Revista de Matemática: Teoría y Aplicaciones 2010 **17**(2) : 179–197
cimpa – ucr issn: 1409-2433

# AN ALTERNATIVE TO CHAID SEGMENTATION ALGORITHM BASED ON ENTROPY

# UNA ALTERNATIVA AL ALGORITMO CHAID DE SEGMENTACIÓN BASADA EN ENTROPÍA

María Purificación Galindo Villardón*
José Luis Vicente Villardón† Ana Dorado Díaz‡
Purificación Vicente Galindo§
María Carmen Patino Alonso¶

*Received: 11 Apr 2009; Revised: 21 Sep 2009; Accepted: 2 Oct 2009*

*Departamento de Estadística de la Universidad de Salamanca, C/Alfonso X El Sabio s/n, Campus Miguel de Unamuno, Facultad de Medicina, 37007 Salamanca, Spain. E-Mail: pgalindo@usal.es

†Misma dirección que/*same address as* M.P. Galindo E-Mail: villardon@usal.es

‡Servicio de Estadística, Consejería de Sanidad, Junta de Castilla y León, Spain. E-Mail: dordiaan@jcyl.es

§Misma dirección que/*same address as* M.P. Galindo E-Mail: purivic@yahoo.com

¶Misma dirección que/*same address as* M.P. Galindo E-Mail: mcarmen_patino@yahoo.es

**Abstract**

The CHAID (Chi-Squared Automatic Interaction Detection) tree-based segmentation technique has been found to be an effective approach for obtaining meaningful segments that are predictive of a K-category (nominal or ordinal) criterion variable. CHAID was designed to detect, in an automatic way, the interaction between several categorical or ordinal predictors in explaining a categorical response, but, this may not be true when Simpson's paradox is present. This is due to the fact that CHAID is a forward selection algorithm based on the marginal counts. In this paper we propose a backwards elimination algorithm that starts with the full set of predictors (or full tree) and eliminates predictors progressively. The elimination procedure is based on Conditional Independence contrasts using the concept of entropy. The proposed procedure is compared to CHAID.

**Keywords:** Segmentation, CHAID, entropy, conditional independence.

**Resumen**

La técnica de segmentación basada en árboles CHAID (Detección Automática de Interacción basada en el Chi Cuadrado, o *Chi-Squared Automatic Interaction Detection*, por sus siglas en inglés) ha mostrado ser útil para obtener segmentos significativos que sean predictivos de una variable criterio de K categorías (nominal u ordinal). CHAID fue diseñado para detectar, de manera automática, la interacción entre varios predictores categóricos u ordinales para explicar una respuesta categórica, pero esto puede no ser cierto cuando se presenta la paradoja de Simpson. Esto se debe al hecho de que CHAID es un algoritmo de selección hacia adelante basado en conteos marginales. En este artículo proponemos un algoritmo de eliminación hacia atrás que empieza con el conjunto completo de predictores (o árbol completo) y elimina progresivamente predictores. El procedimiento de eliminación está basado en contrastes de independencia condicional usando el concepto de entropía. El procedimiento propuesto es comparado con CHAID.

**Palabras clave:** Segmentación, CHAID, entropía, independencia condicional.

**Mathematics Subject Classification:** 62H30, 94A17.

# 1 Introduction

Chi-square automatic interaction detection (CHAID) is a particular case of the Automatic Interaction Detection algorithms. It is a statistically valid response modelling where both, the dependent variable and the predictors are categorical [8]. It has been applied in Direct Marketing [10], in studies

of Mobility and Housing choice [4], in Shopping Centre Market Research [2], in Political Marketing [6] [9], in medicine [11], in education [12], among many others, and still receives a considerable attention in the literature. CHAID uses a sequential "forward selection" procedure for splitting (segmenting) the original set of individuals into subgroups to maximize the differences between sub-groups on their response profiles. A decision tree presents the results of the splitting process. The algorithm has three main steps:

1. For each predictor, merging the categories with similar response patterns. The type of predictor determines the admissible groupings: for a nominal variable (free predictor) any combination is possible, for an ordinal variable (monotonic predictor), only adjacent categories can be merged. A "floating predictor" in which all the categories except one are ordinal, can also be used.

2. Searching for the best predictor based on the chi-square significance tests for the cross-tabulations of the response and each predictor.

3. Splitting the original sample using the categories of the predictor with the lowest p-values.

The CHAID algorithm is based on marginal independence contrasts. It is well known that in presence of interaction between two variables, the effect of each predictor can be confounded in the marginal tables and that when data from several groups are combined to form a single group, a reversal of the direction of association may occur.

That is known as the "Simpson's Paradox" [14]. Ávila [1] showed that the algorithm could not detect interaction just because it is present.

Let us illustrate that with a fictitious example: Suppose that we have one dependent variable "Expectancy of finding a job" and three explicative variables: "Sex", "Race" and "Social Status"; all of them with two categories. Suppose that the theoretical structure is represented on the tree in Figure 1.

