# Responder profiling with CHAID and dependency analysis

Dietrich Eherler, Thomas Lehmann

Lehrstuhl für Wirtschafts- und Sozialstatistik
Friedrich-Schiller-Universität Jena
Carl-Zeiss-Str. 3, 07743 Jena

Email: {T.Lehmann},{D.Eherler}@wiwi.uni-jena.de

## 1. Introduction

The identification of responder profiles is one of the most important issues in marketing research. In this paper two approaches are applied to deal with this problem, illustrated by a case study from the automobile sector. The objective of the direct mailing considered here was the acquisition of new customers. Therefore persons similar to the existing customers of the company had been contacted with a brochure of the products. We will investigate if the achieved response rate could have been increased by the application of the suggested approaches. Starting point is information about this former mailing to potential customers. This information consists of the response behaviour of the addressed individuals as well as demographic, geographic and product specific data.

One widely used method for this kind of problem is CHAID analysis. We compare the results of this approach with those obtained by dependency analysis, a method which is still rarely applied to this issue. It turns out that CHAID and dependency analysis are appropriate tools to classify responders and non responders by their attributes and thus to improve the performance of direct mailing activities.

## 2. Methods for classification

### 2.1 CHAID Analysis

CHAID (Chi Squared Automatic Interaction Detector) is an exploratory method for classifying categorical data. The purpose of the procedure is to split a set of objects in a way, that the subgroups differ significantly with respect to a designated criterion. The criterion matches the dependent variable, while the remaining attributes represent their predictors in the model. The segments derived by CHAID are mutually exclusive and exhaustive which means, that the segments do not overlap and each object of the sample is contained in exactly one segment. Therefore the application of the method approves the classification of new objects by knowing the categories of the predictors (Magidson 1993).

CHAID analyses the instantiations of explanatory variables and the criterion variable for significant differences. Figure 2.1 contains a contingency table with $Y$ as the dependent and $X$ as the predictor, whereas possible interactions between both variables can be identified by their combined frequencies. If no interaction between the variables exists, it can be expected, that the relative frequencies of the criterion variable $Y$ within each category of the predictor $X$ correspond to the marginal frequencies of $Y$. In the example we would expect a conditional frequency of $y_1$ given $x_1$ of 60% under consideration of the marginal distribution of $Y$, which differs from the observed conditional frequency of 40%. Both variables are obviously not independent.

**Fig. 2.1:** Contingency table

| | | ***Y*** | | |
|---|---|---|---|---|
| | | $y_1$ | $y_2$ | |
| ***X*** | $x_1$ | 40 | 60 | 100 |
| | $x_2$ | 80 | 20 | 100 |
| | | 120 | 80 | 200 |

