



## Research Paper

## CHAID algorithm as an appropriate analytical method for tourism market segmentation

Flora M<sup>a</sup> Díaz-Pérez\*, M<sup>a</sup> Bethencourt-Cejas

University of La Laguna, Tenerife – Canary Islands, Spain

## ARTICLE INFO

## Article history:

Received 24 June 2015

Received in revised form

8 December 2015

Accepted 11 January 2016

Available online 29 January 2016

## Keywords:

Market segmentation

Direct expenditure

Multivariate analysis

CHAID algorithm

Criterion methods

## ABSTRACT

Studies of the segmentation of the tourism markets have traditionally been undertaken by regression methods. The need to have a significant number of segments and qualifying variables has led, however, to the use of other procedures of multivariate analysis. CHAID (Chi-square Automatic Interaction Detection), which is more complex than other multivariate techniques, has rarely been used. This study applies the traditional methods of multivariate analysis and CHAID to the same population of tourists visiting a particular destination to compare the quality of the information obtained on tourism market segmentation. The results suggest that the analysis based on CHAID matches the nature of the problem studied better than those provided by discriminant analysis.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Literature about market segmentation has a tradition that goes back several decades. The different contributions in the field of segmentation try, ultimately, to match the expectations of different groups of tourists with offers made from tourist services suppliers. Specifically, researchers consider how to adjust the characteristics of the offers to what homogeneous segments of tourism services users expect. Several studies have been aimed at finding out the particular characteristics of tourists included in a specific segment (Assaker & Hallak, 2012; Chen, 2003; Kim, Timothy, & Hwang 2011; Koc, 2002, 2004; Legohérel, Hsu, & Daucé, 2015; Legohérel & Wong 2006; Mok & Iverson 2000; Nicholson & Pearce 2000; Shoemaker, 1984, 1989, 1994).

Moreover, the literature on segmentation of tourism markets tends to be related to the supply of hotel establishments, where segmentation related to the whole industry has been given scant attention (Bowen, 1998; Dixona, Backman, Backman, & Norman, 2012; Koc & Altınay, 2007; Mak, 2004; Pizam & Reichel, 1979; Snowball, 2004; Spotts & Mahoney, 1991). The reasons could lie in the difficulty of obtaining aggregated data (Chung, Oh, Kim, & Han 2004). In addition, when researchers investigate the implications that different market segments have for the economic development of a destination, the decision on what are the most

appropriate criteria can be a key issue in the analysis. Indeed, the study of the industry is particularly relevant when the measurement of the impact on the local economy is based on direct expenditure (Alegre, Cladera, & Sard, 2011).

With regards to the method used, the study of the segmentation of tourism markets has traditionally been undertaken using regression methods. These methods have been used with different levels of complexity to date (Fredman, 2008; Mok & Iverson, 2000; Ng & Lew, 2009). The need for a significant number of segments and qualifying variables has led, however, to the use of other procedures of multivariate analysis in order to learn about the characteristics of tourism segments: cluster analysis, multiple correspondence analysis and discriminant analysis are the most commonly used. However, as a market segmentation method, CHAID (Chi-square Automatic Interaction Detection) is more sophisticated than other multivariate analysis techniques (McCarty & Hastak 2007). It has, however, rarely been used in the study of the tourism markets. This is despite CHAID having very important advantages, particularly in that it does not require the use of parametric tests for predictive variables.

This paper considers the degree to which CHAID can be applied to tourism market segmentation when direct expenditure in destination is used as criterion variable. In order to facilitate this, the methods of multivariate analysis and CHAID are applied to the same population of tourists. This enable a comparison of the quality of the information obtained on market segmentation offered by the two techniques.

\* Corresponding author.

E-mail address: [fdiazp@ull.edu.es](mailto:fdiazp@ull.edu.es) (F.M. Díaz-Pérez).

## 2. Literature review

The idea that markets in general and tourism markets in particular are segmented is not new in marketing studies. In this respect, various authors (e.g. Cook & Mindak, 1984; Kardes, 2002; Mok & Iverson, 2000; Rhim & Cooper 2005; Solomon, Bamossy, & Askegaard, 2002) have highlighted the benefits of combining a variety of marketing strategies to capture different market segments of the same product. This confirms the need to delve deeper into segmentation to find as many segments as possible statistically speaking.

The literature on tourism market segmentation also contains a significant number of studies that have used tourism expenditure as a segmentation variable. Some of the first studies were carried out by LaPage (1969) and Stynes and Mahoney (1980), but were not very successful in terms of identifying clearly distinct groups of tourism users based on expenditure. However, more recent studies (e.g. Díaz-Pérez, Bethencourt-Cejas & Álvarez-González, 2005; Legohérel 1998; Legohérel & Wong, 2006; Spotts & Mahoney, 1991) have generated accurate information about the composition and characteristics of homogeneous groups of tourists according to their expenditure. Likewise, there are several studies that emphasize the relevance of the expenditure as a variable to provide a useful segmentation of the tourist market (e.g. Alegre et al., 2011; Mok & Iverson, 2000; Pizam & Reichel, 1979; Spotts & Mahoney, 1991) following, among others, the valid patterns of segmentation established by Kotler (2003).