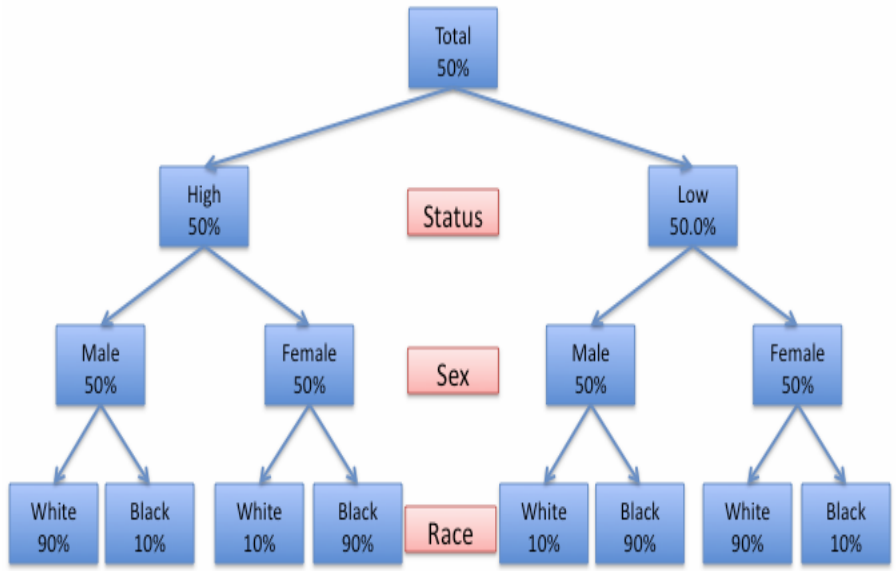


Figure 1: Tree showing a three-way extreme interaction.

Observe that 50% of the Males and 50% of the Females have positive expectations of finding a job.The same percentages apply for individuals with High and Low Status and also for any combination of Sex and Status, for example, "High Status Males" and "High Status Females" have both a 50% chance of having positive expectancy. If we consider Race the chances are quite different, "High Status White Males" have a 90% chance of positive expectancy while "High Status White Females" have only 10% chance. For "High Status Black Males and Females" all the percentages reverse and reverse again when "Low Status" individuals are considered. The terminal nodes of the tree are independent from the ordering of the variables. See figure 2 for a different ordering.

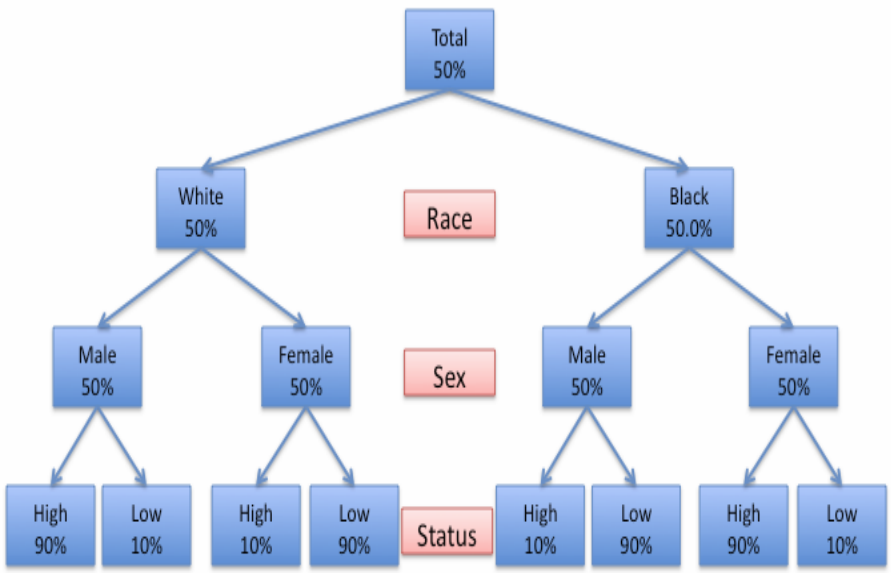


Figure 2: Another ordering of the tree showing a three-way extreme interaction.

Now we see that "Blacks" and "Whites" have the same chance of positive expectancy, moreover, the chance for the combination of the categories of any two variables is always 50%; thus the only tree that describes correctly the data is the one containing all the variables.

If we have, for example, 100 individuals for each combination of categories of the explicative variables and we add some sampling variation to the theoretical structure we would have observed frequencies as in Table 1.

| Variables | | | Expectancy of finding a job | |
|---|---|---|---|---|
| Social Status | Sex | Race | Yes | No |
| | Male | White | 88 | 12 |
| High | | Black | 9 | 91 |
| | Female | White | 10 | 90 |
| | | Black | 85 | 15 |
| | Male | White | 11 | 89 |
| Low | | Black | 90 | 10 |
| | Female | White | 86 | 14 |
| | | Black | 13 | 87 |

Table 1: Table of data constructed to show the problems associated with the CHAID procedure. The marginal totals for all the combinations of explicative variables are fixed.

A simple inspection of the data table shows that the "Expectancy of finding a job" depends on the three explicative variables and that there is a three-way interaction since the percentages of response for each of the values of the different variables are inverted. The results of the independence contrasts for the cross-tabulations of "expectancy of finding a job" and each explicative variable are shown in Table 2.

| Variable | Chi-square | p-value |
|---|---|---|
| Status | 0,320 | 0,5715 |
| Sex | 0,080 | 0,7773 |
| Race | 0,020 | 0,8875 |

Table 2: Marginal associations of Expectancy with the rest of the variables.

With these results, there is no association between the "Expectancy of finding a job" and the rest of variables. It is not possible to make a segmentation of the population. All the groups have similar response patterns. If we apply the CHAID algorithm to this case, none of the response variables will permit splitting of the population, but we do have an interaction. The conclusion of the statistical analysis is in clear contradiction with the visual scrutiny of the multi-way data table and the tree. For example, in the group of individuals with a high social status, male and white, 88% have the expectancy of finding a job as compared with the 10% in the group of individuals with high social status but male and black.

It is clear that CHAID does not detect any interaction even though it exists. A paradox can be seen in that a method of automatic interaction detection is unable to detect the interaction precisely because it is present and interferes with the procedures of analysis.

What it is wrong about the CHAID procedure for this table is that is based on marginal independence. Conditional independence is needed to study this example: for any combination of the categories of two independent variables, the response depends on the third. It seems then reasonable to base the conclusions of any procedure or algorithm on conditional rather than marginal independence.

This paper proposes an alternative backwards algorithm for automatic interaction detection based on conditional rather than marginal independence. We use entropy [13] to test the conditional independence.