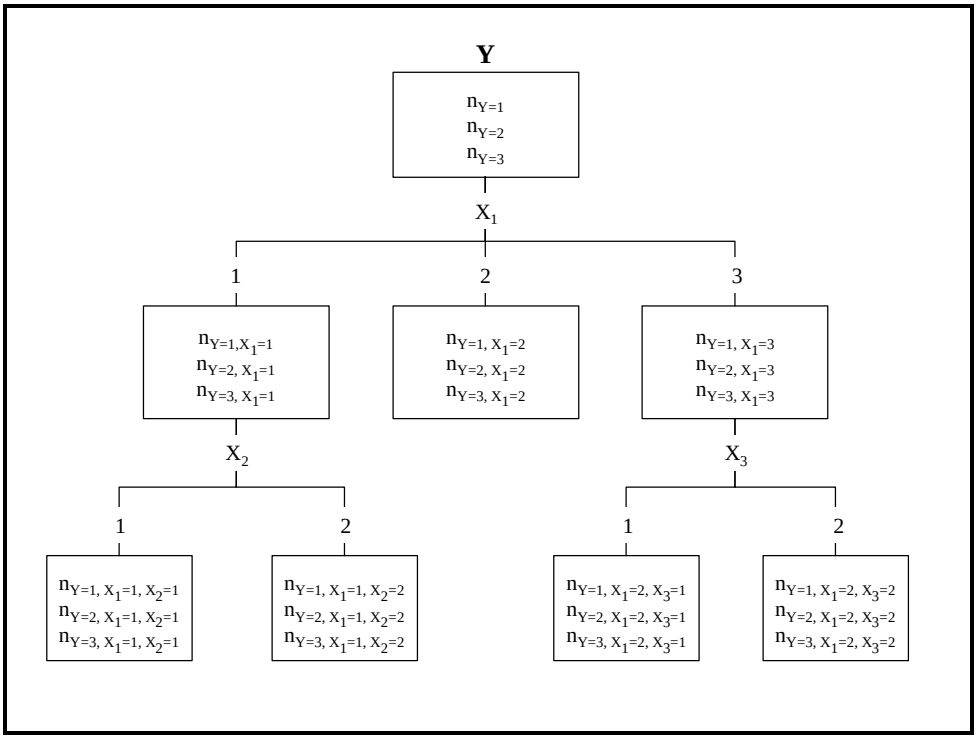
To infer from interactions in the sample possible dependencies in the population, Chi-squared tests of independence will be performed within CHAID (see Magidson 1994). The test statistic of the Chi-squared test cumulates the (standardised) squared deviations between observed and expected frequencies. Therefore large values of the test statistic indicate interactions between the analysed variables. To validate the possible dependencies the test statistic will be compared to the percentile of the distribution of cumulated, normally distributed random variables (Kass 1980). If the test statistic exceeds the value of this Chi-squared distribution defined by significance level, the null hypothesis of independence of the variables has to be rejected.[1] In this case an interaction between both variables, with error probability (p-value) defined by the significance level, exists.

If the instantiations of several explanatory variables differ significantly with respect to the dependent variable, the sample will be split by the predictor with the lowest error probability in CHAID. This implicates, that the conditional frequencies of the criterion variable $Y$ given variable $X_i$ differ the most and therefore $X_i$ is the best predictor of $Y$ (Brosius 1997).

Groups divided of the sample will be tested for further segmentation by performing Chi-squared tests of the criterion variable and the remaining predictors. This method will be proceeded for all subgroups until no more dependencies between the criterion variable and the predictors can be found. In this case the further segmentation of the sample would be not appropriate.

The segmentation process and the results of CHAID are best illustrated by a tree diagram used in Figure 2.2.

**Fig. 2.2:** Tree diagram




[1] The test statistic is only approximately Chi-squared distributed and therefore large sample size is required.

Each node of the tree diagram represents a subgroup of the sample. The root node contains the whole sample and the absolute frequencies $n_i$ for each category of $Y$ are listed. At the next level of the tree the sample is divided by $X_1$ as the best predictor of the dependent variable, while the three child nodes represent the several instantiations of $X_1$. The child nodes contain the information about the frequencies of the criterion variable $Y$ related to the corresponding subgroup. Similarly the sample has been further segmented by the explanatory variables $X_2$ and $X_3$.

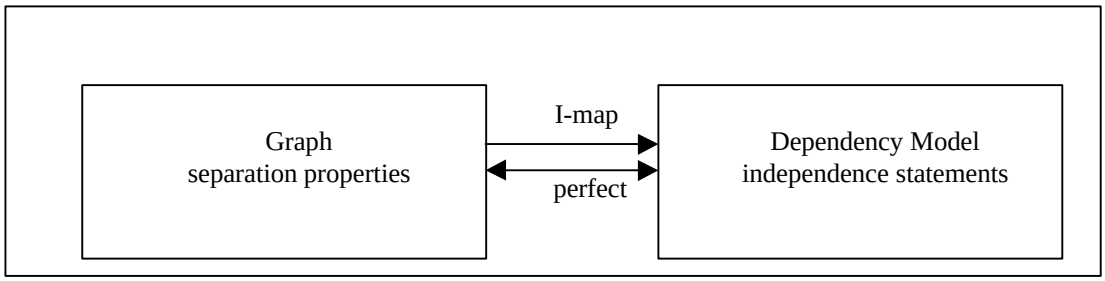
## 2.2 Dependency Analysis and Graphical Models