### 2.1. Chi-square Automatic Interaction Detection (CHAID)

CHAID as a methodological approach appears in the literature under various names, including: Automatic Interaction Detection, Classification and Regression Tree, Artificial Neural Network and Genetic Algorithm. Although the segmentation procedure of the CHAID algorithm was first introduced by Kass in 1975, it has been little used in the segmentation of markets specifically: it has tended to have been applied more to general consumer research (e.g. Haughton & Oulabi, 1997; Levin & Zahav, 2001a; Magidson, 1994; Riquier, Luxton, & Sharp, 1997).

In such consumer research, algorithms cover both descriptive analyses, which are methods not based on a criterion variable, and predictive analyses, based on a criterion variable. Regarding segmentation of the tourism market, researchers have used two types of analysis *a priori* based on data collected (in the tourists' home locations) or *post hoc* (based on data collected when the tourist is leaving the destination). Frequently, both the *a priori* and the *post hoc* analyses have been descriptive in nature: that is, they have not been based on the establishment of a criterion variable. CHAID, however, sets up a predictive analysis establishing a criterion variable associated with the rest of variables that configure the segments as a result of a relation of dependency demonstrated by a significant chi-square.

CHAID is an analysis based on a criterion variable with two or more categories. This allows researchers to determine the segmentation with respect to that variable and in accordance with the combination of a range of independent variables (predictors) (Chen, 2003; Díaz-Pérez et al., 2005; Legohérel et al., 2015). To apply the CHAID procedure, therefore, a dependent variable (criterion) and the independent variables (predictors) have to be previously chosen.

The number of categories of independent variables depends on whether the results of the Chi-square test are significant or not. In the resulting tree, the most significant independent variable appears in the first node of the classification. The process of node formation and segment configuration ends when there is no significant relation between the dependent and independent

**Table 1**  
Strengths of CHAID.

- 
- 1) Chi-square is a non-parametric statistics
  - 2) Nominal type and interval variables can be considered as predictors
  - 3) Continuous variables can be chosen as criterion variables
  - 4) Establishing a criterion variable according to the objectives of an operator of tourist destination
- 

variables. This process is subject to the limitations imposed by the size of the sample: CHAID analysis is restricted by sample size criteria, in particular the sample size required per predictor variable (Chung et al., 2004; Kim et al., 2011; do Valle, Pintassilgo, Matias, & André, 2012). The lack of knowledge and complexity of the method (McCarty & Hastak, 2007) has, however, limited its use to relatively few cases of market segmentation of tourist destinations, and there have been few previous studies (Chen, 2003; Chung et al., 2004; Díaz-Pérez et al., 2005).

#### 2.1.1. Strengths of CHAID

It is important to highlight some of the strengths of CHAID as tourism market segmentation method compared to other methods. These strengths can be summarized by the following four characteristics (see Table 1): (1) Chi-square is a non-parametric statistics, (2) Nominal type and interval variables can be considered as predictors, (3) Continuous variables can be chosen as criterion variables, and (4) A criterion variable can be established according to the objectives of a tourism destination operator.

The main feature of this method is that *no type of distribution of independent variables is assumed a priori*. This is because it relies on the use of the Chi-square statistic. There are several advantages of Chi-square from the market segmentation perspective. First, it is a non-parametric statistical method of free distribution. Second, *the segments can be defined not just by ordinal variables but also by nominal type variables*. As such, any form of variable distribution is accepted in the classification process, rather than exclusively a normal one. Regarding the kinds of variables that can be included in the construction of the tree, the range is considerable, both in terms of the number and diversity of variables. Thus, for example, CHAID allows very useful segmentation variables for tourism markets to be included such as gender, age, household income, nationality, season and category of the establishment. Some of these variables are categorical or nominal, others are ordinal or interval-based. Under such circumstances, a technique that is not subject to the rigidity of the normal distribution and the requirement of ordinal variables will generally be the most appropriate: hence, Chi-square is the ideal statistical method for these cases (Diepen & Franses, 2006). With respect to the dependent variable, CHAID offers, in a natural way, greater flexibility to incorporate continuous criterion variables to the analysis, since continuous variables can always be dichotomized.

When CHAID is compared with non-criterion methods often used in the studies of tourism marketing, such as cluster analysis, it can be noted that the techniques based on the setting of a criterion variable have several advantages. First, they are based on a group of variables that discriminate among various segments, using the condition of significant correlation. In contrast, when applying non-criterion techniques it is not known if all variables will be able to differentiate among segments before their creation. Indeed, any procedure not based on the setting of a criterion variable relies on a set of variables that may not constitute significant descriptors of the segments, which implies that the researcher loses information and makes a worthless analysis with regard to the research objectives. To solve this problem and validate if the crosstab analysis of these variables constitutes a

significant discriminant of segments, researchers have traditionally carried out a discriminant analysis *post-hoc*. Second, CHAID allows the classification of new cases observed in segments that had been obtained, as it builds segments that are mutually exclusive (that is, they do not overlap), which means that each object is included in only one segment (Kass, 1980).

2.2. Expenditure as a criterion variable

Reviewing the literature on the segmentation of tourism markets (see Table 2), it can be seen that not only has limited use been made of CHAID in tourism market segmentation, but also that even fewer studies have implemented CHAID using tourism expenditure in destination as a criterion of segmentation. This is surprising considering the importance of expenditure in a destination as a measure of the impact of the tourist industry on the local economy (Alegre et al., 2011).

