# An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables

Jay Magidson[1] and Jeroen K. Vermunt[2]

[1] Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478, USA
[2] Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, Netherlands

**Abstract.** The CHAID algorithm has proven to be an effective approach for obtaining a quick but meaningful segmentation where segments are defined in terms of demographic or other variables that are predictive of a *single* categorical criterion (dependent) variable. However, response data may contain ratings or purchase history on *several* products, or, in discrete choice experiments, preferences among alternatives in each of *several* choice sets. We propose an efficient hybrid methodology combining features of CHAID and latent class modeling (LCM) to build a classification tree that is predictive of *multiple* criteria. The resulting method provides an alternative to the standard method of profiling latent classes in LCM through the inclusion of (active) covariates.

## 1 Background and summary of approach

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. For example, the dependent variable might be response to a mailing (responders vs. non-responders). Each of the resulting segments, depicted as a terminal node in a tree diagram, is defined as a combination of directly observable categorical predictors such as AGE = 18-24 & INCOME = \$80,000+. Descriptive entries in each tree node consist of the sample size and the corresponding observed distribution on the dependent variable (e.g., associated response rate).

Latent class (LC) models are useful in identifying segments that underlie *multiple* response variables. While the resulting latent classes can be either ordered (ordinal latent variable) or unordered (nominal latent variable), they are not actionable like CHAID segments, because by definition they are unobservable (latent).

In this paper we propose a hybrid methodology that combines strengths of both approaches. After decomposing a set of $M$ response variables into $K$ underlying latent class segments, a modified CHAID algorithm is used with the $K$ latent classes serving as the $K$-category nominal (ordinal) criterion variable. The resulting CHAID segments, derived from selected demographic

or other exogenous variables that are predictive of the classes, should also tend to be predictive of the $M$ criterion variables.

The hybrid method also provides an alternative to the use of covariates in LCM to profile the classes. In practice, one or more demographic or other exogenous variables are included in an LCM to describe/predict the latent classes using a multinomial logit model. The proposed CHAID-based alternative is especially advantageous when the number of covariates is large, when covariate effects are non-linear, or when there are complicated higher-order interactions.

In the next section we provide brief introductions to the standard CHAID algorithm and the standard LC (cluster and factor) models. We then provide the technical details of the hybrid approach, followed by an empirical example from a pre-post survey (Burns et al. (2001)). We conclude with some final remarks.

## 2 The CHAID algorithm

The original CHAID algorithm was introduced by Kass (1980) for nominal dependent variables. CHAID is a recursive partitioning method useful in exploratory analyses that relate a potentially large number of categorical predictor variables to a single categorical nominal dependent variable. It was extended to ordinal dependent variables by Magidson (1993) who illustrated how this extension could be used to take advantage of fixed scores such as profitability, for each category of the dependent variable when such scores are known, as well as how to estimate meaningful scores when category scores are unknown. Chi-squared goodness of fit tests are used to identify significant predictors, and to merge predictor categories that do not differ in their prediction of the dependent variable.

Predictor categories are eligible to be merged according to specified scale types. Any categories of Nominal ("free") predictors can be merged, while only adjacent categories of ordinal or grouped continuous ("monotonic") predictors are allowed to merge. A final scale type ("float") may be used to specify that the variable is to be treated as monotonic except for the final category, often corresponding to a 'don't know' or 'missing' response, which is free to merge with any of the other categories. Technical settings include significance levels associated with merging and splitting and a stopping rule. A case weight and a frequency variable may also be included in the analysis.

As an example, Figure 1 illustrates a CHAID analysis based on data from a post-election survey on 1,051 persons who voted for either Bush or Gore in the 2000 U.S. election. The dependent variable (VOTE) is the candidate voted for and the predictors are 5 demographic variables: 1) MARSTAT (1=married, 2=widowed, 3= separated/divorced, 4= never married, 5= other – "Free"), AGEr (1=18-24, 2=25-34, 3=35-44, 4= 45-54, 5= 55-64, 6=65+, '.' = refused – "Float"), GENDER (1 = male, 2 = female), EDUCATION

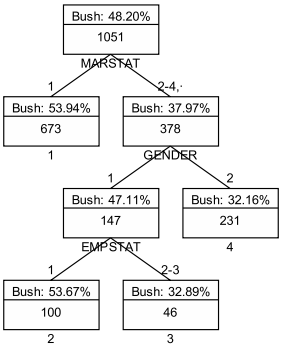


**Fig. 1.** CHAID tree for VOTE.

(1 = less than HS, 2= HS grad, 3= some college, 4=college grad, 5= post grad, 5-refused – "Float"), and EMPLOYED (1 = Yes, 2 = No, 3 = retired – "Free").