When performing regression or classification, we are interested in the conditional probability distribution for an outcome or class variable $Y$ given a set of explanatory variables $X$. At the heart of such an approach is the concept of conditional independence (e.g. Dawid 1979). The objective of dependency analysis is to display significant dependencies of variables in a considered domain (Fayyad et al. 1996). Whereas the notion of stochastic dependency is a symmetric concept, in a multivariate approach there are possibilities to determine uniquely the direction of the influence based on data. This methodology is known as learning graphical models from data. In what follows, a brief introduction of the theoretical background as well as two methods for learning graphs from data are given.

### 2.2.1 Some underlying theoretical aspects

Dependency analysis can be hosted in the wider framework of the so called knowledge discovery in databases (KDD) process. Within this scope, it plays a key role as a data mining technique. The approach of dependency analysis described here rests upon the idea to model probabilistic dependencies as graphs. This approach enables to reduce the complexity of a modelled domain as well as to handle with uncertainty (Jordan 1998). The aim is to map independence of variables as separation of nodes in graphs. Depending on the type of graph, different separation concepts can be used. The following definitions establish the formal link between probabilistic independence and vertice separation in graphs (see Figure 2.3). A dependency model is a list of independence statements derived from a joint probability distribution of a set of variables V. A graph over the set of variables V is an independence map of a dependency model, if separation in the graph implies independence of the corresponding variables. A graph is a perfect map, if separation and independence coincide (Pearl 1988). If the graph is a directed acyclic graph, d-separation (Pearl 1988) is an appropriate separation concept.

**Fig. 2.3:** Link of graph and probability theory




If the task is to learn a directed acyclic graph from data, the assumption has to be made, that there exists a graph that is a perfect map for the distribution that generated the data. The idea is, that the data to be analysed is generated by an unknown recursive equation system. In the context of the considered problem this means, that the explanatory variables determine the states of the outcome variable by a certain function. The task is to discover as many features as possible of this process from the given data. The result is a class of Markov equivalent models, i.e. models that cannot be distinguished by observational data alone.

### 2.2.2 Two procedures for learning graphs

The two algorithms are the PC-algorithm (Spirtes and Glymour 1991) and the K2-algorithm (Cooper and Herskovitz 1991). The first algorithm is independence constraint driven, the latter score driven.

The PC-algorithm works in two steps. In the first step, the adjacencies are generated. From the complete undirected graph, an edge between two variables is eliminated upon finding a conditional independence relation. Details can be found in Spirtes et al. (2001). In the second step, the remaining edges are oriented in the following manner. If $X$ and $Z$ are adjacent, $Z$ and $Y$ are adjacent, but $X$ and $Y$ are not adjacent, then the edges $X$-$Z$-$Y$ are

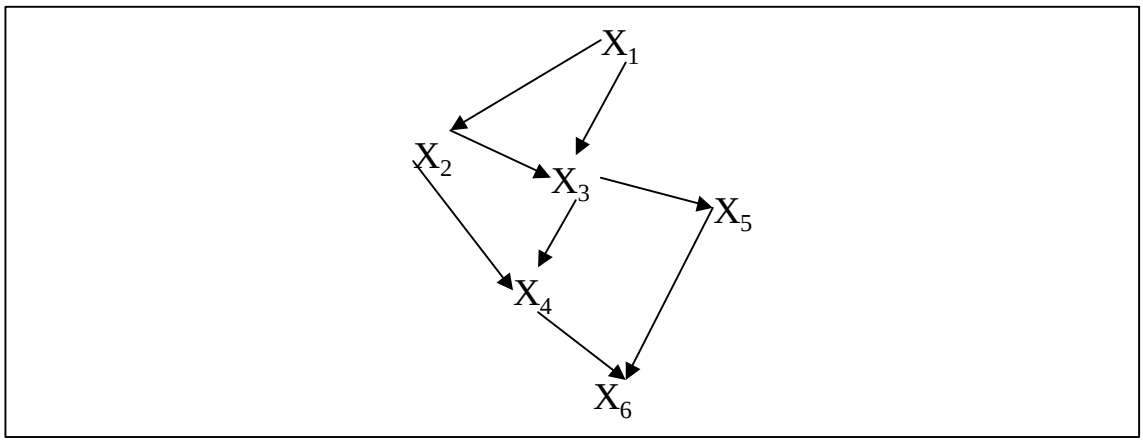
oriented to $X \rightarrow Z \leftarrow Y$, if and only if $Z$ is not a member of the set of variables that rendered $X$ and $Y$ conditionally independent in step 1. More details can again be found in Spirtes et al. (2001).

