International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:5, No:5, 2011

Marketing Segmentation of Students Willing to Study Abroad based on Cluster Analysis

Kamila Tislerova, Marta Zambochova

***Abstract*—**Market segmentation is one of the most fundamental strategic marketing concepts. The better the segment which is chosen for targeting by a particular organisation, the more successful the organisation is assumed to be in the marketplace. Also higher education institutions have to improve their marketing tools for attracting foreign students, particularly when demanding tuition fees. This contribution aims at demonstrating the proper usage of the cluster analysis for segmentation (represented by students´ willingness to study abroad) and also, based on large international survey, offers some practical marketing implications.

***Keywords*—**Market Segmentation; Students´ Preferences; Study Abroad; Cluster Analysis

I. Introduction

MARKET segmentation is one of the most fundamental strategic marketing concepts: grouping people (with the willingness, preferences, fears, purchasing power, etc.) according to their similarity in several dimensions related to the product under consideration.

From a marketing management viewpoint, market segmentation is the act of dividing a market into distinct groups who might be attracted to different products or services. This technique is widely accepted as one of the requirements for successful marketing. By dividing the market into relatively homogenous subgroups or target markets, both strategy formulation and tactical decision making can be more effective [6].

The basis for selecting the optimal market segment to target is the number of segmentation solutions resulting from partitioning empirical data. Therefore the quality of groupings management chooses from is crucial to organisational success and requires professional use of techniques to determine potentially useful market segments. Thus, the methodology applied when constructing or revealing [3] clusters from empiric survey data becomes a discriminating success factor and potential source of competitive advantage [1].

There are many techniques available for grouping individuals into market segments on the basis of multivariate survey information but clustering remains the most popular and most widely applied method [2].

Kamila Tislerova is with the University of Jan Evangelista Purkyne in Usti nad Labem, lecturer of Business Administration. E-mail: kamila.tislerova@ujep.cz

Marta Zambochova is with the University of Jan Evangelista Purkyne in Usti nad Labem, Czech Republic. Research field: Statistics. E-mail: marta.zambochova@ujep.cz

Cluster analysis, the most common method for market segmentation, is an iterative process that requires the researcher to make numerous decisions relating to the creation and interpretation of the clusters. The most important aspect of segmentation is the interpretation of the clusters for the usefulness of the solution [5].

II. Methodology

Both international surveying and the methodology of segments creation (cluster analysis) will be described in this section.

A. Research objectives

- To outline the potential of cluster analysis in market segmentation
- To discuss the concept of cluster analysis and basic ideas and algorithms
- To acquire a reasonable segments of foreign students (according to their willingness to study abroad)
- To formulate the practical implication for marketing activities of higher education institutions.

B. International survey

In the period of April – December 2010 a large international survey was carried out. The main purpose of the survey was to identify students' expectations and fears and to acquire enough data for high quality segmentation. Issues such as current study results, family income, influencers, the acceptable amount of tuition fees, willingness to learn a foreign language and many other factors were examined.

Questionnaires of 26 issues were distributed in several countries by Czech researchers with the help of foreign universities. Also many interviews were conducted in order to determine the foreign students' attitude and especially in order to find an explanation for certain behaviour patterns. In this paper only qualitative data are processed and interpreted.

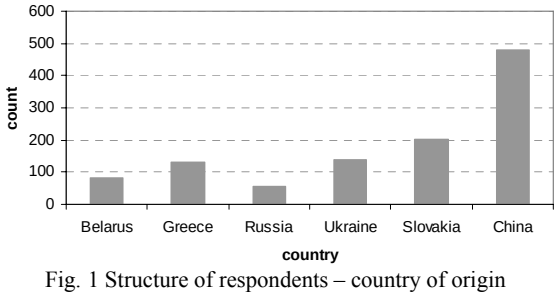

Fig. 1 Structure of respondents – country of origin

Students from six countries (see Fig. 1) were questioned in total number of 1093 respondents. The

average age of students is 20.87 years. 62% of the respondents are females.