Section 2 presents some notation used in the rest of the paper, section 3 some results about collapsibility and conditional independence that justify the algorithm, section 4 the marginal and conditional contrasts based on entropy, section 5 presents a forward algorithm equivalent to the classical CHAID method, but based on entropy, and section 6 the proposed backwards algorithm. The *forward algorithm* makes up the segmentation tree contrasting reduction in the entropy. The backwards algorithm starts with the full tree (with all the variables interactively coded), and contrasts conditional independence (based on "entropy") in order to search for the predictors not providing a significant increment in the entropy. These variables are eliminated globally or in partial branches of the tree.

## 2 Notation

Let us start with some notation. Suppose that we have now a set $\Gamma = \{\mathbf{i}, \mathbf{j}, \mathbf{k}, \ldots\}$ of several variables in which $\mathbf{i}$ is the response and denote with $V = \{\mathbf{j}, \mathbf{k}, \mathbf{l}, \ldots\}$ the set of predictors. We will denote with $V \backslash \{\mathbf{j}\}$ the set of all predictors except $\mathbf{j}$. Let us suppose that we have observed frequencies for the multi-way table given by $f_{ijkl\ldots}(i = 1, \ldots, I; j = 1, \ldots, J, k = 1, \ldots, K, l = 1, \ldots, L, \ldots)$. Marginal tables of any dimension can be defined by collapsing on the appropriate variables, for example, $f_{i\bullet\bullet\bullet\ldots\bullet}$ are the marginal frequencies of the categories of the response, $f_{ij\bullet\bullet\ldots\bullet}$ are the frequencies of the marginal two way table for variables $\mathbf{i}$ and $\mathbf{j}$ collapsing on the rest), $f_{\bullet\bullet\bullet\bullet\ldots\bullet} = N$ is the grand total or sample size, and so on. Observe that the points indicate "collapse" or "sum" over the categories of that variable to obtain the corresponding marginal tables.

In the same way we can define probabilities, for example $p_{ijkl\ldots} = P(\mathbf{i} = i, \mathbf{j} = j, \mathbf{k} = k, \mathbf{l} = l, \ldots)$ are the joint probabilities, $p_{i\bullet\bullet\bullet\ldots\bullet} = P(\mathbf{i} = i)$ are

the marginal probabilities for $\mathbf{i}$, $p_{ij\bullet\bullet\cdots\bullet} = P(\mathbf{i} = i, \mathbf{j} = j)$ are the marginal probabilities for the two-way table with $\boldsymbol{i}$ and $\mathbf{j}$, $p_{i/jkl\ldots} = P(\mathbf{i} = i/\mathbf{j} = j, \mathbf{k} = k, \mathbf{l} = l, \ldots)$ are the conditional probabilities of the category $i$ of $\mathbf{i}$ given particular values of the rest of the variables, $p_{i/j\bullet\bullet\cdots\bullet} = P(\mathbf{i} = i|\mathbf{j} = j)$ are the conditional probabilities of the category $i$ of $\mathbf{i}$ given the category $j$ of $\mathbf{j}$ in the two-way table collapsed on the rest of the predictors, and so on. The estimation of the probabilities is made in the usual way, for example, $\hat{p}_{ijkl\ldots} = \frac{f_{ijkl\ldots}}{N}, \hat{p}_{ij\bullet\bullet\cdots\bullet} = \frac{f_{ij\bullet\bullet\cdots\bullet}}{N}, \hat{p}_{i/jkl\ldots} = \frac{f_{ijkl\ldots}}{f_{\bullet jkl\ldots}}, \hat{p}_{i/j\bullet\bullet\cdots\bullet} = \frac{f_{ijkl\ldots}}{f_{\bullet jkl\cdots\bullet}}$.

## 3 Some results about collapsibility and conditional independence

In the analysis of a multi-way table, it is helpful to reduce the dimension of the table or convenient to look at the condensed (summed over certain variables) table. As we have seen at the introduction, the multi-way tables may be affected by what it is called " Simpson's Paradox" when the marginal tables are studied. Simpson's Paradox refers to the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group. The reversals are surprising because summing over a variable appears to be a form of average, but the attributes of the components are not preserved in the average. However, there are certain tables that do not exhibit this problem, in such cases it is advantageous to collapse the original table, especially when the observed frequencies are small in many of the cells of the table.

The condition to use the marginal tables to test for the association between variables is that the table should be collapsible over the variables that we have summed. There are many definitions of collapsibility in different contexts, but in this paper we refer to collapsibility in the sense that the relationship between the response and a set of predictors can be studied in the marginal tables collapsing (or summing) over the rest. This collapsibility can be characterized in terms of odds ratios as follows: A multi-way table is said to be *collapsible* over a set of factors if the marginal odds ratios are identical to the odds ratios in the complete table. For example, suppose that in the table $\Gamma = \{\mathbf{i}, \mathbf{j}, \mathbf{k}, \mathbf{l} \ldots\}$ we want to study the relation between the response $\mathbf{i}$ and a predictor $\mathbf{j}$ collapsing over the rest, the table is *collapsible* over $V\backslash\{j\}$ if

$$\frac{P_{ijklm\ldots}P_{i'j'klm\ldots}}{P_{i'jklm\ldots}P_{ij'klm\ldots}} = \frac{P_{ij\bullet\bullet\cdots\bullet}P_{i'j'\bullet\bullet\cdots\bullet}}{P_{i'j\bullet\bullet\cdots\bullet}P_{ij'\bullet\bullet\cdots\bullet}} \quad \forall i, i', j, j', klm\ldots$$

The condition can be characterized in terms of conditional independence: If the response $\mathbf{i}$ is conditionally independent of $V\backslash\{\mathbf{j}\}$ given

$\mathbf{j}(\mathbf{i} \amalg V\backslash\{\mathbf{j}\}/\mathbf{j})$, then it is possible to collapse over $V\backslash\{\mathbf{j}\}$ to study the relation between $\mathbf{i}$ and $\mathbf{j}$ (see for example, [3], [1]). The same result holds for any partition into two subsets of the set of predictors $V$.

