History and Potential of Binary Segmentation for Exploratory Data Analysis

James N. Morgan
University of Michigan

Abstract: Exploratory data analysis has become more important as large rich data sets become available, with many explanatory variables representing competing theoretical constructs. The restrictive assumptions of linearity and additivity of effects as in regression are no longer necessary to save degrees of freedom. Where there is a clear criterion (dependent) variable or classification, sequential binary segmentation (tree) programs are being used. We explain why, using the current enhanced version (SEARCH) of the original Automatic Interaction Detector program as an illustration. Even the simple example uncovers an interaction that might well have been missed with the usual multivariate regression. We then suggest some promising uses and provide one simple example.

Key words: Decision trees, interaction detection, sequential binary segmentation.

1. Introduction

Thomas Kuhn argues that science progresses by running into problems and then making a major shift in its paradigms or models. (Kuhn, 1996). We may be seeing such a shift as the available data become richer and the possible explanations multiply. The conventional wisdom instructed us to form a theoretical model, test it for statistical significance, whether the probability was less than .01 that the null hypothesis that it had no explanatory power might be true. Once we left designed experiments with their decomposition of variance and started dealing with human beings, the number of explanatory variables increased dramatically. In order to save "degrees of freedom" restrictive assumptions were made, that the effects of each predictor were linear, and that their individual effects were independent and additive. So we came to the linear models of multiple regression. Such analysis fails Karl Popper's criterion of falsibility, since a predictor might easily have large and significant effects on some subgroup of the population (Popper, 1965).

The objectives of data analysis, in order of their importance, were:

1. Tests of significance; probability that the model doesn't represent the population.
2. Estimates of THE size of the effect of each predictor,
3. How much unexplained variance (error) was reduced by each predictor,
4. Whether effects were linear and additive.

With these priorities, multiple regression or other similar linear models were appropriate, and have been widely used.

We restrict ourselves to searching for the optimal structure of explanation where there is a criterion or dependent variable, a rich assortment of plausible explanatory variables, and a substantial sized sample (500 or more). There is a variety of other statistical approaches for different problems. Multistage linear models worry about different error variances for data at different levels of aggregation, as when we have characteristics of the student, the class, the school, and the state. However, with large samples, error variances are not the issue, but how well we can predict the dependent variable, and when we put all our explanatory variables into a few classes (losing very little of their explanatory power), the level of aggregation in one sense does not vary that much across predictors (Raudenbush and Bryk, 2002; Luke, 2004).

Sometimes the term multilevel linear models is used, usually meaning that some interaction terms have been added, but nothing is said about their selection. (Goldstein, 1995)

The term "data mining" sometimes refers to searching for patterns or combinations that predict cheaters using credit cards or phone services, but sometimes merely a search for unusual patterns in large data sets. Instead of focusing on a criterion variable and operating sequentially, many programs simply look for oddities.

Log linear models were proposed when the criterion was a classification, but instead of starting with the safest, full sample searching, they create the detailed multiway table and see how it can be simplified by omitting of truncating explanatory classes. Hence the decisions can be idiosyncratic from the beginning. Search, using the chi option, reverses the sequence, producing a tree (root) the first splits of which at least are quite stable and reproducible (Chan and Loh, 2004).

Neural networks use tree diagrams like those SEARCH produces, but their goal is to investigate dynamic learning, where the weights given to each splitting decision depend on whether it paid off in earlier trials (Abdi *et al.*, 1999). Cognitive psychologists think this is how we learn, and applied engineers use such

learning systems to develop programs that recognize faces, decide which e-mail is spam, etc. The tree is assumed, not searched for.