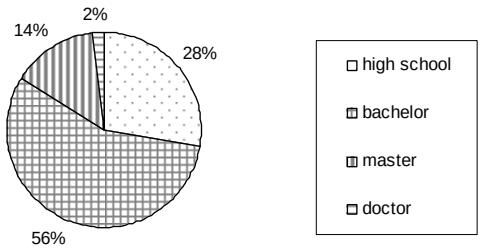

Fig. 2 Structure of respondents – level of study

In order to formalize the current field of study, four main streams were defined: Social Studies, Technical Sciences, Natural Sciences and Art. Respondents come from all of the fields proportionally. The structure of the current level of study is shown in Fig. 2

C. Statistics generally

This contribution deals with the classification of two basic ways of learning – namely supervised learning and unsupervised learning. In the first case the decision rules for assigning objects to the groups are created according to the training set.

The decision trees are representatives of this group of methods. In the second case the selected objective function due to its minimization divides objects into categories, so that the objects belonging to one category are more similar to each other than data from different categories. Cluster analysis belongs to this group. See [15].

D. Cluster Analysis

Cluster analysis, see [4], [8], deals with data objects similarity. It solves the set of objects splitting into several previously non-specified groups (clusters) so that the objects in the single clusters are the most similar to each other as possible and the objects outside of the different clusters should bear the most similarity. Cluster analysis can be realized by many different methods.

The statistical program systems usually include both the hierarchical algorithm result which is usually displayed in the form of a dendogram and non hierarchical iterative algorithm *k*-means and very often also a two-way combination. In the statistical system SPSS there is a two-step method implemented starting at the 11.5 version.

The choice of a hierarchical method was not suitable for this survey due to the relatively large number of subjects. The algorithm *k*-means is designed for the clustering of objects which are described with the use of quantitative variables and it was not the case in this research. The usage of this method would require pre-processing the data with the help of binarization; it means that each variable transfers into several binary variables (the variable of the value 0 and 1). The most suitable method for data proceeding in this survey seems to be the two-step method.

The principles of the two-step cluster method are described for example in [7]. This method uses the algorithm of BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), which is explained in more detail in [11], or [12]. The algorithm creates a so-called CF-tree, which is progressively fulfilled by incoming data. The advantage of this principle is that it goes through the data file only once. The disadvantage is the sensitivity of the entry data ordering.

CF-trees work with so-called CF-characteristic – Clustering Feature of the cluster. Data collected in CF-characteristic are sufficient for the calculation of centroids, inter-group proximity measures and compactness of clusters. This characteristic creates an organized triad of $CF = (N, LS, SS)$, where N means the number of objects in a cluster, LS represents a vector sum of all cluster objects and SS states these objects Square coordinates, e.g.

$$LS = \sum_{i=1}^{N} X_i \,,\; SS = \sum_{i=1}^{N} X_i^2 \,. \qquad (1)$$

CF-trees are highly balanced trees of two parameters. The first parameter is the threshold P and the second one is the branching factor (F, L). Each internal node of CF-tree applies in that it contains maximally F descents. The task of the internal nodes is to allow the finding of the proper leaf for new subject categorization. Each leaf contains maximally L entries. Every leaf node represents a cluster created by all the sub-clusters constituted by the single entries of the leaf. However, the threshold rule has to be valid for every leaf entry that the entry radius is smaller than the threshold P.

The clustering algorithm is realized in three main phases. In the first phase the CF-tree is created and the entering objects are progressively organized. In the second phase the CF-tree condensates and optimizes itself due to its threshold adjustment and with the help of the proper tree re-designing the outliers is eliminated. In the third phase the impact of entry data order sensitivity is minimized. The algorithm clusters together with the leave´s tops using the agglomerative hierarchy cluster algorithm.

E. Decision Trees

Various types of decision trees are widely used in data models. The decision trees can be regarded as the structures which recursively separate surveyed data according to certain decision criteria. The root represents the population file. The inner nodes demonstrate the sub-systems of the population set. The values of the dependent variable are explained in the tree leaves. Two types of decision trees have been used: the classification trees (every leaf contains a category) and regression trees (every leaf contains a constant – the estimation of a dependent variable).

The decision tree has been recursively created by the space division of independent variable values and has been based on searching for the question (splitting condition), which is best of all for dividing the surveyed data space into sub-sets, it means which one maximizes the splitting criterion. The splitting procedure is finished as soon as the cessation rule is reached. There are two possible ways to set up the quality of the generated tree: the system of training and test data and the other way is the cross validation.

A large number of algorithms were developed for the decision trees creation. CART, ID3, C4.5, AID, CHAID

and QUEST algorithms are the most frequent ones, see [8], [13] or [14]. This contribution treats three algorithm types implemented in the statistical system – CART, CHAID and QUEST.