CHAID analysis has been used in tourism marketing from the 2000s to achieve several objectives. First it has been used to identify the preferences when choosing hotel establishments using demographic variables (Chung et al., 2004) or to clarify the preferences in the choice of hotels and restaurants (Legohérel et al., 2015). Second, it has been used to obtain information on the level of expenditure, using both demographic and other variables regarding trip characteristics (Díaz-Pérez et al., 2005) and to describe expenditure habits (Legohérel & Wong, 2006). Third, some studies have used it to identify the future recommendations using satisfaction with the products, price rises and poor service as independent variables (Chen, 2003), to find out the return probability (Assaker & Hallak, 2012; Hsu & Kang, 2007) and to obtain information about the intentions of recommending and visiting the destination in the future (Vassiliadis, 2008). In spite, therefore, of some authors (Alegre et al., 2011) pointing out the importance of using the tourism expenditure as a segmentation variable of tourism markets associated with a destination, the studies that use the direct expenditure in destination as a segmentation criterion are limited.

If we look at the use of CHAID as a segmentation procedure, it is even more difficult to find studies applied to the entire industry. A thorough search was carried out as part of the present paper of previous studies that have used CHAID as a segmentation method in tourism marketing. The first step was a classification according to the criterion variable chosen. Second, the grouping criterion considered was whether studies had resorted to other techniques of multi-variant analyses apart from CHAID. The results are depicted in the following four tables. Table 2.1 shows the most relevant studies using CHAID analysis to segment tourism markets of a destination according to expenditure. Table 2.2 groups those studies that have chosen criterion variables linked to the idea of tourist loyalty. Table 2.3 contains studies that have used variety-seeking behavior as the criterion variable. Finally, Table 2.4 shows other criterion variables.

Studying the information contained in the papers included in the above tables leads us to the following observations:

1. Table 2.1 contains the few studies in which the tourism market has been segmented using CHAID and using expenditure as the

**Table 2.1**  
Expenditure as criterion variable/applying CHAID.

Authors	Journal	Number of segments
Díaz-Pérez et al. (2005)	Tourism Management	14
Legohérel and Wong (2006)	Journal of Travel and Tourism Market	5, but several variables inside each one

**Table 2.2**  
Criterion variables related to loyalty/applying CHAID and factor analysis.

Authors	Journal	Factor analysis	Number of segments
Hsu and Kang (2007)	Journal of Travel Research	Not applied	9
Vassiliadis (2008)	International Journal of Tourism Research	Not applied	Not specified
Kim et al. (2011)	Tourism Management	Main component factor analysis	17
Assaker and Hallak (2012)	Journal of Vacation Marketing	Not applied	5

**Table 2.3**  
Variety-seeking behavior as criterion variable/applying CHAID.

Author	Journal	Criterion variable	Number of segments
Legohérel et al. (2015)	Tourism Management	Variety-seeking behavior for restaurants and hotels	5, but several variables inside each one

- criterion variable. In these studies, a common aspect is that their final aim is the analysis of the effect that each market segment has on local economic development.
2. It is instructive to consider how far the researchers that resorted to CHAID as their segmentation method may have fully appreciated the advantages that this market segmentation tool has over other methods. Table 2.4 contains various studies ranging from those that only proposed using vCHAID as a last resort, or as an afterthought (Chen, 2003), to those whose authors were aware of the potential of CHAID as a segmentation method and who attempt to identify market segments that the policy maker has more chance of influencing (Chung et al., 2004). It is worth noting that the early studies using CHAID were not aware of its full potential regarding its accuracy of segmentation. The CHAID method is, in fact, one of the most advanced from the perspective of the objectives of market segmentation.
  3. Some studies have focused on comparing the results obtained using CHAID with those achieved with other multi-variant methods. In studies in which two or more multi-variant methods have been used to segment the market, CHAID has demonstrated itself to be superior from the point of view of segmentation objectives. Specifically, all of the papers in Table 2.4, as well as those corresponding to Kim et al. (2011), contained in Table 2.2, demonstrate this. A specific comment is required for the work of McCarty and Hastak (2007), as not only does it identify the segments, but also aims to demonstrate the superiority of CHAID analysis when the sample size is small.

To sum up, considering the scarcity of analyses that have chosen expenditure as the criterion variable to build the tree, the analysis in this paper is novel in demonstrating the superiority of the CHAID segmentation method over other multi-variant analysis techniques.

