5], the image perceived by different groups of interest, or target groups, on an organization is determinant for achieving success, independently from organization management. In this context, and given the long experience in other sectors, corporate identity is known as a powerful instrument of competitive advantage [2,37].

Conscious about the new situation, many universities have begun to develop and implement programmes of corporate identity as part of their growth and expansion strategies. This identity has many diverse dimensions, defined by factors which are both external and internal to the institution, composing the so-called corporate identity mix [37]. However, corporate image is the dimension with a higher weight in the new marketing strategies developed by universities. This relevance is based upon works [36] showing that elements such as logo, colours and objects influence the image that any entity is willing to transmit to both external and internal agents.

In this context, we consider the case of the University of the Basque Country (UPV/EHU) that responds to the majority of the demand for higher education in the Autonomous Community of the Basque Country in Spain. This is a public young university which is in competition with the traditional private Deusto university. The institution, as part of a large project, the main aim of which is revamping its corporate image, is launching a corporate shop (also named gift or souvenir shop). It is evident in this process that it is very important to know about the starting image concept. However, the image concept is not unique as the different target groups, depending on the relationship maintained with the university, perceive a different image and are more sensitive to different dimensions of such a concept.

In this work, we analyse the behaviour of three internal strata of the institution: students, teaching staff and administrative staff. The interest in considering these internal agents is evident because, as pointed out in [35], they are receptors of the identity signals emitted by the institution and, at the same time, they are also agents being participants inside it who become emission instruments of such signals.

Furthermore, university members' attitudes (whatever the kind of members considered) with respect to the university or corporate shop are very interesting, given that such attitudes are a reflection of their identification with the principles, values, targets and other institutional characteristics. These can be measured through members' intention to get items featuring the logo representing the university [1,45].

This work is developed with respect to two basic targets. First, it aims to perform an analysis leading to recommend policies related to the kind of products to be offered more successfully in the university shop. Thus, it is of interest to know the main characteristics that any interviewed person thinks that an article featuring the university logo should possess in order to know the opinion on the articles being shown. It is also of interest to establish the personal characteristics of individuals most likely to buy logo articles. Secondly, some hypotheses coming from corporate marketing hypotheses are assumed, conforming to a testing objective using the available data:

(1) We start from the well-established assumption on organization identification stating that the degree of identification with an institution for an internal agent is mirrored in his/her attitude towards it and in a behaviour consistent with its objectives. Given that, it is believed that there should be a strong relationship between the satisfaction degree about institution membership and the will to possess articles featuring the institutional logo.

    In the current context, we understand that the institution identification degree is translated into the satisfaction degree about institutional membership. We then consider that the way to test that a greater identification implies a positive attitude to the university image is confirming a greater intentionality about possessing articles featuring the institutional logo, as a show of membership pride.

    One of the hypotheses to be tested in this work is thus the dependency of the intentionality to buy articles featuring the university logo on the degree of its membership satisfaction. The evidence about this relationship will arise along the work, through the different statistical techniques used.

(2) The fact is considered that the identification degree varies across the three strata considered in the analysis due to the different relationships of each of them with the institution. In this sense, several works show analyses on the impact of corporate image on a given institution's workers, considering a list with different categories [13,21].

    In this line, we find works related to higher education in Spain. In [50], an analysis is performed on the internal image of a university institution restricted to the student stratum, while in [49], a similar work refers only to the administrative staff stratum.

    In this work, we test, through different methods, if there exist differences in the attitudes shown by the three considered strata with respect to the articles displaying the university logo.

(3) Finally, according to the previous working hypothesis, we expect to find three different attitudes for the individuals of interest in this analysis as a function of the stratum they belong to. We establish the assumption that the individuals more implied in the institution image revamping policy are those showing a greater probability to buy articles featuring the logo.

    Regarding student behaviour, it must be said that there exists some controversy regarding the role played by university students. First, the student–university relationship is not the usual customer–corporation relationship. As pointed out in [50], students are both customers requiring higher education and the product to be demanded later by employers. On the other hand, it is evident that the link to the university is shorter in time when compared with that established with the other strata. Additionally, the service that students perceive is not only about education as it also extends to other aspects such as campus lifestyle [6].

    With respect to administrative and service staff, some studies [43] reveal the lack of integration/identification of this segment in the university organizational culture, which is organized around teaching and research. In this sense, in [49], when analysing the internal image that

Figure 1. The UPV/EHU logo.

> this stratum has on the university, the conclusion reached is that, generally speaking, the students see the university as a mere employment agent.
>
> With all these facts in mind, and in order to be consequent with the statements exposed on the different strata, the hypothesis to be checked is that the stratum most implied in this revamping policy is the teaching/research group, probably because given its characteristics it is precisely the most concerned about the prestige of the institution it is a part of.

After profiling these objectives, an on-line survey is carried out at the UPV/EHU to better know its members' perceptions about the institution and to know the acceptability and the potential success of the corporate shop selling products provided with the university logo. This logo is shown in Figure 1.

A large data set is obtained from this survey, which also includes a large number of questions. Therefore, it is necessary to extract the information and the relationships inherent to the data in an ordered and effective way. The data are a mixture of subsets of quantitative, categorical (closed questions) and frequency (open-ended) questions.

Such different kinds of variables are to be analysed by both different and complementary methods of analysis and to be divided up into two categories: symmetrical and non-symmetrical methods. Symmetrical methods are those where all variables are equally considered from the processing point of view making no distinctions between dependent and independent variables. Non-symmetrical techniques are those where a selected variable is considered as dependent on others.

The symmetrical methods will be composed of adequate exploratory multivariate techniques: particularly, factor methods complemented with cluster analysis. However, we do not perform a pure exploration; we start from previous hypotheses and assumptions taken from the corporate marketing literature as detailed before in this section.

The non-symmetrical methods, namely, partial least squares (PLS) path modelling and logit models, both feature the target variable 'propensity to buy' playing a privileged role and which has to be estimated. These methods complement and validate the structures previously observed, allowing us to extract interesting conclusions from the point of view of the improvement of the institutional image, related to the establishment of commercial policies in the corporate shop.

This paper is organized as follows. In Section 2, we describe briefly the characteristics of the survey from which the data are originated. In Section 3, we argue for the data mining methods to be used in the analysis. In Sections 4 and 5, the exact methods to be used for each case are justified and explained and the obtained results on the data are given. In Section 6, conclusions are given.

2. Case study: the data

In the context pointed out in Section 1, an on-line survey to collect the needed information on the desired subject has been set up. The survey is addressed to the members of the research and teaching staff, administrative staff and the students of the university.

Table 1. Technical characteristics of the on-line survey.

	Students	Administration staff	Research and teaching
Population	48,995	1128	3982
Sample size	2289	768	1499
Response (%)	547 (23.9)	444 (57.81)	754 (50.30)
Sampling error	0.042	0.036	0.032
Confidence level	0.95	0.95	0.95

The access to filling in the survey was possible only by invitation, and there was a period of 1 month for doing so. In order to get a good response rate such invitation was sent in advance by surface mail followed by an invitation e-mail including a link to the on-line questionnaire, plus a later reminder. This procedure limits the adverse effects of on-line surveys on response rates [24], which include technical, anonymity and spam consideration issues [44], when compared with mail surveys.