2. The Innovation

With the advent of data sets containing thousands of cases, and increasing concern with the fact that testing one model after another verged on ransacking, an old statistical fact reappeared: The explanatory power of any one predictor is largely exhausted by using a few categories instead of the numerical detail. (In the statistical literature it is called the "loss from grouping"). (Yule and Kendall, 1937; Kalton, 1967). So one could abandon the assumption of linear effects by converting each predictor into a set of 1-0 "dummy variables" one for each category. Of course, the multiple regression cannot handle a perfect correlation among predictors, the computer balking at dividing by zero, so one class of each predictor had to be left out. This made the first two tests, significance and importance, chancy, depending on the excluded class, since the coefficients were all in terms of differences from that class. But it was possible to transform the coefficients of each predictor into a complete set with a weighted average of 0, even if it left the significance tests trickier.(Suits, 1957) The Institute for Social Research's Multiple Classification Analysis program did that, and the comparison of the original subgroup means for each class of each predictor with the same means adjusted by the regression coefficients provided useful information on the effects of the multivariate regression. (Andrews ,1967)

But the implications of the small loss from grouping went further. One could simulate what a researcher did in searching a data set for an appropriate model among a large set of potentially misspecified models, rather than imposing one model after another. The trick was sequential binary segmentation. We explain it when there is a numerical dependent (criterion) variable, but will discuss dependent variables that are dichotomies, classifications, rankings, and simple covariances later. One sweeps through the predictors, checking the explanatory power (reduction in error variance) of $k-1$ splits, group 1 versus the rest, $1+2$ and the rest, etc. (Or if the order is meaningless, each of the k subgroups versus the rest; or, more dangerously the $k-1$ tests after the classes have been reordered on the dependent variable.). Selecting the best split across all the predictors, the data are actually divided, and the process repeated on each sample subgroup thus developed. This abandons the assumptions both of linear effects, and of independent additivity of effects. Eash split is reported along with the percent reduction of the original unexplained (error) variance. These percentages are additive, the sum being equivalent to the r-squared of regression.

The focus of the selection of splits is on reduction in error variance. The summed squared deviations from the mean of a group is commonly called unex-

plained variance or error variance. In the analysis of designed experiments, the effects of various explanatory classifications in reducing error variance are measured by adding the variances within the subgroups, and subtracting that from the variance of the parent group. If there are two predictors, then a two-way table of subgroups provides a sum of variances around the subgroup means to indicate additional reductions in error from an interaction effect. The procedure is known as an analysis of components of variance. With binary segmentation programs, each possible split provides an error reduction, and the largest is selected for each predictor, then the largest of those across all the predictors is selected. For any one predictor, focusing on significance tests would give the same split, since there are always 1 by $n-1$ degrees of freedom, but across all the predictors, a large group with a large difference will be favored by focusing on error reduction. Actually the reduction in error, with some simple algebra, reduces to the squared means of each of the two subgroups times the number of cases (or weight sum) for that group.

When two predictors are correlated, or represent the same basic cause, splitting on one will reduce the potential of the other. A table is displayed in the output showing the remainlng potential of each predictor in each subgroup. Regression, in contrast, divides up the credit, or with high intercorrelations can vastly increase the estimated errors, and even reverse the direction of one estimated effect.

The world is full of "interaction effects", which were built into the experimental models in agriculture and elsewhere, estimated for explanatory power and tested for significance, but with multiple regression with many predictors, the number of possible cross product and triple-product terms explodes. With binary segmentation, interaction effects appear whenever each of two subgroups from a prior split is then split using a different predictor. Studying days in the hospital might first separate men from women, then find a totally different age effect with young women having children and old men with prostate problems. But other interactions are less obvious.

A crucial advantage of this is that it fits with Karl Popper's notion of falsibility, since if a predictor never manages to be used in any of the splits, then one can say that it not only doesn't matter overall, but also doesn't matter for any meaningful subgroup of the population (Popper,1965).

