Market modelling · 10
Data mining models - 2

Chuck Chakrapani

Problems with AID
Although the classic AID technique had several attractive features it hadseveral problems too. These had been documented in papers dating back to the mid-seventies (See The Pitfalls of AID Analysis by Doyle and Fenwick, Journal of Marketing Research, 12, 408-413, 1975).

Multiway classification
One major limitation of AID is its binary split procedure. At each level, the technique simply split the group into two. This made the analysis quite cumbersome. Consider an example in which usage of a product varies by three different income groups: low, medium and high. AID groups two of the three groups together, since the split can only be binary. It might split the larger group at the next level. However, this is not always guaranteed. If some other variable was related to usage after the first split, this other variable will take precedence over income. The result is that we may never know that there are three distinct usage patterns that are related to income.

Optimal recoding
There is another advantage to multiway classification. Suppose our data contains 10 different income categories. Multiway classification splits the 10 groups into three income categories, such as: Group 1: Income categories 1 to 4; Group 2: Income categories 5 to 8; Group 3: Income categories 9 and 10. In this case we can, instead of arbitrarily deciding the low, medium and high categories (as we would do in regular cross-tabulations), define what the three income categories are on the basis of the usage patterns of these groups. This is equivalent to rescaling and optimally recoding the income variable such that it can be related to our marketing objective.

CHAID and KnowledgeSeeker (KS)
Kass introduced the CHAID technique (now available as a part of an optional module in SPSS) in 1980 to identify the best multiway grouping of data on the basis of statistical significance tests. This technique was further refined and improvements are incorporated in a commercial program known as the KnowledgeSeeker. (The similarities and differences of these two techniques are beyond the scope of this article. Here we will simply deal with what these techniques do.)

Data mining applications
CHAID/KS techniques has several applications to data mining. Let us start with the normal way we analyse data. We cross tabulate data using certain variables as banner points (the independent variables). Suppose that one such variable is income, which we have defined as follows: Bottom one-third = Low Income; Middle one-third = Middle Income; Top one-third = High Income, and obtain the following results:

Income group	Usage
Low (1/3)	20%
Medium (1/3)	25
High (1/3)	27

The figures indicate that income is somewhat related to usage and that the higher the income the higher the usage. But as we saw earlier, the scale we chose to group customers (bottom one-third, middle-one-third and top one-third) is purely arbitrary and there is no reason why usage should follow these groupings. The real changes in usage occurs at some levels of income and consumers are distributed as follows:

Income group	Usage
Low (Bottom 50%)	20%
Medium (40%)	30
High (Top 10%)	20

It is obvious that income is very strongly related to usage. The usage is substantially higher in the middle income group than in the top or thebottom income groups. This fact is effectively hidden in the crosstabsbecause we have no way of knowing that usage level changes at the median level of income rather than at the bottom 1/3 level. Similarly, another change in usage occurs at the top 10% level and not at the 2/3 level that we used in the cross-tabs. (see Exhibit 1A and 1B. Exhibit 2 shows the results of an analysis carried out by Dr John Liefeld of the University of Guelph. Note how usage of CSBs jumps from 18% to 28% when income exceeds $15,000. If income groups were arbitrarily defined, they will be subject to distortions illustrated in Exhibit 1A and 1B) CHAID/KS techniques are designed to spot this hidden pattern in the data.

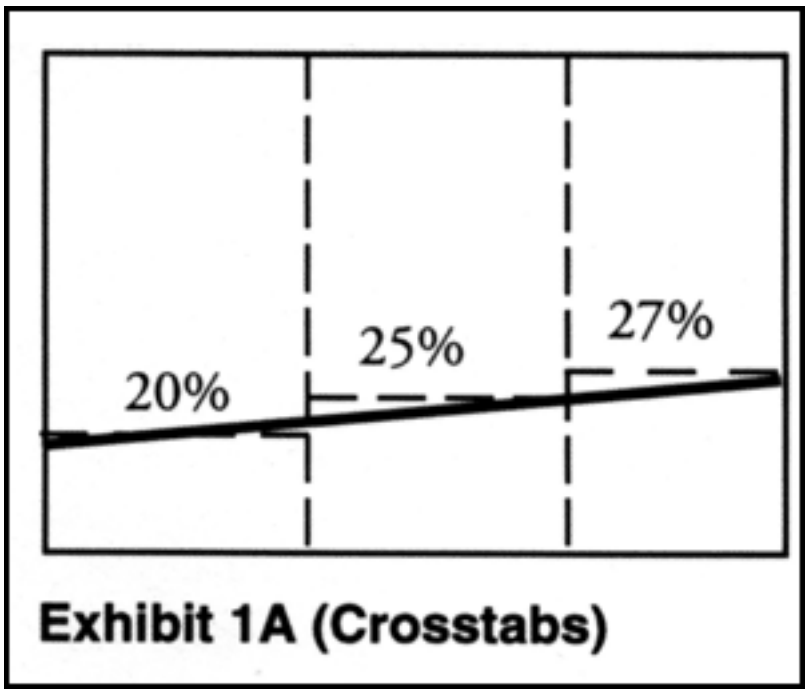


Exhibit 1A (Crosstabs)