The above review of the literature on the application of the CHAID also shows the different uses of this technique with various segmentation aims. Chung et al. (2004), for example, applied CHAID analysis with the sole purpose of suggesting the most appropriate technique to select the most meaningful and appropriate variable as an intermediate step to take advantage of the benefits of segmentation. Therefore, they performed a disjointed analysis procedure, wherein the core of the analysis was based on the cluster analysis. McCarty and Hastak (2007) performed a

**Table 2.4**  
Others criterion variables/applying CHAID and others multi-variant analyses techniques.

Authors	Journal	Others multi-variant analyses techniques	Aims	Number of segments
Chen (2003)	Annals of Tourism Research	ANOVA to test the homogeneity	The actionable and non-actionable segments	7 Segments
Chung et al. (2004)	Tourism Management	Cluster analysis	Applied to know the variable with the greatest segmenting power	9 Segments
McCarty and Hastak (2007)	Journal of Business Research	RFM. Logistic regression	Likely responders to a particular mailing	Not specified

comparative analysis of CHAID with two other methods: RFM (Regency, Frequency, and Monetary Value) and Logistic Regression. However, this comparison focused on the characteristics and cost of the information collection, rather than on the quality of the results according to the main objective of the segmentation. Specifically, the results indicated that “RFM may have difficulties when the response rate is low and the database marketer desires to send an offering to a relatively small portion of the entire file (30% or less)” (McCarty & Hastak, 2007, p. 661). Kim et al. (2011) applied CHAID to different variables than the ones selected to implement the factorial analysis; therefore, the working data was also different. Hence, no study was found that, in order to compare the utility of the analysis, applied the same set of data to the two segmentation methods considered at the same time. More particularly, no comparative analysis has yet been performed using direct expenditure of the individual tourist in destination as the criterion variable.

This paper therefore aims to compare the results obtained using another multi-variant segmentation technique (discriminant analysis) with those of CHAID. One of the aims is to determine whether the CHAID segmentation technique is the most appropriate method to segment tourism destination markets. In particular, it is hoped to demonstrate that in those cases in which the aim is to analyze the segmentation based on a set of diverse variables (categorical, nominal, ordinal, by intervals) or linked to different objectives (economic, demographic, psychological) of tourism markets CHAID technique is the most suitable.

### 3. Hypotheses

Traditional studies on tourism market segmentation have only focused on a few variables. This can be explained partly by the limitation that the sample size imposes on this analytical technique and by the characteristics and possibilities of the segmentation technique applied. The costs of enlarging the sample mean that it is advisable to search for the most accurate and efficient techniques regarding the relations and number of segmentation variables.

The above reasoning leads to the following hypotheses in this paper.

**H1.** - The comparison of the information obtained from applying the two techniques, CHAID and discriminant analysis, to the same dataset will show that a greater number of classifying variables can be obtained with the application of CHAID than with discriminant analysis.

**H2.** - If other techniques of multivariate analysis, such as Discriminant Analysis, are applied to the same dataset, is it possible to find different relevant variables for segmentation other than daily expenditure in the destination under consideration.

### 4. Applying CHAID and discriminant analysis to the same population of tourists

To test the hypotheses, CHAID analysis and then discriminant analysis were applied to the same dataset relating to tourists who had visited the Spanish island of La Palma in 2002. The database was obtained from the *Plan of Competitiveness for the Tourism Product: Isla de La Palma*, supported financially by the Canarian Department of Education, Culture and Sport, Government of the Canary Islands. This dataset is rather old but it represents the most complete and most recent empirical study available of the destination (La Palma Island) collected before the onset of the 2007 financial crisis. This crisis has still not been overcome in the Canary Islands and data from the current context of economic instability could well lead to erroneous results. Moreover, it should be borne in mind that the main aim of this paper is the comparison between two methodologies that analyze information using quantitative research techniques.

In particular, the procedure followed to demonstrate the hypotheses noted above was as follows: Firstly, the study took as its starting point the results obtained by applying the CHAID analysis to a representative sample of the total number of tourists that visited the Spanish island of La Palma in a particular year. Secondly, the more traditional multivariate analysis, in the form of discriminant analysis was applied to the same dataset, aiming to compare the results obtained and to be able to test the adopted hypotheses.

#### 4.1. Data and variables

##### 4.1.1. Data

The analysis was performed with the data taken from an empirical study executed in 2002 (Bethencourt, Díaz, Alvarez & González, 2002), which used a representative sample of the total number of tourists who visited the Spanish island of La Palma ( $n=324$ ). The island of La Palma is one of the seven islands that make up the Canarian archipelago. In the year of the survey, the islands received over ten million tourists. For La Palma, the total number of tourists in 2002 was 129,907 (ISTAC (Instituto Canario de Estadística), 2002), this is the population for this study and corresponds to 1.22% of the Canary Island tourism market.

##### 4.1.2. Variables

The variables studied are shown in Table 3.

Season – Tourism in La Palma can be divided into two seasons based on the number of tourist arrivals. High season corresponds to the greatest influx of tourists, which is between November and April, and low season, which runs from May to October, has the fewest number of tourists. It should be noted that high and low season is linked to number of tourists rather than their expenditure in the destination.

Expenditure – This is the main variable under consideration and is the expenditure per tourist per day.

**Table 3**  
Variables.

Season	Expenditure
Nationality	Profession
Gender	Income
Age	Services contracted
Traveling alone or accompanied	

**Nationality** – This refers to the country of origin of the tourist. European countries were included in the questionnaire: Germany, Holland, Austria, Italy and England, among others. These are the most frequent countries of origin for tourists who visit the destination other than domestic tourists, who are mainly tourists from the Spanish mainland and the other Canary Islands.

**Profession** – Interviewees' professions were classified into the following groups on the questionnaire: homemaker, unemployed, retired/pensioner, student, owner/landlord, farmer (differentiating between small or large farm). There were also self-employed or business owner with five employees or fewer, or with more than six, management positions with five subordinates or fewer, and those with six subordinates or more, other administrative/office workers, skilled and unskilled laborers (employees).

