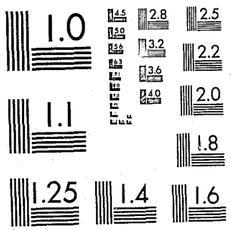
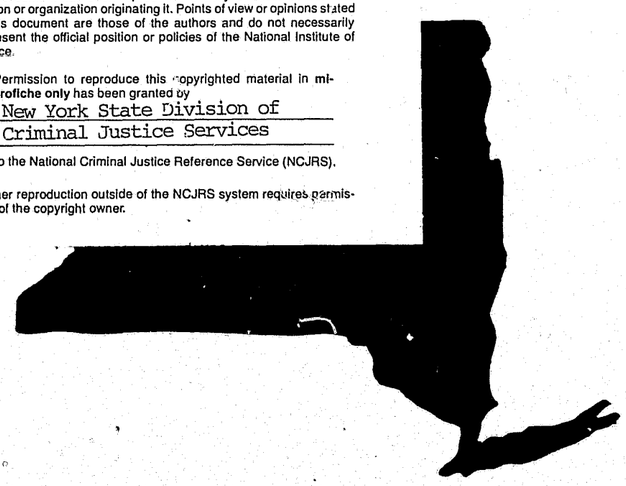


OFFICE OF POLICY ANALYSIS, RESEARCH & STATISTICAL SERVICES

NEW YORK STATE DIVISION of CRIMINAL JUSTICE SERVICES

PREDICTIVE ATTRIBUTE ANALYSIS: A TECHNICAL REPORT ON THE VALIDITY AND RELIABILITY OF THE METHOD

February, 1985



NEW YORK STATE
DIVISION OF CRIMINAL JUSTICE SERVICES
Richard J. Condon
Commissioner

OFFICE OF POLICY ANALYSIS, RESEARCH AND STATISTICAL SERVICES
Sherwood E. Zimmerman
Deputy Commissioner

PREDICTIVE ATTRIBUTE ANALYSIS: A TECHNICAL REPORT ON THE VALIDITY AND RELIABILITY OF THE METHOD

February, 1985

Bureau of Research and Evaluation
Bruce Frederick
Chief

Prepared by:
Newton F. Walker

Predictive Attribute Analysis: Contents of the Technical Report

I

INTRODUCTION AND BACKGROUND

INTRODUCTION AND BACKGROUND

Purpose

The purpose of this Technical Report is to provide criminal justice researchers with a better understanding of Predictive Attribute Analysis. We have found PAA to be a method encumbered by complex statistical undercurrents that are veiled by the simplistic tree diagrams and subgroup categorizations that result from an analysis. To the extent that users of methods like PAA become aware of these issues, however, the interpretations drawn can be substantially improved.

The characteristics of PAA are examined in this report by (1) considering explicit and implicit analytic assumptions and their implications, and by (2) conducting a series of empirical studies using PAA methods with both real and artificial data systems. A computer program has been developed to facilitate the conduct and evaluation of PAA studies. At the conclusion of the report we present some suggestions for appropriate research applications of PAA.

Scope of the Project

This project[1] originated as a consequence of validity and reliability questions that arose from the use of the Predictive Attribute Analysis method by our agency. The primary objectives were (first) to develop for ourselves a clear understanding of the Predictive Attribute Analysis method, (second) to use this background to build a flexible computer program with which reseachers could both conduct an analysis and have a sense of the confidence appropriate to the results, and (third) to produce a comprehensive report detailing these activities.

The Technical Report for the project provides a description of the PAA method and a discussion of important methodological issues which bear on its

use. Readers of the report are expected to have at least an intermediate-level background in statistics, but an attempt has been made throughout the report to supply appropriate references for readers not familiar with particular topics. Potential users of the computer program produced by the project are strongly encouraged to read the entire report before attempting to use PAA in substantive applications.

The contents of this Technical Report include: (1) a description of the features of the Predictive Attribute Analysis method, (2) a review of some published applications and evaluations of PAA, (3) a statement of methodological issues affecting the use and interpretations of PAA, (4) a discussion of statistical issues relevant to an understanding of these methodological issues, (5) a description of the PAAVE computer program, (6) a discussion of a series of simulation analyses performed to investigate remaining questions about PAA, and (7) a summary statement providing conclusions and recommendations concerning the proper use of Predictive Attribute Analysis.

The computer program was written not only to provide a means for the computational processing of a PAA, but also as a vehicle through which to study some of the issues which have been found to affect the validity and reliability of PAA results. Requisite computational features discussed in the PAA literature, as well as features deemed useful from our present investigation of the method, form the basis of the program. Where theory-based answers to questions about the credibility of an analysis are not easily provided, we have tried to provide appropriate empirical feedback through the program.

One of the most interesting and novel features of the program was not a part of the original conceptualization. Because of our concern for the potential variety of causes of instability within the analysis of a particular data system, it became clear midway through the project that some means would be necessary to evaluate the strength of such effects. Our solution -- bootstrap resampling -- provided such valuable insights for our own applications that we

decided to automate and incorporate this cross-validation capability directly into the version of the program we created for general distribution.

The program is intended for use by researchers who are familiar with Predictive Attribute Analysis. As such, the program does not necessarily prohibit some of the procedures or applications that we will caution against in this report. We urge potential program users to read this report thoroughly before using the PAAVE computer program for serious analytic work.

The User's Guide provides a detailed description of the various functions incorporated into the PAAVE computer program. Guidelines for the operation of the program and examples illustrating different parameter specifications are presented. The Guide is intended to be used in conjunction with this report, as it contains little discussion of the logical, methodological, or statistical issues to be presented here. A complete FORTRAN program listing with internal documentation is included, as well as examples of program output.

Predictive Attribute Analysis

Predictive Attribute Analysis is a quasi-statistical technique for the sequential subdivision of groups on the basis of characteristics of those groups that predict a criterion attribute well.

Briefly, PAA processing proceeds as follow: Given a set of predictor variables for a criterion of interest, the analysis begins by selecting a best predictor attribute for the criterion variable. The total group is then divided into subgroups on the basis of the presence or absence of this predictor attribute. The analysis then proceeds to find the best predictor attribute within each of the subgroups and to define new subgroups at this level, continuing in this manner until specified stopping criteria are met.

The PAA method has often been used to explore interaction effects among a given set of predictor variables, as well as to define characteristics of

important subgroups of a particular population. Results are usually summarized in the form of a PAA 'tree' (dendograph), where branches represent the sequence of sub-categorizations and where nodes show the particular attributes found to best predict the criterion. For example, the observed rate for an incarcerate/not-incarcerate decision might be expected to differ depending on the race, sex, and prior-history characteristics of individuals; a PAA might be used to see if particular combinations of these attributes have especially high or low incarceration rates.

Characteristics of the particular subgroups might then be studied further, or a prediction table or equation might be constructed. Because a PAA provides a logical decision process for selecting predictors, the (laborious) alternative of examining all possible combinations of contingency tables is avoided.

The PAA technique is often used to highlight the interaction between the relative usefulness of prediction information and the subgroup membership of an individual -- an analysis typically yields a different succession of predictor variables for each different subgroup.

Predictive Attribute Analysis is one of the class of categorical data analysis methods which derive from the mechanization of a set of decision rules rather than from the implementation of a formal statistical argument. More generally, PAA can be viewed as one of the many prediction methods intended for use with restricted-value dependent measures. The PAA method has been considered useful by some practitioners because it is a relatively efficient automation of an otherwise complex and cumbersome decision process.

Statistical considerations are a secondary aspect of the PAA prediction process. They provide the means for the selection of particular variables at particular decision points, but they are pointwise to the extent that information on either horizontal or vertical planes of the analysis is not incorporated into the decision. For instance, the selection of a predictor

attribute for a given subgroup is not affected by that variable's correlation with any previously-selected predictors for that PAA branch. One consequence of this limited depth of field is the poor reproducability of PAA results which has been observed, but PAA is certainly not alone among prediction methods in exhibiting a characteristic of poor reproducability.

Example of the PAA Method

Predictive Attribute Analysis is perhaps best introduced through a simplified example: Suppose one is interested in answering questions about the relation of violent and property crimes to certain known characteristics of large metropolitan areas. Information is available for ten such areas (courtesy of the 1982 Information Please Almanac) regarding the following characteristics: number of violent crimes reported, metropolitan population, number of law enforcement officers, change in number of officers since the previous year, and won/lost percentage for that area's National Football League franchise (Table 1.1a).