Table 1 contains the sampling technical characteristics. The number of invitations or sample size was fixed per stratum and chosen in order to keep a maximum margin of error of 2% for a 95% confidence level in the estimation of the success ratios for binary items, assuming $p = 0.5$. The sampling was proportionally random according to gender, campus, age (for non-students) and degree cycle (for students) and the results were encouraging, with a global response of 1742 effective answers (around 40%), though not equally distributed.

The response rate for the students is not very low in absolute terms, but it is so relative to the other university strata. We think that this is due to a persistent preference of the students for e-mail accounts in the *cloud* (Hotmail, Yahoo or Gmail, for instance) over the university domain accounts (often unchecked for long periods), which were those used for e-mail invitations to the questionnaire.

The relationship between response rate and representativeness is not always clear. In any case, we make the general assumption in this work that respondents' opinions are similar to non-respondents' based on the considerations mentioned above and the fact that this survey does not refer to sensitive personal attitudes as those referred to, for example, sex, drugs or personal income would do.

Table 2 shows us the three most relevant questions included in the sample. They are followed by 26 questions on the valuation (from 1 to 4) of the same number of products (shown in a photo), the valuation (from 1 to 7) of eight proposed desirable characteristics of products (sober, traditional,

Table 2. Questions, possible answers and type of variables.

Question	Possible answer	Type
I am satisfied about being a member of this university	1 = completely unsatisfied 2 = rather unsatisfied 3 = neither satisfied or unsatisfied 4 = rather satisfied 5 = very satisfied	Categorical
Would you be interested in buying a product featuring the university logo?	1 = yes 2 = no	Binary
$\longrightarrow$ Could you state why?	(*open answer*)	Textual data
Gender	Male, female	Categorical
Age	1 = 17–22, 2 = 23–29, 3 = 30–44, 4 = +44	Categorical
Campus	Araba, Bizkaia, Gipuzkoa	Categorical
Status	Student, administration, teaching–research	Categorical

stylish, modern, practical, artistic, daring and original) and finally some personal information (gender, age, status and campus – up to three possible), as also shown in Table 2.

We were particularly interested in getting information on preferences on the 26 products, so we intentionally dropped the middle point in product valuation questions. These questions are those which we analyse by means of a knowledge extraction process using data mining techniques. The computer programs used have been varied, including SPAD, SPSS and R.

3. Data mining: symmetrical and non-symmetrical methods

Symmetrical methods (Section 1) are core techniques within data mining applications [20,30,51]. They are mostly multidimensional exploratory statistical methods allowing for visualizations based on geometric and algebraic principles under the control of flexible and robust inference methods, also known as geometric data analysis [26]. In classic inference statistics, the scientific general hypotheses on a phenomenon precede the observation or statistical experimentation stage. However, in multidimensional exploratory statistics, the technical hypotheses intervene in the practice of statistical methods.

The work on qualitative variables, frequent in survey analysis, is by nature more complex than that of the continuous numeric variables, supported by the normal distribution and by criteria such as maximum likelihood or least squares. However, the advances in computing have fuelled the development of, on the one hand, factor and segmentation or classification analysis and, on the other hand, log-linear, discriminant and logistic models, which we refer to as non-symmetrical methods.

According to [30], there are two kinds of approximations to multidimensional statistics: descriptive or exploratory (such as non-supervised approximations in learning theory) and confirmatory and inference approximations (such as supervised ones); they are complementary. In [20], a finer partition of data mining types of methods is made taking into account the different practitioners' objectives; however, it is not contradictory to the one considered here.

Exploratory methods provide a first idea of the nature of the relationships between the variables preceding an explanatory and predictive stage(s). However, these approximations are not always easily distinguished or identified. Pure exploration is of little plausibility due to the existence of prior knowledge about the phenomenon subject to analysis. That prior knowledge can be seen as general hypotheses forming users' predictions and expectations [see 30]. Thus, in this case, we have considered three starting hypotheses, as stated in Section 1.

Confirmatory, inference or predictive methods depending on the classification used are of greater appeal to many data analysts. The spirit of model building allows for the exhibition of summary equations that are easy to understand, containing parameters with useful interpretations, where theoretical hypotheses can be often easily tested in a comprehensive way and, particularly, given their non-symmetrical nature, where predictions can be obtained naturally. Possible model failures can often be overcome by considering departures from the initial assumptions or different specifications.

The data available and the questions to be answered in this work do not strictly need predictive techniques in order to be successfully analysed. We are just interested in a large multidimensional sample understanding. Inference techniques do not possess a definitive advantage in this case, so we have decided to use additionally exploratory methods in order to extract the most relevant information from this data set.

4. Symmetrical methods: exploratory multivariate techniques

The data available, as explained in Section 2, are composed of closed and open-ended questions leading to the consideration of different kinds of variables. The different nature of these needs the use of different statistical treatment techniques. In this section, however, all variables will be treated in a symmetrical way in the sense that no variable is considered as dependent on any other.

The closed questions lead to quantitative or categorical individuals × variables tables. Depending upon which kinds of variables are to be considered as active (i.e. those variables for which some dimensionality reduction technique will extract the main factors or components), we can consider, as principal axis methods, a principal component analysis (PCA) applied to quantitative variables [31], a multiple correspondence analysis (MCA) applied to categorical variables (see, e.g. [27] or [31]), or any of the variants.[1]

The open-ended questions lead to an individuals × words frequency table that should be treated by two-way CA in the framework of text analysis (see, e.g. [33] or [38]). After performing a factor method, we want to identify the various patterns with respect to attitudes and opinions about the university corporate products from the answers given to the closed and open-ended questions corresponding to these topics.

We follow a two-step approach as proposed in, for example [29] or [33], which combines principal axis and clustering methods (see also, e.g. [9,10,17]). A principal axis method is used as a pre-processing step, and clustering considers a number of the first principal coordinate vectors accounting for a significant percentage of the initial variability. Clustering is performed over the principal axes, which are, in fact, quantitative variables, and then a suitable algorithm can be used [40]. The elimination of the last axes, far from being a drawback, acts as a filter for the random fluctuations which could mask important features. Finally, the clusters are described by the significant over- and under-represented characteristics of the gathered individuals.

4.1 *PCA of quantitative variables and clustering*

We first consider as active variables the scores given to questions asking for certain desirable characteristics of products (original, sober, . . .), which are measured on a seven-point scale and provide variables considered as quantitative. The idea behind the introduction of such questions is to consider desirable characteristics in the same way done in a semiometric questionnaire [32,42,46]. The respondents must rate these words according to a seven-point scale, the lowest level (mark = 1) coding a *most unpleasant feeling* about the word and the highest level (mark = 7) coding a *most pleasant feeling* about the word. The variables regarding personal characteristics, such as gender or age, are considered as supplementary variables (i.e. variables that are projected on the axes obtained with the active variables, but not considered for factor extraction), and the variables reflecting satisfaction with the institution and the interest in buying are also treated in this way. In order to assist in axis interpretation, test values (see, e.g. [30, p. 114]) are calculated. A test value for a given supplementary category on a given axis is obtained as a standardized normal t-test under the null hypothesis that this category taken at random is placed on the origin, thus not being characteristic for that given axis. Large test values in absolute terms imply that the category does characterize the axis of interest.