Algorithm CART

This algorithm was originally described by its authors Breiman, Freidman, Olshen and Stone in 1984 in the article „Classification and Regression trees“. The algorithm (see [9], [10]) can be applicable in the case that there are one or more independent variables. These variables can be continuous or categorical (both ordinal and nominal). There is also one dependent variable, which can be also categorical (both ordinal and nominal) or continuous.

Because only YES/NO questions (condition of splitting) are permitted, the algorithm result can be composed only of the form of the binary tree (it means that every node is divided into two child nodes). In every algorithm step the algorithm goes through all the potential splitting with the help of all permissible values of all variables and the best solution is looked for. The increasing of data purity serves as the measurement. It means that one splitting is better than the other one if two more homogenous (according to independent variables) data files are acquired compared to another way of splitting. Algorithm splitting differs for classification trees and for regression trees.

The child node homogeneity is in the case of the classification trees measured by the impurity function $i(t)$. The maximal homogeneity of two newly built child nodes is constructed as the maximal purity reduction $\Delta i(t)$.

$$\Delta i(t) = i(t_r) - E(i(t_d)), \tag{2}$$

where t_r represents the parent node, t_d the child node. In order to set up the child node t_p, the probability of the child node P_p and the left child node t_l, the probability of the left child node P_l the expected value formula should be applied as follows:

$$\Delta i(t) = i(t_r) - P_l \cdot i(t_l) - P_p \cdot i(t_p). \tag{3}$$

For each node the CART algorithm solves the maximization problem for the $\Delta i(t)$ function going through all the potential splitting. The $\Delta i(t)$ function can be defined in different ways. The most frequent is the Gini index method.

The regression trees are used in the case when the dependent variable is not categorical. The algorithm looks for the best splitting based on the sum of variance minimizing in the terms of two newly built child nodes in this case. This algorithm works on the basis of the algorithm of minimizing the sum of squares.

CHAID Method

The method of CHAID (Chi-squared Automatic Interaction Detector) was developed in 1980 by G.V. Kass. This method, see [10], has arisen by the modification of the AID method for the categorical dependent variable. The non-binary trees can be regarded as a result of this modification. The method uses the chi-square test. The splitting algorithm is realized as follows: In terms of one leaf node the contingency table (sized $m \times k$) values of independent variable (m categories) is created. After that the pair of the category of independent variable predictor is found and the sub-table sized $2 \times k$ has the less important value of chi-square test. These two categories are merged. By this operation the new contingency table is created – sized $(m-1) \times k$.

The merge procedure is repeated until the significance of chi-square test declines under the pre-scribed value. Having reached this, the splitting procedure of one parent node to several child nodes has been finished. The process continues in this way for each of the leaf nodes until the insignificant result of the chi-square test is reached.

Algorithm QUEST

The method is described in an article from 1997 written by W.Y. Loh and Y.S. Shih, "Split Selection Methods for Classification Trees". The algorithm is applicable only for a nominal dependent variable. Similarly, as in the case of CART, binary trees are created. Unlike the CART method, the QUEST method realizes during the tree building the separate selection of variables for node splitting and the split point selection. The QUEST method (for for Quick, Unbiased, Efficient, Statistical Tree) removes some of the disadvantages of using exhaustive search algorithms (eg CART), such as the difficulty of processing and universality results in reduction.

In the first step, the algorithm converts all categorical independent variables to "ordinal" ones by CRIMCOORD transformation. Furthermore, ANOVA F-test is conducted for each variable performed in each leaf node; if this test fails, the Leven´s F-test is used. To separate the node such explanatory variable is selected, which is most associated with the explained variable. To search for the split point of selected independent variable the method of Quadratic discriminant analysis (QDA) is used, unlike the FACT algorithm, where the method of linear discriminant analysis (LDA) applies.

The procedure is repeated recursively until it stops (on the basis of the criteria for stopping).

III. Segments Creation

On the basis of variables No. 13 and 18-24 the cluster analysis was performed, namely the procedure from SPSS TwoStep Analysis, which has created four clusters. Question 13 explores the intentions of respondents to study abroad, questions 18-20 are focused on the intended length of stay, the required level of study (Bachelor, Master, Doctoral) and field of study (Social, Technical, Natural Sciences and Art). In questions 21-23 the issues concerning the preferable language of study is examined including the willingness and condition for learning the Czech language. Question 24 asks the requirement for the type of certification on their studies.