Overall, 48.2% voted for Bush. This is displayed in the top (root node) of the tree. Among the 5 demographic predictors included in this analysis, only 2 were significant at the root node – MARSTAT ($p<.00001$), and GENDER ($p<.01$). The CHAID analysis resulted in 4 segments. The best segments for Bush are S1, consisting of the 673 married voters (53.94% for Bush) and S2 consisting of 100 unmarried employed males (53.67% for Bush). The remaining segments – S3 (unmarried unemployed males) and S4 (unmarried females) – voted more than 2:1 in favor of Gore over Bush.

One limitation of CHAID is that segments are defined based on a single criterion variable. Given situations where multiple criteria exist, it is not clear how one should go about obtaining a single common segmentation. Using one dependent variable as the criterion may result in one set of segments, while use of an alternative dependent variable will likely yield a different set of segments. Moreover, the categories of a predictor may merge in different ways depending upon which dependent variable is used, again leading to different segments.

In addition, when multiple dependent variables do exist, they may be of different scale types (nominal, ordinal, continuous, count, etc.). Using a 3-category response variable as an example Magidson (1993) showed that CHAID segments resulting from treating the dependent variable as ordinal (using profitability scores for the categories) differed substantially from segments derived from the nominal algorithm which ignored the scores. The hybrid approach resolves the need to chose between different segmentations because indicators with differing scale types can be used in extended LCMs, yielding a single LC solution. An important advantage of this hybrid approach over approaches based on specific measures for node homogeneity rather than a model (e.g., Kim and Lee (2003)) is that the LC model used here can handle dependent variables of different scale types.

## 3 Latent class modeling

A LC model postulates a nominal $K$-category latent (unobservable) variable $X$ to explain the associations/correlations between the observed response variables (multiple criteria; Lazarsfeld and Henry (1968); Goodman (1974)). Each category of $X$ is called a latent class. Let $Y_m$ denote one of $M$ nominal response variables, $m = 1, 2..., M$; $j_m$ is a particular response category and $J_m$ the number of categories of variable $Y_m$. Notation $\mathbf{Y}$ and $\mathbf{j}$ is used to refer to a full response vector and a full set of response categories. The LC model for $M$ response variables is defined as

$$P(\mathbf{Y} = \mathbf{j}) = \sum_{k=1}^{K} P(X = k, \mathbf{Y} = \mathbf{j}) = \sum_{k=1}^{K} P(X = k) P(\mathbf{Y} = \mathbf{j} | X = k)$$
$$= \sum_{k=1}^{K} P(X = k) \prod_{m=1}^{M} P(Y_m = j_m | X = k), \quad (1)$$

where $P(X = k)$ denotes the probability of being in latent class $k$, $k = 1, 2, ..., K$, and $P(Y_m = j_m | X = k)$ denotes the conditional probability of obtaining the $j_m$th response to item $Y_m$, from members of class $k$, $j_m = 1, 2, ..., J_m$.

Cases with response pattern $\mathbf{j}$ are typically classified into the latent class for which the posterior membership probability $P(X = k|\mathbf{Y} = \mathbf{j})$ is highest. Estimates for the posterior membership probabilities – for $k = 1, 2, ..., K$ – can be obtained using Bayes theorem as follows:

$$P(X = k|\mathbf{Y} = \mathbf{j}) = \frac{P(X = k, \mathbf{Y} = \mathbf{j})}{P(\mathbf{Y} = \mathbf{j})}. \quad (2)$$

The numerator and denominator were defined in equation (1).

Recent advances allow for dependent variables (indicators) of varying scale types to be used – including mixing categorical, continuous, and count variables – by specifying the appropriate probability densities $P(Y_m = j_m|X = k)$ (Vermunt and Magidson (2002)). By expressiong the mean of these densities in terms of a generalized linear model (GLM), one can include direct effects between 2 or more indicators, multiple categorical latent variables, continuous latent factors and/or other terms into the model (see Magidson and Vermunt (2001); Vermunt and Magidson, (2005)).

It is also possible to include one or more exogenous variables called covariates in a LCM, allowing one to explore the relationship between exogenous variables and the latent classes and assess the significance of such relationships in a formal way. However, the covariates included in LCM influence the estimates of the parameters in the original measurement model. If the covariate part of the model holds true, inclusion of the covariates improves the efficiency of the estimates. However, if it is misspecified, the estimates

may become somewhat biased. In addition, profiling latent classes in terms of many covariates may cause the solution to become unstable. As an alternative, Magidson and Vermunt (2001) allow covariates to be treated in an *inactive* manner – providing appropriate cross-tabulations but not influencing the original measurement model. But this approach comes at the expense of no longer being able to assess statistical significance.