The first factor is a size factor which distinguishes persons who select higher scores for all or most of such desirable characteristics from those who select lower values. Those who give higher marks are also people who manifest a greater satisfaction, interest in buying and are over 44 years. The positive side of the second factor corresponds to higher scores given to *sober*, *traditional*, *stylish* and *artistic* features and to respondents over 44, teaching–research staff and males, whereas the negative side corresponds to higher scores given to daring, original and modern characteristics. Finally, the third factor highlights individuals scoring highly the term *practical*, who are mostly students and under 30.

After performing a hierarchical clustering on the PCA first five axes, using Ward's criterion, three clusters are produced. For each quantitative variable, the global (full sample) and cluster means are compared using a statistic taking the form of the classic t-test (see, e.g. [30, pp. 291–292]). Among the elements characterizing a cluster, it is possible to find categories which have not been used for the formation of this cluster, but which can be useful for its description under the same principle of supplementary categories in factor analysis. A supplementary category can be considered relevant in a class if its presence in this class is significantly higher than that in the whole set of individuals. This level of significance is determined by a *test value* associated with a comparison of the proportion of individuals of that category in each cluster compared to the proportion in the whole sample in the framework of a hypergeometric law (see, e.g. [33, pp. 130–134]). Thus, for every cluster, we associate each category with a significance level (or p-value) that allows the categories to be ranked from the most over-represented to the most under-represented.

The first cluster (46%) corresponds exactly to individuals on the positive side of the first factor (over 44, fully satisfied, declaring buying interest, high scores to all characteristics). The second one (31%) corresponds to individuals who rank high the characteristics of original, daring, modern and practical and who are students, under 30, neither satisfied nor dissatisfied and who do not manifest buying interest. This is a group which might be attracted to the first group, composed of potential buyers, by improving the characteristics of the products in the way that they consider important, as given by the semiometric characteristics. The last cluster (23%) give low scores to most of the characteristics and manifest no interest in buying while being indifferent towards the institution. This group seems a difficult one to reach in market research terms.

The first analysis provides three main directions of variability by means of a PCA. The clustering over the main factors helps to group individuals into homogeneous families, where each cluster represents a market segment with different characteristics and reachable through different marketing strategies or perhaps other products not considered here.

4.2 *MCA of categorical variables and clustering*

For a second factor method, we choose the categorical variables referring to the valuation of the 26 articles (after seeing a displayed photo) on a scale 1–4 as the active variables of a MCA. As supplementary variables, we choose the product characteristics, the satisfaction variable, the intention to buy and the individuals' personal data.

Figure 2 shows the projection of the active categories on the MCA main plane. It shows how the first factor represents a global propensity to buy, roughly ordering categories from left to right with respect to their probability to buy, from lower to higher. The plane shows a typical Guttman effect [14, pp. 147–148] with the second factor reflecting differences between extreme and centred opinions.

With respect to the projections of the most relevant supplementary categories, it is shown in Figure 3 that the first factor is positively related to the satisfaction with the institution and the declared propensity to buy. This shows the relationship of these variables with the overall propensity to buy individually the 26 products.

A mixed clustering in three steps (see [31, pp. 130–145], based on [54]) is carried out on eight MCA first principal axes, accounting for a significant proportion of the inertia contained in the data table. This process starts by choosing a partition in a number of clusters with random initial centres and then updating those centres calculating the centroids of the groups of individuals nearest to the centres (K-means algorithm); the process is repeated until the clusters are stable. We reduce further the number of clusters by means of a hierarchical algorithm (Ward's method) and refine the resulting partition through a consolidation step with re-assignment (testing moving centres with convergence achieved in seven iterations). This results in a partition of six clusters with an inter-inertia over total inertia ratio of 55.62%. The positions of the final cluster centres on the plane are shown in Figure 4, and they follow the pattern set by the active categories on the same plane.

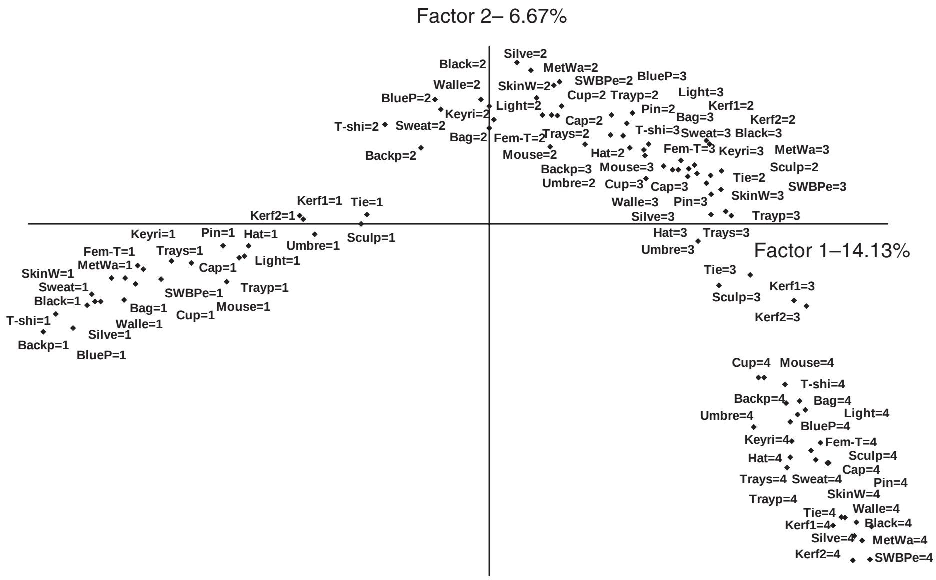

Figure 2. MCA: active categories on plane (1, 2).

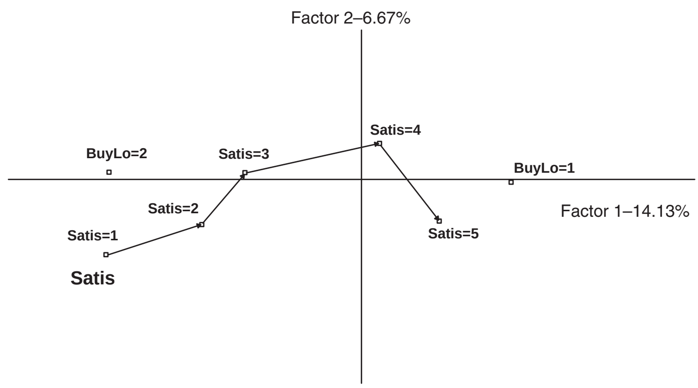

Figure 3. MCA: supplementary categories on plane (1, 2).