We have chosen for this example to do a Predictive Attribute Analysis in order to 1) find characteristics of an area which best predict crime rate, and 2) see if there are meaningful combinations of characteristics having markedly low or high predicted crime rates. Note that our illustrative analysis is exploratory in the sense that we are stating no hypotheses prior to examining the data.

Given our stated research questions and available data, we proceed with the PAA method. The steps to be followed in this example are outlined in Table 1.2.

As a type of 'attribute' analysis, PAA requires that all measures be dichotomous -- coded as the presence or absence of the characteristic. For the purpose of this example we choose simply to use a mean split on all variables (although, in general, justification would be required for such a choice, since the cutpoints for binary splits do affect PAA outcomes). Table 1.1b displays the recoded data.

TABLE 1.1a

Raw Data for PAA Example

Population (x 1ØØØ)	Law Enforcement Officers	Change in Number of LEOs	Football Percentage	Violent Crimes	Property Crime
7Ø71	28Ø12	- 4.9	.25Ø	1862	6874
3ØØ5	15242	- 6.4	.438	9Ø9	5192
2967	8998	- 6.7	.688	1515	7518
1688	8748	- 5.5	.75Ø	827	3872
1594	3831	- 1.1	.688	878	7874
12Ø3	5613	-11.9	.563	167Ø	7125
9Ø4	2593	Ø.7	.75Ø	1298	933Ø
787	3726	- 4.8	.438	1963	7361
679	2121	12.5	.375	1675	9Ø57
638	4583	- 1.5	.375	16Ø9	6993

TABLE 1.1b

Dichotomized Data for PAA Example

POP	LEO	CHG	FTB	VIO	PRO
1	1	Ø	Ø	1	Ø
1	1	Ø	1	Ø	Ø
1	1	Ø	1	1	1
Ø	1	Ø	1	Ø	Ø
Ø	Ø	1	1	Ø	1
Ø	Ø	Ø	1	1	1
Ø	Ø	1	Ø	Ø	1
Ø	Ø	Ø	Ø	1	1
Ø	Ø	1	Ø	1	1
Ø	Ø	1	Ø	1	Ø

TABLE 1.2

Steps Comprising a Predictive Attribute Analysis.

Ø - Determine research questions, assemble data, etc.

---

1 - Dichotomize variables

2 - Determine parameters to control PAA processing algorithm

---

3 - Construct 2x2 contingency tables for each combination of (dep var) x (indep var)

4 - Select the best (indep var) predictor of the (dep var) by the a priori statistical criterion

5 - Split the sample observations into two subgroups based on the two categories of this selected predictor

---

6 - Construct two sets of 2x2 tables (one for each subgroup) for (dep var) x (remaining indep vars)

7 - Select, for each subgroup, the best predictor from the (remaining indep vars)

8 - Split each subgroup into two further subgroups based on the four possible categories of the two selected predictor variables in each branch

---

9 - Proceed as above until specified stopping criteria for the analysis are met

Next one must select the parameters that control the execution of the PAA algorithm. Of primary importance is the statistical criterion used to measure relative predictability; the results of a PAA can be quite different depending on the type of predictive or associative relationship being measured. For now we somewhat arbitrarily choose the chi-square statistic, although in general an explicit rationale would be required.

Other parameters must also be supplied to determine at what point the analysis will terminate. These can be in the form of statistical significance tests (for the chi-square values), or they can be minimum cell or marginal counts for the individual contingency tables. Because of the artificial nature and small sample size for this example, we choose simply to let the analysis proceed through two levels and then terminate.

At this point we begin the first level of the analysis. First, 2x2 contingency tables are constructed for each pairing of the criterion (dependent) variable (VIO) with the four predictor (independent) variables (POP, LEO, CHG, FTB). Chi-square statistics are then calculated for each of the four tables. These tables and statistics are presented in Figure 1.1a. As can be seen, the chi-square statistic for the FTB variable is the largest, so FTB is chosen as the overall best predictor for number of violent crimes.

Next we divide the ten metropolitan areas into two subgroups based on each area's FTB characteristic; there are five areas with values of zero (losing seasons) and five areas with values of one (winning seasons). We then repeat the process of contingency table construction for each of the subgroups by generating 2x2 tables for each of the three remaining predictor/criterion combinations and then calculating the appropriate chi-square statistics (see Figure 1.1b). At this second level, given FTB=Ø, both POP and LEO are equally-best predictors. Given FTB=1, CHG is the best predictor.

Since we have reached our prespecified stopping criterion of two levels, we terminate the analysis at this point.

FIGURE 1.1a

Contingency Tables for PAA Example
Level Ø: Original Tables

POP

	Vio Crime Ø	Vio Crime 1
Ø	3	4
1	1	2

Chi-Square=.Ø8

LEO

	Vio Crime Ø	Vio Crime 1
Ø	2	4
1	2	2

Chi-Square=.28

CHG

	Vio Crime Ø	Vio Crime 1
Ø	2	4
1	2	2

Chi-Square=.28

FTB

	Vio Crime Ø	Vio Crime 1
Ø	1	4
1	3	2

Chi-Square=1.9

FIGURE 1.1b

Contingency Tables for PAA Example
Level 1: Conditional on Value of FTB

FTB=Ø

POP

	Vio Crime Ø	Vio Crime 1
Ø	0	3
1	1	1
	1	4

Chi-Square=1.88

LEO

	Vio Crime Ø	Vio Crime 1
Ø	0	3
1	1	1
	1	4

Chi-Square=1.88

CHG

	Vio Crime Ø	Vio Crime 1
Ø	1	2
1	0	2
	1	4

Chi-Square=.56

FTB=1

POP

	Vio Crime Ø	Vio Crime 1
Ø	3	1
1	0	1
	3	2

Chi-Square=2.81

LEO

	Vio Crime Ø	Vio Crime 1
Ø	2	1
1	1	1
	3	2

Chi-Square=.14

CHG

	Vio Crime Ø	Vio Crime 1
Ø	1	2
1	2	0
	3	2

Chi-Square=3.33

In reporting the results of the analysis, one would typically say something to the effect that "for areas with poor football records, population and number of law enforcement officers are the most useful predictors of violent crime rate, whereas for areas with good football records, relative change in the number of law enforcement officers is the most useful predictor of violent crime rate." This answers our first research question, and reinforces our use of the PAA method because of the different combinations of predictors found for the different subgroups.

With respect to our second research question, it is evident from the marginals of the second set of contingency tables (Figure 1.1b) that the conditional probability of a high number of crimes is different for the four terminal subgroups. For areas with poor football records and either small populations or small numbers of officers (top row), the probability of a high (versus low) number of violent crimes is 1.0; for areas with good football records and an increasing number of officers (bottom row), the probability of a high number of violent crimes is .00. Again we are pleased with our choice of the PAA method, since we can find intuitive support for these different expectations for crime frequency given the differing community characteristics.

We will withhold judgmental comments on the interpretation given for this example until we have discussed some of the methodological and statistical issues relevant to this type of classification analysis.

Overview of PAA Literature

The review of literature accompanying this report is intended to serve as an overview of the background and application of the PAA approach and as a context for subsequent discussion of some methodological and statistical issues. We look at PAA from the perspectives of both statistical literature and criminal justice applications literature, and we attempt to relate aspects of the method to analytical techniques which are most likely familiar to the reader. Because of the number of tangential issues that are raised as a result of a careful

study of the PAA approach, we encourage readers to consult the references cited.

Predictive Attribute Analysis is one of the class of qualitative data analysis techniques that began to appear in the 1960's. The method was developed by MacNaughton-Smith (1963, 1965) as an attempt to formalize the ad hoc approach for detecting interaction effects in the analysis of multivariate contingency tables. A number of other dendographic (or "tree") methods were also developed at this time, notably the Automatic Interaction Detection (AID) family of methods (Sonquist & Morgan, 1964). The AID methods have evolved considerably from the original model. See, for example, Morgan & Messanger (1973) for THAID, Perreault & Barksdale (1980) for CHAID, and Breiman et al (1984) for CART.