In contrast to the PC-algorithm, the K2-algorithm requires a complete ordering of the variables. If the primary objective is to find the direct influences of a response variable, the order can be made by grouping the variables into explanatory and response variables. Given the order, the procedure constructs a directed acyclic graph by searching iteratively for the best set of parents within the set of the preceding variables. The decision is made by the evaluation of a score. There is a wide variety of scores that can be used such as Bayesian scores or information criteria (see Castillo et al. 1997).

### 2.2.3 Interpreting the results

In the application below, both methods are applied. The interpretation of the resulting graph is straight forward. A directed edge between two variables $X \rightarrow Y$ indicates a direct influence of $X$ on $Y$. A directed path, i.e. a sequence of directed edges, displays an indirect influence. An undirected edge $X$-$Y$ means, that the algorithm cannot decide, whether there is an edge from $X$ to $Y$ or from $Y$ to $X$.

**Fig. 2.4:** A graph




The structure of the graph also portrays a decomposition of the joint probability distribution of the involved variables. The distribution can be factorised by exploiting the independence relations implied by d-separation as

$$P(X_1=x_1,...,X_6=x_6)=\prod_{i=1}^{n} P(X_i=x_i|\mathbf{PA}_i=\mathbf{pa}_i)=P(X_1=x_1)\cdot...\cdot P(X_6=x_6|X_4=x_4,X_5=x_5) \quad (\mathbf{1})$$

where $\mathbf{PA}_i$ is the set of parents of $X_i$. This distribution as well as the topology of the graph can be used to evaluate the influence between variables. The set of parents also allows to identify segments of the considered sample. Therefore, the goal is to find an appropriate set of parents of the target variable.

## 3. Comparison of the methods

### 3.1 Description of the data

The data analysed in this case study consists of demographic, geographic and product specific data as well as the response behaviour.

**Table 3.1:** Description of the variables

| Variable | Description | Range |
|---|---|---|
| $X_1$ | Composed variable describing affluence | 1,...,9 {very low,...,very high} |
| $X_2$ | Age | 1,...,7 {up to 35 years,...,61 years and above} |
| $X_3$ | Index for anonymity requirement | 0,...,7 {very low,...,very high} |
| $X_4$ | Proportion of commercial activity in the area | 0,...,8 { very low,...,very high} |
| $X_5$ | Micro-geographic type variable | 1,...,38 {attractive urban,...,younger rural} |
| $X_6$ | Index of family structure | 1,...,9 {mainly singles,...,mainly families} |
| $X_7$ | Proportion of younger inhabitants | 1,...,9 {very low,...,very high} |
| $X_8$ | Index of car size | small, medium, big |
| $X_9$ | Risk from non-payment | 1,...,9 {very low,..., very high} |
| $X_{10}$-$X_{17}$ | Other micro-geographic and product specific variables | |
| $X_{18}$ | Ownership of advertised product | Yes, no |

