CHAID and Earlier Supervised Tree Methods

Chapter · November 2013

CITATIONS
59

READS
1,209

1 author:

Gilbert Ritschard
University of Geneva
203 PUBLICATIONS **3,652** CITATIONS

SEE PROFILE

Please cite as: Ritschard, G. (2013). CHAID and Earlier Supervised Tree Methods. In J.J. McArdle & G. Ritschard (eds), *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences*, Routeledge, New York, pages 48–74

CHAID and Earlier Supervised Tree Methods

Gilbert Ritschard

Abstract

The aim of this paper is twofold. First we discuss the origin of tree methods. Essentially we survey earlier methods that led to CHAID (Kass, 1980; Biggs et al., 1991). The second goal is then to explain in details the functioning of CHAID, especially the differences between the original method as described in Kass (1980) and the nowadays currently implemented extension that was proposed by Biggs et al. (1991).

Keywords: Classification tree, regression tree, recursive partitioning, CHAID, AID, THAID, ELISEE.

Classification and regression trees, also known as recursive partitioning, segmentation trees or decision trees, are nowadays widely used either as prediction tools or simply as exploratory tools. Their interest lies mainly in

their capacity to detect and account for non linear effects on the response variable, and specially of even high order interactions between predictors.

The aim of this paper is to explain in details the functioning of the CHAID tree growing algorithm as it is implemented for instance in SPSS (2001) and to draw the history of tree methods that led to it. We start with this latter point.

1 AID, THAID, ELISEE and Other Earlier Tree Growing Algorithms

The first methods for inducing trees from data appeared in the survey analysis domain and were mainly developed by statisticians. We give here a short presentation of these earlier methods. See also the nice survey by Fielding and O'Muircheartaigh (1977).

Tree growing, also known as "hierarchical splitting", "partitioning", "group dividing" or "segmentation" , finds its origin in the analysis of survey data. Perhaps the first published proposal is the one by Belson (1959). He addresses a matching issue that is in fact just a predictive one: Predicting the outcome for a second group given the outcome observed for the first one. Predictors as well as the outcome variable are dichotomized and the growing criterion used is the difference (for one of the two outcome categories) between the observed count and the number expected under the no association assumption. Other earlier proposals include those by Morgan and Sonquist (1963) who

proposed the AID (Automatic Interaction Detector) algorithm for growing a binary regression tree, i.e., one in which the outcome variable is quantitative, and by Cellard, Labbé, and Savitsky (1967) who proposed ELISEE (Exploration of Links and Interactions through Segmentation of an Experimental Ensemble), which is a binary method for categorical dependent variables. The former regression tree method was popularized thanks to the AID computer programme developed at Ann Arbor by Sonquist, Baker, and Morgan (1971), while the latter segmentation method was popularized by Bouroche and Tenenhaus (1970, 1972). AID was certainly the most famous of these earlier programmes and Sonquist (1969) showed its interest as a complementary tool with multiple correlation analysis. See Thompson (1972) for a thorough comparison of AID and Belson's method. Messenger and Mandell (1972) and Morgan and Messenger (1973) extended AID for categorical outcome using a so called theta criterion, which resulted in THAID (THeta AID). Gillo (Gillo, 1972; Gillo and Shelly, 1974) extended AID for multivariate quantitative outcome variables (MAID). Press, Rogers, and Shure (1969) developed an interactive tree growing tool allowing multibranching, IDEA (Interactive Data Exploration and Analysis). Independently from those developments in the survey data analysis framework, Hunt (see Hunt, Marin, and Stone, 1966) has proposed a series of decision tree induction algorithms called *Concept Learning Systems* (CLS-1 to CLS-9). Those algorithms were explicitly developed with an Artificial Intelligence perspective and are primarily intended for doing classification, i.e., for categorical response variables. CLS-1

to CLS-8 build binary trees, while the latter CLS-9 allows multibranching.

1.1 Motivation of the earlier methods

The motivation behind these first approaches is essentially to discover how the outcome variable is linked to the potential explanatory factors and more specifically to special configurations of factor values. If we except Hunt, Authors are mainly interested in finding alternatives to the restrictions of the linear model, in which the effect of the explanatory variables are basically additive, i.e the effect of any variable is independent of the value taken by other ones. The primary concern is thus to detect important interactions, not for improving prediction, but just to gain better knowledge about how the outcome variable is linked to the explanatory factors.

> "Particularly in the social sciences, there are two powerful reasons for believing that it is a mistake to assume that the various influences are additive. In the first place, there are already many instances known of powerful interaction effects—advanced education helps a man more than it does a woman when it comes to making money, [...] Second, the measured classifications are only proxies for more than one construct. [...] We may have interaction effects not because the world is full of interactions, but because our variables have to interact to produce the theoretical constructs that really matter." (Morgan and Sonquist, 1963, p

416.)

> "Whenever these interrelationships become very complex—containing non linearity and interaction—the usefulness of classical approaches is limited." (Press, Rogers, and Shure, 1969, p 364.)

It is interesting to look at the application domains considered in these earlier works. Indeed the need for such tree approaches followed the spread of survey analyses and the increasing place taken by empirical analyses within social sciences in the late 50's. Belson (1959), for instance, developed his method for analysing survey data collected by the BBC for studying "the effects of exposure to television upon the degree to which individuals participate with others in the home in their various household activities." He grows thus a tree for "High" and "Low" degree of joint activity. Later, in Belson (1978) he used a tree approach for investigating causal hypotheses concerning delinquent behavior. Morgan and Sonquist (1963) illustrate their AID method with an analysis of the differences in living expenses of families, on which "age and education can obviously not operate additively with race, retired status, and whether the individual is a farmer or not." Ross and Bang (1966) resort to AID for highlighting differences in family profiles that may explain differences in their chances of adoption. In the same vein, Orr (1972) makes an in depth study of transitions proportions between successive academic stages using AID on data from the National Survey of Health and Development. Interestingly in these latter applications, the response is

a binary variable that is treated as quantitative. AID was also applied for instance in psychology (Tanofsky et al., 1969) and marketing (Armstrong and Andress, 1970; Assael, 1970). Press et al. (1969) illustrate their IDEA interactive technique with an example of questionnaire data about participation in an uprising. Gillo and Shelly (1974) discuss their multivariate tree method with an example where the outcome vector is the pattern of overall, job and leisure satisfactions. Cellard et al. (1967) is an exception in that their illustrations are not in the social science field. They present two applications, one in engineering where they are interested in the effects of different technical and environmental factors on the gripping of locomotive engines and one in marketing where they attempt at characterizing standards expected by foreign hosts of French hotels. Nonetheless, their concern is still describing links and interactions and not making prediction or classification.

1.2 Splitting Criteria

The focus being on effects and interaction effects, the aim of these earlier methods was primarily to segment the data into groups with as much different as possible distributions of the outcome variable. Therefore, the splitting criteria considered naturally are measures of the association strength between the outcome and split variables. This contrasts with more recent methods such as CART and C4.5 for instance, which are primarily oriented towards classification and prediction and attempt therefore to maximize the homogeneity of each group by means of purity measures. Note that the clas-

sification concern is, nevertheless, somehow implicitly present behind some of these earlier methods. The θ measure used in THAID is for instance just the classification error rate, and two (Shannon's entropy and Light and Margolin (1971)'s variance of a categorical variable) among the three alternatives mentioned by Messenger and Mandell (1972) also are some kinds of purity measures.