Observe that this is the condition for the first step of CHAID to be applicable, for all $\mathbf{j} \in V$ the response has to be conditionally independent of $V\backslash\{\mathbf{j}\}$ given $\mathbf{j}$; this condition is very difficult to satisfy in practice.

Moreover, once the first split has been done, for example using predictor $\mathbf{j}$, the condition has to be fulfilled for every branch of the tree

$$\Gamma_{\mathbf{j}=j} = \{\mathbf{i}, \mathbf{j} = j, \mathbf{k}, \mathbf{l}, \mathbf{m}, \ldots\}, j = 1, \ldots, J.$$

## 4 Tests for marginal and conditional independence based on entropy

In information theory, entropy [13] is a measure of the uncertainty associated to a random variable in the sense that quantifies the amount of information contained in it. The unconditional entropy of the response is defined as

$$H(\mathbf{i}) = H(p_{1\bullet\bullet\bullet\cdots\bullet,\cdots,}p_{I\bullet\bullet\bullet\cdots\bullet}) = -\sum_{i=1}^{I} p_{i\bullet\bullet\bullet\cdots\bullet} \log p_{i\bullet\bullet\bullet\cdots\bullet}$$

and can be estimated using the observed marginal probabilities

$$\hat{H}(\mathbf{i}) = \hat{H}(\hat{p}_{1\bullet\bullet\bullet\cdots\bullet,\cdots,}\hat{p}_{I\bullet\bullet\bullet\cdots\bullet}) = -\sum_{i=1}^{I} \hat{p}_{i\bullet\bullet\bullet\cdots\bullet} \log \hat{p}_{i\bullet\bullet\bullet\cdots\bullet}$$

where $\hat{p}_{i\bullet\bullet\bullet\cdots\bullet} = \frac{f_{i\bullet\bullet\bullet\cdots\bullet}}{N}$. This is the target amount of information or uncertainty that has to be explained by the predictor $\mathbf{j}$.

Let $p_{i/j\bullet\bullet\cdots\bullet} = P(\mathbf{i} = i/\mathbf{j} = j)$ the conditional probability of the category $i$ of $\mathbf{i}$ given the category $j$ of $\mathbf{j}$. The conditional entropy of the response $\mathbf{i}$ given the predictor $\mathbf{j} = j$, is defined as

$$H(\mathbf{i}/\mathbf{j} = j) = -\sum_{i=1}^{I} p_{i/j\bullet\bullet\cdots\bullet} \log p_{i/j\bullet\bullet\cdots\bullet}$$

on average for all the categories of $\mathbf{j}$

$$H(\mathbf{i}/\mathbf{j}) = -\sum_{j=1}^{J} p_{\bullet j\bullet\bullet\cdots\bullet} \sum_{i=1}^{I} p_{i/j\bullet\bullet\cdots\bullet} \log p_{i/j\bullet\bullet\cdots\bullet}$$

and can be estimated by

$$\hat{H}(\mathbf{i}/\mathbf{j}) = -\sum_{j=1}^{J} \hat{p}_{\bullet j \bullet\bullet\cdots\bullet} \sum_{i=1}^{I} \hat{p}_{i/j\bullet\bullet\cdots\bullet} \log \hat{p}_{i/j\bullet\bullet\cdots\bullet}$$

where $\hat{p}_{\bullet j\bullet\bullet\cdots\bullet} = \frac{f_{\bullet j\bullet\bullet\cdots\bullet}}{N}$ and $\hat{p}_{i/j\bullet\bullet\cdots\bullet} = \frac{f_{ij\bullet\bullet\cdots\bullet}}{\hat{p}_{\bullet j\bullet\bullet\cdots\bullet}}$.

The difference between the unconditional and conditional entropies,

$$I(\mathbf{i}/\mathbf{j}) = H(\mathbf{i}) - H(\mathbf{i}/\mathbf{j})$$

is the amount of information explained by the variable $\mathbf{j}$, and can be estimated as

$$\hat{I}(\mathbf{i}/\mathbf{j}) = \hat{H}(\mathbf{i}) - \hat{H}(\mathbf{i}/\mathbf{j}).$$

In other words it is the amount of uncertainty reduced when the subjects are split into groups defined by the categories of $\mathbf{j}$. If the predictor explains the response perfectly, the conditional entropy is 0; that occurs when all the split groups, segments or nodes are " pure", i.e., when all the cases in a segment have identical values of the response.

It is known that CHAID operates on the two-way marginal tables using the chi-squared goodness of fit statistic. An alternative to test for marginal independence of the response and any of the variables ($\mathbf{i} \amalg \mathbf{j}$) would be to use the likelihood ratio test given by

$$G_M = 2\sum_{ij} f_{ij\bullet\bullet\cdots\bullet} \log \frac{f_{ij\bullet\bullet\cdots\bullet}}{\hat{f}_{ij\bullet\bullet\cdots\bullet}}$$

where $\hat{f}_{ij\bullet\bullet\cdots\bullet} = \frac{f_{i\bullet\bullet\cdots\bullet} f_{\bullet j\bullet\bullet\cdots\bullet}}{N}$ is the expected frequency under the hypothesis of marginal independence. The $G$ statistic is closely related to the reduction in the entropy