Resulting clusters of TwoStep Analysis can be described as follows:

Cluster No. 1 (348 respondents)

- Ambivalent about whether to study abroad;
- If they have decided to study abroad, the length is from one semester to one academic year;
- They prefer to study at the Bachelor level;

- Most of them are current students of Social Sciences;
- They are interested in studying in Czech language or a combination of Czech and English languages;
- They would like to pass Czech language course (six months);
- The amount of 1500 Euro for a six month course does not seem beneficial to them (maybe considered);
- They require a Czech Diploma with a Diploma Supplement which is recognized in all of Europe.

Cluster No. 2 (382 respondents)

- Ambivalent about whether to study abroad;
- If they have decided to study abroad, the length would be one academic year;
- Preferable level of study is the Master level;
- Current students of Art, Social and Technical Studies;
- They are interested in studying in English;
- They did not take into account the rate for the language course (1500 Euro);
- Interested in a Czech Diploma with a Diploma Supplement which is recognized all over Europe – Master´s Degree (eventually Bachelor´s Degree).

Cluster No. 3 (27 respondents)

- They are not interested in studying abroad;
- Most questions were not answered.

Cluster No. 4 (336 respondents)

- They are considering studying abroad but later;
- If they study abroad, it would be for longer period than one academic year;
- Most of them would like to attend Doctoral Studies;
- Mostly current students of Natural and Technical Sciences;
- They are interested in studying programs delivered in the language combination Czech and English;
- They are willing to pass a Czech language course of one year duration;
- The rate of 3000 Euro for one year language course seems to be reasonable (may be they would consider it);
- They would welcome a Czech Diploma with a Diploma Supplement valid all over Europe at all levels of study

Considering the country of origin – the first cluster is created mostly from students from Slovakia, in the second cluster the Greek students are in the majority and the fourth cluster is created mostly from Chinese students. The third cluster is equable.

In the next step it was with the usage of a chi-square test of independence and examined which question (1-17) correlates with the established division into clusters. In questions 1-17 the issues concerning the identification of students, willingness to study abroad and fears about studying abroad. The null hypothesis of these tests was the independence of current variables and the variables which are classified as belonging to a given cluster.

Only a few variables released a p-value higher than 0,05. This means that these variables are not affected by belonging to individual clusters. This group includes the fears of the security situation in the country (p-value = 0.586), fears of the financial costs associated with the stay (p-value = 0.575), fear aroused from different religious persuasion (p-value = 0.206), language barriers (p-value = 0.168) and also expectations to receive a Diploma with entire Europe validity (p-value = 0.278), expectations to settle down in the country after graduation (p-value = 0.219) and expectation of better chances on the labor market when come returning to the home country (p-value = 0.133).

For other variables, the p-value reaches maximally 0.038, even with most variables p-value approaches nearly zero, indicating a rejection of the null hypothesis of independence. Therefore, these variables correlate with the distribution into the clusters; it is possible to determine what the prevailing value of the variable is for the cluster.

In the next step, the crucial decision trees we constructed using three different algorithms implemented in SPSS, namely CART, CHAID, and QUEST. As the target variable, the newly created variable "belonging to the cluster" was chosen and as input variables were those variables selected which belong to each issue of the questionnaire and have demonstrated their independence in previous chi-square tests. The value of Risk estimate was recorded similarly for all trees created (0,282-0,325); this result is not ideal but it can be considered a sufficient level. The worst results were obtained by using the algorithm QUEST.

According to the above mentioned algorithms the members of given clusters can be characterized in terms of responses to selected questions as follows (the third cluster is not sufficiently specified; therefore it is not placed in the summarization):

Algorithm CHAID

Cluster No. 1

- Students of Social Sciences with strong fears concerning separation from the family (answers 1 and 2 on five-point scale) mostly recruited from families with low incomes;

Cluster No. 2

- Students at the bachelor level with an average or below the average study results; without greater fears concerning family separation (answers 3 to 5 on five-point scale);

Cluster No. 4

- Students of Natural and Technical Sciences exhibit very strong fears of separation from the family (answers 1 and 2 on five-point scale).