The splitting criterion used depends indeed on the nature of the variables considered. Belson (1959) dichotomizes all variables. He looks at the 2×2 contingency table that cross tabulates each explanatory factor with the outcome variable and compares this table with the one expected under the independence hypothesis. Except for its sign, the deviation is the same for all 4 cells. Belson's association criterion consists thus just in one of these deviance.

[Table 1 about here.]

In AID, the dependent variable is quantitative and the splitting criterion proposed in Morgan and Sonquist (1963) is the largest reduction in unexplained sum of squares. The latter is commonly known as the residual or within sum of squares, WSS, in analysis of variance. It reads

$$\text{WSS} \;=\; \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

where $\bar{y}_j$ is the mean value of the y_{ij}'s in node j. Here it is the WSS for the $g = 2$ groups that would be produced by the split. Maximizing this reduction

is equivalent to maximize the η^2 coefficient, i.e., the ratio BSS/TSS, where TSS is the total sum of squares (before the split)

$$\text{TSS} \;=\; \sum_{j=1}^{g}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2 \;\;,$$

which is independent of the split variable, and BSS = TSS − WSS the resulting between sum of squares. Hence, it is some sort of R^2 association measure. MAID (Gillo and Shelly, 1974) uses a generalized version of this criterion applicable in the multivariate case. The proposed generalization is indeed a variant of Wilks' Λ, namely $1 - k/\operatorname{tr}(\mathbf{TW}^{-1}) = \operatorname{tr}(\mathbf{BW}^{-1})/\operatorname{tr}(\mathbf{TW}^{-1})$, where $\mathbf{T}$, $\mathbf{W}$ and $\mathbf{B} = \mathbf{T} - \mathbf{W}$ are respectively the total, within and between cross product matrices among the k dependent variables.

Kass (1975) introduces a statistical significance criteria for AID, namely the p-value of the BSS/TSS ratio that he evaluates through a distribution-free permutation test. A Chi-square approximation of this test is proposed in Scott and Knott (1976).

In the interactive IDEA system, Press et al. (1969) consider also a scaled indicator of the proportion of explained variation for the case of quantitative outcome. For categorical outcome variables they resort to the significance of the independence Pearson Chi-square for the table that cross-tabulates the predictor with the outcome discrete variable. ELISEE is also based on the Chi-square. Its authors consider indeed the squared ϕ, which is some normalized Chi-square obtained by dividing it by the number of cases. Since

ELISEE considers only binary splits, the table that cross-tabulates at each node the split with the outcome has always only two columns. The ϕ criterion is in such cases equivalent to Cramer's v.

CHAID (Chi-square Automatic Interaction Detector), introduced by Kass (1980) as an evolution of AID and THAID, is certainly nowadays the most popular among these earlier statistical supervised tree growing techniques. We describe its functioning in details hereafter.

2 CHAID

As indicated by its name, CHAID uses a Chi-square splitting criterion. More specifically, it uses the p-value of the Chi-square. In his 1980 paper in *Applied Statistics*, Kass discusses only the case of a categorical dependent variable. The method is, nevertheless, most often implemented with an option for handling also quantitative dependent variables. The criteria is in that case the p-value of the F statistic for the difference in mean values between the g nodes generated by the split:

$$F = \frac{\mathrm{BSS}/(g-1)}{\mathrm{WSS}/(n-g)} \quad \sim \quad F_{(g-1),(n-g)} \quad .$$

An alternative could be using the Kass (1975)'s permutation test or its χ^2 approximation (Scott and Knott, 1976).

The main characteristics of CHAID that contributed to its popularity are:

1. At each node, CHAID determines for each potential predictor the optimal n-ary split it would produce, and selects the predictor on the basis of these optimal splits.

2. CHAID uses p-values with a Bonferroni correction as splitting criteria.

Resorting to p-values as growing criteria provides stopping rules that automatically account for statistical significance. Thresholds are naturally set to usual critical values considered for statistical significance, namely 1%, 5% or 10%. Such p-value criteria are sensitive to the number of cases involved in the split and tend to avoid splitting into too small groups. Note that though CHAID popularized the idea of using p-values for selecting predictors, it was not the first attempt to do so. As mentioned above, p-values were also used by Press et al. (1969) in their IDEA system. The originality of the method proposed by Kass is, however, that for evaluating the significance of each split it applies first a Bonferroni correction on the p-value. We will explain later in what this consists.

The most original contribution of CHAID is no doubt the first point, i.e., the idea of looking at the optimal n-ary split for each predictor. Firstly, as shown from Table 1, most methods considered only binary splits. Those that allow for n-ary splits, such as the method by (Hunt et al., 1966), set the number of splits to the number of categories of the potential predictor, which could be particularly unsuited for predictors with many categories.

We explain how CHAID works by means of a real world example data

set. Let us first describe these data.

2.1 Illustrative data

We consider administrative data about the 762 first year students who were enrolled in fall 1998 at the Faculty of Economic and Social Sciences (ESS) of the University of Geneva (Petroff et al., 2001). The data will be used to find out how the situation (1. eliminated, 2. repeating first year, 3. passed) of each student after the first year is linked to her/his available personal characteristics. The response variable is thus the student situation in October 1999. The predictors retained are birth year, year when first time registered at University of Geneva, chosen orientation (Social Sciences or Business and Economics), type of secondary diploma achieved (classic/Latin, scientific, economics, modern, technical, non Swiss, none, missing), place where secondary diploma was obtained (Geneva, Switzerland outside Geneva, Abroad), age when secondary diploma was obtained, nationality (Geneva, Swiss except Geneva, Europe, Non Europe) and mother's living place (Geneva, Switzerland outside Geneva, Abroad).

2.2 Optimal n-ary split

Consider the contingency Table 2 between the dependent variable (situation after 1st year) and the type of secondary diploma. The latter predictor has $c = 8$ categories, namely: classic or Latin, scientific, economics, modern,

technical, non Swiss, no secondary diploma and missing. Ideally, we would like to look at all possibilities of segmenting the population by means of these 8 categories. The number of such possibilities is given by the number of ways of splitting into 2 groups, plus the number of ways of splitting into 3 groups, and so on until the one way of splitting into c groups.

[Table 2 about here.]

When, as in our case, the predictor is purely nominal, i.e., there is no restriction as to how to partition the c categories, the total number of ways of partitioning them is known as Bell (1938)'s number $B(c)$ and may be obtained through the recursive formula:

$$B(c) = \sum_{g=0}^{c-1} \binom{c-1}{g} B(g)$$

with $B(0)$ set equal to 1. Alternatively, $B(c)$ may be expressed as

$$B(c) = \sum_{g=1}^{c} S(c,g)$$

where $S(c,g)$ is the Stirling number of the second kind giving the number of ways of splitting c values into g groups

$$S(c,g) = \sum_{i=0}^{g-1} \frac{(-1)^i (g-i)^c}{i!(g-i)!} \quad . \tag{1}$$

For our 8 categories, this gives $B(8) = 4140$ possibilities. We get the number of segmentation possibilities by subtracting 1 to this number since grouping all 8 categories into a single class would not be a split. This gives 4139 possibilities.

In the case of an ordinal predictor where the categories are naturally ordered, the groups should be constituted by contiguous categories only. The number of such partitions is

$$G(c) \;=\; \sum_{g=0}^{c-1} \binom{c-1}{g} \;=\; 2^{(c-1)} \;\; .$$

For $c = 8$, we would have $G(8) = 128$, and hence $128-1 = 127$ possibilities of segmentation. This is much less than in the nominal case, but still increases exponentially with the number of categories c.