$$\begin{aligned}
G_M &= 2\sum_{ij} f_{ij\bullet\bullet\cdots\bullet} \log \frac{f_{ij\bullet\bullet\cdots\bullet}}{\frac{f_{i\bullet\bullet\cdots\bullet} f_{\bullet j\bullet\bullet\cdots\bullet}}{N}} = 2\sum_{ij} f_{ij\bullet\bullet\cdots\bullet} \log \frac{\frac{f_{ij\bullet\bullet\cdots\bullet}}{f_{\bullet j\bullet\bullet\cdots\bullet}}}{\frac{f_{i\bullet\bullet\cdots\bullet}}{N}} \\
&= 2\sum_{ij} f_{ij\bullet\bullet\cdots\bullet} \log \frac{f_{ij\bullet\bullet\cdots\bullet}}{f_{\bullet j\bullet\bullet\cdots\bullet}} - 2\sum_{ij} f_{ij\bullet\bullet\cdots\bullet} \log \frac{f_{i\bullet\bullet\cdots\bullet}}{N} \\
&= 2N\left[\sum_{j} \hat{p}_{\bullet j\bullet\bullet\cdots\bullet} \sum_{i} \hat{p}_{i/j\bullet\bullet\cdots\bullet} \log \hat{p}_{i/j\bullet\bullet\cdots\bullet} - \sum_{i} \hat{p}_{i\bullet\bullet\cdots\bullet} \log \hat{p}_{i\bullet\bullet\cdots\bullet}\right] \\
&= 2N\left[\hat{H}(\mathbf{i}) - \hat{H}(\mathbf{i}/\mathbf{j})\right] = 2N\hat{I}(\mathbf{i}/\mathbf{j}).
\end{aligned}$$

It is known that if $\mathbf{i}$ and $\mathbf{j}$ are independent $\mathbf{i} \amalg \mathbf{j}$, the statistic $G_M = 2N\hat{I}(\mathbf{i}/\mathbf{j})$ has a chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.

We can obtain an analogue result to test for conditional independence when we have a multi-way table. The entropy for the multi-way table considering the full tree is

$$H(\mathbf{i}/\mathbf{jkl}\ldots) = -\sum_{j=1}^{J} p_{\bullet jkl\ldots} \sum_{i=1}^{I} p_{i/jkl\ldots} \log p_{i/jkl\ldots}$$

that can be estimated as

$$\hat{H}(\mathbf{i}/\mathbf{jkl}\ldots) = -\sum_{j=1}^{J} \hat{p}_{\bullet jkl\ldots} \sum_{i=1}^{I} \hat{p}_{i/jkl\ldots} \log \hat{p}_{i/jkl\ldots}$$

where $\hat{p}_{\bullet jkl\ldots} = \frac{f_{\bullet jkl\ldots}}{N}$ and $\hat{p}_{i/jkl\ldots} = \frac{f_{ijkl\ldots}}{f_{\bullet jkl\ldots}}$. This is the smallest uncertainty we can get when all the predictors are present.

If we collapse (or sum) over the categories of a predictor, for example $\mathbf{j}$, the average entropy for the branches collapsed tree is

$$H(\mathbf{i}/\mathbf{kl}\ldots) = -\sum_{j=1}^{J} p_{\bullet\bullet kl\ldots} \sum_{i=1}^{I} p_{i/\bullet kl\ldots} \log p_{i/\bullet kl\ldots}$$

that can be estimated as

$$\hat{H}(\mathbf{i}/\mathbf{kl}\ldots) = -\sum_{j=1}^{J} \hat{p}_{\bullet\bullet kl\ldots} \sum_{i=1}^{I} \hat{p}_{i/\bullet kl\ldots} \log \hat{p}_{i/\bullet kl\ldots}$$

where $\hat{p}_{\bullet\bullet kl\ldots} = \frac{f_{\bullet\bullet kl\ldots}}{N}$ and $\hat{p}_{i/\bullet kl\ldots} = \frac{f_{i\bullet kl\ldots}}{f_{\bullet\bullet kl\ldots}}$. This is the amount of uncertainty left after eliminating the predictor $\mathbf{j}$ and should be higher than the uncertainty for the complete table because the information of $\mathbf{j}$ is no longer available. The amount of information lost by collapsing is then

$$\hat{I}(\mathbf{i}/V\backslash\{\mathbf{j}\}) = \hat{H}(\mathbf{i}/\mathbf{kl}\ldots) - \hat{H}(\mathbf{i}/\mathbf{jkl}\ldots)$$

Testing for conditional independence between the response and one of the predictors given the rest ($\mathbf{i} \amalg \mathbf{j}/V\backslash\{\mathbf{j}\}$) can help us to decide if the lost is statistically significant. The likelihood ratio statistics to test the hypothesis is

$$G_C = 2\sum_{ijkl} f_{ijkl} \log \frac{f_{ijkl\ldots}}{f_{ijkl\ldots}}$$

where $f_{ijkl\ldots} = \frac{f_{i\bullet kl\ldots} f_{\bullet jkl\ldots}}{f_{\bullet\bullet kl\ldots}}$ is the expected frequency under the conditional independence model, i.e., the expected frequencies in the two-way tables

for **i** and **j** resulting for each combination of the categories of the rest of the variables. The G statistic is closely related to the change in the entropy