**Income** – This variable refers to the average monthly income of the family unit of the person interviewed. There were seven income intervals.

**Traveling alone or accompanied** – This variable includes the following options: traveling alone, accompanied by partner, accompanied by partner and children and accompanied by friends.

**Contracted services** – These include the following: hotel (without board, with half-board or with full board), tourism house, cottage, rented apartment and time-sharing.

**Demographic variables** – Gender and age of tourists were considered, the latter was included by age intervals in the questionnaire.

**Data collection procedure** – The size of the sample was 324 interviews, with an error of 5.5% and a reliability of 95.5%. The survey was carried out on the island of La Palma, Canary Islands (Spain). Data collection occurred on two occasions, once in the high season and once in the low season to ensure it was representative of seasonality. Stratified sampling was carried out each time, with random selection in each stratum. The type of accommodation and the nationality were used as the stratification criteria in each stratum.

## 5. Results

### 5.1. Results of CHAID analysis

To perform the segmentation of La Palma's tourism market, an analysis based on the CHAID procedure was first carried out. The variables explained above were the same in the case of the discriminant analysis, which is presented in the next section.

As indicated above, the aim of the research was to find which of the sample's characteristic variables most accurately predict expenditure on the island. Following the CHAID procedure, a criterion variable was established: 'daily spending of tourist in destination' and the sample was subsequently divided into two, the mathematical mean of sample established the line of segmentation. This resulted in one variable with two values (high cost and low cost) that was dependent on nominal variables. This technique provides information about the relation between the criterion variable and diverse predictor variables, and the model uses the Chi-squared as a division criterion. The direct daily expenditure of the tourists used as the *criterion variable* was dichotomized (the

dividing line being the arithmetical mean of the surveyed sample) to give a variable with two separate values (high expenditure and low expenditure). *Predictive variables* included nationality, gender, age, traveling alone or with others, profession, family income level, intention to return in the future, destination services and season during which the individual had visited the island. The results obtained enabled us to identify homogenous groups of tourists based on their direct daily expenditure during the stay. These groups were mainly identified by *nationality, average family income and season*, the total mean expenditure being 31.20 euros.

Based on the decision tree generated by the CHAID procedure, tourist nationality was identified as the first significant segmentation ( $p \leq 0.005$ ). The result shows two opposed segments. On the one hand, there are German tourists and a small group of Dutch tourists whose recorded expenditure was below the general mean, while on the other a second group exists consisting mainly of Spanish tourists (although Austrian and Italian tourists also fall into this category) whose expenditure is above the average level.

The group of tourists whose expenditure was below the general mean (mainly Germans) was further segmented by the CHAID into two subgroups with significant values. These subgroups was defined according to family income level: those with incomes lower tend to spend an average of no more than 30.5 euros, while those with higher income levels tend to spend an average of 31.5 euros.

Taking these variables (nationality and income) into consideration, the below-average expenditure segment had a correct classification power of 88%. In trying to find a predictive pattern for above-average expenditure, however, it was noted that, in addition to nationality (Spanish, Austrian and Italian), the season (high or low) during which the tourist visited the island was also a determining factor, reaching a group discrimination power of 87%. This was partly due to the subdivision of the group and the number of nationalities included in this segment, which also revealed significant differences with regard to spending criteria.

Despite the limitations imposed by the size of the sample, the results obtained during the study on La Palma were fairly similar to those obtained during previous studies (Díaz-Pérez et al., 2005) focusing on the entire region of Canary Islands as a whole. The variables: *nationality, high-low season, profession or average family unit income and destination service* recur in both studies, indicating a certain consistency with regard to results. The results obtained in the present study are expressed in Fig. 1.

### 5.2. Discriminant analysis

To compare the CHAID analysis results, two groups were created dividing the total sample between 'low expenditure' (Group 1) and 'high expenditure' (Group 2) tourists, using the arithmetical mean.

#### 5.2.1. Variables that discriminate

The variables used in this analysis were the same as those used for CHAID. To identify which variables differentiated the most the two groups mentioned above, a one-factor ANOVA was performed. The variables that initially showed differences between the two groups were: nationality, hotel without board and gender.

#### 5.2.2. Correlation matrix

In the correlations matrix, there is only one strong negative correlation between the variable 'nationality' and 'season'.

Variables were then introduced into the analysis using a 'step-by-step' procedure (see Table 4). First, the variable that best discriminated the groups was entered in the discriminant function, that is, the one that maximizes the Mahalanobis distance among the closest groups. The next function that discriminates the most was then introduced; and so on until none of the remaining