To avoid scanning all possibilities, Kass (1980, p 121) suggests the following heuristic:

Step 1 "Find the pair of categories of the predictor (only considering allowable pairs as determined by the type of the predictor) whose $2 \times r$ sub-table (r being the number of categories of the dependent variable) is least significantly different. If this significance does not reach a critical value, merge the two categories, consider this merger as a single compound category, and repeat this step."

Step 2 "For each compound category consisting of three or more of the original categories, find the most significant binary split (constrained by

the type of the predictor) into which the merger may be resolved. If the significance is beyond a critical value, implement the split and return to Step 1."

To illustrate on our example, we provide in Table 3 the Pearson Chi-square computed for each pair of the original $c = 8$ categories.

[Table 3 about here.]

The smaller Chi-square of .06 is obtained for the pair $\{5, 7\}$, i.e., for technical diploma and no secondary diploma. The corresponding 3×2 table is shown as Table 4. The degrees of freedom are $df = (3-1)(2-1) = 2$, and the p-value of the obtained Chi-square is $p(\chi^2_2 \geq = .06) = 96.7\%$. In the general case where the dependent variable has r categories, the p-values are obtained from a Chi-square distribution with $r - 1$ degrees of freedom.

[Table 4 about here.]

We repeat then the same process on the 3×7 table obtained by replacing columns 5 and 7 of Table 2 with the single merged column {5,7}. The pairwise Chi-squares and their p-values are given in Table 5. The last significant Chi-square is .34 (p-value = 84.6%) and corresponds to the pair {3,8}, i.e., scientific and missing. We merge these two categories and iterate until we get significant Chi-squares for all remaining pairs.

[Table 5 about here.]

The successive merges are recapitulated in Table 6. The process ends at iteration 6, with the Chi-squares and p-values depicted in Table 7.

[Table 6 about here.]

[Table 7 about here.]

The next step (*Step 2*) is to check that none of the compound categories formed by more than 2 categories ({1,3,8} and {5,6,7}) can be significantly dichotomized. This is a top-down way of testing the homogeneity of the groups obtained in Step 1 in a bottom-up manner. For group {1,3,8}, we have to check that splitting it into {1,3} and {8} is not significant. The other possibility has already been considered in the third row of Table 6. Likewise, for group {5,6,7}, we have to check that splitting into {5,6} versus {7} is not significant, the other possibility corresponding to the last row in Table 6. The results are

Split	Chi-square	p-value	
{1,3} versus {8}	.39	82.1%	non significant
{5,6} versus {7}	2.28	19.4%	non significant

Hence, there is no significant way of splitting the three groups obtained. The best way of segmenting the whole data by means of the type of secondary diploma is thus according to the heuristic: {classic/Latin, scientific, missing}, {modern, economics} and {technical, non Swiss, no secondary diploma}.

2.3 Ordinal and floating predictors

We do not detail here the grouping process for all other predictors. It is worth mentioning, however, that 'Birth year', 'Year when 1st registered' and 'Age at secondary diploma' have ordered categories. Such predictors, are called *ordinal* or *monotonic* and receive a special treatment allowing only merges of adjacent categories. In fact, beneath purely *nominal* (also called *free*) and *ordinal* (*monotonic*) predictors, Kass (1980) distinguishes also a third type of variable which he calls *floating predictor*. This refers to variables that have ordered categories except for one of them (usually the missing value), which cannot be positioned on the ordinal scale. Such predictors are dealt with like ordinal ones except for the floating category which is allowed to merge with any other category or compound category. 'Year when first registered', which has one missing value, is such a floating predictor. It takes 11 different ordered year values and one missing category. 'Age at secondary diploma' would also have been such a floating predictor because of those few students that were exceptionally accepted at the ESS Faculty without having obtained their secondary diploma. However, it was already recoded into 4 categories in the data set we used, students 'without secondary diploma' being grouped with the older age class. Hence, this 4 category predictor is simply considered has an ordinal one.

For illustrating how floating predictor are handled, let us detail the merging process for the 'Year when first registered' variable. Its cross tabulation with the target variable is shown in Table 8. For the first four older enrollment

dates the distribution is the same: all concerned students were eliminated at the end of the first year. The four categories being adjacent, we merge them together. After this merge, the mere Chi-squares to be considered are shown in Table 9. Empty cells correspond to non adjacent categories, and hence to unfeasible merges. Note that there are no such empty cells in the row and column corresponding to the 'missing' category. The table clearly exhibits the reduction in the number of pairs to examine that is achieved by taking the order of categories into consideration.

[Table 8 about here.]

[Table 9 about here.]

The first suggested merge is to put the missing value with the group of the older registration dates (≤ 91). The Chi-square is 0, since the distribution is the same for 'missing' and '≤ 91'. Without showing here the Chi-squares recomputed for accounting for the previous merged categories, it seems quite obvious that the next merges will be '92' with '93', then the resulting group with '94', and so on. We leave to the reader to check that the merging process ends up with two groups, namely '≤ 97 or missing' and '98', the latter being the group of those students who enrolled for the first time when they started their first year at the ESS Faculty. The final Chi-square is 18.72 for 2 degrees of freedom and its p-value 0.00863%, making it clear that the two final groups have dissimilar distributions.

2.4 Growing the tree

Having explained how the categories of each predictor are optimally merged at each node, we can now describe the tree growing process. We start with a simple growing scheme based on the classical non adjusted p-value of the independence Chi-square for the table that cross tabulates, at the concerned node, the target variable with the predictor. We will afterwards explain the Bonferroni adjustment proposed by Kass (1980) and show that using it may generate a different tree.

[Table 10 about here.]

Table 10 summarizes the results obtained at the root node by applying the merging heuristic to the eight considered predictors. The Chi-square reported is the one for the cross-tabulation between the dependent target variable (*Status after 1st year*) and the concerned predictor with optimally merged categories. The number of the latter ones is indicated under #*splits*. Since we have three statuses (eliminated, repeating, passed) for the dependent variable, the number of degrees of freedom df is in each case $(3-1)(\#\text{splits}-1)$. This information is the one considered for choosing the best first split in the tree growing process.

Table 10 reveals that the *Type of secondary diploma* grouped into the three groups found above has the smallest p-value. It is thus the most significantly linked with the *Status after 1st year* and hence the most discriminating factor. We select it for generating the first split.

[Figure 1 about here.]

We iterate then the process at each of the resulting nodes. Figure 1 shows the tree we obtain. At level 1 we notice that a different splitting variable is used for each node. The best splitting variable for the group formed by the 248 students that have a secondary diploma in the classic/Latin or scientific domain is the ordinal birth year variable with its values merged into three birth year classes, namely 1976 and before, 1977-78, and after 1978. These birth year classes correspond to age classes '22 and more', '20-21' and 'less than 20' for students in 1998 when they started their first year at the ESS Faculty. The 303 students with a modern or economic oriented secondary diploma are split according to their chosen orientation in the ESS Faculty, which takes only two values. Finally, the remaining 211 students are distinguished according to their enrollment date. Indeed, the distinction is between those for whom the first year at the ESS Faculty was the first year at the University of Geneva from those who spent already at least one year at the University.