An early review of PAA is found in Simon's (1971) monograph. Simon's review emphasized applications of prediction methods rather than their theoretical or statistical justification. The work is particularly informative because comparisons are drawn for a half-dozen different types of association and prediction methods, including PAA.

The validation studies which were conducted for each of the predictive techniques Simon considered readily show the difficulty of producing generalizable results from any of these types of prediction techniques (given the selected data systems under analysis). Predictive power, defined here as the Pearson product-moment correlation between predicted and observed outcomes for a validation sample, was minimal for all techniques. In summarizing the series of studies, Simon concluded "In failing to produce an instrument of high power, the study shared the general fate of criminological predictor studies... (that) although small groups of good or bad risks can be distinguished, for many of the cases little discrimination is achieved." Possible explanations which were discussed included the omission of potentially important variables and the over-simplification caused by the use of dichotomized variables where continuous measures are available.

Sutton (1978) used the PAA method as one analytical component of a report addressing variations in federal criminal sentences, using regression analysis as a framework to discuss predictive power and PAA as a framework to discuss interaction. Sutton reported poor predictive power with both regression and PAA techniques, and he concluded that decisions were affected "largely on factors that were not included in this analysis." Neither the PAA nor regression results were cross-validated, however, and Sutton's results should therefore be interpreted cautiously.

Perreault and Barksdale (1980) have approached the attribute analysis paradigm from the perspective of marketing research. They discuss the use of the original Automatic Interaction Detection (AID) model (Sonquist & Morgan, 1964) and a modification (Chi-Square Automatic Interaction Detection) which provides certain improvements over the original method. The CHAID procedure is similar to the AID procedure in that it is a hierarchial search procedure used to identify tested interaction. Unlike the PAA procedures, it does not require a priori dichotomization of all measures. Chi-square tests for predictor variables selection are modified to reflect the number of comparisons being made at a particular decision point. Perreault and Barksdale were attracted to the AID/CHAID procedures because of the non-metric assumptions and the ability to combine binary, nominal, and ordinal levels of information. Noting the problems of misuse and misinterpretation that often occur with the use of AID-type procedures, they suggest using linear model or logit procedures to cross-validate the preliminary hypotheses suggested by an AID analysis.

Research Issues

No comprehensive and systematic study or evaluation of the PAA method could be found in the literature. This is especially unfortunate because of the many conjectures and implicit assumptions accompanying most PAA applications. Some of these points can be addressed directly through logical and statistical arguments, but other issues are more subtle and appear to interrelate with the types of data systems under analysis. These research questions form the basis

for the set of simulation analyses presented in this report. Besides providing very useful insights into the general performance characteristics of PAA, the simulations also led to major modifications in the PAAVE computer program to provide the user with better feedback regarding the level of confidence appropriate to the results of a given analysis.

Cited Issues

Predictive Attribute Analysis has some often-cited advantages and disadvantages, as well as some unknown characteristics. These have been suggested by the applications literature and our own use of the methods and are summarized in Table 1.3. We have attempted to approach each of these characteristics from a neutral perspective, looking for either supporting or disconfirming evidence. In addition, we have supplemented our original list of characteristics to be investigated beyond those which have usually been considered in the context of evaluating a methodology such as PAA.

We often found, in informal discussions, a reluctance on the part of practicing analysts to use PAA, especially as a singular analytic device. Most, however, could not express particular methodological or statistical reasons for their beliefs. We have tried in our evaluation to develop either theoretical or empirical grounds for these expressed areas of concern.

There are three generally cited advantages for the PAA method. Primary among these is that PAA is especially appropriate for detecting interactions among the predictor attributes. Whereas many analytic procedures require the a priori specification of a model, PAA can be used in an exploratory fashion to perhaps discover unanticipated effects.

A second cited advantage is the less-restrictive set of statistical assumptions that a PAA imposes on the data. There is no assumption of normality or even of continuously-distributed measures; the analysis proceeds using

TABLE 1.3

Some Cited Characteristics of PAA.

Cited Advantages

1) Identification of interaction effects.
2) Few assumptions on underlying distributions.
3) Clear display of pattern of relationships.
4) Non symmetrical

Cited Disadvantages

1) Difficult to implement using canned packages.
2) Requires dichotomized variables.
3) Potential masking of non-primary predictors.
4) Requires relatively large sample size.

Uninvestigated Characteristics

1) Sensitivity to sampling variability.
2) Sensitivity to rank order variation in samples.
3) Sensitivity to correlations among predictors.
4) Sensitivity to type of association measure used.
5) Sensitivity to stopping criteria.
6) Sensitivity to sample size.

dichotomous variables (frequently encountered in criminal justice research problems).

A third advantage is the clear display of the pattern of relationships that is the byproduct of a PAA. The tree-like branching diagrams clearly depict the subgroups and their characteristics and emphasize the stepwise nature of the analysis as it proceeds from the most significant effects downward.

A fourth cited advantage is that the analysis can be used in a conceptually non-symmetrical manner. The PAA method can be directed toward answering questions about the predictability of one attribute by other attributes rather than questions about the (symmetrical) association of the attributes.

Four disadvantages of the PAA method are generally noted. Perhaps most importantly from a practical point of view, the analyses have been very laborious to conduct with generally available statistical software (such as SPSS or BMDP). No dedicated procedures exist in these packages; the analyst is forced to program each level of the analysis separately, examining printed results for all tables and then designing the specifications for the next level of the analysis. This can be quite cumbersome -- at the fourth level of an analysis with 20 predictor variables, 600 contingency tables must have been produced and analyzed.

A second disadvantage is the complement to the 'advantage' of being able to use dichotomous variables; if a greater detail of information is available, the analysis cannot take advantage of it. Only binary information is processed.

A third disadvantage is the tendency at the level of interpretation to misrepresent the importance of non-primary predictors; that is, at the terminal subgroup level, one tends to equate all attributes of that group, despite the fact that some are more (statistically) pronounced discriminators.

A fourth disadvantage is that relatively large samples are required to conduct a PAA analysis to any depth. For example, at the fourth level of a PAA analysis, if all splits were 50%-50%, there would be an average of 156 cases per cell for an initial sample of 10,000. However, if the splits were 30%-70%, some cells would be expected to average 20 cases per cell; for 20%-80% splits, 4 cases. At the fifth level, even with a sample of 100,000, 20%-80% splits along a branch would yield cells with only 8 cases expected.

Theoretical and Empirical Issues

A number of questions have been raised about the use of PAA as a general-purpose prediction method and about its behavior under certain conditions. These issues concern the reproducibility of results and the validity of the associated interpretations. Reliability and validity are affected by both the structural relations within the data under analysis as well as the parameters which control the analytic processing itself. Previous literature has primarily addressed questions of reproducibility and not questions of validity; this report addresses both.

There are three questions related primarily to validity. First, does PAA have the ability to recover a known structure from a given system of data? Second, does PAA recover the best (most efficient and parsimonious) model that represents the data? Third, how do interrelationships among the predictors affect the analysis?

Two questions relate primarily to reliability. Do the results of a particular analysis replicate either (1) in terms of the predictor variables selected or (2) in terms of the individuals who comprise each of the terminal subgroups?

Two further questions relate to issues of both validity and reliability; they involve sensitivity of the analysis to (1) the statistical criteria used to select predictor variables at each step and (2) the stopping criteria

(statistical and otherwise) to terminate processing for particular branches. The choice of criteria affects both the results (and hence the interpretation) of a particular analysis and the susceptibility of that analysis to poor reproducibility.

Footnotes

[1]Support for this project was provided through the Bureau of Justice Statistics, Cooperative Agreement # 82-BJ-CX-K017.

II
METHODOLOGICAL AND STATISTICAL ISSUES

a sample of individuals is used to generate a statement applicable to other persons selected from an equivalent sample. The idea is to select a set of the most useful variables from all available information, yielding a set of measures that have, in some predefined sense, the best predictive power and then to use a method or logic to combine these measures in as efficient and coherent a manner as possible. The goal is to afford the best possible prediction capability for the problem at hand.