3. Some History

An automatic interaction detection program was described in an article in the Journal of the American Statistical Association (Morgan and Sonquist, 1963). Earlier, William Belson in England had proposed a segmentation approach in marketing (Belson, 1959). An actual program, called the Automatic Interaction

Detector, was developed with funds from the National Science Foundation and described in a monograph (Sonquist and Morgan, 1964). It was later improved and described in another monograph (Sonquist, Baker and Morgan, 1974) and then improved again and renamed SEARCH. It was heavily used in a research project published as Productive Americans, (Morgan, Sirageldin, and Baerwaldt, 1966). It was also described in a chapter of a book on data analysis (Fielding, 1977). An overview of AID and other such appeared in 1992 (McKenzie and Low, 1992).

The original program was called the Automatic Interaction Detector, and required a large computer so that the data could all be easily accessible. Since then, personal computers can handle substantial data sets and the extensive calculations necessary with blinding speed. A version updated and enhanced for the PC was developed by Peter Solenberger and Pauline Nagara at the Institute for Social Research. Called SEARCH, it is available free as a stand-alone program assuming that SAS is available for data management, filtering and recoding, at www.isr.umich.edu/src/smp/search (Sonquist *et al.* 1973). And a complete set of software with SEARCH and Multiple Classification Analysis, and data management, called Microsiris is available from Neal Van Eck free (http://www.microsiris.com).

Many similar programs, vaying in complexity, are available at varying prices. We have found that they tend to produce similar results. Others focused on categorical dependent variables (Breiman *et al.*, 1984). A study of many such programs (classification trees) also showed only small differeances (Lim, Loh and Shih, 2000). SPSS has a Tree Program, Salford Systems has CART. A variety of data-mining or neural network programs is also on the market, with apparently different objectives, though the explanations of what they do leave much to be desired (see http://www.kdnuggets.com).

4. The Enhanced SEARCH Program

The splitting criterion in SEARCH is the reduction in error variance from each single binary split. And the focus on reduction in error variance appears in the main stopping criterion, the minimum reduction in variance relative to the original total. The reason is another statistical fact, namely that as sample sizes get larger and larger, test of significance become less important. All formulas for standard errors have in their denominator the square root of the number of cases. Anything with effects worth looking at will be significant, or if it is not, would be if you doubled the sample size. But as subgroup sizes get smaller, a highly significant difference may be totally unimportant for predicting over the whole population. We have conducted tests of the results, using a genuinely independent part sample from the Panel Study of Income Dynamics (a random subsample of

a clustered sample is not really independent). The attrition in testing the results on a fresh sample was minimal for means, the variance explained (equivalent to an R-squared) only dropped from 15% to 13%. Of course, that was with 4500 cases to search and 1500 cases to test. Attrition was also small for covariances, but substantial for a "slopes only" option, which was droppped from the program.

Sometimes the criterion (dependent) variable is not a number, but a dichotomy, a ranking, a set of categories, or even a covariance with a dominant explanatory variable, like the effects of race or gender or age or education. The measure of explanatory power varies, but the principle is the same. For rankings, a rank correlation; for categories, a likelihood ratio chi square. A variety of rules for stopping include a minimum reduction in error, a pseudo significance test, the size of final groups, and the number of splits.

The results of a SEARCH run can be presented as a "tree" (really the roots), but if the criterion is more complex than a simple average, this gets difficult, so SEARCH has added a simple hierarchical table that, with a little editing, is publishable. An example is appended, along with a "tree". (The "tree" looks better in this simple example.)

Each set of predictor classes can be kept in order, or (dangerously) reordered, or used to compare each class against the rest combined. Predictors can be ranked, exhausting one set before entering the next. Recode can be saved to produce expected values or residuals. A premium can be set for splits that maintain symmetry for the second split of a pair. Splits can be predefined. Extreme outlyers at any stage can be omitted and identified. Default options allow analysis with minimal specifications. A wide variety of output is available, but automatically one gets the overall percent of variance explained, a warning if weights were used, a split summary table, a table of the best binary split of each group by each predictor (to see what almost made it), a distribution of a categorical dependent variable for each group, and a group tree structure, a hierarchical table easily edited into printable form (see example at end). Using item weights leads to underestimates of error variances, and the warning in the program provides an approximate estimate of how much.