We may wonder why the tree growing stopped after level 3. This is because each leaf (terminal node) met at least one stopping criterion. With CHAID, four stopping rules are classically considered:

- an α_{split} threshold for the splitting p-value above which CHAID does not split the node (was set to 5%);
- a maximal number of levels $\max_{level}$ (was set to 5);

- a minimal parent node size $\min_{parent}$ (was set to 100), meaning that CHAID does not try to split a node with less than $\min_{parent}$ cases;
- a minimal node size $\min_{node}$ (was set to 50), meaning that CHAID considers only splits into groups with each at least $\min_{node}$ cases.

In the tree of Figure 1 there are seven out of nine leaves with less than 100 cases, which meet thus our $\min_{node} = 100$ constraint. For the two remaining ones, CHAID found no further significant split, all p-values exceeding the $\alpha_{split} = 5\%$ threshold.

2.5 Bonferroni adjustement

As already mentioned in the introduction of this section 2, one of the main characteristics of CHAID is the use of Bonferroni adjusted p-values. The aim of the Bonferroni adjustment is to account for multiple testing. For instance, at level 1 we have seen that the p-value for the optimally merged 'Type of secondary diploma' predictor is 0.000000000035. Since this corresponds to the optimal n-ary grouping of the categories, the p-value for any other possible partition should be greater than this value. This supposes that we have also tested these other solutions, hence the multiple tests. For the best split to be non statistically significant, all other possibilities should also be non significant. Assuming independence between the m tests and a same type I error probability α for each of them, the total type I error probability is $m\alpha$. To ensure a total probability of α, the Bonferroni correction consists thus in

lowering the critical value α for the sole test considered by dividing it by the number m of underlying tests, or alternatively, what CHAID does, by multiplying the p-value of the optimal solution by m. Since it ignores dependences between tests, the Bonferroni correction is conservative and known to be often much too restrictive (Abdi, 2007). Furthermore, for our splitting issue, it is not so evident to determine the number of tests to take into consideration.

Kass (1980) proposes as an approximation to set the Bonferroni multiplier m to the number of ways a c category predictor can be reduced to g groups, g being the final number of optimally merged categories. The way of calculating m depends indeed on the nature of the predictor. We give hereafter the formulae used by the classical CHAID algorithm for each of the three types of predictors.

Purely nominal (free) predictors. The number m of ways of partitioning c categories into g groups is given by the Stirling number of the second kind (1)

$$m \;=\; \sum_{i=0}^{g-1} \frac{(-1)^i (g-i)^c}{i!(g-i)!} \;\; .$$

For instance for partitioning $c = 4$ categories a, b, c, d into $g = 2$ groups, there are $m = (2^4/2) + (-1) = 7$ possibilities, namely the seven ones depicted in Table 11

[Table 11 about here.]

Ordinal (monotonic) predictors. Here the groups are non overlapping subsequences of categories. The first such subsequence starts necessarily with the first category. Hence m is number of ways we can chose the starting points of the other $g-1$ subsequences among the $c-1$ remaining categories, that is

$$m = \binom{c-1}{g-1} .$$

For $c = 4$ ordered categories a, b, c, d, there are thus $m = 3!/2 = 3$ possibilities to partition them into 2 groups, namely $\{a, bcd\}$, $\{ab, cd\}$ and $\{abc, d\}$.

Floating predictors

$$m = \binom{c-2}{g-2} + g\binom{c-2}{g-1} = \frac{g-1+g(c-g)}{c-1}\binom{c-1}{g-1} .$$

Thus 3 ordered categories a, b, c and one floating value f can be grouped in $m = (1 + 2 \cdot 2)/3)3 = 5$ ways into 2 groups, namely the 5 possibilities shown in Table 12.

[Table 12 about here.]

Table 13 gives the Bonferroni adjusted p-value for each of our 8 predictors at the root node. The table should be compared with the values in Table 10 page 52. The *rank* column in Table 13 reveals changes in the ranking of the

predictors as compared to what resulted from the unadjusted p-values. With a multiplier of 966 the 'Type of secondary diploma' is now only the third best predictor, while the retained splitting variable will here be 'Where secondary diploma', which has 3 categories only.

[Table 13 about here.]

[Figure 2 about here.]

This example shows clearly that predictors with many categories tend to be more penalized by this Bonferroni correction than predictors with few categories. Also, nominal predictors are more penalized than ordinal or floating predictors. This is somehow justified since the more there are split possibilities, the greater the chance to find the best split among them. Nevertheless, the correction seems to be often excessive, and hence reverses the bias in favor of predictor offering few splitting alternatives.

Using Kass proposition of Bonferroni correction, we get the grown tree shown in Figure 2. The tree is quite different from that (Figure 1) obtained with uncorrected p-values. First, it looks less complex. This is indeed a direct consequence of the implicit reinforcement of the α_{split} stopping rule that follows from the Bonferroni adjustment. Secondly, as depicted by the detailed table for the first split (Table 13), applying the correction may change at each node the ranking of the predictors, hence the used splitting predictor.

3 Exhaustive CHAID

Biggs, De Ville, and Suen (1991) propose two important improvements to the CHAID method.

1. A more thorough heuristic for finding, at each node, the optimal way of grouping the categories of each predictor.

2. A better suited approximation for the Bonferroni correction factor that avoids discriminating nominal variables with a large number of categories.

The improved CHAID method is commonly known as Exhaustive CHAID. The name is confusing however since though it is more thorough, the search is still not exhaustive. Furthermore, it refers only to the first of the two improvements, while differences between CHAID and Exhaustive CHAID trees result more often from the alternative Bonferroni adjustment.

3.1 Extended search for the optimal g-way split

Kass (1980)'s heuristic merges iteratively pairs of similar distributed columns, considering indeed only pairs allowed by the nominal, ordinal or floating nature of the predictor, until no pair can be found with a non significant associated Chi-square. Biggs et al. (1991) propose to pursue the merging process until we obtain only two groups. The best g-way split is then determined by seeking among the such determined successive best $c - 1$ groupings into c,

$c-1$, ..., 2 classes, the one for which the cross-tabulation with the response variable has the most significant Chi-square.

The floating category of a floating predictor is also handled differently. Kass's heuristic considers merges involving the floating category at each iteration, while Biggs et al's proposition is to first seek for the best merge of the ordered categories and only afterwards chose between merging the floating category with the most alike group or keep it as a category per se.

[Table 14 about here.]

To illustrate, consider again our example of the "Type of secondary diploma" at the root node. With Kass's heuristic we stopped with the Table 7 page 49 in which all Chi-squares are statistically significant. Biggs et al's heuristic goes one step further and merges $\{2, 4\}$ with $\{5, 6, 7\}$, that is it considers also the best grouping into two classes, namely {classic/Latin, scientific, miss.} and {modern, economics, technical, non Swiss, none}. It looks then at the p-value for the cross tabulation of each grouping with the response variable. There are $c-1$ such tables with respectively c, $c-1$, ..., 2 columns. Table 14 reports these values. The lowest p-value is attained for $g = 4$ groups, which differs from the partitioning into 3 groups suggested by Kass's heuristic. The difference concerns the group {technical, non Swiss, none}, which is in this 4 group solution broken down into {non Swiss} and {technical, none}.

[Figure 3 about here.]

Note that the best g-way grouping must indeed comply the minimal node size constraint if we want effectively use it for splitting the node. From Table 2 page 44 it is readily shown, for example, that the {technical, none} group of the 4 group solution contains only 21 cases and does not fit our $\min_{node}$ constraint that was set at 50. Hence this solution cannot be retained. Furthermore, since the solution into g groups is derived from that into $g-1$ ones, this same stopping rule would indeed also preclude using any of the 'best' solutions into $g \geq 4$ groups. Hence, in our example, the only choice is to split into 2 or 3 groups. The solution into 2 groups has a much higher p-value, which makes it less interesting. Eventually, we end up again with the same partition into 3 groups as before.