The partition description is as follows. Cluster 1 (accounting for 15.73% of the full sample) contains those who would buy prior to watching any product and who answer that they are *very likely to buy* many products; in addition, they are over 44, fully satisfied, females, members of the teaching and research staff and give high scores to stylish and traditional product characteristics. Cluster 2 (17.91% of sample) is formed by those who are likely to buy; they are over 44, would buy prior to watching the product and rank highly stylish, traditional and sober features. Cluster 3 (17.74%) is composed of those who manifest that they are unlikely to buy sober and stylish products (metallic) but are likely to buy original, modern and practical products (textiles and bags). Cluster 4 (12.80%) groups individuals unlikely to buy anything with low scores for stylish products. Cluster 5 (18.66%) is composed of individuals very unlikely to buy, aged between 18 and 22, students, from Gipuzkoa campus, neither satisfied nor dissatisfied and with low scores on traditional, sober or stylish products. Finally, cluster 6 (17.16%) consists of those who are very unlikely to buy, between 30 and 44, males and with low marks for all characteristics of the products.

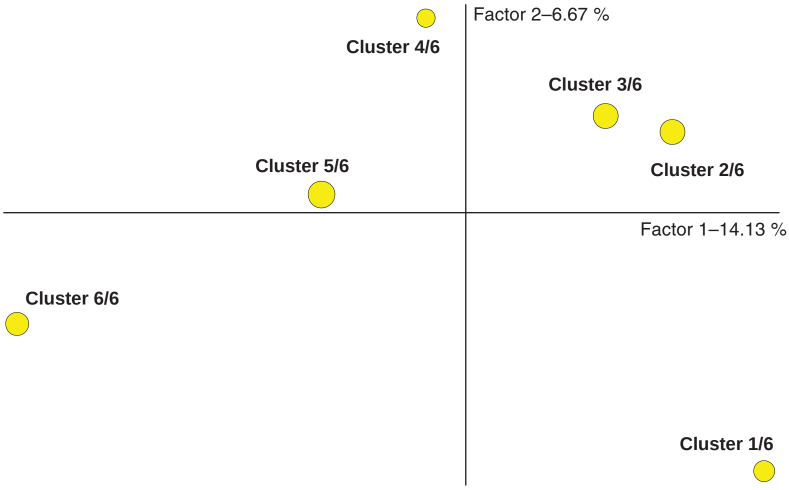

Figure 4. Clustering on MCA factors. Cluster centres and relative sizes represented by circle diameters.

This MCA confirms the tight relationship between the interest to buy articles featuring the logo (before visualization), the degree of satisfaction about the institution and the scores given to the proposed desirable characteristics of the products. The clustering process shows marketing implications on the buyers' and non-buyers' personal characteristics and on the articles, as some are perceived as stylish, traditional and sober and others as modern, original and practical. Furthermore, the parabolic path appearing in Figure 2 is similar to those shown in Figures 3 and 4, reinforcing its interpretation as an indicator of the propensity to buy the displayed products.

4.3 *Textual CA*

Survey questionnaires frequently include closed and open-ended questions about the same topic. The closed questions lead to quantitative or categorical individuals × variables tables and the open-ended questions lead to individuals × words frequency tables, also called lexical tables. Table 2 shows the closed question: *Would you be interested in buying a product featuring the university logo?* and the following open-ended question: *Could you state why?* The introduction of an open question such as this has the objective of enriching the data beyond the classic information available from closed questions. In addition, at the UPV/EHU, many members are bilingual and respondents could answer in any of the two official languages, Basque or Spanish, at their convenience. A previous work [7] shows no differences between Basque and Spanish responses. Only the 1243 respondents having answered the open-ended question in Spanish are kept to perform this text analysis. Other questions are projected on the factor axes as supplementary variables. The question referred to personal overall satisfaction with respect to the university has been recoded due to the small frequencies obtained in the three lowest categories resulting in the next three-point scale variable: 1 = completely unsatisfied, rather unsatisfied or neither satisfied nor unsatisfied, 2 = rather satisfied and 3 = very satisfied. Thus, the total number of categories of the variables in Table 2 is reduced from 19 to 17.

After a process of lemmatization, that is, the grouping of words arising from the different inflections of one lemma, we have decided to keep not only words but also repeated segments (see, e.g. [33, pp. 35–38]) and use the same frequency threshold for both lexical units, equal to 15 appearances. Thus, we obtain 239 total lexical units in the text corpus, resulting in a 1243 individuals × 239 lexical units frequency table.

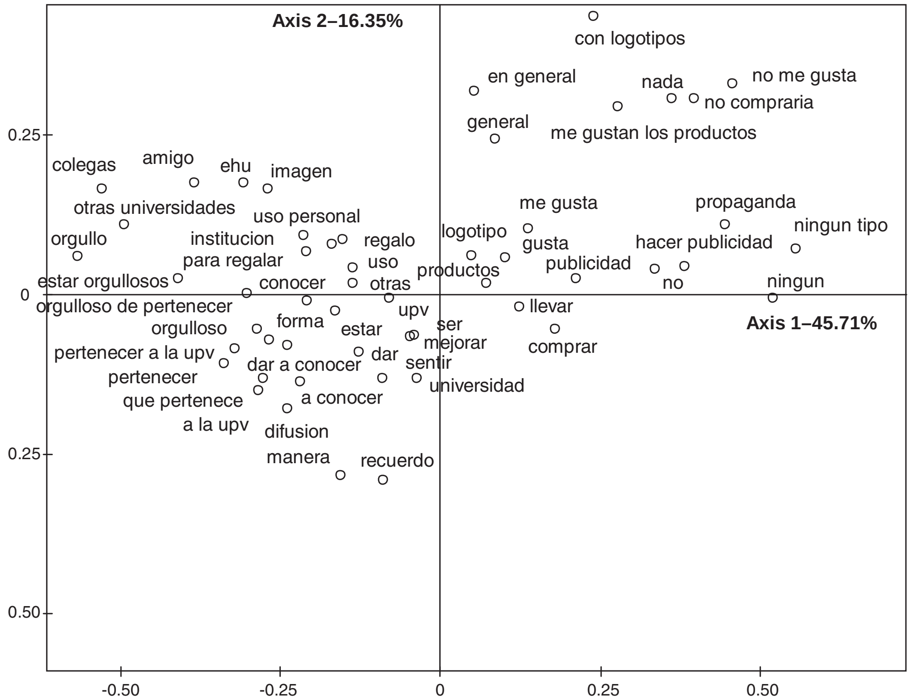

Figure 5. CA-grouped table: active words and segments on plane (1, 2).

Individuals' isolated responses are very sparse (see [28]). In order to diminish the effect of too many zeroes in the frequency table, we proceed to treat grouped responses. A new table is formed containing words used by categories of closed questions and personal characteristics, which are given in Table 2. Individuals are thus grouped into 17 categories, resulting in a 17 categories × 239 lexical units frequency table.

Figure 5 shows the result of the textual CA of the grouped table. It shows how the first factor reflects different attitudes aligning words and segments from right to left expressing negative ideas on the right side: *no me gusta (I don't like), no compraría (I wouldn't buy), ningún tipo (none),* and positive ones on the left side: *pertenecer a la upv (be a member of the UPV/EHU) orgulloso (proud), orgulloso de pertenecer (proud of being a member)*, all related to the idea of being very satisfied and proud of their university.