5. Applications

Here are some examples of difficulties that can be handled by SEARCH better than by the prevailing methods: First a dichotomous dependent variable in a multiple regression is seen as difficult partly because of disparate variances — the standard deviation of a percentage is the square root of pq/n which is maximal at 50% and very small at the extremes. That only matters for significance testing. Second, the regression can produce predictions greater than 1 or less than 0, which are impossible. The reason of course is the assumption of independent additive

linear effects. In other words, the model is mis-specified. Transformations such as log-odds simply hide the problem. If we wanted a predictive model that is optimal, the SEARCH segmentation will produce one, and whether one uses variance reduction or chi square as the splitting criterion will not matter much. So one could predict whether an individual is likely to vote and adjust the number intending to vote for each candidate to a proper reduction for not voting. A more important use takes us to the next problem, selection bias.

Analysis of a data set or subset can be distorted if it excludes some individuals, such as non-employed women. The conventional solution is to estimate the probability that an individual is in the sub-sample being used, and introduce into the regression the inverse of the log odds. This assumes that the effect of selection is linear in logs and uniform across all subgroups. Sampling statisticians have for years used a more flexible method, namely simply weighting each case by the inverse of the probability it would be included. Again the lure of the conventional method is an easy test of significance for an overall selection bias, but a researcher might well want to see what the effect was on the estimated effects of other predictors, and on different subgroups. So an obvious solution is a SEARCH run to develop subgroups with widely different probabilities of inclusion, using the recode developed to assign a weight to each individual in the selected group. One can then run the analysis weighted and unweighted so see what changes. If nothing changes, the weights were "uniformative", and the weights were unnecessary, but it can well be that the effects are concentrated on some subgroups.

A third advantage uses the chi option for a categorical criterion variable when one suspects that different factors cause increases from decreases and from no change. Take saving, for example. Both substantial savers and dissavers differ from those with little or no saving, the latter mostly having no assets and little credit. And large dissavers are often engaged in capital transactions or bequests. So converting saving to five categories and treating it as a classification allows a flexible initial search for structure.

If information is missing for some cases on the dependent variable, SEARCH can be used to estimate it, but a more flexible procedure would be to exclude those cases and weight by the inverse of the probability of good data. Missing data on explanatory variables can be assigned, again using a search run on the rest of the cases. Those concerned with variances and tests of significance point to the smaller variances of these subgroup means used in the assignments, but with large samples and a focus on explanatory power, this seems of little concern. Where it is, one can add random errors to the assignments. Such a program can also assign missing categorical data either using the modes or distributing the assignments according to a subgroup distribution.

Sometimes our concern is with sequences of causation, from clearly exogenous background variables like age, gender, race, where grew up, through actions like geographic mobility, educational achievement, occupational selection, to current variables where the causation could easily go the other way, e.g. marital status, attitudes and behaviors, local conditions, or work hours. It is easy to save the recode to generate residuals from one analysis to be used in a subsequent analysis. This was done extensively in Productive Americans (Morgan *et al.*, 1966). The same two-stage procedure can be used to analyze first whether and then how much, or whether up, down or no change, and how much.

Particularly in policy-relevant areas, there is a pressing demand for aggregate estimates, such as the effect of an income increase on aggregate consumption. But in fact, the effects of X on Y even "net of other things" as in regression, can be quite deceptive. The aggregate consumption function so estimated would be based on the current distribution of income, but dramatic changes in income distribution could alter the estimate.

Another advantage of the sequential approach is that in using the results for diagnosis or decision making, one can combine the error reduction of each split with the cost of securing the information necessary, and economize on the use of information in decision making. This would seem particularly important in medical diagnosis where tests are expensive.

6. Implications

So we come to reconsider the priorities. Instead of testing a model using restrictive assumption, we are searching for the best among a large set of potentially misspecified models using large data sets. Our priorities are reversed: We want :

1. To see whether there are interaction effects, non-additivies across predictors, where each of a pair is then split on different predictors
2. To see how we can best reduce the predictive error variance,
3. To see the extent of the effects, and their variability among subgroups.
4. To be sure we can extrapolate the results at least to the parent population, if not to all populations at all times.