In the next section, we show how the hybrid algorithm provides an alternative treatment to the use of both active and inactive covariates in LC models. The new approach provides an assessment of statistical significance for *selected* covariates included within the LCM framework, whether the covariate is specified as active or inactive. Those covariates specified as inactive do not alter the estimates obtained from the LCM.

## 4 The hybrid CHAID algorithm

Our hybrid CHAID algorithm involves 3 steps.

1. Perform an LC cluster analysis on $M$ response variables to obtain $K$ latent classes.
2. Perform a CHAID analysis using the $K$ classes as a nominal dependent variable.
3. Obtain predictions for each of $M$ response variables based on the resulting CHAID segments and/or on any preliminary set of CHAID segments.

Step 1 yields class-specific predicted probabilities for each category of the $m$-th dependent variable[1], as well as posterior membership probabilities for each case.

Step 2 yields a set of CHAID segments that differ with respect to their average posterior membership probabilities for each class. We use the posterior membership probabilities defined in equation (2) as fixed case weights as opposed to the modal assignment into one of the $K$ classes. This weighting eliminates bias due to the misclassification error that occurs if cases were equated (with probability one) to that segment having the highest posterior probability. Specifically, each case contributes $K$ records to the data, the $k$th record of which contains the value $k$ for the dependent variable, and contains a case weight of $P(X = k|\mathbf{Y} = \mathbf{j})$, the posterior membership probability associated with that case. Thus, as opposed to the original algorithm where chi-square is calculated on observed 2-way tables, in the hybrid algorithm, the chi-squared statistic is computed on 2-way tables of *weighted* cell counts.[2]

If as an alternative to performing a standard LC analysis, one performs an LC factor analysis in step 1, in step 2 the CHAID *ordinal* algorithm can

[1] When one or more of the dependent variables are quantitative, for each class this step also yields predicted means for the quantitative dependent variables.

[2] The new algorithm also incorporates sampling weights, if present, using an efficient ML algorithm proposed by Vermunt and Magidson (2001).

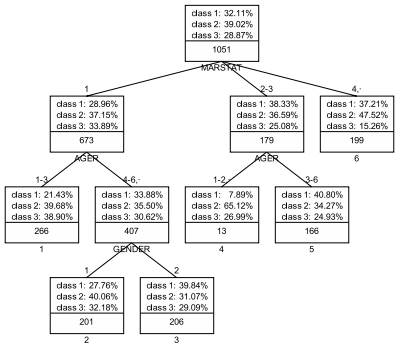


**Fig. 2.** Hybrid CHAID tree for 11 dependent variables.

be used to obtain segments based on the use of any of the LC factors as the ordinal dependent variable, or a single segmentation can be obtained using the nominal algorithm to identify segments based on the single joint latent variable defined as a combination of two or more identified LC factors.

Step 3 involves obtaining predictions for any or all of the $M$ dependent variables for each of the $I$ CHAID segments by cross-tabulating the resulting CHAID segments by the desired dependent variable(s). An alternative is to obtain predictions as follows

$$P(Y_m = j|i) = \sum_{k=1}^{K} P(Y_m = j|X = k)P(X = k|i).$$

As can be seen, we compute a weighted average of the class-specific distributions for dependent variable $Y_m$ obtained in step 1 $[P(Y_m = j|X = k)]$, with the average posterior membership probabilities obtained in step 2 for segment $i$ being used as the weights $[P(X = k|i)]$.

## 5 Empirical example

Among other questions, the pre-election survey solicited ratings for each candidate on 5 attributes – leadership, caring, knowledge, honesty and morality. A LCM was fit to these data, using VOTE as an active covariate, and the 5 demographics as inactive covariates. This model may be viewed as a kind of unsupervised regression with 11 dependent variables – VOTE, plus the 10 attribute ratings. This LCM yielded 3 segments. The first segment (32%) favored Gore, the second (39%) was neutral and the third favored Bush with respect to the attribute ratings and in their votes. These percentages are displayed in the root node of the hybrid CHAID tree in Figure 2.

The hybrid CHAID used the 3-category latent variable (segments) as the dependent variable and again utilized the 5 demographics as the predictors.