Implicit in this definition are some further assumptions which should be noted. First, the individual characteristics or attributes under study, both as independent and dependent measures, must be both adequately defined and properly measured. Second, the formal methodology for evaluating the available information must be a coherent and logical process that is appropriate to the task at hand. Third, where statistical assumptions are made, they must be shown to be appropriate given the properties of the data under analysis. Interpretations must be couched in an understanding of limitations imposed by a failure to meet these conditions.

Historically, the task of prediction has been accomplished through a variety of logical arithmetical and mathematical techniques. At one end of this continuum is a simple "unit-weighting" accumulation of points for the presence or absence of particular attributes; at the other end are the complex statistical methods which take into account not only the association of the predictive measures with the dependent measure, but also the relationships among the prediction measures themselves. (The family of multiple regression methods is a well-known example.)

Predictive Attribute Analysis is a multivariable, not a multivariate, method. We say this to emphasize the fact that a PAA examines a number of prediction measures at each point in the sequential process, but it does not

examine the interrelationships among these measures at either horizontal or vertical decision planes (in contrast, consider stepwise regression procedures). Many of the observed deficiencies of PAA can be traced to this characteristic.

Models and Parameters

Inherent in mathematical representations of relationships is the use of a MODEL with associated PARAMETERS. For example, $Y = a + bX$ is the familiar model specifying Y as a linear function of X; a and b are the parameters of y-axis intercept and slope, respectively. Alternative model specifications are $Y = a + bX_1 + cX_2$ or $Y = a + bX_1 + cX_2 + dX_1X_2$. Both of these models are also linear, but the second includes a term for interaction between X_1 and X_2. A model of the form $Y = aX^b$ is, of course, not linear (except where b = 1). See Winer (1971) for a thorough discussion.

It is important to remember, however, that the same sample data can be approximated by any number of different models (with consequently different parameters). Generally speaking, the more parameters in the model, the better the approximation to a particular data sample. However, an equation of order [N-1] can always be found to connect N points. The goal, therefore, is to be parsimonious and to select the model with the fewest parameters that adequately reproduces the relationships in question. In general, this approach will provide a more stable model, better resistant to random sample variations.

Model specification is a difficult task. Magidson (1982) discusses some behavioral characteristics of log-linear and attribute analysis methods including the consequences of such practices as (1) omitting influential variables, (2) omitting interaction effects, and (3) misspecification of the correct model. The point stressed is that "no analytic technique can compensate for lack of theory in deciding which variables to include in an analysis or how to interpret results."

Interaction Effects

Too often interaction among variables is treated as a phenomenon to be either avoided at all costs, or at best, apologized for. This approach to the analysis of complex data systems can lead to serious misinterpretations of the important forces that are active in the system under study, especially in cases where the analysis is exploratory rather than hypothesis-driven.

Interaction effects are to be expected. A common-sense approach to the study of factors affecting groups of individuals must lead one to expect that external influences do not impact equally on all individuals, nor do internal characteristics have equal influence in all settings.

One of the reasons for the propagation of dendographic procedures has been that they seemed to offer an analytic approach suited for the detection of interaction effects. Although this appears to be the case, there are a number of reasons that the results of a particular PAA, especially one presented in the absence of independent confirmatory support, should be interpreted with great caution. There are a number of different types of interaction effects, some of which a PAA will not detect (Magidson, 1982). In addition, as we will see, often one cannot distinguish between the detection of a main effect and an interaction effect by using PAA. Lewis (1962) presents a general review of various technical analytic approaches to the unravelling of interaction effects in multidimensional tables.

Quantitative data analysis procedures typically hypothesize a model and then proceed to estimate the parameters. A saturated model would consist of all main effects (of the form A, B, C, ...) and all interaction effects (of the form AB, AC, BC, ABC, ...) for each independant variable. The detection of interaction effects, however, is usually not emphasized, and the methods generally espoused for the detection of interactions (eg. Winer, 1971; Kirk, 1968) are not the most efficient available.

Qualitative data analytic procedures have tended to be more encouraging and supportive of the analysis of interaction effects. The AID and CHAID procedures mentioned previously, as well as the PAA procedure, were orginally advocated for their claimed interaction detection capabilities. Log-linear methods easily allow tests of interaction at any level of complexity, and in addition allow the omnibus test for the presence of any significant unspecified (residual) effects.

In practice, however, many researchers do not investigate interactions in a thorough manner, taking into account the different types of interaction and the alternative methods for their detection. When only main effects are postulated within a model, it is often due to an inability on the part of the researcher to explicitly create and define the proper representation of an interaction effect.[1] Even when interactions are hypothesized, they may be couched in assumptions that are inappropriate to the data system under analysis.

Usually models are assumed to be hierarchial in nature, where the higher-order effects for any model, in which a particular embedded effect is missing, are defined to be absent (eg., absence of an AB interaction implies absence of an ABC interaction). Models in nature are not necessarily hierarchical, however, and this assumption can lead to erroneous conclusions (see Magidson, 1982). A justification for the adoption of a hierarchical model comes from the assumption of multivariate normality: in the special case where the dependent variable and all the independent variables jointly fit a multivariate-normal distribution, a hierarchical model is appropriate (see Anderson, 1958). In practice, this assumption should be questioned -- particularly when the variables are categorical or binary in specification.

Model Detection by PAA

An important difference between PAA and techniques such as the log-linear methods is that PAA is often used as an exploratory procedure to discover a model for the data system under analysis, whereas log-linear methods,

must be supplied with a model (for which goodness-of-fit measures are obtained). From this perspective, some analysts have suggested using PAA as the exploratory method to suggest a preliminary model, then using log-linear methods to validate that hypothesized model. One should examine, however, how suitable PAA is for this task.

One procedure for investigating the ability of a Predictive Attribute Analysis to detect the actual effects imbedded in a defined model is to consider the possible combinations of effects for a two-factor (A and B) model; i.e., A, B, AB, A & AB, B & AB, A & B, A & B & AB effects. Where observations are limited to the set (Ø,1), the possible values of an observed criterion Y for all combinations of A and B (and hence AB) values for the model Y=B are given in Figure 2.1. Using the lambda statistic as a selection criterion, we find (by application of the PAA method) only an effect for B (as expected).

As we look at the results of this exercise for the other possible models, however, we find some unsettling results. As can be seen from Table 2.1, the model inferred from a PAA is not necessarily the model from which the observations were generated. Specifically, an interaction effect (AB) cannot be distinguished from two independent main effects (A & B) or those main effects plus an interaction to (A & B & AB). Also, an interaction with only one component having a main effect (A & AB) cannot be distinguished from the simple main effect (A).

These problems become compounded, of course, as the model expands to incorporate more than three variables. The net effect, then, is to cast doubt on the utility of using PAA as an independent model-generation technique. Because an attempt to validate a prespecified model would be subject to the same concerns, a PAA would also be inappropriate for hypothesis testing applications. Even as an exploratory technique, the results of any particular PAA might

FIGURE 2.1

Example of Model Recovery Procedure

Model: Y=bB

Given dichotomous data, possibilities are as follows:

A	B	AB	Y
Ø	Ø	Ø	Ø
Ø	1	Ø	1
1	Ø	Ø	Ø
1	1	1	1

Note: Y simply equals B, regardless of the values of A or AB.

Contingency Tables:

A \ Y	Ø	1
Ø	1	1
1	1	1

Lambda = Ø.Ø

B \ Y	Ø	1
Ø	2	0
1	0	2

Lambda = 1.Ø

Y Given B=Ø

A \ Y	Ø	1
Ø	1	0
1	1	0

Lambda = Ø.Ø

Y=Given B=1

A \ Y	Ø	1
Ø	0	1
2	0	1

Lambda = Ø.Ø

PAA Tree:

TABLE 2.1

Models and Their PAA-Recovered Counterparts.

MODEL	PAA TREE	PAA MODEL(s)
Y=aA (A main-effect)	A - -	Y=aA
Y=bB (B main-effect)	B - -	Y=bB
Y=xAB (AB interaction)	A - B	Y=aA+bB or Y=xAB
Y=aA+xAB (A main-effect & AB interaction)	A - -	Y=aA
Y=bB+xAB (B main-effect & AB interaction)	B - -	Y=bB
Y=aA+bB (A & B main-effects)	A - B	Y=aA+bB or Y=xAB
Y=aA+bB+xAB (A & B main-effects & AB interaction)	A - B	Y=aA+bB or Y=xAB

produce dozens of potential models for evaluation by subsequent confirmatory analysis.