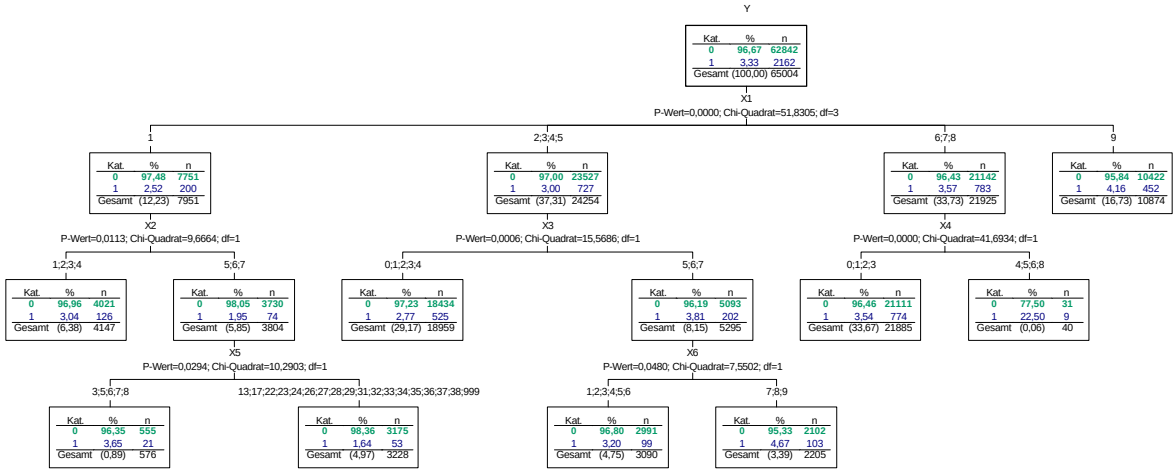
The assumption is made that the response behaviour can be explained by the variables above. In a first step the relevant variables explaining the target variable have to be extracted. The instantiations of these variables will be used to identify subgroups with the expected highest response rates.

### 3.2 Results of CHAID

A typical field of application of CHAID is the identification of target groups for direct mailing, because the segmentation derived by the algorithm is predictive for the dependent variable (see Magidson 1994 p. 125). The dependent variable corresponds to the behaviour of the addressed persons, while the predictors record their attributes. In this way target groups with expected high proportions of responding individuals can be identified.

The tree diagram derived in the above described way of the response analysis is represented in Figure 3.1. At the root node a response rate of 3,33% could be achieved by contacting persons in the former mailing. The sample will be first segmented by the variable $X_1$, because the test of independence between $X_1$ and the dependent variable yields the lowest error probability. Furthermore separate instantiations of $X_1$ are merged, as there are no significant differences between them related to the values of the criterion variable $Y$. Thereby it has to be considered, that the predictors within the case study are in most cases ordinal and therefore only adjacent categories can be merged. Generally it turns out, that the rate of response increases with higher instantiations of the variable $X_1$. The sample is not divided at the child node with the highest value of $X_1$, as the p-values of the tests of independence are not below the significance level respectively the size of the subgroups built in case of rejection of the null hypothesis is below the defined minimum size. The nodes of the classification tree grouped by $X_1$=1, $X_1$={2,3,4,5} and $X_1$={6,7,8} are separated by the variables $X_2$, $X_3$ and $X_4$ and, at the last level, by $X_5$ und $X_6$. This indicates, that the remaining variables not listed in the classification tree have no significant influence to the behaviour of response in the model.

**Fig. 3.1:** Classification Tree




The goodness of the segmentation can be evaluated by the comparison of response rate of the whole sample and the response rate of the terminal nodes. Therefore a response index was created for the best four terminal nodes to clarify the gains of addressing selected subgroups instead of addressing the sample. These subgroups are characterised by higher response rates than the average rate.