Growing the tree on our data with this refined best g-way grouping method — but without using Bonferroni adjustments — we get the tree shown in figure 3. Note that it is the same as the one (Figure 1 page 38) obtained with Kass's best g-way grouping heuristic and no Bonferroni adjustment. This illustrates the small impact of this first Biggs et al's refinement.

3.2 Revised Bonferroni correction

We have seen that the Bonferroni multiplier proposed by Kass excessively penalizes predictors with many categories. The penalization is at its highest when the number of formed groups is close to $c/2$ and favors for instance splits into two groups rather than three, since the multiplier is much smaller for $g = 2$ than for $g = 3$. Moreover, we may also wonder why Kass considers

only the alternative ways of grouping into g groups, and disregards those into $g+1$, $g+2$, ..., c groups. These are the main weaknesses addressed by Biggs et al. (1991).

Those authors propose a Bonferroni correction taking explicitly into account the fact that when looking for the best split into $k-1$ groups one only explores groupings that result from the optimal solution into k groups. For non floating predictors with c categories, their Bonferroni multiplier reads thus:

$$m_B(c) \quad = \quad 1 + \sum_{k=2}^{c} m(k, k-1) \tag{2}$$

where $m(k, k-1)$ denotes the number of ways of grouping k categories into $k-1$ groups and the '1' standing for the trivial partitioning into c values. Unlike Kass' solution, this formula is related to the iterative working of the merging heuristic and is therefore better sounded. In formula (2) the value of $m(k, k-1)$ depends indeed on the either ordinal or nominal nature of the predictor.

- For nominal (free) predictors it is given by the corresponding Stirling number of the second kind

$$m(k, k-1) \quad = \quad S(k, k-1) = \sum_{i=0}^{k-2} (-1)^i \frac{(k-1-i)^k}{i!(k-1-i)!} \quad .$$

- For ordinal (monotonic) predictors it is just

$$m(k, k-1) \;\; = \;\; k-1$$

The Bonferroni multiplier $m_B^{float}(c)$ for a *floating predictor* with c values is the one $m_B^{ord}(c-1)$ for an ordinal variable with $c-1$ categories plus the number of ways of placing the floating category. There are at most $c-1$ groups into which the floating category can be placed or it can be kept as a separate group. Hence, the multiplier for floating predictors is

$$m_B^{float}(c) \;\; = \;\; m_B^{ord}(c-1) + (c-1) + 1 = m_B^{ord}(c) \;\; ,$$

that is the same as for ordinal (monotonic) predictors.

[Table 15 about here.]

An important point is that whatever the nature of the predictor, Biggs et al's proposition is, unlike Kass' solution, independent of the final number of groups retained. It depends upon c only. Table 15 reports the values of the multiplier until $c = 10$ categories. Though increasing exponentially with c, the correction factors grow much more slowly than Kass's solution. For the "Type of secondary diploma" that is selected as best splitting attribute at the root node, the multiplier is for instance 84 with Biggs et al's proposition, while it was 966 with Kass's solution for the retained partition into 3 groups. This is more than a factor 10 difference.

[Figure 4 about here.]

To illustrate the use of Biggs et al's Bonferroni adjustment, we consider again the possibilities of splitting the root node by using the best k-way groupings of each predictor. The latter groupings are the same as those obtained with Kass's methods, which allows us to start again from the information gathered in Table 10 page 52. The p-values without adjustment and the adjusted ones are reported in Table 16. We may notice that the ranking of the predictor remains unchanged, meaning that here and unlike what did happen with Kass's Bonferroni adjustment, the same predictor "Type of secondary diploma" is retained with and without Bonferroni adjustment.

[Table 16 about here.]

Iterating the same process at each new obtained node, we end up with the tree shown in Figure 4. Comparing this tree with the one in Figure 3 reveals that the use of Biggs et al's Bonferroni adjustment did not change at any node the retained splitting predictor. We end up however with a slightly less complex tree. The first level node corresponding to those students who had none or a technical or non Swiss secondary diploma is not split, while it was in the first tree. This is because there is now at this node no possible split for which the critical adjusted p-value is below the critical value (set at 5%).

4 Conclusion

Despite their many advantages and obvious efficiency for detecting interactions that matter and more generally as exploratory tools, tree methods were also more or less severely criticized. Einhorn (1972, 1973) for instance warned about easy tree misuse. He draw attention to the nowadays well known over fitting problem, i.e., on the danger to get a too complex tree when focusing only on the training sample and suggested to resort to cross-validation. More generally, he looked sceptical on the reliability of tree outcomes. The discussion with Morgan and Andrews (1973) hold mainly on the meaning of exploratory approaches, Einhorn claiming that only model building based on clearly enounced assumptions may usefully exploit quantitative data. Doyle (1973) advanced also a series of criticisms that were more specifically addressed to the application of AID that Heald (1972) ran on a sample of only 70 data. He recalls that trees are mainly intended for large samples (a minimum of about 1000 cases). He stresses that no general goodness of fit measure is provided with trees and that trees cannot determine the global importance of the factors that intervene in the tree. Indeed, the tree does not provide effects controlled by all other covariates. Instead, it shows the additional information brought by the covariate conditional to the already made splits. The tree structure obtained from a sample data set is known to be unstable in the sense that a very small change in the data may considerably affect it.

Most of these criticisms have since then received efficient answers. Cross-

validation and test of generalization on test data are nowadays currently used with trees. Ensemble methods, and especially forest trees introduced by Breiman (2001) offer solutions to the robustness issue when the concern is prediction. The measures of the variable importance in forest trees proposed for instance by Strobl, Malley, and Tutz (2009) also answer one of Doyle's criticism. As for the goodness of fit of trees, we proposed ourself (Ritschard and Zighed, 2003; Ritschard, 2006) deviance-based measures for investigating statistical information brought by the obtained segmentation. Nonetheless, trees remain exploratory tools. Although they demonstrated good prediction capacities, they are not intended for testing causality hypotheses.

It also is worth mentioning that, although we focused in this paper on early tree methods, their development did not stop with CHAID. Important more recent milestones are among others the CART method by Breiman, Friedman, Olshen, and Stone (1984), which boosted the use of trees in the 80's, and the ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993) algorithms. These methods are clearly oriented towards prediction and have in common to use split criteria based on entropy reduction (Gini, i.e., quadratic entropy for CART and Shannon's entropy for ID3 and C4.5) aiming at finding pure nodes. In the statistical area, recent developments were oriented towards solving the bias favoring predictors with many different values. Milestones are here QUEST (Loh and Shih, 1997), GUIDE (Loh, 2007) and 'party' (Hothorn et al., 2006b,a).

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Armstrong, J. S. and J. G. Andress (1970). Exploratory analysis of marketing data: Tree vs regression. *Journal of Marketing Research 7*, 487–492.

Assael, H. (1970). Segementing markets by group purchasing behavior. *Journal of Marketing Research 7*, 153–158.

Bell, E. T. (1938). The iterated exponential numbers. *Ann. Math. 39*, 539–557.

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Applied Statistics 8*(2), 65–75.

Belson, W. A. (1978). Investigating causal hypotheses concerning delinquent behaviour, with special reference to new strategies in data collection and analysis. *The Statistician 27*(1), 1–25.

Biggs, D., B. De Ville, and E. Suen (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics 18*(1), 49–62.