The last of these is reflected in one of the alternative stopping rules, which applies an illegitimate significance test before allowing the split. More important, a careful researcher will hold out a genuinely independent small subsample and force the subgrouping from the SEARCH run to see whether the total error reduction is seriously attenuated. In many samples there is clustering at several levels which means that a randomly selected subset is not independent and only

the sample designer can specify a small independent sub-sample. Large samples are required for the searching process, but significance testing almost by definition does not require a large sample.

Searching is not mindless ransacking but a reproducible process with pre-stated procedures. Once the predictors are chosen, and a decision made on each whether one tries the $k-1$ in order, or each of k against the rest, or the $k-1$ after reordering, and a set of stopping rules selected, the computer searches for the best model in terms of predictive power. And since the early splits are based on larger numbers of cases, they are most dependable A doctor, desiring to have guidance as to diagnosis would like to find a subgroup where one diagnosis dominated, but he would also like to know what other diagnoses were also common for that group. The approach is appropriate when there is a criterion variable or classification, many explanatory characteristics, and a substantial sample size. Other problems in the broader data mining field call for different approaches. Approaches which start with elaborate detail, like the k-way tables of long-linear models, face the likelihood of idiosyncratic choices when early decisions are based on very few cases.

7. Conclusion

The shift from model testing to searching for the best model is indeed a paradigm shift, and the resistance from those accustomed to the conventional approach has been intense. But we are seeing more and more cases of interaction effects, difficult to find with linear regression models. The example appended surprised us, with its indication that the effects of mobility and location on earnings differ depending on one's education. It is becoming clear that it is combinations of genes and even combinations of genes with environment that lead to medical outcomes. Any one of several things can make you poor, and only a combination of skill, luck, and perseverance can make you rich. Any new approach can be overdone, and marketers "segmenting" mailing lists have often pushed the process beyond where it is reproducible. The notion of holding out a small sub-sample to test for attrition (a sign the splitting went beyond reason) is not easy to sell. One remedy lies in the fact that SEARCH demands an initial selection of variables that makes it reproducible on the same or a new set of data. It is not mindless ransacking whose artistic details get lost. Indeed, we need a convention that demands not only that the data be made available to others for analysis, but the SEARCH specifications as well.

Baker. Neal VanEck converted it to the PC, and Peter Solenberger and Pauline Nagara developed the enhanced program for the website.

Appendix: An Example.

Below is part of the output, a hierarchical table. The data are from the 1989 Wave of the Panel Study of Income Dynamics reporting for 1988. The documentation and full data set are available at http://psidonline.isr.umich.edu¡http://psidonline.isr.umich.edu¿. The hourly earnings come from questions on annual earnings, weeks worked, and hours per week and only with those with earnings between \$.50 and \$99.99 were included.

Group Tree Structure

```
Group 1: All Cases
    Group 2: V17545: V17545:
        Group 10: V16973: V16973:
             Group 12: V17545: V17545:
             Group 13: V17545: V17545:
                  Group 14 V16631: V16631:
                  Group 15 V16631: V16631:
                       Group 18: V16306: V16306:
                       Group 19: V16306: V16306:
        Group 11: V16973: V16973:
    Group 3: V16974: V16987


         ...          ...          ...
```

Figure 1: An example of the output, a hierarchical table. Descriptions for each group are omitted to save printing space. A typical example is:
Group 14 V16631: V16631: AGE OF 1989 HEAD, Codes 1-3
$N = 703$, Sum$(WT) = 9154.10$, Mean$(Y) = 12.5821$.
In the above output, groups of the same indentation are of the same "rank" in the hierarchy. For example, Group 14 and Group 15 together form Group 13 and their N values add up to the N value of Group 13.