**Table 3.2:** Response index of selected subgroups with expected high response rates

| Segment-No. | Attributes of the subgroups | Response rate | Response index |
|---|---|---|---|
| 1 | $X_1 = \{6,7,8\}$, $X_4 = \{4,5,6,8\}$ | 22,5% | 676 |
| 2 | $X_1 = \{2,3,4,5\}$, $X_3 = \{5,6,7\}$, $X_6 = \{7,8,9\}$ | 4,67% | 140 |
| 3 | $X_1 = 9$ | 4,16% | 125 |
| 4 | $X_1 = 1$, $X_2 = \{5,6,7\}$, $X_5 = \{3,5,6,7,8\}$ | 3,65% | 110 |

Generally the target group of segment 1 characterised by the specific attributes should be addressed in the next mailing. However it has to be considered, that this result is only based on a small size of the subgroup. Due to their segment size the response rates of the remaining three groups are more profound, whereas the addressing of these subgroups should be clearly preferred to the random selection of addresses.

## 3.3 Results of dependency analysis

### 3.3.1 Preselection of variables

Due to the fact, that a variable can be only adjacent to the target variable, if it is not independent given any subset of the remaining variables, a preselection of the predictor variables can be made. Variables, that are not marginally dependent of the outcome variable, are excluded a priori from the learning process. This leads to higher efficiency. As benchmark for this decision the p-value of the Chi-squared independence test is chosen. The results are given in table 3.3.

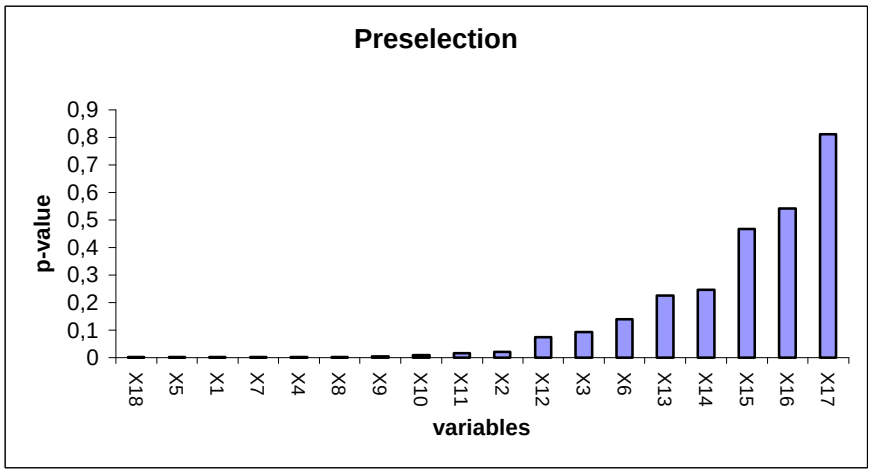
**Table 3.3:** Preselection of explanatory variables

| Variable | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ |
|---|---|---|---|---|---|---|---|---|---|
| p-value | 0,001 | 0,023 | 0,094 | 0,002 | 0,001 | 0,141 | 0,001 | 0,002 | 0,006 |

| Variable | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $X_{17}$ | $X_{18}$ |
|---|---|---|---|---|---|---|---|---|---|
| p-value | 0,01 | 0,018 | 0,075 | 0,226 | 0,246 | 0,468 | 0,543 | 0,812 | 0,001 |

The critical value is set to 1%. Therefore, the variables $X_1$, $X_4$, $X_5$, $X_7$, $X_8$ and $X_9$ were chosen as input for the learning algorithm. Although highly significant, the variable $X_{18}$ was left out of the analysis. This variable indicates the ownership of the product. However, the goal of the campaign was to reach new customers.