Subgroup Classification

An alternative use of the PAA procedures, beyond the discovery of important predictor variables, is the definition of subgroups within a population. The premise is that, for each subgroup, the variables used to define the groups are, in some statistical sense, the most important.

Each subgroup is defined through a series of sequentially-dependent selection decisions. Although this process can systematically exclude certain model effects, the final predictor composite may nonetheless contain the same variables that would have been selected by a more valid model specification approach. (If one considers only the final subgroups and their defining attributes, PAA provides a system for classification that is "model independent" in the sense that different models could require the inclusion of the same variables.) One would then proceed to validate these groupings by alternative analyses or by replication.

However, to the extent that any subgroups are selected by statistical criteria, they are subject to sampling variability, raising the issue of reliability in the findings. While order of selection may not be important for an analysis that always proceeds to the exhaustion of all independent variables, it does become important once stopping criteria are introduced and the analysis does not proceed to consider subsequent predictors. Thus, the interpretation of PAA results (in terms of any "defining" characteristics of subgroups) rapidly becomes a very complex issue. Furthermore, many of the parameters for assessing reliability are data-specific (e.g., interrelationships among predictors), requiring unique validation studies to be able to make statements about the validity of particular results. The issue of subgroup classification will be investigated empirically in the next chapter.

Validation Methods

Validation, in the sense of the generalizability of results from one sample to another, is a central issue in the context of prediction methods. Most prediction techniques strive to minimize the unpredictable variance within a particular sample; while this tends to optimize the predictability for that sample, it is often to the detriment of the overall generalizability of the results. One needs to be aware of the potential limitations of the results on any one particular analysis when application is made to other samples. This section describes two means for assessing this type of validity - traditional cross-validation methods and bootstrap resampling methods. Because many of the uncertainties surrounding the interpretation of a specific Predictive Attribute Analysis are difficult to resolve by a strictly theoretical approach, we will suggest an empirical approach for assessing the relative confidence we wish to place in the results of a particular analysis.

A general discussion of validation issues in measurement can be found in Cronbach (1971); more technical presentations are given by Stone (1974) and Efron (1983). The focus here is on the ability to make accurate generalizations from the analysis of one sample to a larger population, hence the general approach is to compare the results of a number of analyses of (simulated) samplings from a given population.

It should be mentioned that 'robust' and 'resistant' analytical procedures are sometimes recommended for continuous-variable prediction problems. Although a number of methods for pre-and post-processing of data have been developed (Huber, 1981; Mosteller & Tukey, 1977), these methods can often contribute their own shortcomings and, in any case, are often not well suited for categorical or dichotomous data systems. Some non-parametric methods do offer a reduced sensitivity to these problems, but at the risk of not finding effects that are indeed present in the population system.

The Cross-Validation Method

The empirical assessment of validity has traditionally been done using the family of cross-validation methods. The simplest form of this procedure uses two samples (or one subdivided sample) from the data system of interest: the prediction instrument (or table or equation) is constructed using the first sample and then validated by using that instrument to predict the dependent measure in the second sample. A measure of predictive accuracy is then obtained by comparing predicted and observed values of this dependent measure in the second sample. A common extension of this cross-validation approach uses N-1 cases in the sample to develop a prediction model for the remaining case, then averages the prediction errors for the N replications of this procedure to obtain a measure of prediction error.

The Bootstrap Method for Resampling

The bootstrap procedure is a particularly elegant and easily applied member of the family of cross-validation methods. Efron (1979) developed the bootstrap as an alternative to and extension of existing cross-validation procedures. The method requires minimal effort toward model specification, distributional assumptions, and analytic effort, and as a sampling procedure it is applicable in an automated form to variety of situations over a broad range of complexity (see especially Efron & Gong, 1983; also Efron, 1982). The efficacy of the bootstrap is demonstrated by the clarity it lends to the interpretation of PAA results.

The procedural definition of the bootstrap sampling method is straightforward: for some arbitrarily large number of replications[2], a random sample of size N is drawn with replacement from the original data set of size N. The analysis of interest is then conducted once for each of these samples; if desired, nonparametric estimates of statistical attributes of the data can be calculated.

The rationale behind bootstrap sampling is as follows: if one believes that every type of multivariate data point (or "case") present in the "population" is represented in the sample (even if not proportionally to its frequency in the population), then one can generate approximations to random samples from this unknown population by using repeated sampling-with-replacement procedures. One thereby expects that at least some of the bootstrap samples will be distributed very similarly to the unknown population. If desired, one can then use nonparametric reasoning to generate appropriate statistical measures and associated confidence intervals.

For the purpose of investigating PAA validity and generalizability issues, we have used the bootstrap method to generate repeated samplings from selected data sets. A PAA was then performed on each of those samples and the results compared. Section III examines two aspects of the results in particular: (1) how consistently the pattern of selection for predictor variables was replicated, and (2) how consistently the membership of cases in the terminal subgroups defined by the analysis was replicated.

Analysis of 2x2 Contingency Tables

This section discusses some characteristics of two-way contingency tables that are dichotomous in both variables. Presented first are some considerations involved in the use of binary-coded information, whether that is the level at which the information is observed or whether there is an intermediary recoding of ordinal-or interval-level information. Second, statistical measures of association and prediction that are appropriate for 2x2 tables are considered. Finally, this section examines how the use of these different statistical measures can affect the results of a Predictive Attribute Analysis.

Dichotomous Variables

The analysis of criminal justice data often requires the consideration of dichotomous information. Frequently data are available only at the level of binary coding (incarcerated/not-incarcerated, prison/jail, felony/misdemeanor),

even where there might be evidence for an underlying ordered succession of categories. Quite often the dependent measure in a predictive analysis is binary, representing the final decision (however uncertain or qualified) resulting from a complex decision process.

Typically, measures available to an analysis are not all binary, however. In general, it is preferable to use an analytic method which can take advantage of the greatest detail present in the data. Logistic regression techniques, for example, allow the use of dichotomous, ordinal, or interval measures in the prediction of a dichotomous criterion, where PAA requires that all variables be dichotomous.

It is preferable to choose a method of analysis for which the assumptions are most appropriate to the characteristics of the data at hand rather than forcing the data to fit the model (by dichotomization, for example). This approach provides greater power for the detection of real effects within the data and also avoids the problems of deciding at what point to create artificial categorizations from a more detailed data representation. Bishop, Feinberg, and Holland (1975, p. 371) note that "... different choices of boundaries (based on collapsing of categories) can lead to different conclusions regarding the dependence or independence of variables. Little guidance is available to help the investigator make such choices."

In the case where PAA is determined to be an appropriate analysis but where a number of important predictors require dichotomization, our recommendation is to select several alternative cutpoints for each predictor and examine the results of the analyses. An alternative would be to create a set of binary 'dummy' variables for each category of an original variable, where (Ø or 1) would represent the (absence or presence) of that category for that variable. These devices should be used only for variables for which there is no clear theoretical or empirical guidance for imposing cutpoints -- theory should take precedence over exploration.

Statistical Measures

A useful categorization for the various measures of association for 2x2 tables is presented by Bishop, Feinberg, and Holland (1975). There are four general classes: 1) measures based on the ordinary chi-square, 2) measures based on the cross-product ratio, 3) proportional-reduction-of-error measures, and 4) proportion-of-explained-variance measures.

For the purposes of this report, four statistics are considered which represent three of the cateogries described above. Cross-product (odds) ratio measures are not considered since (1) they have not been a part of the PAA applications literature or much of the modern contingency table literature, and (2) these measures possess some properties that make them appear less-than-optimally interpretable for PAA-type applications.[3] This report does examine the Chi-Square coefficient, the Phi-coefficient, Goodman & Kruskal's Lambda coefficient, and the Uncertainty coefficient.

Each of these statistical measures is discussed with respect to several characteristics of the measure that affect their use in a Predictive Attribute Analysis, such as whether or not they are independent of sample size, whether they assume a dependent/independent distinction, or whether they measure symmetric 'association' or asymmetric 'prediction.'