Bouroche, J.-M. and M. Tenenhaus (1970). Quelques méthodes de segementation. *Revue française d'informatique et de recherche opérationnelle 4*(2), 29–42.

Bouroche, J.-M. and M. Tenenhaus (1972). Some segmentation methods.

Metra 7, 407–418.

Breiman, L. (2001). Random forest. *Machine Learning 45*, 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Cellard, J. C., B. Labbé, and G. Savitsky (1967). Le programme ELISEE, présentation et application. *Metra 3*(6), 511–519.

Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly (1970-1977) 24*(3), 465–467.

Einhorn, H. J. (1972). Alchemy in the behavioral sciences. *The Public Opinion Quarterly 36*(3), 367–378.

Einhorn, H. J. (1973). Reply to Morgan and Andrews. *The Public Opinion Quarterly 37*(1), 129–131.

Fielding, A. and C. A. O'Muircheartaigh (1977). Binary segmentation in survey analysis with particular reference to AID. *The Statistician 26*(1), 17–28.

Gillo, M. W. (1972). MAID, a Honeywell 600 program for an automatized survey analysis. *Behaviorial Science 17*(2), 251–252.

Gillo, M. W. and M. W. Shelly (1974). Predictive modeling of multivariable and multivariate data. *Journal of the American Statistical Association 69*(347), 646–653.

Heald, G. I. (1972). The application of the automatic interaction detector programme and multiple regression techniques to the assessment of store performance and site selection. *Operational Research Quarterly 23*(4),

445–457.

Hothorn, T., K. Hornik, and A. Zeileis (2006a). party: A laboratory for recursive part(y)itioning. User's manual.

Hothorn, T., K. Hornik, and A. Zeileis (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Hunt, E. B., J. Marin, and P. J. Stone (1966). *Experiments in induction*. New York and London: Academic Press.

Kass, G. V. (1975). Significance testing in automatic interaction detection (A.I.D.). *Applied Statistics 24*(2), 178–189.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Light, R. J. and B. H. Margolin (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association 66*(335), 534–544.

Loh, W.-Y. (2007). GUIDE (version 5) User manual. Technical report, Department of Statistics, University of Wisconsin, Madison.

Loh, W.-Y. and Y.-S. Shih (1997). Split selection methods for classification trees. *Statistica Sinica 7*, 815–840.

Messenger, R. and L. Mandell (1972). A modal search technique for predictibe nominal scale multivariate analys. *Journal of the American Statistical Association 67*(340), 768–772.

Morgan, J. N. and F. M. Andrews (1973). A comment on Einhorn's "Alchemy in the behavioral sciences". *The Public Opinion Quarterly 37*(1), 127–129.

Morgan, J. N. and R. C. Messenger (1973). THAID a sequential analysis program for analysis of nominal scale dependent variables. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Morgan, J. N. and J. A. Sonquist (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association 58*, 415–434.

Orr, L. (1972). The dependence of transition proportions in the education system on observed social factors and school characteristics. *Journal of the Royal Statistical Society. Series A (General) 135*(1), 74–95.

Petroff, C., A.-M. Bettex, and A. Korffy (2001). Itinéraires d'étudiants à la Faculté des sciences économiques et sociales: le premier cycle. Technical report, Université de Genève, Faculté SES.

Press, L. I., M. S. Rogers, and G. H. Shure (1969). An interactive technique for the analysis of multivariate data. *Behavioral Science 14*(5), 364–370.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Ritschard, G. (2006). Computing and using the deviance with classification trees. In A. Rizzi and M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 55–66. Berlin: Springer.

Ritschard, G. and D. A. Zighed (2003). Goodness-of-fit measures for induction trees. In N. Zhong, Z. Ras, S. Tsumo, and E. Suzuki (Eds.), *Foun-*

dations of Intelligent Systems, ISMIS03, Volume LNAI 2871, pp. 57–64. Berlin: Springer.

Ross, J. A. and S. Bang (1966). The AID computer programme, used to predict adoption of family planning in Koyang. *Population Studies 20*(1), 61–75.

Scott, A. J. and M. Knott (1976). An approximate test for use with AID. *Applied Statistics 25*(2), 103–106.

Sonquist, J. A. (1969). Finding variables that work. *The Public Opinion Quarterly 33*(1), 83–95.

Sonquist, J. A., E. L. Baker, and J. N. Morgan (1971). Searching for structure (Alias–AID–III). Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago: SPSS Inc.

Strobl, C., J. Malley, and G. Tutz (2009). An introduction to recursive partitioning:. *Psychological Methods 14*(4), 323–348.

Tanofsky, R., R. R. Shepps, and P. J. O'Neill (1969). Pattern analysis of biographical predictors of success as an insurance salesman. *Journal of Applied Psychology 53*(2, Part 1), 136 – 139.

Thompson, V. R. (1972). Sequential dichotomisation: Two techniques. *The Statistician 21*(3), 181–194.