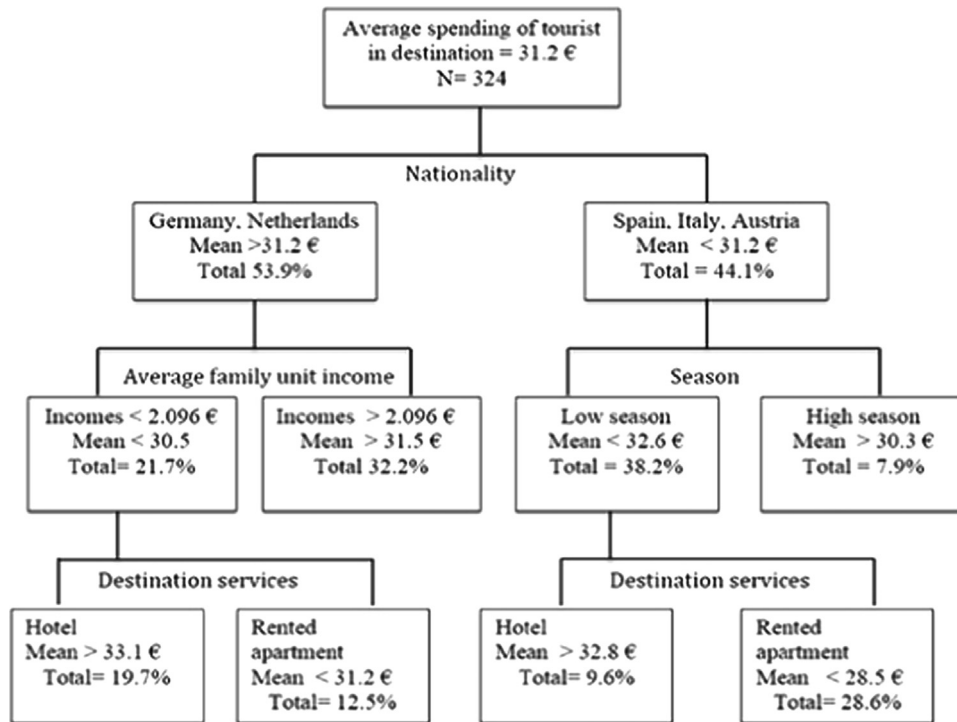


Fig. 1. Results for CHAID analysis.

variables complied with the conditions required by the model ( $F$  minimum to enter=3.84.  $F$  maximum to eliminate=2.71).

The variables that will be part of the discriminant function and that, consequently, discriminate best were: hotel without board (room-only) and gender. Nationality, that at first seemed to discriminate between the two groups, offered an  $F$  of entry of 3.84 (3.69). It was thus not included as part of the discriminant function.

5.2.3. Comparison of groups of pairs

Applying a comparison of pair groups, and considering the variables previously mentioned, the results of a hypothesis of average equality between Groups 1 and 2 indicated that significant differences exist between both, as shown in Table 5.

5.2.4. Coefficients of classification function: Fisher's linear discriminant functions

With respect to the coefficients from the classification function, Fisher's linear discriminant functions, the coefficient of the classification of each group is provided in Table 6. In the event that a tourist comes to La Palma, he or she will be included in Group 1 or 2, depending on whether he or she spends more or less than average, keeping in mind that he or she has been staying in a hotel on a room-only basis, and considering gender as well.

Regarding the classification results, 60.9% of the cases were well classified using the expenditure variable to create the 'group' variable, with a higher percentage of correct ones being in Group 1 (63,4%) than in Group 2 (57%).

Table 4 Variables that best discriminate the groups.

Step	Introduced	Min. D square	F exact
1	Hotel without board (room-only)	0.148	11.838
2	Gender	0.283	11.292

Table 5 Results of a hypothesis of average equality.

Group		1	2
1	F		11.292
	Sig.		0.000
2	F	11.292	
	Sig.	0.000	

Table 6 Fisher's linear discriminant functions.

	Group	
	1	2
Destination services: hotel without board (room-only)	0.2235	1.7164
Gender	7.0147	6.2587
Constant	-6.4044	-5.5927

6. Conclusions

In comparing of the results of the discriminant analysis with respect to those obtained with CHAID, it is possible to draw the following conclusions. First, it is assumed that the discriminant analysis is a solid test, that is, it is little affected by the nature of the independent variables, which try to create the classification function of the cases in one of the values of the dependent variable, in this case higher or lower expenditure than average. According to the results provided by discriminant analysis, it can be concluded that the most discriminating variables between the two expenditure groups were destination services and gender. According to this classification, nationality and income variables were included since the discriminant analysis did not indicate them to be predictors. In order to clarify the reasons for these differences, a series of exceptions that are at the basis of both methodologies needs to be evaluated. Discriminant analysis found gender to be a significant demographic variable, which was not

shown by the CHAID analysis. Additionally, discriminant analysis did not generate any information whatsoever on the important segmentation power of nationality: a variable that is usually associated with the cultural traits of communities.

Second, nominal or interval variables were being used systematically in the case of the independent variables, when the basic recommendation for the use of discriminant analysis is that the independent variables are continuous or interval type, paying close attention to the fact that their distribution adapt to the normal curve. Using the same data in both analyses, there are some assumptions about the distribution of the variables that do not really adjust to the requirements of discriminant analysis, which is a parametric model, and hence, its results strongly depend on the inviolability of the restrictive principles on which it is based, mainly, the adjustment to the normality of their data. This sharply contrasts the CHAID test, since it is based on the iterative use of Chi-square statistic, the non-parametric statistical test par excellence, where no type of distribution is assumed regarding the independent variables involved in the analysis.