With respect to the projections of the 17 active categories, they are mostly well represented in the main plane. It is shown in Figure 6 that the first factor is negatively related to the satisfaction with the institution and the declared propensity to buy (before product visualization). Thus, we can relate the negative textual expressions with a low degree of satisfaction about university membership and with the negative propensity to buy, whereas the positive textual expressions are associated with high satisfaction and positive propensity to buy. On the negative side, we can indeed see those who are proud of being members of the university and, at the same time, are very satisfied and manifest interest in buying products with logo; with respect to their personal characteristics, they are respondents over 44 and teaching–research staff, while the positive side corresponds to non-buyers, little satisfied, under 30 and student-type respondents.

The results obtained from the text analysis point of view show the same structure previously observed except for the categories of the gender variable. Their position is neutral (Figure 6) and both their contributions to the axis inertia and their poor quality of representation lead to the conclusion that they are of little relevance when considering text expressions. It is shown

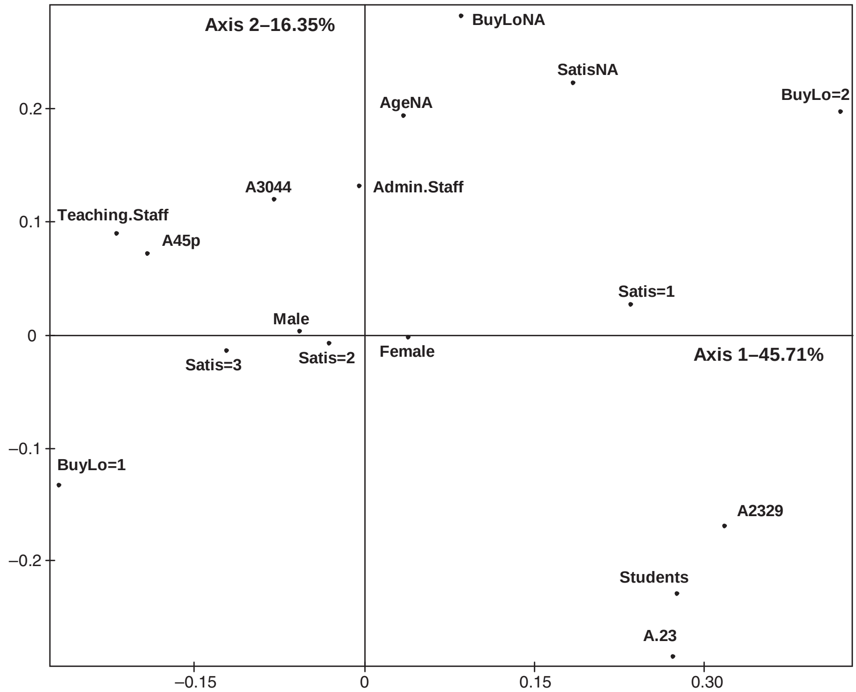

Figure 6. CA-grouped table: respondents' active categories on plane (1, 2).

that the text analysis reinforces the relationship between satisfaction and the prior willingness to purchase articles featuring the logo for members of the institution with the same characteristics already obtained in the previous analyses. It should be noted that this analysis is not directly comparable due to the fact that only the text responses of the 1243 individuals answering in Spanish have been considered. However, we have not found significant differences in the answers in the two languages, apart from expressions that are not directly comparable due to language structural differences. In any case, we think that it is interesting to see how the higher frequency words or segments obtained from an open question can be plotted in the same plane of the categories obtained from the closed questions, thus integrating the open question in the analysis. This procedure advocates the consideration of open questions for survey analysis, which are often discarded in quantitative analysis or simply treated in a qualitative, even subjective way.

5. Non-symmetrical methods: regression-related techniques

In this section, we consider only the variables obtained from the closed questions of the survey. In this case, we choose the modelling of the relationship among the variables in such a way that a variable is depending upon others and which is related somewhat to regression analysis. This asymmetrical treatment of the variables makes us use the non-symmetrical term for the methods exposed here.

After the analyses presented in Section 4, a latent structure that we name *propensity to buy* becomes apparent and is amenable to modelling and treatment by different alternatives of non-symmetrical modelling. We select, among them, PLS path modelling and logit models.

5.1 *PLS path modelling*

PLS path modelling (see, e.g. [8,53] or [47]) or the PLS approach to path models is a technique based on the relationships between latent variables (LVs) in a regression framework, where such variables are constructed with underlying manifest variables (MVs). It is a procedure similar to structural equation modelling (SEM) with respect to the interest in constructing LVs and establishing relationships between them. Differences are both theoretical and empirical, though not always excessive in practical terms. On the theoretical side, SEM relies on heavy distributional assumptions, much less needed for the PLS path modelling. On the practical side, SEM normally requires many available observations (at least a few hundreds) compared with PLS. Comparisons are given in [12,23]. In addition, there are occasional convergence difficulties in SEM, much less frequent in the case of PLS (further details and a comparison example are given in [47]). Often, the numerical results are not very different, being more optimistic in terms of fit those obtained by SEM. In this work, we have chosen the PLS path model to analyse the variables obtained from the questions of the survey.

We are going to construct a global propensity to buy using all MVs defined by the respondents' valuation of the 26 products shown in the survey questionnaire, resulting in a global LV. At the same time, we want unidimensional partial propensities to buy groups of products, and in order to be auto-selected by the data, we do not want to impose any additional structure other than that imposed by the model itself. These will also have the form of LVs and will be sought with a previous PCA of the valuations of all the 26 products displayed in the survey.

In the way explained above, eight groups of products are obtained providing the same number of partial LVs, $\xi_1, \ldots, \xi_8$. The composition of such groups, that is, product valuation variables considered here as MVs, is given in Table 3, along with their labels.

Selecting all products valuations, we construct the global propensity to buy. Finally, we formulate the external model given in Equation (1):

$$\xi = \sum_{j=1}^{8} \beta_j \xi_j + \nu. \tag{1}$$

Figure 7 shows the path model specified. The numbers are correlations and show relatively high values between the partial LVs and the global one. We can also see the pairwise correlations between individual MVs and the LVs.

The actual estimates of the external model parameters are given in Equation (2). These show higher values for textiles, bags and pens products groups (corresponding to the partial LVs ξ_3, ξ_4 and ξ_8, see Table 3), which are those with a higher acceptability among the respondents:

$$\begin{aligned} E(\xi) = {} & 0.0865 * \text{umbh} + 0.1335 * \text{tie} + 0.2041 * \text{textiles} + 0.2114 * \text{bag} \\ & + 0.1791 * \text{wat} + 0.1292 * \text{mous} + 0.0881 * \text{scul} + 0.2322 * \text{pens}. \end{aligned} \tag{2}$$

Table 3. Groups of products to be considered as LV.