$$\begin{aligned} G_C &= 2\sum_{ijkl\ldots} f_{ijkl\ldots}\log\left(\frac{\frac{f_{ijkl\ldots}}{f_{\bullet jkl\ldots}}}{\frac{f_{i\bullet kl\ldots}}{f_{\bullet\bullet kl\ldots}}}\right) \\ &= 2\left[\sum_{ijkl\ldots} f_{ijkl\ldots}\log\frac{f_{ijkl\ldots}}{f_{\bullet jkl\ldots}} - \sum_{ijkl\ldots} f_{ijkl\ldots}\log\frac{f_{i\bullet kl\ldots}}{f_{\bullet\bullet kl\ldots}}\right] \\ &= 2N\left(\sum_{j=1}^{J}\hat{p}_{\bullet jkl\ldots}\sum_{i=1}^{I}\hat{p}_{\frac{i}{jkl}\ldots}\log\hat{p}_{\frac{i}{jkl}\ldots}\right. \\ &\quad \left. -\sum_{j=1}^{J}\hat{p}_{\bullet jkl\ldots}\sum_{i=1}^{I}\hat{p}_{\frac{i}{jkl}\ldots}\log\ \hat{p}_{\frac{i}{jkl}\ldots}\right) \\ &= 2N\left(\hat{H}\left(\mathbf{i}\backslash\mathbf{lk}\ldots\right) - \hat{H}\left(\mathbf{i}\backslash\mathbf{lk}\ldots\right)\right) \\ &= 2N\hat{I}\left(\mathbf{i}/V\backslash\{\mathbf{j}\}\right). \end{aligned}$$

It is known that $G_C$ has a chi-squared distribution with $(I-1)(J-1)KL\ldots$ degrees of freedom.

## 5 Forward algorithm based on the concept of entropy

Using the results about the marginal independence, a forward algorithm can be defined. The algorithm is similar to the CHAID procedure except that uses the likelihood ratio to test for independence and entropy to evaluate the changes. The algorithm stars with the marginal distribution of the response and tries to explain the uncertainty of the response $\hat{H}(\mathbf{i}) = -\sum_{i=1}^{I}\ \hat{p}_{i\bullet\bullet\bullet\ldots\bullet}\log\ \hat{p}_{i\bullet\bullet\bullet\ldots\bullet}$ adding a predictor at each step. The procedure could be described as follows:

**Step 0:** The first node of segment includes all the subjects and the whole set $V$ of predictors.

**Step 1: Merging the categories**. Merge the categories according to the type of predictor as in CHAID. Use the likelihood ratio test rather than the goodness of fit test.

**Step 2: Searching for the best predictors to explain the response.** For each segment or node, we calculate the amount of uncertainty explained by each available predictor $\mathbf{j} \in V$

$$\hat{I}(\mathbf{i}\backslash\mathbf{j}) = \hat{H}(\mathbf{i}) - \hat{H}(\mathbf{i}\backslash\mathbf{j})$$

and an associated p-value from the statistic $G_M = 2N\hat{I}(\mathbf{i}\backslash\mathbf{j})$. Some corrections for multiple comparisons may be needed as in CHAID.

Select the significant predictor with highest $\hat{I}(\mathbf{i}\backslash\mathbf{j})$ as the best predictor. If there are no significant predictors stop, otherwise select the significant predictor with highest $\hat{I}(\mathbf{i}\backslash\mathbf{j})$ as the best predictor.

**Step 3: Splitting the node.** Split the node according to the categories of the best predictor.

**Step 4:** For each new segment, reduce the set of predictors in one and go to Step 1 until no significant contribution is found.

It has to be noted that this forward algorithm has the same problems as classical CHAID concerning the collapsibility conditions. It has been described here as a basis for comparison to the proposed backward algorithm, detailed in the next section. For more details, see Dorado (1998) [5].

## 6 Backwards algorithm based on conditional independence and entropy

Using the previous results it seems to be adequate to use conditional rather than marginal independence to study a multi-way contingency table with a response variable. We describe the algorithm, and some of the results used to construct it, in the following paragraphs.

**Step 1:** The starting point is the whole multi-way table represented by the full tree with all the variables concatenated. Remember that the concatenation order is irrelevant, so the right concatenation will be used for each conditional independence contrast.

**Step 2: Searching for the predictors independent from the response**. The variables, not providing a significant increment in the entropy $\hat{I}(\mathbf{i}\backslash V\{\mathbf{j}\})$, are eliminated. The significance is established using the likelihood ratio test for conditional independence. In case there are some candidates, the one with a bigger p-value is eliminated first and the process is repeated collapsing over the eliminated variable. Repeat the elimination until no predictor can be

removed. This is based in the fact that if the response **i** is conditionally independent of **j** given $V\backslash\{\mathbf{j}\}$ ($\mathbf{i} \amalg \mathbf{j}/V\backslash\{\mathbf{j}\}$), then it is possible to collapse over **j** to study the relation between **i** and V $\backslash\{\mathbf{j}\}$, then **j** can be removed.

**Step 3: Searching for the best predictors**. All the values for which the conditional independence is rejected are kept for analysis. The best predictor is the one having a higher significant increment of the entropy.

**Step 4: Setting the order of the split.** The order of the concatenation of the tree can be done according to the order of the increments, i.e., the best predictor is used for the first split, the next for the second and so on.

**Step 5: Pruning the tree in partial branches.** Having a conditional dependence for a particular predictor, means that the structure of each brand in the tree may not be the same. In this step we study each brand of the first split separately using the same procedure starting in Step 1. If the best predictor is j we have J new tables or branches

$$\Gamma_{\mathbf{j}=j} = \{\mathbf{i}, \mathbf{j} = j, \mathbf{k}, \mathbf{l}, \mathbf{m} \ldots\}, j = \mathbf{1}, \ldots, \mathbf{J}$$

to study. The procedure ends when no more pruning is possible.

**Step 6: Studying the symmetric branches in the final tree.** After repeating the previous steps as many times as necessary, we get a final tree where branches with the same structure can be found. In this case, simplifying by joining branches with a similar structure can be considered.

For full details, see Dorado (1998) [5].

## 7 Example

We have applied the proposed algorithm to identify some socio-economic profiles of women handling irregular jobs, i.e., work in the underground economy which is not (totally or partially) declared for tax and social security purposes and does not observe the pay and conditions laid down by sectorial collective agreements. Due to the difficulties to identify people with irregular works, because it's clandestine, and the lack of previous studies in the area, a snowball sampling has been used.