Algorithm CART

Cluster No. 1

- Students of Social Sciences, coming from families of income lower than average;

Cluster No. 2

- Students from families having average or above average incomes; study results are average or lower than the average; at present paying tuition fees but they do not suffer because of it;

Cluster No. 4

- Students of natural, technical and of art disciplines currently not paying tuition, or if paying, it causes them great difficulties; from studying abroad they expect particularly the opportunity to acquire foreign work experience.

Algorithm QUEST

Cluster No. 1

- Students of Social sciences;

Cluster No. 2

- Students of secondary schools of art or higher education students currently paying tuition fee;

Cluster No. 4

- Students of Natural and Technical programs; they do not pay tuition fee in current time.

IV. Marketing Implications

- Remarkable results are provided by Two-Step analysis. For marketing implication this segmentation is the easiest one to apply as it almost covers students according to their country of origin;
- Only those segments are reasonable with a sufficient number of students (to invest in);
- The most prospective segment seems to be Cluster No. 4 with the majority of Chinese students;
- Algorithm CHAID proved how important issue the separation from the family is for the students. This is an important point for a marketing strategy creation (to be build on "family replacement");
- Algorithm CART outlined the specific segment (Cluster No. 4) of low-income students and their interest in work experience. If the study program is designed as "sandwich" and paid internship is offered, no large marketing activities are needed for attracting such a segment of students;
- Also higher educational institutions of regional importance might be highly successful in attracting foreign students – Cluster No. 2. These students probably will accept even non-prestigious universities and will appreciate any additional activities towards their self-development (on self-funding basis).

Acknowledgment

The search is financed by the internal grant agency of the J.E. Purkyne University in Usti nad Labem, the Czech Republic (IGA 4320115000401).

References

[1] P.Arabie, L.J. Hubert, Cluster Analysis in Marketing Research, In *Advanced methods in marketing research. Ed. R.P. Bagozzi. Blackwell*: Oxford, 1994, pp. 160–189.
[2] S. Dolnicar, *Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement*, 2003. http://ro.uow.edu.au/commpapers/139
[3] S. Dolnicar, F. Leisch, Behavioral Market Segmentation Using the Bagged Clustering Approach Based on Binary Guest Survey Data: Exploring and Visualizing Unobserved Heterogeneity. *Tourism Analysis*, Vol. 5, Iss. 2, 2000, pp. 163–170.
[4] B.S. Everit, S. Landau, M. Leese, *Cluster Analysis*, 4. Edition, Hodder Arnold, London, 2001.
[5] N. Gupta, *Cluster analysis for market segmentation*, 1st Edition, 2005.
[6] P. Kotler, *Marketing Management, Prentice Hall*; 11th Edition, 2002, 768 p.
[7] H. Řezanková, Shlukování a velké soubory dat, Lázně Bohdaneč 29.11.2004 – 01.12.2004. In: KUPKA, Karel (ed). *Analýza dat 2004/II* Pardubice: TriloByte Statistical Software, 2005, pp. 7–19.
[8] H. Řezanková, D. Húsek, V. Snášel, *Shluková analýza* dat, 2. Edition, Professional Publishing, Praha, 2009.
[9] R. Timofeev, Classification and Regression Trees (CART) Theory and Applications, *CASE–Center of Applied Statistics and Economics*, Humboldt University, Berlin, 2004.
[10] L. Wilkinson, Tree Structured Data Analysis: AID, CHAID and CART, Sun Valley, ID, *Sawtooth/SYSTAT Joint Software Conference*, 1992.
[11] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases, *ACM SIGMOD Record*, Vol. 25. No. 2, 1996, pp. 103–114.
[12] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: A New Data Clustering Algorithms and Its Applications, *Journal of Data Mining and Knowledge Discovery*, Vol. 1, No. 2, 1997, pp. 141–182.
[13] M. Žambochová, Jak na rozhodovací stromy, *Informační Bulletin*, Praha, Vol. 19. No. 3, 2008, pp. 1–12.
[14] M. Žambochová, Data Mining Methods with Trees, *E+M – Ekonomika a Management*, Iss.1, Liberec, 2008, pp. 126 – 132.
[15] M. Žambochová, K. Tišlerová, Classification of Individuals: Willigness to Start their own Business based on Franchise system, *Proceedings – Aplimat 2011*, [CD-ROM], Bratislava: Slovak University of Technology, 2011, pp. 1647 – 1655.