The Chi-Square coefficient can be used both as a measure of association and as a test for independence; the latter will not be considered here since it is a hypothesis test rather than an assessment of relationship. Table 2.2 presents a definitional formula for Chi-Square (and the other statistics discussed here). On the basis of the applications literature, Chi-Square seems to be the statistical criterion of choice for PAA and the other interaction-detection procedures (Phi being an "alternative"). However, lack of discussion in these references as to the justification for using Chi-Square as opposed to alternatives suggests that the choice may not have been based on careful consideration of the properties of the available statistics.

As a measure of association, Chi-Square is symmetric in the sense that no distinction is made for a 'dependent' versus 'independent' variable. It is dependent on the sample size and has no upper bound; the lower bound is zero. Probability statements can be made easily by reference to tabled Chi-Square values for a chosen significance level, where the number of degrees of freedom for a 2x2 table is always one. An advantage that the Chi-Square statistic has is the general familiarity it enjoys from a wide audience.

Although the computational mechanics of the Chi-Square test are easily carried out, their derivation depends on a mathematical rationale that requires important assumptions (1) that observations are independent of each other, and (2) that each observation represents a single joint-event (or cell location in the contingency table). Probabilities obtained from a chi-square table are estimates and approximate (barring an infinite sample size). The accuracy of this approximation depends not only on overall sample size, but on such factors as the significance level employed, the total number of cells in the contingency table, and the true marginal distributions in the population. Based on these considerations, a conservative guideline that has been generally endorsed is to require a minimum cell frequency of 5 for tables with more than 1 degree of freedom and 1Ø for tables with a single degree of freedom (as have 2x2 tables).

The Phi coefficient for a 2x2 contingency table is equivalent to the Pearson product-moment correlation coefficient. Phi is also directly obtainable from the chi-square coefficient as the square-root of (Chi-Square divided by the sample size). The Phi coefficient has the resultant advantages of 1) a standardized range of zero to one and (2) an interpretation directly analogous to that of an ordinary correlation coefficient.

Goodman and Kruskal's (1954) Lambda coefficient is a measure of predictive association (asymmetric) rather than of simple association (symmetric). Specifically, lambda measures the proportional reduction in error of predicting the one variable by having knowledge of the other variable. This approach is

fundamentally different from that of the measures of association such as Chi-Square or Phi.

It is possible (and frequently occurs with criminal justice measures) for association to exist where predictability does not: Chi-Square is greater than zero while Lambda equals zero. This does not imply that Lambda is unsatisfactory, but rather that Lambda is a measure of a different property of the data. In practice, PAA predictor variables selected by Lambda are typically different from those selected by the other statistics.

Lambda ranges from zero to one, being zero only if knowledge of one variable is of no help in predicting the other and being one only if knowledge of an individual's category on one variable determines the category of the other variable. Lambda requires information in the diagonals of the 2x2 table to be parallel (i.e., both maximums on the same diagonal), otherwise Lambda is zero. As a measure developed for nominal-level data, Lambda considers only the mode of the distribution. It should be noted that it is possible, even where there is perfect predictability of a criterion variable, for Lambda to be zero if the predictor variable contributes no new information (beyond that available from the marginal distributions).

The Uncertainty coefficient (Theil, 1967) is derived from an information-theoretic approach. Like Lambda, it is asymmetric; it indicates the proportion by which uncertainty on the dependent variable is reduced by knowledge of the independent variable. Unlike Lambda, it considers the entire distribution of observations, not just the modal category. The Uncertainty coefficient ranges from zero to one and is independent of the number of observations.

Use of the Various Statistics for Predictive Inference

The information available in the format of a contingency table can be thought of in a number of research contexts. Hypotheses may relate to questions of association of variables, to goodness-of-fit of one variable's distribution

to a proposed model, or to questions of predictability of one measure from another.

The a priori design model brought to an analysis like PAA can substantially affect the results and interpretations. Different statistical measures have been developed to meet different needs of researchers, and the assumptions of each must be considered carefully. The use of a prediction versus an association statistical measure in a PAA implies a different set of hypotheses is to be tested, and hence the interpretation of results should not be the same.

The Chi-Square and Phi coefficients are inherently tests of association, symmetrical in nature (like a correlation coefficient). The Lambda and Uncertainty coefficients are, on the other hand, asymmetric measures of the predictability of one measure by the other (like a regression coefficient). Chi-Square values for AxB and BxA tables are necessarily the same; lambda values for AxB and BxA tables are not necessarily the same.

The difference between an associative measure and a predictive measure can be clarified by reference to Chi-Square and Lambda as specific examples. Chi-Square is a technique that evaluates differences between observed and expected observations (expected either via a hypothisized distribution or from tne marginal distribution which was observed). This is a symmetrical conceptualization; the distributions either match or they do not. Lambda, however, is an index develcped to measure the proportional reduction in the probability of error in predicting B by having knowledge of A. If the information contained in A does not reduce the probability of error in predicting B at all, Lambda equals zero; if the information completely determines the prediction of B from A, then the index is one.

With reflection it can be seen that the idea of prediction is not equivalent to the idea of association. It is possible, as noted above, that statistical association exists even though the predictability measure equals zero.

It may be the case that A and B are not independent, but that the relationship is such that knowledge of A does not change one's expectation about B.

Which of these types of statistics is preferable for Predictive Attribute Analysis? In discussing the choice of measures for contingency tables, Bishop, Feinberg, and Holland (1975, p. 393) comment: "No single measure is better than all others in every circumstance. Different measures have different purposes, and our selection must depend on our objective in studying a set of data. If the focus is on departures from bivariate independence, [chi-square based measures] are useful, while [predictive association measures] may mislead. If the focus is on prediction, the reverse is true, and we may profitably choose [predictive association measures]."

Logical consideration of the properties of each of these statistics would suggest the use of one of the predictive association measures (Lambda or Uncertainty) rather than a measure of independence (Chi-Square or Phi). Choosing between Lambda and Uncertainty would then depend on whether the variables are considered to be strictly nominal or have ordinal/interval characteristics. This decision becomes problematic where measures of both types are present in an analysis.

In practice, Chi-Square, Phi, and the Uncertainty coefficient perform similarly in selecting predictor variables at particular PAA nodes. Lambda very frequently selects a different sequence of predictors. None of the measures provides a significantly greater level of reliability across samples when the depth of the analysis is held constant. These comments will be discussed more fully in the sections of this report that discuss the simulation studies and suggest general recommendations. At this point, however, we can suggest no all-inclusive guideline for a "best" PAA statistic.

Alternatives to PAA

The kinds of research questions which are addressed by a Predictive

Attribute Analysis can also be addressed by a number of other analytic techniques. If one visualizes a continuum ranging in complexity from singular contingency table analysis, to ad hoc contingency table analyses, to logit and probit analysis, to log-linear models, and through general-model categorical analysis techniques, PAA would fall into the domain of the ad hoc methods. While there are theoretical reasons for prefering the methods at the multivariate end of the continuum, there are often practical considerations which impede the successful application of these methods. PAA is often chosen as a compromise method.

Some general references that provide a background for issues affecting contingency table analysis are Bishop et al (1975): Discrete Multivariate Analysis: Theory and Practice, Fleiss (1981): Statistical Methods for Rates and Proportions, Davis (1974): "Hierarchial Models for Significance Tests in Multivariate Contingency Tables: An Exegesis of Goodman's Recent Papers," and Goodman and Kruskal's (1979) Measures of Association for Cross Classifications. Brieman et al. (1984) have developed some interesting extensions in the area of regression-tree procedures.

In reviewing some analytical methods for the type of predicition problem where the dependant variable (at least) is binary, it is interesting to consider the historical evolution of those methods. The problem was initially: given that Ordinary Least Squares (OLS) methods are not appropriate to the case when the dependent measure is not of a continuous nature, what modifications of OLS might prove servicable? This perspective led first to studies attempting to delineate more clearly the conditions under which a conventional regression analysis would provide a reasonable appropriation. The 'conventional wisdom' emerging from this work held that OLS could be used in situations where the dependent measure was expected to fall midway between the extreme probabilities of zero and one.

Modification of the OLS algebra produced the probit and the logit methods (see Finney, 1952), both explicitly designed for predicition of a dichotomous

dependent variable. The emphasis here is on the relative weights for the set of predictor variables. The subsequently developed log-linear methods, due primarily to Goodman and collegues (eg, 1978), emphasized the model development aspect of analysis; the analysis was geared to uncover the combination of main effects and interaction effects that could be properly used to represent the underlying forces acting on the observations.