Group 1: All Cases, $N = 5235$, mean(Y)=13.0220

|Group 2: Education 1989 H, no college degree, $N = 4186$, mean(Y) = \$10.81
| |Group 10: Married men, $N = 2535$, mean(Y) = 11.01*
| | |**Group 12: Education 12 grades of less, $N = 1420$,**
| | | **mean(Y) = \$11.01***
| | |Group 13: Education 13+, no college degree, $N = 1115$,
| | | mean(Y) = \$11.45
| | |**Group 14: Age=18-34, $N = 703$, mean(Y) = \$12.58***
| | |Group 15: Age=35+, $N = 412$, mean(Y) = \$16.24
| | |**Group 18: City of 25,000, nearby, $N = 287$,**
| | | **mean(Y) = \$17.83***
| | |**Group 19: No city of 25,000+, nearby, $N = 125$,**
| | | **mean(Y) = \$12.85***
| |**Group 11: Single man or woman, $N = 1651$, mean(Y) = \$8.55***
|Group 3: College grad or more, $N = 1049$, mean(Y) = \$19.48

|**Group 4: Age 18-29, $N = 374$, mean(Y) = \$14.17**
|Group 5: Age 30+, $N = 675$, mean(Y) = \$22.00

|Group 6: Married men, $N = 530$, mean(Y) = \$24.34
| |Group 8: Age 30-39, $N = 329$, mean(Y) = \$20.77
| | |**Group 16: Living in same state, group,**
| | | $N = 174$, **mean(Y) = \$16.86***
| | |**Group 17: Living in different state,**
| | | $N = 155$, **mean(Y) = \$25.30***
| |**Group 9: Age 40+, $N = 201$, mean(Y) = \$28.04***
|**Group 7: Single men, women, $N = 145$, mean(Y) = \$16.38***

Figure 2: Hourly earnings of family heads in 1984. With some easy editing, we obtain the same hierarchical structure by using proper bars. This is a presentable hierarchical table since the groups are more clear. Again, the detailed description of each group is omitted. Implications: College grads need to move to different state, while the rest need to be in or near a large city. Age matters more to college grads. **Ten final groups accounts for 23.2% of the variance.**
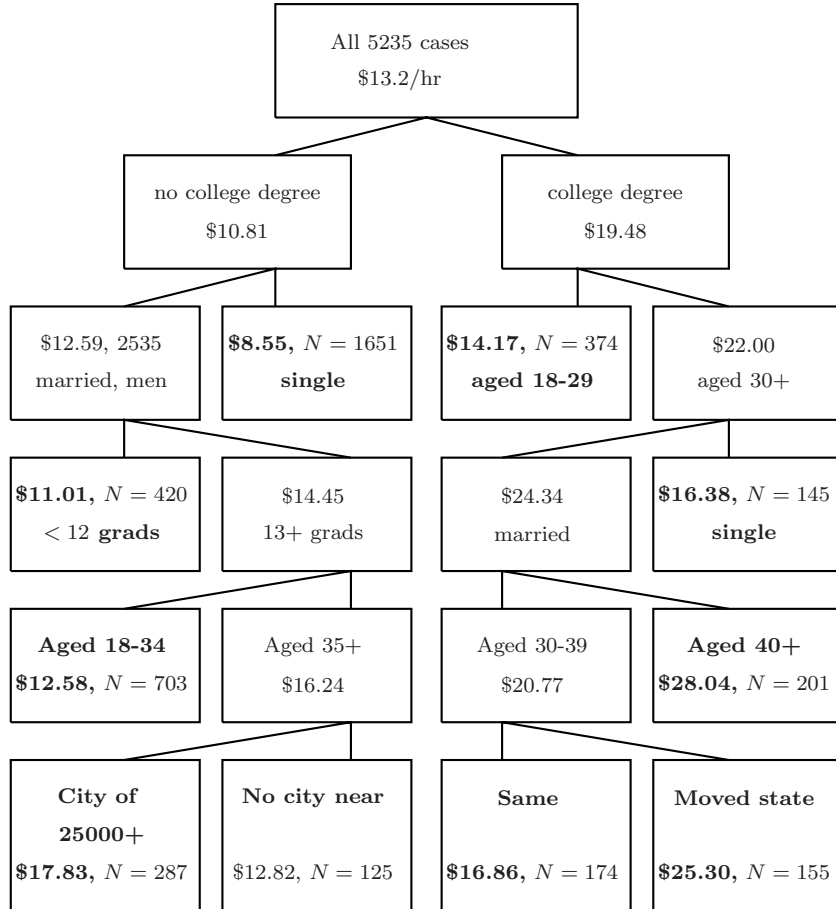
Figure 3: The final hierarchical graph. The implications are: College grads need to move to a different state, while the rest need to be in or near a large city. Age matters more to college grads. With 5235 cases, ten final groups account for 23.2% of the variance, equivalent to an R^2, but not testable for significance because of the searching.