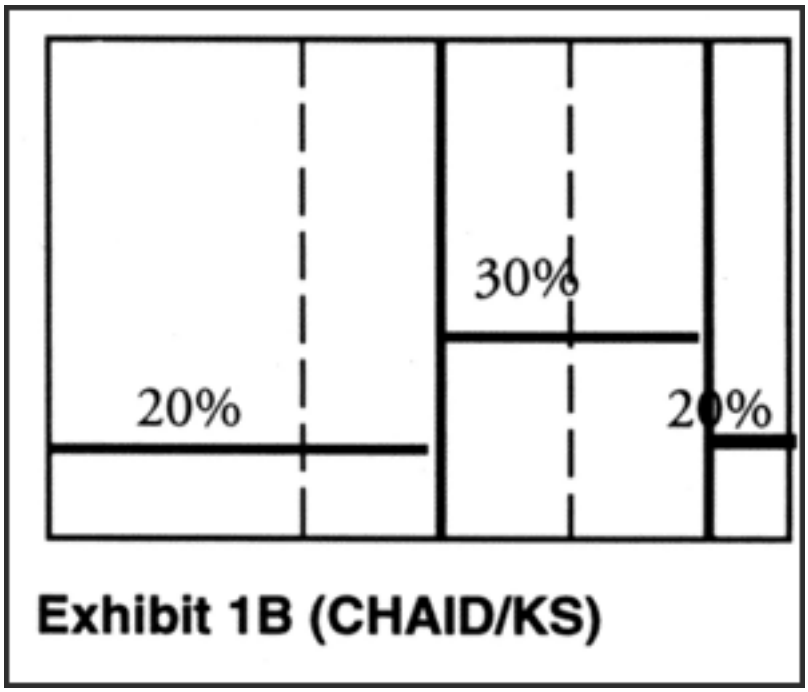


Exhibit 1B (CHAID/KS)

How does a combination of variables influence...?
Sometimes we work with a combination of variables. Suppose we are working with a database for a financial institution. We may want to know how a combination of products held in your institution by a customer affects your profitability. In this case your dependent variable will be profitability and the independent variables will be a Ometa-variableO which indicates whether a person owns each of the six products offered by the institution. The results might indicate that (1) customers with a single product are the least profitable; (2) customers with 3 or 4 products are the most profitable: (3) customers with 5 or more products are not any more profitable than those who hold only 2 products. These meta-variable analyses are difficult and practically impossible in many cases to carryout manually or through straight cross-tabs.

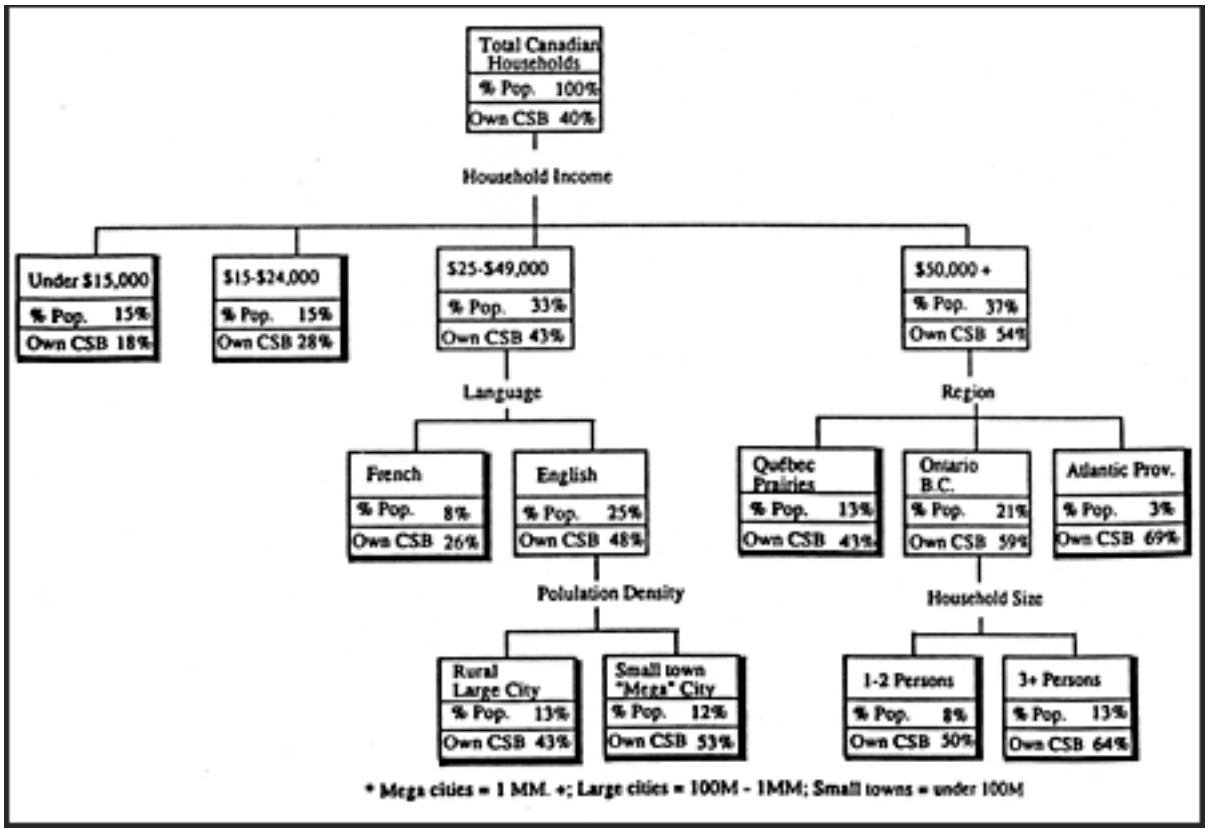
What if ...?
Suppose your analysis indicates that a 10% offer increases your returns.