List of Figures

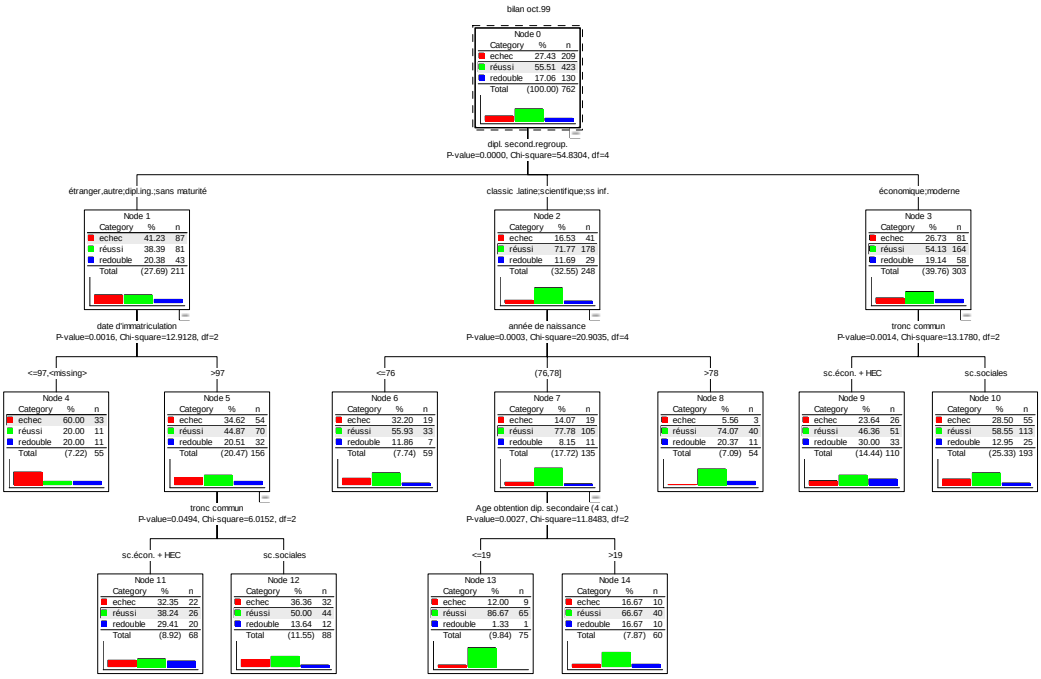

Figure 1: CHAID, without Bonferroni adjustment

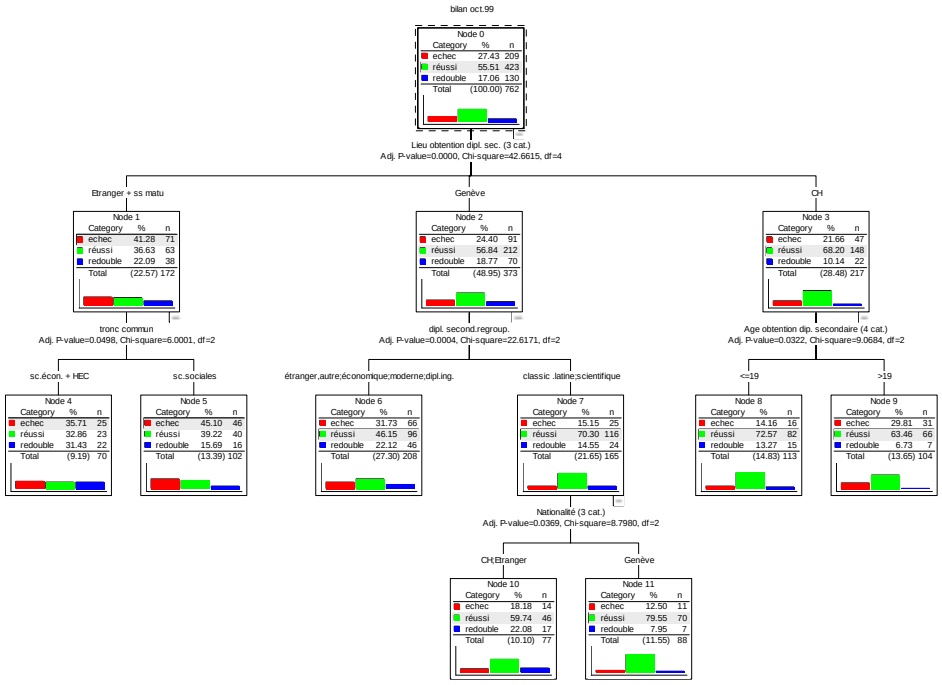

Figure 2: CHAID, with Bonferroni adjustment

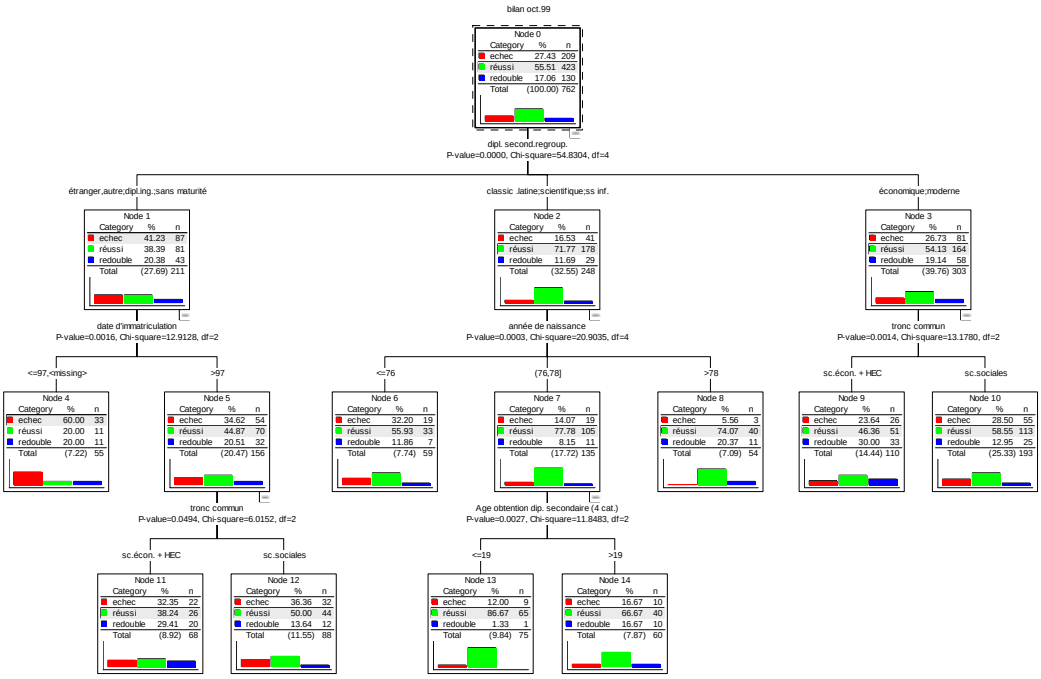

Figure 3: XCHAID, without Bonferroni adjustment

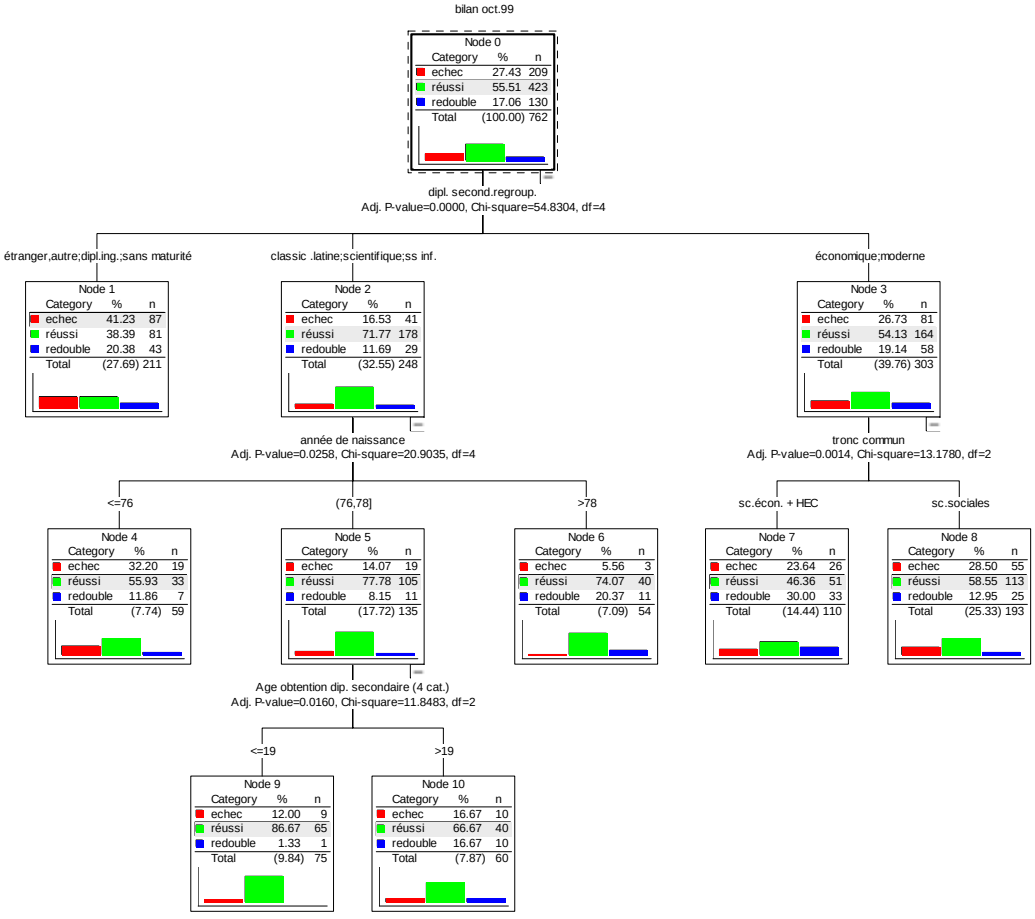

Figure 4: XCHAID, with Bonferroni adjustment

List of Tables

Table 1: Main earlier tree growing algorithms

	Local	Dependent variable		Splitting criterion		
Algorithm	split	quantitative	categorical	association	purity	p-value
Belson	binary		x	x		
AID	binary	x		x		
MAID	binary	x		x		
THAID	binary		x	x	x	
Hunt et al.	n-ary		x	x		
ELISEE	binary		x	x		
IDEA	n-ary	x	x	x		x
CHAID	n-ary	x	x	x		x