Finally, by using CHAID analysis, a greater number of significant segments and variables emerge. Without a doubt, this last result shows the superiority of CHAID as a segmentation method applicable to tourism markets, allowing more precise and correct decisions for all levels of the real segmentation of markets. It is therefore possible to infer that the results of the analysis based on CHAID matches the nature of the problem studied better than those provided by the discriminant analysis.

## 7. Discussion

This study provides a theoretical analysis of tourism market segmentation based on an empirical study applied to the same database. The results validate the superiority of CHAID compared to the multivariate analysis techniques that have been traditionally used. Thus, for instance, although gender as variable of segmentation is significant with the use of the basic multivariate analysis, the role of this demographic variable can be avoided from the point of view of the objectives pursued with the tourism markets segmentation by a destination operator. In other words, when the demographic variables are not relevant from the point of view of the objectives pursued by the analysis, knowing the role of others such as the country of origin, type of establishment, season and the income level of the family unit, among others, can be more useful in tourism marketing. For the case of non-criterion methods, in order to make a cross-validation study of whether these variables are significant in the segment description, a discriminant analysis is usually applied *post-hoc*. However, to the extent that a *non-criterion* analysis uses a set of variables that may not be significant descriptors to explain the configuration of the segments, it can result in non-optimal homogeneity from the point of view of the classifiers.

Regarding the different forms of regression analysis, CHAID is presented as a more rigorous technique of analysis to the extent that the researcher does not incorporate any value judgment when selecting independent variables. Said selection is executed by this automatic statistical procedure depending on the classificatory power of significant variables. The only knowledge established *a priori* is the definition of the criterion variable, which belongs to the researcher, based on the particular objectives of the research. Although the selection of independent variables is also made *a priori*, the procedure can consider a substantially higher number of possible independent variables, whose association or not with the criterion variable will be in hands of CHAID by determining it automatically.

When creating segments, the nature of the analyzed variables

is largely qualitative. Others are continuous but easily transformable into qualitative variables: including age (young adults) and high incomes, among others. Certain restrictive principles are required in the multivariate analyses such as the normality of the independent variables or the homoscedasticity that in many cases makes extremely difficult being able to use these models in a consequent manner. For all the above, Chi-square is the basic statistic involved in CHAID, which can discern between segments (of a qualitative nature) in a natural and more powerful way than other alternative statistical analyses models, since Chi-square is designed for discrete and free-distribution variables.

Finally, the hierarchy between variables due to their segmentation power, establishing a ranking and specifying which has the greatest power to segment the population, provides very useful information in tourism marketing. From the point of view of the destination operators, knowing which variables segment the most in regards to the tourist expenditure in destination, and obtaining a ranking by decreasing power of segmentation, is an essential tool when the goal is to aim the promotion towards the tourism segments with the largest contribution to the local economy.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

## Funding

This paper is based on data gathered for a study entitled Plan of Competiveness for the Tourism Product: Isla de La Palma, supported financially by the Canarian Department of Education, Culture and Sport, Government of the Canary Islands, PI 2000/127.

## References

- Alegre, J., Cladera, M., & Sard, M. (2011). Analysing the influence of tourist motivations on tourist expenditure at a sun-and-sand destination. *Tourism Economics*, 17(4), 813–832.
- Assaker, G., & Hallak, R. (2012). European travelers' return likelihood and satisfaction with mediterranean sun-and-sand destinations: A chi-square automatic identification detector-based segmentation approach. *Journal of Vacation Marketing*, 18(2), 105–120.
- Bethencourt, M., Díaz, F., Alvarez, J.A. & González, O. (2002). The segmentation of the Canarian tourism market with regard to expenditure: An empirical study of La Palma, Paper 251, Industry to Advanced Services. Dortmund: 42nd Congress of the European Regional Science Association.
- Bowen, J. (1998). Market segmentation in hospitality research: No longer a sequential process. *International Journal of Contemporary Hospitality Management*, 10(7), 289–296.
- Chen, J. S. (2003). Market segmentation by tourists' sentiments. *Annals of Tourism Research*, 30(1), 178–193.
- Chung, K. Y., Oh, S. Y., Kim, S. S., & Han, S. Y. (2004). Three representative market segmentation methodologies for hotel guest room customers. *Tourism Management*, 25(4), 429–444.
- Cook, V. J., & Mindak, W. A. (1984). A search for constants: The heavy user revisited. *Journal of Consumer Marketing* (pp. 79–81), 79–81.
- Díaz-Pérez, F. M., Bethencourt-Cejas, M., & Álvarez-González, J. A. (2005). The segmentation of Canary island tourism markets by expenditure: Implication for tourism policy. *Tourism Management*, 26(6), 961–964.
- Diepen, M. van, & Franses, P. H. (2006). Evaluating chi-squared automatic interaction detection. *Information Systems*, 31, 814–831.
- Dixona, A. W., Backman, S., Backman, K., & Norman, W. (2012). Expenditure-based segmentation of sport tourists. *Journal of Sport Tourism*, 17(1), 5–21.
- Fredman, P. (2008). Determinants of visitor expenditures in mountain tourism. *Tourism Economics*, 14(2), 297–311.
- Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with CART and CHAID. *Journal of Interactive Marketing*, 11(4), 42–53.
- Hsu, H. C., & Kang, S. K. (2007). CHAID-based segmentation: International visitors' trip characteristics and perceptions. *Journal of Travel Research*, 46, 207–216.
- ISTAC (Instituto Canario de Estadística). (2002). Demanda turística: Turistas y