But so do many other factors. If would like to know the effect of 10% off, after having identified all other factors that contribute to usage, you can use CHAID/KS procedures. To do this, we first carry out the analysis using all other variables of interest. Then, we can use the 10% off offer to identify its residual effect. By doing this we may also be able to identify the segments for which the offer has appeal and those for which it doesnot. If the groups for which the offer is appealing are not profitable,then we may decide not to offer the discount, since customers are reacting favourably.

CHAID/KS techniques make data mining easy because they explore voluminous data with relative ease (I recently used KS to analyse 40,000 records and the program went through the analysis in a short period of time.) They show us how to regroup our variables such that they are scaled optimally. They enable us to explore how variables act in combination with other variables as well as singly. They enable us to perform what-if operations.

Another advantage of this technique is its ability to handle inherently non-linear relationships. If we examine Exhibit 1B, it is clear that the relationship is not linear. Usage level is low at either end of the income continuum but high in the middle. Common multivariate techniques we use in marketing research are linear. Consequently, a technique like multiple regression analysis would not have been able to identify the relationship between usage and income in that example correctly. These advantages make CHAID/KS very useful tools for data mining.

Exhibit 2: Key Demographic Splits for Canada Savings Bond Ownership



Book Review

Correspondence Analysis in the Social Sciences

Edited by Michael Greenacre and Jorg Blasius
Published by Academic Press, Harcourt Brace Publishers, London (UK), 370p. £45.

One main problem faced by many analysts is the scarcity of materials that deal with the interpretation of computer outputs obtained by using special analytic techniques. Most textbooks on multivariate analysis usually provide one uncomplicated output and explain what the output means. What happens when the data are a bit more complex and the results do not neatly fit the example provided? In these cases, the analyst is usually on his/her own. That this is true of so many techniques can be seen by the low level explanation (or exaggerated claims) that normally accompanies the analysis of multivariate analysis.

Therefore, any book that deals with the interpretation of the results generated by applying a technique is welcome. When a book covers as much ground as this one , it is particularly appreciated.

Three chapters - 1,3 and 7 - provide an introduction to simple, multiple and joint correspondence analysis and its computational aspect. While an even simpler introduction to the topic can be found in Greenacres Correspondence Analysis in Practice, there are many actual examples of correspondence analysis that can be invaluable to marketing researchers.

Some of the cases provided in the book deals with longitudinal data such as measuring changes in household panel data. There is also a chapter on visualizing as structural changes by means of correspondence analysis. Another chapter deals with event history data by cluster analysis and multiple correspondence analysis.

Other chapters deal with some unusual (at least in the U.S. and Canada) applications of correspondence analysis, such as the analysis of textual data from personal advertisements. A chapter that may be of particular interest to marketing researchers is entitled Product Perception and Preference in Consumer Decision Making.

Correspondence analysis is becoming increasingly popular in marketing research. While techniques such as perceptual mapping and conjoint analysis do not have many books devoted to their applications, correspondence analysis seems to have attracted full length books that discuss its applications. Even when the technique is rich, if we don't know how to use it to its fullest potential, it is of limited use.

It is, therefore, heartening to see books such as these that show how to use a technique and how to interpret information after we have carried out the analysis.

Chuck Chakrapani

Dr. Chuck Chakrapani of Standard Research Systems is a Toronto-based consultant, author and seminar leader. He works internationally.