**Snowball sampling** is a special non-probability method used when the desired sample characteristic is rare. It may be extremely difficult or

cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. While this technique can dramatically lower search costs, it comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

The study has been carried out in Salamanca (Spain). Although Spanish government holds a large-scale study of the labour market, it does not have enough level of desegregation to study small areas. More details about the object and design of the study can be found in [7]

The response variable is the regularity situation (regular or irregular). The predictors considered are Immigration (Immigrant or National), Activity Sector (Agriculture, Domestic Services, Clothes Manufacture, Small Shops, Hotel and Catering, Other), Marital Status (Single, Partnered, Divorced/Widowed), Age ($< 30, 30 - 45, > 45$) and Education (Primary, Secondary and College).

The multidimensional table containing all the predictors and the response has 324 combinations of categories of the predictors from which 214 (66.05%) have been observed. This is usual when dealing with high dimensional tables in which many of the combinations have a very low probability of been observed. The entropy for the initial table was 0.537.

The first step of the procedure consists in searching for the predictors that are independent from the response. The only variable that produces a non-significant increment in the entropy after collapsing is the marital status and consequently it is removed. The initial tree has then 4 predictors (Immigration, Activity Sector, Age and Education). The entropy of the table is 0.568. The second step is splitting the tree according to the categories of the best predictor. The following table shows the increment in the entropy, the likelihood ratio statistics and the associated p-values for the conditional independence tests after removing each predictor.

| Variable | Entropy | G-statistic | d.f. | p-value |
|---|---|---|---|---|
| Immigration | 0.0204 | 110.20 | 54 | $1.0018 \times 10^{-05}$ |
| Activity Sector | 0.0424 | 229.9 | 90 | $4.1411 \times 10^{-14}$ |
| Age | 0.0290 | 156.91 | 72 | $2.9770 \times 10^{-08}$ |
| Education | 0.0253 | 136.93 | 72 | $6.19 \times 10^{-6}$ |

Table 3: Increment of the entropies and conditional independence tests for the initial tree.

The highest significant increment is for "Activity Sector", and this predictor is selected for the first split of the tree. The same procedure is

applied to every branch of the tree. We do not describe the whole process because its length. Figure 3 shows the final tree.

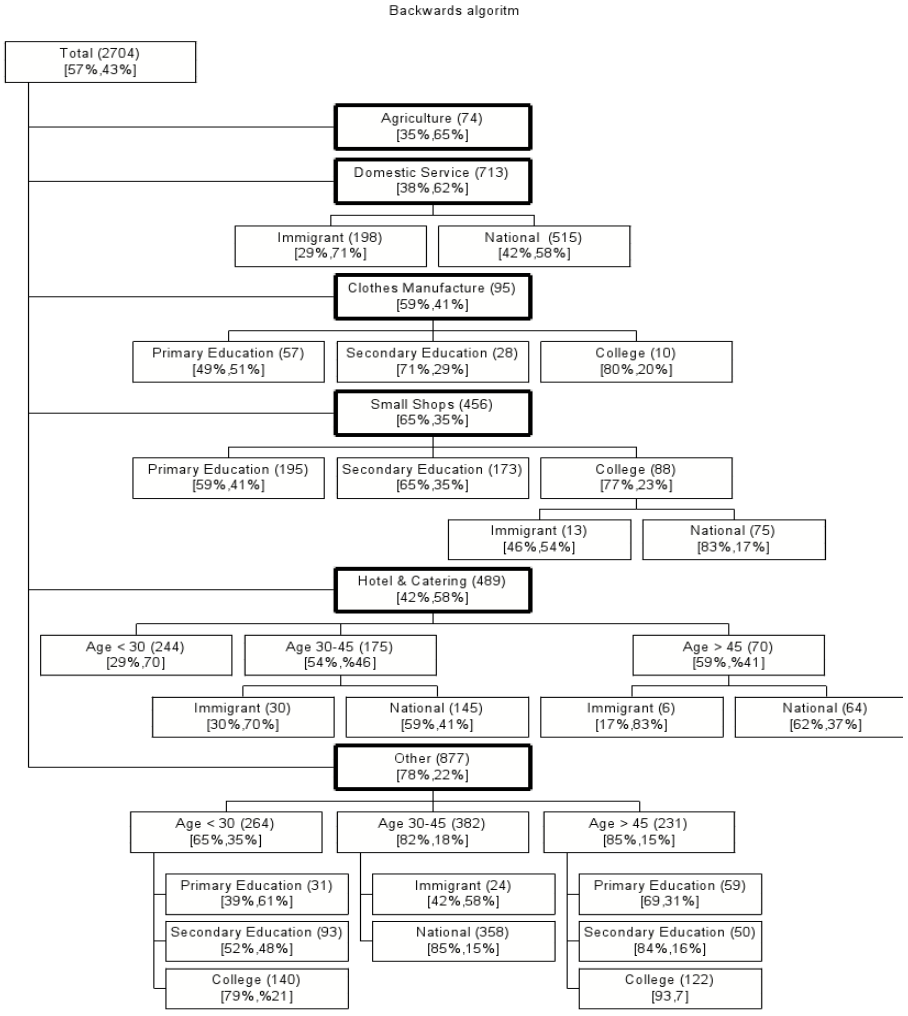


Figure 3: Final tree obtained from the backward procedure. Sample size is between parenthesis and percentages of regular and irregular workers are between brackets.

The entropy of the reduced tree, calculated as the weighted average of the entropy at each branch using the marginal probabilities as weights, is 0.586, i.e., we have an increment of 0,018 after reducing the tree from the

original 108 to the 23 terminal nodes.

Observe, for example that the Activity sector "Agriculture" has produced a terminal node meaning that in this sector no other variable can significantly explain the behaviour of the response. In the "Domestic service", the Immigration Status provides additional information about the response: the immigrants have a percent of irregular jobs (71%) higher than the nationals (58%). The rest of the tree can be interpreted in the same way (see Figure 3).