- pasajeros. Gobierno de Canarias. (<http://www.gobiernodecanarias.org/8/03/2015>).
- Kardes, F. R. (2002). *Consumer behavior and managerial decision making*. Upper Saddle River, NJ: Prentice-Hall.
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119–127.
- Kim, S. S., Timothy, D. J., & Hwang, J. (2011). Understanding Japanese tourists' shopping preferences using the decision tree analysis method. *Tourism Management*, 32(3), 544–554.
- Koc, E. (2002). The impact of gender in marketing communications: The role of cognitive and affective cues. *Journal of Marketing Communications*, 8(4), 257–275.
- Koc, E. (2004). The role of family members in the family holiday purchase decision-making process. *Journal of International Hospitality and Tourism Administration*, 5(2), 85–102.
- Koc, E., & Altınay, G. (2007). An analysis of seasonality in monthly per person tourist spending in Turkish inbound tourism from a market segmentation perspective. *Tourism Management*, 28(1), 227–237.
- Kotler, P. (2003). *Marketing management* (11th ed.). Upper Saddle River: Prentice-Hall.
- LaPage, W. F. (1969). *Campground marketing: The heavy-half strategy*. Upper Darby, PA: Northeastern Forest Experiment Station.
- Legohérel, P. (1998). Toward a market segmentation of the tourism trade: Expenditure levels and consumer behavior instability. *Journal of Travel Research*, 29(4), 12–18.
- Legohérel, P., & Wong, K. F. (2006). Market segmentation in the tourism industry and consumers' spending: What about direct expenditure? *Journal of Travel and Tourism Marketing*, 20(2), 15–30.
- Legohérel, P., Hsu, C. H. C., & Daucé, B. (2015). Variety-seeking: Using the CHAID segmentation approach in analyzing the international traveler market. *Tourism Management*, 46, 359–366.
- Levin, N., & Zahav, J. (2001a). Predictive modeling using segmentation. *Journal of Interactive Marketing*, 20(2), 3–22.
- Magidson, J. (1994). The CHAID approach to segmentation modeling: Chi-square automatic interaction detection. In: R. Bagozzi (Ed.), *Advanced method of marketing research* (pp. 118–159). Cambridge, MA: Blackwell.
- Mak, J. (2004). *Tourism and the economy: Understanding the economics of tourism*. Honolulu, HI: University of Hawaii Press.
- McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, 60, 656–662.
- Mok, C., & Iverson, L. G. (2000). Expenditure-based segmentation: Taiwanese tourists to Guam. *Tourism Management*, 21(3), 299–305.
- Ng, P. T., & Lew, A. A. (2009). *Quantile regression analysis of visitor spending: an example of mainland chinese tourists in Hong Kong*. United States: Northern Arizona University.
- Nicholson, R., & Pearce, D. (2000). Who goes to events: A comparative analysis of the profile characteristics of visitors to four south island events in New Zealand. *Journal of Vacation Marketing*, 6, 236–253.
- Pizam, A., & Reichel, A. (1979). Big spenders and little spenders in US tourism. *Journal of Travel Research*, 18(1), 42–43.
- Rhim, H., & Cooper, L. G. (2005). Assessing potential threats to incumbent brands: New product positioning under price competition in a multisegmented market. *International Journal of Research in Marketing*, 22(2), 159–182.
- Riquier, C., Luxton, S., & Sharp, B. (1997). Probabilistic segmentation modelling. *Journal of the Market Research Society*, 39(4), 571–588.
- Shoemaker, S. (1984). Marketing to older travelers. *The Cornell Hotel and Restaurant Administration Quarterly*, 75(2), 84–91.
- Shoemaker, S. (1989). Segmentation of the senior pleasure travel market. *Journal of Travel Research*, 27(3), 14–21.
- Shoemaker, S. (1994). Segmenting the US travel market according to benefit realized. *Journal of Travel Research*, 32(3), 8–21.
- Snowball, J. D. (2004). Interpreting economic impact study results: Spending patterns, visitor numbers and festival aims. *South African Journal of Economics*, 72(5), 1075–1083.
- Solomon, M., Bamossy, G., & Askegaard, S. (2002). *Consumer behaviour: A European perspective*. Essex: Pears.
- Spotts, D. M., & Mahoney, E. M. (1991). Segmenting visitors to a destination region based on the volume of their expenditure. *Journal of Travel Research*, 29(4), 24–31.
- Stynes, D. M., & Mahoney, E. M. (1980). *Michigan downhill ski market study: Segmenting active skiers*. East Lansing, MI: Michigan University Agricultural Experiment Station.
- do Valle, P. O., Pintassilgo, P., Matias, A., & André, F. (2012). Tourist attitude towards an accommodation tax earmarked for environmental protection: A survey in Algarve. *Tourism Management*, 33, 1408–1416.
- Vassiliadis, C. A. (2008). Destination product characteristics as useful predictors for repeat visiting and recommendation segmentation variables in tourism: A CHAID exhaustive analysis. *International Journal of Tourism Research*, 10, 439–452.