References

Abdi, H., Valentin, D. and Edelman, B. (2004). *Neural Networks*, Sage Publications.

Andrews, Frank (1967). *Multiple Classification Analysis: A Report on a Computer Program for Multiple Regression Using Categorical Predictors.* Survey Reseach Center, University of Michigan.

Belsen, William A. (1959). Matching and prediction on the principle of biological classification. *Applied Statistics* **8**, 65-75.

Breiman, L., Friedman J., Olshen R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Chan, K-Y., and Loh, W-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of the Computational and Graphical Statistics* **13**, 826-852.

Fielding, A. (1977). Binary segmentation: The automatic interaction detector and related techniques for exploring data structures. In *Analysis of Survey Data, Vol I, Exploring Data Structures* (Edited by. C. A. O'Muircheartaigh and C. Payne). 1977.

Goldstein, H. (2002). *Multilevel Statistical Models*. Halsted.

Kalton, Graham (1967). *Techniques for Choosing the Number of Alternative Response Categories to Provide in Order to Coerce an Individual's Position on a Continuum*. Institute for Social Research, University of Michigan, Ann Arbor.

Kuhn, Thomas (1996). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lim, Tjen-Sien, Loh, Wei-Yin, and Shih, Yu-Shan (2000). A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. *Machine Learning* **10**, 203-228.

Luke, D. A. (2004). *Multilevel Modeling*. Sage Publications.

Mckenzie, D. P., and Low, L. H. (1992). The Construction of computerized classification systems using machine learning algorithms: An overview. *Computers in Human Behavior* **8**, 155-167.

Morgan, James N. (1985). Comparing static and dynamic estimates of behavioral responses to changes in family composition or income. *Journal of Consumer Research* **12**, 83-89.

Morgan, James N. (1990). A conditional analysis of movers housing responses. *Journal of Economic Behavior and Organization* **14**, 79-95.

Morgan, James N. and Sonquist, John A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association* **58**, 415-435.

Morgan, James N., Sirageldin, Ismail and Baerwaldt, Nancy (1966). *Productive Americans*. Institute for Social Research.

Popper, Karl (1965). *The Logic of Scientific Discovery*. Harpers.

Raudenbush, S. and Bryk, A. (2002). *Hierarchical Linear Models*. Sage Publications.

Sonquist, J. A., Morgan, J. N. (1964). *The Detection of Interaction Effects*. Survey Research Center, University of Michigan.

Sonquist, John A., Baker, Elizabeth L. and N. Morgan, James N. (1973). *Searching for Structure* (Rev. Ed.). Institute for Social Research.

Suits, Daniel (1957). The use of dummy variables in regression equations. *Journal of the American Statistical Association* **52**, 548-551.

Yule, G. Udny, and Kendall, M. G. (1937). *Introduction to the Theory of Statistics.* Griffin.

James N. Morgan
Senior Research Scientist, Emeritus
Institute for Social Research
The University of Michigan
Ann Arbor, Michigan 48106
jnmorgan@umich.edu