Both the logit and probit methods assume a dichotomous dependent variable with categorical and/or continuous independent variables. The techniques are conceptually similar approaches to the problem of prediction, the major difference being that the probit model is based on the normal function while the logit model is based on the logistic function. (These distributions are practically identical for our purposes.) From a pragmatic point of view, the techniques are "no more than a convenient mathematical device for solving certain equations" (Finney, 1971), and should be treated much as other general transformational procedures. Other transformations have indeed been proposed, but they seem essentially indistinguishable from logits and probits in their performance over a wide variety of applications (Finney, 1971).

Log-linear methods emphasize the evaluation of alternative model specifications. The method for specifying these effects has an appealing parallelism to the conventional analysis of variance methods known to most researchers. Given a hypothesis that any effect (or group of effects) is a component of the model, estimates can be derived for the expected cell frequencies in the contingency table system. Given these expected frequencies, a likelihood ratio test can be performed to assess the goodness-of-fit of the data to the specified model.

In principle, the procedure can be extended to tables of any dimension; any effect or combination of effects can be considered. The statistical assumptions are those of a standard Chi-Square test-- observations are independent and arise from multinomial sampling of some population. A practical restriction is that large samples are required in order to have a reasonable chance of expected

frequencies greater than zero in all cells. Because the number of potential hypotheses increases rapidly (due to the number of potential interactions), the procedure is somewhat inappropriate as an exploratory tool for complex tables. Log-linear models, used for the level of complexity of contingency tables typically analyzed by a PAA, would have to be stated very explicitly beforehand because of the depth of potential interaction effects. Log-linear models in general are inappropriate for exploring questions for which little previous analytical work had been done.

In addition to these widely-known and generally available analytic procedures, there are also other alternative approaches to the analysis of data systems of the type considered here. Examples are clustering procedures, profile-comparison procedures, multidimensional scaling methods, discrimination analysis, and general structural-model categorical methods (Pruzek & Lehrer, 1980). These methods are not addressed here, but they (and others) are certainly not to be dismissed from consideration by the analyst.

The literature gives numerous examples of how certain methods are not well suited to particular data systems. In the case where nested interactions are expected to exist (that is, different interaction effects for different subgroups), PAA-like procedures are generally more likely to detect effects than log-linear methods. Because PAA is a conditionally-oriented method, it can uncover effects that would otherwise be averaged out at other levels of specification. Another type of interaction is the symmetric interaction, where the effect is approximately equal and opposite for two subgroups. The near-zero average makes the effect difficult to detect. These effects are generally not detected by a PAA procedure because they fail to pass a main-effects statistical test. Magidson (1982) shows how a log-linear analysis can detect these effects, but only if the model is specified as a nested interaction model and not a hierarchical model. One is thereby cautioned against limiting the use of log-linear procedures to hierarchical models if theory suggests otherwise.

With respect to the selection of particular analytic methods our recommended strategy is, succinctly, a multi-method approach oriented toward developing a consensus in conclusions.

Footnotes

[1]One solution to this problem of the effective specification of complex effects is presented in Pruzek & Walker (1982).

[2]Typically 100-200.

[3]See Bishop et. al. (1975), p. 383 especially.

III
EMPIRICAL STUDIES OF THE PAA METHOD

EMPIRICAL STUDIES OF THE PAA METHOD

The two major components of this part of the project were (1) the development and implementation of a mainframe computer program, required for (2) the design and conduct of a series of simulation studies. The computer program was developed with attention to the known characteristics of PAA as discussed in the previous chapters with the intent of also providing means for the control of parameters of the analysis (type of statistic used, branch termination criteria, etc.) that could alter the validity/reliability characteristics of the method. The group of simulation analyses which are reported here was aimed at further defining strengths and weakness of PAA and providing guidelines for the use and interpretation of the method.

Introduction

The need for further study of the empirical performance characteristics of PAA is evident from the set of unresolved research questions that remains after the theoretical considerations presented in the previous chapter. Although some of the inherent characteristics of the method were described, there remain other aspects that call for investigation using real data systems. These aspects are especially difficult to study because of their dependence on particular interrelationships found within given data systems.

An illustration may serve to clarify the kinds of questions that remain. It is obvious, for example, that the first split in a PAA tree will indicate the source of the strongest main effect present in the data. This effect should be found over sampling with a consistency that relates to the relative magnitudes of the correlations between the dependant variable and the set of independent variables. (If C is predicted by A and B, given a correlation r_{cb}=.3Ø and r_{ca}=.1Ø, we would expect B to be consistently selected. If r_{ca}=.3Ø and r_{ca}=.28, we would expect random sampling variation to impact

significantly on the statistical selection process, with B having a slight edge over a large number of samplings.) What is not known a priori, however, is how to derive expectations for particular variables at successive levels of the PAA tree, nor how the terminal subgroups defined by the complete PAA analysis might be affected by predictor interrelationships. To questions of this nature we directed the efforts of the simulation studies.

PAAVE: Predictive Attribute Analysis with Validation Extensions

Presented below is a brief description of the PAAVE[1] computer program, its specifications, capabilities, and limitations, and the principles that guided its development. Potential users of the program will want to refer to the more detailed User's Guide which accompanies this technical report.

Guidelines for Program Development

The mainframe FORTRAN computer program for Predictive Attribute Analysis was developed to (1) carry out the computational efforts required to perform a PAA analysis (with support for data input and dichotomization), (2) allow flexible user specification of parameters controlling the PAA processing, and (3) provide several kinds of feedback to the user regarding the confidence appropriate to the results of the analysis. It was decided early in the project not to attempt to incorporate methodological extensions into the program, such as allowing categorical rather than dichotomous data or attempting look-ahead computations in the manner available in AID/CHAID programs. There were two reasons for this decision: first, computer routines already exist to perform these manipulations, and second, the fundamental issues for this kind of categorical analysis procedure can be adequately addressed by a study of the generic PAA processing logic. The program therefore adheres to the definitions originally put forth by MacNaughton-Smith (1965).

It was known at the outset of the project that the PAA computer program to be developed would need to provide a variety of different kinds of feedback to

the analyst. Enhancements to the otherwise straightforward PAA processing computations thus took the form of 1) availability of information about the competing predictor variables in addition to the selected predictor variable, and 2) addition of alternative processing controls, such as the ability to force the selection of particular variables at specified nodes or to select various combinations of stopping criteria.