Label	LV	Products
umbh	ξ_1	umbrella, hat
tie	ξ_2	tie, kerchief no. 1, kerchief no. 2
textiles	ξ_3	T-shirt, T-shirt-V, sweater, cap
bag	ξ_4	plastic tray, leather tray, backpack, bag, cup
wat	ξ_5	leather-strapped watch, metallic-strapped watch, wallet
mous	ξ_6	key ring, lighter, mousepad
scul	ξ_7	pin, sculpture
pens	ξ_8	blue pen, black pen, silver pen, silver pen in wooden case

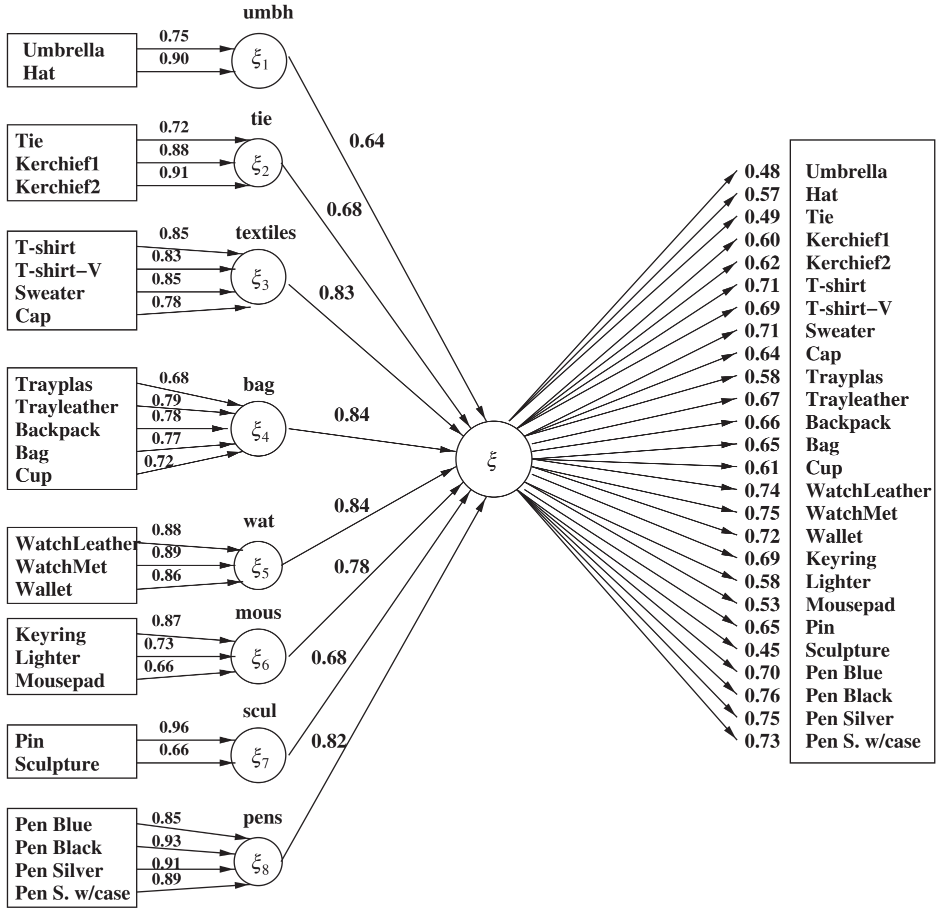

Figure 7. PLS path diagram for products to be sold at the university shop.

In order to get potential buyers' characterization (similar to the projection of supplementary variables in a factor analysis), we perform a regression on the desirable characteristics of the products and the respondents' personal characteristics. This is actually a principal component regression, since the desirable characteristics are highly correlated, selecting two main components out of the seven original variables. The results are given in Equation (3):

$$\begin{aligned} E(\xi) = {} & -0.85 + 0.07 * \text{F1 (original, daring, practical, artistic, modern)} \\ & + 0.11 * \text{F2 (traditional, sober, stylish)} - 0.25 * \text{male} \\ & + 0.15 * \text{satisfied} + 0.26 * \text{very satisfied} + 0.07 * \text{age(+44)} \\ & + 0.06 * \text{teaching–research staff} - 0.10 * \text{higher education} \\ & + 1.18 * \text{overall propensity to buy a logo product} \\ & + 0.14 * \text{campus: Araba} + 0.12 * \text{campus: Bizkaia}, \qquad (3) \\ R^2 = {} & 0.4848. \end{aligned}$$

All parameters whose estimates are shown are significant at the 5% level, using both bootstrap confidence intervals and usual t-test statistics. These estimates show that the individuals most satisfied with the university are more likely to buy. This is also the case for those who have a prior intention to buy, females, members of teaching and research staff, older age and those proceeding from the campuses of Bizkaia and Araba when compared with those from Gipuzkoa. With respect to product characteristics, the respondents marking as more important the terms traditional, sober and stylish are more likely to buy than individuals giving more importance to aspects such as modern and practical.

With respect to the image of the university, a significant higher propensity to buy products featuring the university logo in members with a high regard of their university and of its image is observed.

5.2 *Logit models*

Finally, we apply a discrete choice model, a logit model (see, e.g. [15,22,34]), to the survey data in order to explain the individuals' possible purchase decisions on the products featuring the UPV/EHU logo offered in the corporate shop.

In this case, the goal of the analysis is to establish a model allowing the estimation of the probability of buying logo articles for the university members depending upon their personal characteristics, preferences about product characteristics and the variable referring to the level of satisfaction about the institution. The interest is in the detection of the characteristics of individuals most prone to buy these kinds of products.

According to logit modelling, the probability that the ith individual decides to buy a product (by choosing option 1) is obtained through the expression:

$$P(y_i = 1) = F\left(\sum_j x_{ij}\beta_j\right) = \frac{e^{\sum_j x_{ij}\beta_j}}{1 + e^{\sum_j x_{ij}\beta_j}}, \quad (4)$$

which, after estimating the regression coefficients β_j, allows for the estimation of the probability that a given individual does buy or not.

Given that any individual had the choice for the likelihood of buying measured on a four-point scale in the survey, from 1 to 4, in order to define the binary variable *buy* or *don't buy*, we consider, respectively, if he or she considers the purchase to be likely or very likely (3 and 4) or unlikely or very unlikely (1 and 2). In addition, to establish the variable global propensity to buy products with the university logo, we consider that this does make sense only if the purchase is likely or very likely for a minimum amount of the articles available (as certainly one is not enough). We have considered this minimum to be at least the 25% of the displayed articles, that is, seven products.

The variables considered here are exactly those mentioned in Section 5.1 (PLS path model), so the desirable characteristics of the products are highly correlated. We have overcome this issue by substituting them with their two PCA factors (after performing a varimax rotation). We end up with the estimated model given in Table 4.

This model allows us to extract the following conclusions. With respect to the strata, both students and academic staff (mostly these) show a greater willingness to buy the products. In fact, this result reinforces the idea appearing earlier that administrative staff is usually less involved institutionally. Gender is also a relevant variable, with females being readier to buy such products. Regarding product characteristics, those who prefer a stylish line are more willing to buy. This result indicates that, generally, respondents think that the products shown in the survey have a classic appearance, given that those more eager to buy them are just those who like this feature most. When focusing on the variables referring to the degree of satisfaction with respect to the institution, we obtain that the probability increases with the degree of satisfaction. So, those with the highest probability to buy are the most satisfied, followed by the just satisfied. Finally, the initial or prior buying interest is the most important variable.