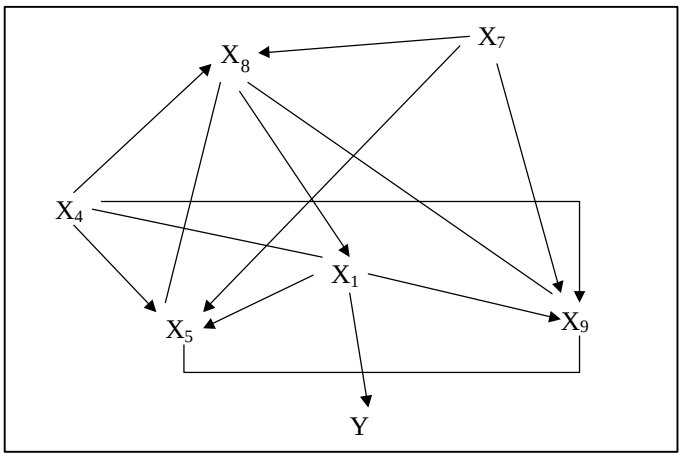
**Fig. 3.2:** Preselection




### 3.3.2 Learning a graph

The instantiations of the selected variables are used as input for the PC-algorithm as it is implemented in the software package Tetrad 3 (Spirtes et al. 1999). As result the graph depicted in Figure 3.3 is obtained. The significance level was set to be 1%.

**Fig. 3.3:** Learned graph



It can be read off from the graph that there is only one parent variable of the target variable. This is variable $X_1$. Defined by the different values of the variable $X_1$, it is possible to make a first but coarse segmentation of the responders. To do this, one can consider the conditional probability distribution estimated from the data. The results are mirrored in Table 3.4.

**Table 3.4:** First segmentation by variable $X_1$

| $X_1$ categories | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| cases | 7951 | 7523 | 5560 | 6210 | 4961 | 5764 | 7999 | 8162 | 10874 |
| rate | 2,52 | 2,9 | 2,82 | 3,13 | 3,2 | 3,34 | 3,69 | 3,58 | 4,16 |
| index | 76 | 88 | 85 | 95 | 97 | 101 | 112 | 108 | 126 |

The average respond rate is 3,33%. Therefore, the interesting groups with a response rate higher than the average are those, where variable $X_1$ takes the values 6, 7, 8, or 9.

### 3.3.3 Refining the result

Because the first analysis led to only one significant explanatory variable, in a second step the objective is to refine the result of the first analysis. To do so, the K2-algorithm is applied with focus on the target variable *Y*. To apply this algorithm that is characterised by a greedy parents search, one has to specify a score function to evaluate the considered graph. As score for the evaluation, the p-value is chosen of the test of conditional independence of the potential parent variable and the target variable given $X_1$. The variable $X_1$ is held fixed as a parent variable derived from the analysis with the PC-algorithm. Candidates for additional variables yielding the best scores are $X_5$ (p-value 0,011) and $X_9$ (p-value 0,042). For the identification of promising target groups in a next step the conditional probability distribution of the outcome variable *Y* can be estimated given the other variables. Some examples are given in the following tables.

Given the state of $X_1$=7, the response rate defined by this instantiation is 3,69% (see Table 3.3). A further segmentation with the variable $X_5$ leads to the following interesting groups displayed in Table 3.5.

**Table 3.5:** Segmentation by variable $X_1$, and $X_5$

| instantiation of $X_1$ | 7 | | | | |
|---|---|---|---|---|---|
| instantiation of $X_5$ | 1 | 2 | 3 | 8 | 9 |
| cases | 378 | 2867 | 1249 | 611 | 2605 |
| rate | 6,35 | 3,91 | 3,44 | 4,75 | 3,03 |
| index | 192 | 118 | 104 | 144 | 92 |

If the group is considered where $X_1$=8 and the response rate is 3,58%, the following responder clusters can be identified (see Table 3.6).

**Table 3.6:** Segmentation by variable $X_1$, and $X_5$

| instantiation of $X_1$ | 8 | | |
|---|---|---|---|
| instantiation of $X_5$ | 1 | 2 | 9 |
| cases | 639 | 3012 | 2375 |
| rate | 3,76 | 3,95 | 3,24 |
| index | 114 | 120 | 98 |

In the next Table 3.7, the results are displayed for the state $X_1$=9. This yielded the highest response rate of 4,15% in the first approach. A further refinement with the variable $X_9$ illustrates further potential for better future response rates (see Table 3.7).