Using the terminal nodes as a rule for classification of the workers we obtain a 72% of the regular workers correctly classified and a 69% of the irregular; the overall percent of correct classification is 69% .

The tree permits obtaining a typology of the irregular workers: Works in the *Agriculture* and *Domestic Services* ; when she works in the *Clothes Manufacture* she has *No or Primary Education*; when she works in a small shop she has *College Education* and is an *Immigrant*; when she works in the *Hotel & Catering Business*, is young ($< 30$), and when is older ($> 30$) is also an Immigrant; for Other jobs the irregular worker is young ($< 30$) and with *Primary Education or medium age* (30-45) and Immigrant. In order to compare the proposed procedure to the classical, CHAID method has been used. Figure 4 shows the resulting tree.

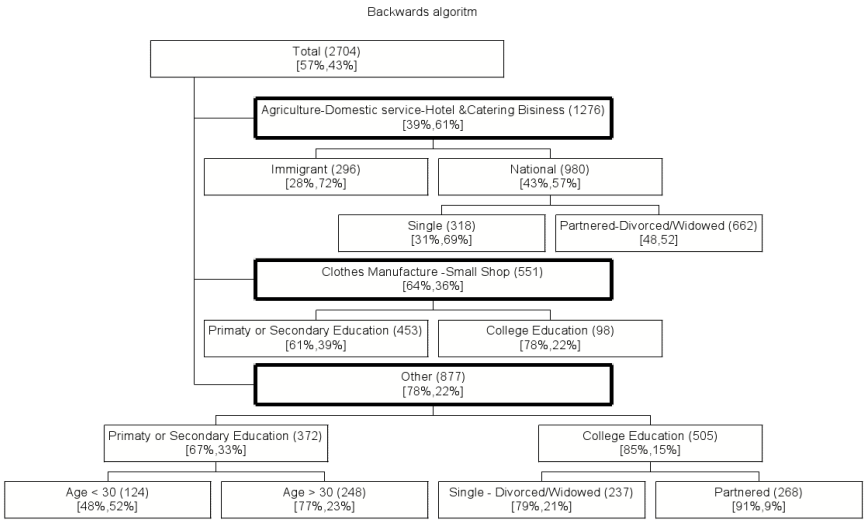


Figure 4: Final tree obtained from the CHAID procedure. Sample size is between parenthesis and percentages of regular and irregular workers is beween brackets.

The entropy of the reduced tree is 0.596, we have an increment of 0,026 after reducing the tree from the original. The increment is higher than the increment for the backward procedure. The percent of correctly classified workers is now 67%.

The resulting tree (see Figure 4) is quite different from the previous one, it includes the variable Marital Status that was collapsed in the backwards procedure. The indices of goodness of fit (entropy lost and classification rate) are slightly worse, but the trees are not completely comparable because have different number of branches. Some further investigation to make the two trees comparable. Van Diepen & Franses [15] use bootstrap methods in CHAID.

## References

[1] Ávila, C.A. (1996) *Una Alternativa al Análisis de Segmentación Basada en el Análisis de Hipótesis de Independencia Condicionada*. Tesis Doctoral, Universidad de Salamanca.

[2] Baron, S.; Phillips, D. (1994) "Attitude survey data reduction using CHAID: an example in shopping centre market research", *Journal of Marketing Management* **10**: 75–88.

[3] Christensen R. (1990) *Log-Linear Models*. Springer-Verlag, New York.

[4] Clark, W.A.V.; Duerloo, M.C.; Dieleman, F.M. (1991) "Modeling categorical data with chi square automatic interaction detection and correspondence analysis ", *Geographical Analysis* **23**: 332–345.

[5] Dorado, A. (1998) *Métodos de Búsqueda de Variables Relevantes en Análisis de Segmentación: Aportaciones desde una Perspectiva Multivariante*. Tesis Doctoral, Universidad de Salamanca.

[6] Dorado, A.; Galindo. P.; Vicente, J.L.; Vicente-Tavera, S. (2002) "El CHAID como herramienta de marketing politico", *Esic Market* **111**: 129–140.

[7] Galindo, M. P.; Vicente-Galindo, P.; Patino-Alonso,C ; Vicente-Villardón, J. L. (2007) "Caracterización multivariante de los perfiles de las mujeres en situación laboral irregular: el caso de Salamanca", *Pecunia* **4**: 49–79.

[8] Kass, G.V. (1980) "An exploratory technique for investigating large quantiles of categorical data", *Applied Statistics* **29**: 119–127.

[9] Malchow, H. (1997) "The targeting revolution in political direct contact", *Campaigns & Elections* **18**: 51–66.

[10] Magidson, J. (1990) "CHAID, LOGIT and Log-linear Modelling", *Marketing Information Systems.* Datrapo Report **IM11-130**: 101–115.

[11] Marques, P.; Tippetts, A.; Voas, R.; Beirness, D. (2001) "Predicting repeat DUI offenses with the alcohol interlock recorder", *Accident–Analysis–and–Prevention* **33(5)**: 609–619.

[12] Mckenney, C. (2000) *Women Chief Academic Officers of Public Community Colleges: Career Paths and Mobility Factors*. Ed. Texas Tech University.

[13] Shannon, C.E.; Weaver, W. (1949/1963) *The Mathematical Theory of Communication*. University Illinois Press, Urbana and Chicago.

[14] Simpson, E.H. (1951) "The interpretation of interaction in contingence tables", *Royal Statistical Association* **13B**: 238–241.

[15] Van Diepen, M.; Franses, P.H. (2006) "Evaluating chi-squared automatic interaction detection", *Information Systems* **31**(8): 814–831.