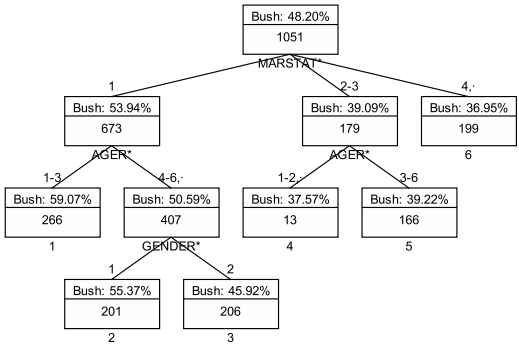


**Fig. 3.** Hybrid CHAID tree for VOTE.

At the root node 3 of the 5 predictors were found to be significant – MARSTAT ($p < .00002$), AGEr ($p<.001$), and GENDER ($p=.01$). Compared to our earlier CHAID, age is more important than when VOTE was the only dependent variable. The hybrid CHAID analysis resulted in 6 segments (Figure 2). Since the attributes are now included as additional dependent variables (the latent classes are a proxy for these dependent variables) we might expect that the resulting segments might predict any single dependent variable less well than CHAID based on only that dependent variable.

Figure 3 shows how the 6 hybrid segments predict VOTE. To compare this to the predictions based on our original segments (Figure 1) we first compare those segments favorable to Bush. Our previous analysis identified segments S1 and S2 as favorable to Bush. The hybrid CHAID (Figure 3) identifies 3 segments most likely to vote for Bush – segments 1, 2 and 3. Note that these 3 segments combined, are equivalent to the original segment S1. Since the hybrid CHAID fails to yield any additional segments that prefer Bush such as S2, it appears that the hybrid segmentation predicts VOTE less well than the original CHAID. Similarly, focusing on segments most favorable to Gore, our previous CHAID identified S3 and S4 (n= 277 cases) as favoring Gore by more than 2:1. The hybrid CHAID finds segments 4, 5 and 6 as favoring Gore, but not by as much as 2:1 over Bush.

## 6 Final comments

In this paper, we introduced a hybrid CHAID algorithm[3] as an extension of CHAID to multiple dependent variables of possibly differing scale types. Alternatively, this hybrid algorithm could be described as an alternative to the standard treatment of active and/or inactive covariates in LCM. The

[3] The extended CHAID algorithm has been implemented in a commercially available computer program called SI-CHAID 4.0, and works in conjunction with the latent class programs Latent GOLD 4.0 and Latent GOLD Choice 4.0.

CHAID-type output can simplify the process of examining the relationship between the demographics and/or other exogenous variables and the latent segments by 1) ranking the covariates from most to least significant and 2) for each covariate, merging categories that are not significantly different. This new output is especially valuable when the number of covariates is large.

We illustrated the hybrid algorithm here with dependent variables consisting of favorability ratings of Bush and Gore on 5 attributes plus the actual vote among 1,051 voters in the 2000 U.S. election. We showed how the hybrid CHAID provides a unique segmentation. We showed how it compares with a segmentation obtained using the traditional CHAID algorithm for a single dependent variable – VOTE. The results suggest that the segments resulting from the hybrid CHAID may fall somewhat short of predictability of any single dependent variable in comparison to the original algorithm, but makes up for this by providing a single unique set of segments that are predictive of all dependent variables.

## References

BURNS, N., KINDER, D.R., ROSENSTONE, S.J., SAPIRO, V., and the National Election Studies (2001): National Election Studies, 2000: Pre-/Post- Election Study [dataset id:2000.T]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

GOODMAN, L.A. (1974): Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61, 215–231.*

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

KIM, S.J. and Lee, K.B. (2003): Constructing decision trees with multiple response variables. *International Journal of Management and Decision Making, 4, 289 – 311.*

LAZARSFELD, P. F. and HENRY, N.W. (1968): *Latent structure analysis.* Houghton Mifflin, Boston.

MAGIDSON, J. (1993): The use of the new ordinal algorithm in CHAID to target profitable segments. *The Journal of Database Marketing, 1, 29–48.*

MAGIDSON, J. and VERMUNT, J.K. (2001): Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology, 31, 223–264.*

VERMUNT, J.K. and MAGIDSON, J. (2001): *Latent class analysis with sampling weights.* Paper presented at the 6th annual meeting of the Methodology Section of the American Sociological Association, University of Minnesota, May 4-5, 2001.

VERMUNT, J.K. and MAGIDSON, J. (2002): Latent class cluster analysis. In: J.A. Hagenaars and A.L. McCutcheon (Eds.): *Applied latent class analysis.* Cambridge University Press, Cambridge, 89–106.

VERMUNT, J. K. and MAGIDSON, J. (2005): *Latent GOLD 4.0 User Manual.* Statistical Innovations Inc, Belmont MA.