**Table 3.7:** Segmentation by variable $X_1$, and $X_9$

| instantiation of $X_1$ | 9 | | |
|---|---|---|---|
| instantiation of $X_9$ | 1 | 2 | 3 |
| cases | 4950 | 2342 | 977 |
| rate | 4,14 | 4,7 | 4,2 |
| Index | 125 | 142 | 127 |

Overall, the groups defined by the parent variables can be studied in more detail by consideration of the (estimated) conditional probability distribution of the target variable, given the states of the parents.

### 3.4 Comparison of the results

Both methods are appropriate to determine the relevant variables. In the considered study the variable $X_1$ turned out to be the best and most important predictor of the dependent variable. This can be derived by the direct influence of this variable in the graph in Figure 3.3 obtained by the dependency analysis as well as the segmentation at the first level by $X_1$ in CHAID. The dependency analysis approach does not pursue to analyse subgroups at the level of a specific variable. In comparison to the dependency analysis CHAID can derive group-specific predictors which is an advantage in the analysis of the considered problem. Therefore, the interesting segments detected by CHAID differ slightly from the results obtained by dependency analysis. In contrast to CHAID the dependency analysis reveals interdependencies between the explanatory variables which can be an advantage in case of several direct predictors, i.e. parent variables. In future response analyses it is recommended to apply both approaches due to their specifics complementarily to yield profound results.

## 4. Discussion

As demonstrated by the case study CHAID as well as dependency analysis are useful for the task of identifying interesting clusters of the considered problem. The visualisation of the results derived by both approaches is intuitive and easy to communicate. Especially CHAID is characterised by high transparency of the segmentation process, which enables the user to modify the classification tree in accordance to his prior knowledge. By merging categories of predictors CHAID provides information for scaling variables in future analyses. Non of the methods requires any special functional form of relationship, which makes both approaches applicable for a wide range of problems. CHAID and dependency analysis require large sample sizes in order to obtain reliable results, but this was not restrictive in the case study analysed in this paper. Therefore, we conclude that Data Mining problems in marketing applications based on a designated criterion can be solved by any of the described and applied approaches.

## References


Brosius, F.: SPSS CHAID – Statistische Datenanalyse für Segmentierungsmodelle und Database Marketing, Internat. Thomson Publ., Bonn 1997.

Castillo, E. Gutierrez, J.M., Hadi, A.S.: Expert Systems and Probabilistic Network Models, Springer, New York 1997.

Cooper, G.F., Herskovitz, E.: A Bayesian Method for Construction Bayesian Belief Networks from Databases, in B.D. D'Ambrosio, P.P. Bonissone (eds.) Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo 1991, 86-94.

Dawid, A.P.: Conditional Independence in Statistical Theory, Journal of the Royal Statistical Society, Ser. B, 41, 1979, 1-31.

Fayyad, U.M., Piatetsky-Shapiro,G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth (eds.) Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park 1996, 1-36.

Jordan, M.I. (ed.): Learning in Graphical Models, The MIT Press, Cambridge (Massachusetts) 1998.

Kass, G.: An exploratory technique for investigating large quantities of categorical data, in Applied statistics, 29 (2), 1980, 119-127.

Magidson, J.: SPSS for Windows CHAID Release 6.0, SPSS Inc., Chicago 1993.

Magidson, J.: The CHAID approach to segmentation modeling: Chi-squared automatic interaction detection, in Bagozzi, R. P. (ed.), Advanced Methods of Marketing Research, Cambridge (Massachusetts) 1994, 118-159.

Pearl, J.: Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo 1988.

Spirtes, P., Glymour, C.: An Algorithm for Fast Recovery of Sparse Causal Graphs, Social Science Computer Review, 9 (1), 1991, 62-72.

Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd Edition, The MIT Press, Cambridge (Massachusetts) 2001.

Spirtes, P, Scheines, R., Meek, C., Richardson, T., Glymour, C., Hoijtink, H., Boomsma, A.: Tetrad 3, http://hss.cmu.edu/html/departments/philosophy/TETRAD/tet3/master.htm, 1999.