Table 4. Logit model estimation for buying probability.

Variable	Coefficient	Standard deviation	Significance level
Constant	−1.298	0.211	0.000
Male	−0.794	0.158	0.000
Student	0.537	0.199	0.007
Teaching and research	0.584	0.200	0.003
Buying initial interest	2.979	0.156	0.000
Satisfied	0.367	0.183	0.045
Very satisfied	0.710	0.210	0.001
Stylish	0.339	0.078	0.000

Nagelkerke $R^2 = 0.502$ (see [39]).

6. Conclusions

This work is part of an ambitious project of a Spanish public university aimed at the improvement and dissemination of its corporate image, particularly, among its members (teaching, research and administrative staff and students), as an important step to integration in the new EHEA.

One of the elements of the project is the establishment of a university shop in order to improve the university image. As a previous step, the convenience of surveying the university members was considered, targeting individuals' perceptions on different issues related to the university. The survey was carried out and analysed aiming at different objectives. Only two have been selected for this work: on the one side, learning about members' perception on the institution and, on the other side, analysing the chances for success of a shop selling products featuring the university logo. Given the diversity of both the information obtained and the objectives, there are different appropriate analysis techniques.

Each technique used leads to specific, though related, conclusions, given its different objectives. The symmetrical methods (PCA, MCA and text analysis) combined with the cluster analysis help to learn what is contained in the data, including relationships and clusterings of similar individuals. On the other hand, non-symmetrical methods such as PLS or logit regressions allow for modelling individuals' global and partial (group) behaviour using inference tools to select a better model with a good fit to the data.

The methods exposed above extract some facts from these particular data consistently in an ordered and effective way. This allows, on the one hand, confirmation of the suitability of these methodologies and, on the other hand, the validation of the working hypotheses which underlie our work. Thus, the important relationship existing between the intention to buy articles displaying the university logo and the degree of satisfaction on its membership has been confirmed. The differences existing among the attitudes showed by the three considered strata towards the offered articles also emerge clearly. The teaching and research staff shows the highest satisfaction degree and intention to buy. We think that as long as they are more satisfied and prone to buy the articles, they should also be the best contributors to the university revamping policy.

In this sense, and more precisely, the results obtained show that, as expected, the willingness to buy articles with the institutional logo reflects to a great extent an individual's identification with the institution he or she is related to. When a person buys an article with the logo, he or she is not simply acquiring the product, but also the added value of an institutional article. The decision to buy involves a feeling of pride regarding the institution, as it is always possible to buy the same or similar products without the logo. The measure of such degree of identification is in the prior interest to buy articles featuring the logo, even before actually seeing the products. This effect is further reinforced by the testimony about the degree of satisfaction with the institution.

On the other hand, it is also clear which are the general characteristics of the articles shown (traditional, stylish, etc.) and the sort of characteristics of possible successful articles not covered in current product line (practical, original or modern). It seems that a better, more modern, design is needed to reach other market segments. The marketing implications obtained have been somewhat conditioned upon the actual articles displayed with photographs in the on-line questionnaire. It has been observed that many have been perceived as stylish and traditional (generally of a metallic aspect) and hold little appeal for younger respondents. As a general issue, this work concludes with the recommendation of the promotion of articles with the characteristics mentioned above and, particularly, belonging to the groups of textiles, bags and desktop articles which would yield a better acceptance for this target public in opening of the university gift shop.

As a result of the previous recommendation, the university has opened the shop offering a more attractive product line. Customers' acceptance has been satisfactory for these products. We think that it would be interesting at this point to perform a new survey similar to this one but considering also sales information. We would be thus able to evaluate if the image improvement has been actually produced.

As a summary, it can be said that these data mining techniques yield useful directions for the university marketing policy, regarding the corporate shop. Two complementary approximations, including symmetrical and non-symmetrical methods, have been successfully used, thus reinforcing the confidence on the results and the implications obtained from the marketing point of view.

Acknowledgements

Financial support from Grupos de investigación del sistema universitario vasco IT-313-07 and IT-321-07 is gratefully acknowledged. We are grateful to the reviewers whose comments have undoubtedly contributed to the improvement of the final version of this paper. We would also like to thank our colleague Fernando Tusell.

Note

1. The variants of MCA are considered to be the methods of joint correspondence analysis (CA) [16,18], homogeneity analysis or nonlinear [14] and dual scaling [41]. A comparative analysis of these possibilities can be found in [48] and a recent reference providing the state of the art can be found in [19].

References

[1] B.E. Ashforth and F. Mael, *Social identity theory and the organization*, Acad. Manage. Rev. 14(1) (1989), pp. 20–39.
[2] M.J. Baker and J.M.T. Balmer, *Visual identity: Trappings of substance*, Eur. J. Mark. 31(5/6) (1997), pp. 366–382.
[3] J.M.T. Balmer, *Corporate identity and the advent of corporate marketing*, J. Mark. Manage. 14(8) (1998), pp. 963–996.
[4] J.M.T. Balmer and S.A. Greyser, *Revealing the Corporation, Perspectives on Identity, Image, Reputation, Corporate Branding and Corporate-Level Marketing*, Routledge, London, 2003.
[5] H. Barich and V. Srinivasan, *Prioritizing marketing image goals under resource constraints*, Sloan Manage. Rev. 35 (1993), pp. 69–76.
[6] J.P. Bean and R. Bradley, *Understanding the satisfaction–performance relationship for college students*, J. Higher Educ. 57 (1986), pp. 293–312.
[7] M. Bécue-Bertaut, K. Fernández-Aguirre, and J.I. Modroño-Herran, *Analysis of a mixture of closed and open-ended questions in the case of a multilingual survey*, in *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, C.H. Skiadas, ed., Birkhäuser, Boston, MA, 2009, pp. 23–34.
[8] W.W. Chin, *The partial least squares approach to structural equation modelling*, in *Modern Methods for Business Research*, G.A. Marcoulides, ed., Lawrence Erlbaum Associates, London, 1998, pp. 295–336.
[9] K. Fernández-Aguirre, I. Gallastegui-Zulaica, J.I. Modroño-Herrán, and V. Nuñez-Antón, *Clients' characteristics and marketing of products: Some evidence from a financial institution*, Int. J. Bank Mark. 21(5) (2003), pp. 243–255.