Table 2: Situation after 1st year by type of secondary diploma

	classic					non			
	Latin	modern	scient.	econ.	techn.	Swiss	none	miss.	Total
	1	2	3	4	5	6	7	8	
eliminated	23	44	18	37	5	76	6	0	209
repeating	16	22	13	36	0	43	0	0	130
passed	87	68	90	96	4	71	6	1	423
Total	126	134	121	169	9	190	12	1	762

Table 3: Chi-squares and their p-values by pair of categories

	1	2	3	4	5	6	7	8
1	0	9.62	0.87	5.25	7.53	30.64	7.37	0.45
2	0.008	0	15.66	4.79	2.81	5.88	2.92	0.96
3	0.647	0.000	0	9.88	9.84	40.80	9.65	0.34
4	0.073	0.091	0.007	0	6.25	16.65	6.37	0.76
5	0.023	0.245	0.007	0.044	0	2.66	0.06	1.11
6	0.000	0.053	0.000	0.000	0.264	0	3.47	1.66
7	0.025	0.232	0.008	0.041	0.969	0.177	0	0.93
8	0.800	0.618	0.842	0.685	0.574	0.436	0.629	0

Chi-squares are above the diagonal and p-values below the diagonal.

Table 4: 3×2 contingency table with smallest Chi-square

	technical diploma	no secondary diploma	Merged
	5	7	{5,7}
eliminated	5	6	11
repeating	0	0	0
passed	4	6	10
Total	9	12	21

Table 5: Chi-squares and their p-values by pair of categories

	1	2	3	4	{5,7}	6	8
1	0	9.62	0.87	5.25	12.98	30.64	0.45
2	0.008	0	15.66	4.79	5.44	5.88	0.96
3	0.647	0.000	0	9.88	16.40	40.80	0.34
4	0.073	0.091	0.007	0	11.63	16.65	0.76
{5,7}	0.002	0.066	0.000	0.003	0	5.97	1.05
6	0.000	0.053	0.000	0.000	0.0505	0	1.66
8	0.800	0.618	0.842	0.685	0.592	0.436	0

Chi-squares are above the diagonal and p-values below the diagonal.

Table 6: Successive merges

Iteration	Merge	Chi-square	p-value
1	{5,7}	.06	96.7%
2	{3,8}	.34	84.6%
3	{1,{3,8}}	.95	62.3%
4	{2,4}	4.79	9.1%
5	{6,{5,7}}	5.97	5.05%

Table 7: Final Chi-squares/p-values

	classic/Latin scientific, miss. {1,3,8}	modern economics {2,4}	technical, none non Swiss {5,6,7}
{1,3,8}	0	18.04	52.94
{2,4}	0.000	0	14.56
{5,6,7}	0.000	0.001	0

Chi-squares are above the diagonal and p-values below the diagonal.

Table 8: Situation after 1st year by 'Year when first registered'

	77	85	89	91	92	93	94	95	96	97	98	missing	Total
eliminated	1	1	2	2	2	2	3	9	12	31	143	1	209
repeating	0	0	0	0	0	1	0	2	7	10	110	0	130
passed	0	0	0	0	2	3	4	4	24	39	347	0	423
Total	1	1	2	2	4	6	7	15	43	80	600	1	762

Table 9: Chi-squares and their p-values by pair of categories of the floating predictor 'Year when 1st registered'

	≤ 91	92	93	94	95	96	97	98	miss
≤ 91	0	3.75							0
92	0.153	0	0.83						0.83
93		0.659	0	1.27					1.56
94			0.529	0	2.41				1.14
95				0.300	0	5.18			0.64
96					0.075	0	1.50		2.44
97						0.472	0	8.53	1.55
98							0.014	0	3.18
miss	1	0.659	0.459	0.565	0.726	0.295	0.461	0.204	0

Chi-squares are above the diagonal and p-values below the diagonal.

Table 10: Summary of possible first level splits

Predictor	#categories	#splits	Chi-square	*df*	p-value
Type of secondary diploma	8	3	54.83	4	.000000000035
Birth year	25	3	53.01	4	.000000000085
Where secondary diploma	3	3	42.66	4	.0000000122
Mother living place	4	2	27.21	2	.00000123
Nationality	3	2	24.26	2	.00000540
Year when 1st registered	11+1	2	18.72	2	.0000863
Age at secondary diploma	4	4	21.86	6	.00128
Chosen orientation	2	2	1.39	2	.499

Table 11: The 7 ways of grouping a, b, c, d into 2 groups

i	group 1	group 2	i	group 1	group 2
1	a	b, c, d	5	a, b	c, d
2	b	a, c, d	6	a, c	b, d
3	c	a, b, d	7	a, d	b, c
4	d	a, b, c			

Table 12: The 5 ways of grouping 3 ordered values a, b, c and one float f into 2 groups

i	group 1	group 2	i	group 1	group 2
1	a, f	bc	4	ab, f	c
2	a	bc, f	5	ab	c, f
3	abc	f			

Table 13: Bonferroni adjusted p-values, Kass's method

Predictor		p-value	multiplier	adj p-value	rank
Type of secondary diploma	nom	.000000000063	966	.0000000341	3
Birth year	ord	.000000000085	276	.0000000234	2
Where secondary diploma	nom	.0000000122	1	.0000000122	1
Mother living place	nom	.00000123	7	.00000864	4
Nationality	nom	.00000540	3	.0000162	5
Year when 1st registered	float	.0000863	19	.00164	7
Age at secondary diploma	ord	.00128	1	.00128	6
Chosen orientation	nom	.499	1	.499	8

Table 14: p-values for each of the best split into $g = 2, \ldots, c$ groups

g	best χ^2	df	p-value	smallest group size
8	67.85	14	0.000000004725	1
7	67.76	12	0.000000000839	1
6	67.50	10	0.000000000135	21
5	66.72	8	0.000000000022	21
4	61.92	6	0.000000000018	21
3	54.83	4	0.000000000035	211
2	39.64	2	0.000000002473	248

Table 15: Biggs et al's Bonferroni multipliers

c	nominal $m_B^{nom}(c)$	ordinal $m_B^{ord}(c)$
3	4	3
4	10	6
5	20	10
6	35	15
7	56	21
8	84	28
9	120	36
10	165	45

Table 16: Bonferroni adjusted p-values, Biggs et al's method

Predictor		c	p-value	$m_B(c)$	adj p-value	rank
Type of secondary dipl.	nom	8	.000000000035	84	.0000000030	1
Birth year	ord	25	.000000000085	300	.0000000254	2
Where secondary dipl.	nom	3	.0000000122	4	.0000000487	3
Mother living place	nom	4	.00000123	10	.0000123	4
Nationality	nom	3	.00000540	4	.0000216	5
Year when 1st regist.	float	11	.0000863	55	.00475	6
Age at secondary dipl.	ord	4	.00128	6	.00770	7
Chosen orientation	nom	2	.499	1	.499	8