[10] K. Fernández-Aguirre, M.I. Landaluce, A. Martín, and J.I. Modroño-Herrán, *Data mining of an on-line survey. A market research application*, in *Data Analysis, Machine Learning and Applications. Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation*, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, eds., Springer-Verlag, Berlin, 2008, pp. 183–191.
[11] C. Fombrun, *Reputation Realising Value from the Corporate Image*, Harvard Business School Press, Boston, MA, 1996.
[12] C. Fornell and F. Bookstein, *Two structural equation models: LISREL and PLS applied to consumer exit-voice theory*, J. Mark. Res. 19 (1982), pp. 440–452.
[13] J.B. Fuller, B. Hester, K. Barnett, and L.F.C. Relyea, *Perceived organizational support and perceived external prestige: Predicting organizational attachment for university faculty, staff and administrators*, J. Soc. Psychol. 146(3) (2006), pp. 327–347.
[14] A. Gifi, *Non Linear Multivariate Analysis*, Wiley, Chichester, 1990.
[15] C. Gourieroux, *Econometrics of Qualitative Dependent Variables*, Cambridge University Press, Cambridge, 2000.
[16] M.J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.
[17] M.J. Greenacre, Graphical analysis of readership data, ‘ismg correspondence analysis and clustering, *Proceedings of the American Statistical Association's Statistical Graphics Section, 1987 annual meeting in San Francisco*, 1988, pp. 40–48.
[18] M.J. Greenacre, *Correspondence Analysis in Practice*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL, 2007.
[19] M.J. Greenacre and J. Blasius (eds.), *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC, London, 2006.
[20] D. Hand, H. Mannila, and P. Smyt, *Principles of Data Mining*, Adaptive Computation and Machine Learning Series, MIT Press, Cambridge, 2001.
[21] O. Herrbach and K. Mignonac. *How organisational image affects employee attitudes*, Hum. Res. Manage. J. 14 (2004), pp. 76–88.
[22] D.R. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley & Sons Inc, New York, 2000.
[23] K.G. Jöreskog and H. Wold, *The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects*, in *Systems Under Indirect Observation*, Part I, K.G. Jöreskog and H. Wold, eds., North Holland, Amsterdam, 1982, pp. 263–270.
[24] M.D. Kaplowitz, T.D. Hadlock, and R. Levine, *A comparison of web and mail survey response rates*, Public Opin. Q. 68(1) (2004), pp. 94–101.
[25] S.H. Kennedy, *Nurturing corporate images: Total communications or ego trip?* Eur. J. Mark. 11(3) (1977), pp. 24–33.
[26] B. Le Roux and H. Rouanet, *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*, Kluwer Academic Publishers, Dordrecht, 2004.
[27] B. Le Roux and H. Rouanet, *Multiple Correspondence Analysis*, Quantitative Applications in the Social Sciences Series, Vol. 163, SAGE Publications, London, 2010.
[28] L. Lebart, *Exploratory analysis of large sparse matrices, with application to textual data*, COMPSTAT, Physica Verlag, Vienna, 1982.
[29] L. Lebart, *Complementary use of correspondence analysis and cluster analysis*, in *Correspondence Analysis in the Social Sciences*, M.J. Greenacre and J. Blasius, eds., Academic Press, London, 1994.
[30] L. Lebart, A. Morineau, and M. Piron, *Statistique Exploratoire Multidimensionelle*, 4th ed., Dunod, Paris, 2006.
[31] L. Lebart, A. Morineau, and K. Warwick, *Multivariate Descriptive Statistical Analysis*, Wiley, New York, 1984.
[32] L. Lebart, M. Piron, and J.-F. Steiner, *La Sémiométrie*, Dunod, Paris, 2003.
[33] L. Lebart, A. Salem, and L. Berry, *Exploring Textual Data*, Kluwer Academic Publishers, New York, 1998.
[34] G.S. Maddala, *Limited-Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs, Cambridge University Press, New York, 1983.
[35] N. Markwich and C. Fill, *Toward a framework for managing corporate identity*, Eur. J. Mark. 31(5–6) (1997), pp. 396–409.
[36] T.C. Melewar and E. Jenkins, *Defining the corporate identity construct*, Corporate Reputation Rev. 5(1) (2002), pp. 76–91.
[37] T.C. Melewar and E. Karaosmanoglu, *Seven dimensions of corporate identity: A categorisation from the practitioners' perspectives*, Eur. J. Mark. 40(7/8) (2006), pp. 846–869.
[38] F. Murtagh, *Correspondence Analysis and Data Coding with Java and R*, Chapman and Hall/CRC, Boca Raton, FL, 2005.
[39] N.J.D. Nagelkerke, *A note on a general definition of the coefficient of determination*, Biometrika 78 (1991), pp. 691–692.
[40] J.-P. Nakache and J. Confais *Approche Pragmatique de la Classification*, Technip, Paris, 2005.
[41] S. Nishisato, *Analysis of Categorical Data*, University of Toronto Press, Toronto, ON, 1980.

[42] C.E. Osgood, *Cross-cultural comparability in attitude measurement via multilingual semantic differentials*, in *Current Studies in Social Psychology*, I. Steiner and M. Fishbein, eds., Holt, Rinehart and Winston, New York, 1965.

[43] T. Pitman, *Perceptions of academics and students as customers: A survey of administrative staff in higher education?* J. Higher Educ. Policy Manage. 22(2) (2000), pp. 165–176.

[44] S.J. Sills and C. Song, *Innovations in survey research: An application of web-based surveys*, Soc. Sci. Comput. Rev. 20 (2002), pp. 22–30.

[45] A. Smidts, A.Th.H. Pruyn, and C.B.M. van Riel, *The impact of employee communication and perceived external prestige on organizational identification*, Acad. Manage. J. 49(5) (2001), pp. 1051–1062.

[46] J.-F. Steiner and O. Auliard, *La Sémiometrie: un outil de validation des réponses*, in *Quality of Information in Sample Surveys*, L. Lebart, ed., Dunod, Paris, 1992.

[47] M. Tenenhaus, V. Esposito Vinzi, Y.-M. Chatelin, and C. Lauro, *PLS path modeling*, Comput. Stat. Data Anal. 48 (2005), pp. 159–205.

[48] M. Tenenhaus and F.W. Young, *An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data*, Psychometrika 21 (1985), pp. 91–119.

[49] J. Traverso, *Imagen interna de la Institución Universitaria. Modelo para el Personal de Administración y Servicios*, Revista de Economía y Empresa XXIII(54/55) (2005), pp. 95–112.

[50] J. Traverso and M.L. Román, *Análisis de la imagen interna de la institución universitaria. Construcción de un modelo para la obtención de la imagen bajo la perspectiva del alumno*, Investigaciones Europeas de Dirección y Economía de la Empresa 13(1) (2007), pp. 231–246.

[51] S. Tuffery, *Data Mining et Statistique Décisionnelle*, Texhnip, Paris, 2005.

[52] C.B.M. Van Riel and J.M.T. Balmer, *Corporate identity: The concept, its measurement and management*, Eur. J. Mark. 31(5/6) (1997), pp. 340–355.

[53] H. Wold, *Model construction and evaluation when theoretical knowledge is scarce: An example of the use of partial least squares*, Cahier 79.06 de Département d'econométrie, Faculté des Sciences Économiques et Sociales, Université de Genéve, Genéve, 1979.

[54] M.A. Wong, *A hybrid clustering method for identifying high density clusters*, J. Am. Statist. Assoc. 77(380) (1982), pp. 841–847.