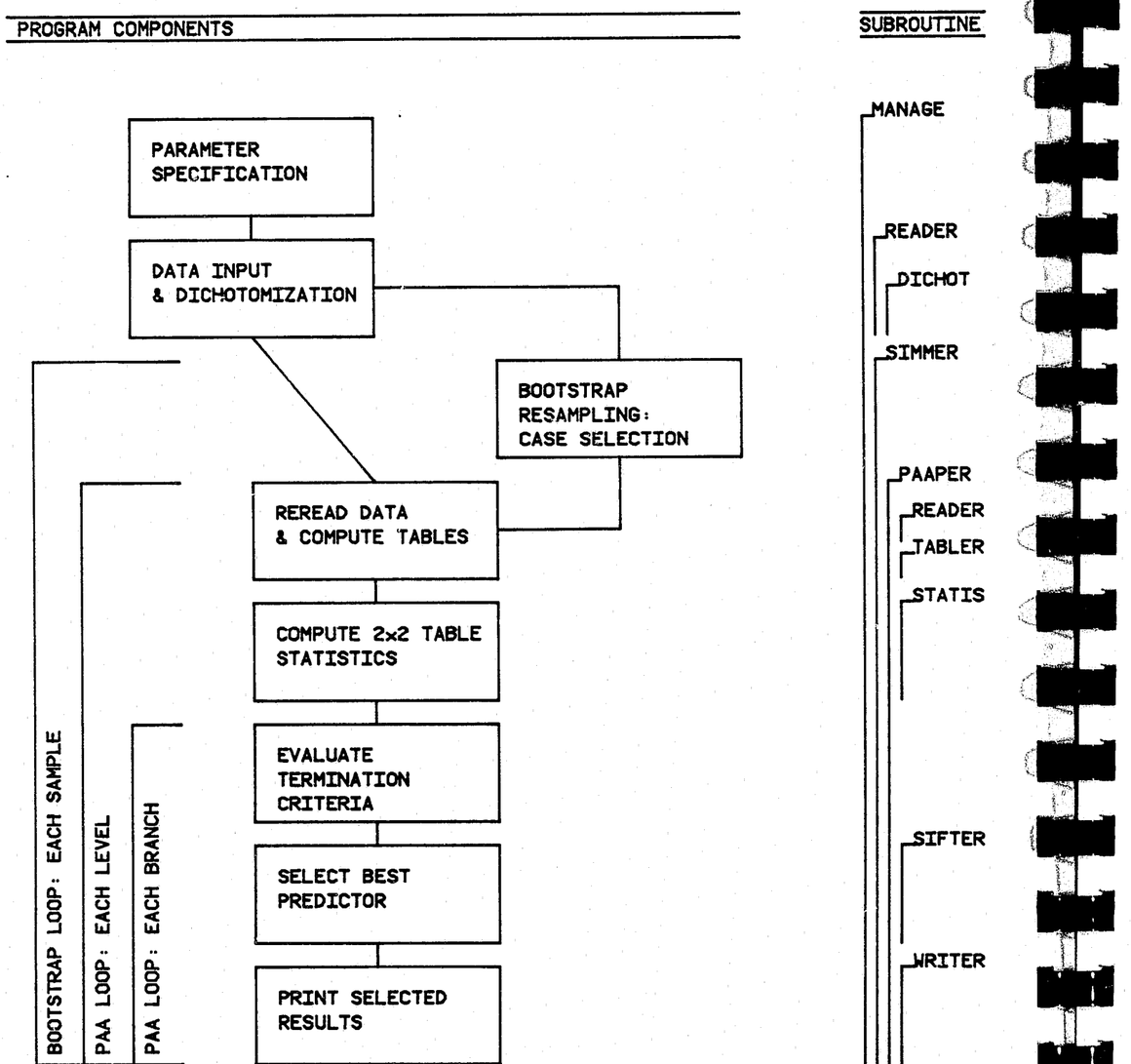
The program was written in a modular format, giving control of specific processing functions to specific FORTRAN subroutines. This programming approach was used both to facilitate development and testing and to allow for future program modification and extension. Users of the program who wish to supplement the existing features with additional processing options should find that the existing routines can be modified without undue effort.

A diagram of the modular components of the program is presented in Figure 3.1. Raw data input, dichotomization, contingency table construction, statistical calculations, and program output are each handled by separate routines. The PAA algorithm itself is the responsibility of a managing routine, and general program execution and parameter specification are controlled by the main routine. The program also has a multilevel trace mode whereby users can follow program execution by a series of markers which note the physical location of program execution[4] and provide intermediate processing output of selected detail.

The program was designed to allow considerable flexibility in the specification of PAA processing options. To control branch termination, several statistical tests are available, as well as absolute and relative cell and marginal frequency criteria. In addition, to allow testing of specific models, the selection of any predictor can be forced at any point in the analysis. Processing can also be limited to a maximum depth (level) for all branches, or can be allowed to proceed regardless of stopping criteria tests (assuming subgroup marginals are not zero).

FIGURE 3.1

ABBREVIATED PAAVE FLOWCHART



Output is available at several levels of detail, depending on the needs of the analyst; at the most voluminous level, all contingency tables and all associated statistics can be produced. The standard mode provides one page of output for each node in the PAA processing tree[2] (see the Appendix for an example). For simulation runs, summary-level output is available, or selected output can be written to online files and post-processed by other programs.

Program Specification and Features

The computer program has been written, with as few exceptions as possible[3], in standard FORTRAN-77 to facilitate conversion to other mainframe systems. It was developed on the Burroughs 6900 using Burroughs FORTRAN-77 running under the CANDE timesharing system.[4] Due to the batch-queue-oriented nature of that system, it was not possible to incorporate a high level of user interactivity into the program.

The current PAAVE program[5] can process 50 variables and an unlimited number of cases in the normal processing mode, 10,000 cases in the bootstrap validation mode. Up to 64 terminal subgroups can be derived through the program (five levels of PAA processing). An abbreviated flowchart is presented in Figure 3.1 to represent the general processing structure for standard and simulation modes.

Four statistics appropriate to 2x2 contingency tables are available within the program: Chi-square, Phi, Uncertainty, and Lambda. Justification for the selection of these statistics and guidelines for their use has been provided previously. For each of these statistics, a branch-termination minimum value can be input such that any observed value below that limit will cause processing in that branch to terminate.

Five additional branch termination criteria are provided besides the statistical criterion: minimum cell size, minimum cell percentage, minimum subgroup size, minimum subgroup percentage, and minimum subgroup ratio. These

may be used in any combination, individually or collectively. Default values are provided in the program, or values can be entered as program parameters. If desired, all branch termination checks (except of course null tables) can be ignored.

An additional program option allows the user to force the selection of a particular variable at any point in the analysis. This option facilitates the testing of hypothesized tree structures as well as the effect of using an alternative predictor in the case where the (statistical) competition is close among predictors at a particular mode.

Numerous modifications to the originally conceived computer program were a byproduct of the series of analyses done in the evaluation component of the study. While we do not, in this report, attempt to exercise all possible variations of these program options, we provide them as an impetus for others to study further the characteristics of the PAA method.

Simulation Analyses

The simulation aspect of the project was undertaken to develop a better understanding of how the PAA algorithm functions in practice. By looking at a series of data analyses, some of which incorporate extensive cross-validation efforts, we can address a series of validity, reliability, and conjoint validity/reliability research questions.

There are two types of data systems for which analyses are presented here. The first uses artificially-generated data; by controlling some of the otherwise variable facets of an analysis, we are better able to investigate the effects of other facets on the results. The second uses data typical of criminal justice databases; by observing the results of a PAA with data for which there is an existing knowledge base, some statements can be made about the proper use and interpretation of results.

There is a concern in these simulations with two general aspects of the results: 1) the magnitude of results relative to an expected outcome, and 2) the potential variability of the results around this expectation. In other words, we are interested not only in factors affecting the results of a single PAA, but also in factors affecting the differences among a sampling of results from a data population.

Research Questions Guiding the Simulation Studies

The previously discussed general research questions which have provided a focus for the project include: 1) PAA's ability to recover a known structure from a data system, 2) PAA's ability to determine a best (most efficient and parsimonious) model which represents the data, 3) the effect on a PAA of interrelationships among the predictor variables, 4) the replicability of PAA results with respect to sequences of predictor variables, 5) the replicability of PAA results with respect to the composition of terminal subgroups, 6) the effect of different statistical criteria on stability of results, and 7) the effect of different branch-termination criteria on stability of results.

The precision with which these research questions can be addressed is enhanced by the use of bootstrap validation procedures. Bootstrapping allows one, through repeated sampling, to derive from a sample of data an approximation to the sampling distribution of the population system from which the sample originated. Therefore, even in situations where there is only a single working sample, one can make extrapolations to the population from which the particular sample was drawn (as if the actual population were known, and one sampled directly from that population).

We have made extensive use of the bootstrap procedures in our simulation work to provide a basis for conclusions about the empirical validity and reliability of a PAA. Because of consistant findings regarding the presence of instability in the analysis of data, we decided to incorporate the means for computer program users to automatically conduct a validation analysis on their

own particular data using the bootstrap procedure. This capability provides analysts with a means to directly assess the level of confidence appropriate to the results of a particular analysis.

The series of studies reported here examined aspects of PAA processing ranging from the statistical properties of individual 2x2 contingency tables having predefined characteristics through the replicability characteristics of complete analyses using OBTS criminal justice database information. Table 3.1 presents the three general categories into which the analyses have been grouped: data systems with 1) defined interrelationships among predictors and criterion, 2) low-strength interrelationships among prediction measures, and 3) complex (naturally occurring) interrelationships among predictors. The data systems with defined relationships (generated from selected models with known characteristics) were used to examine the table-wise selection of predictors by the different statistical measures. The low-order relationship system was used to provide a better understanding of how the PAA method functions in the (worst-case) absence of reliable information. The data systems with naturally-occurring complex relationships, drawn from the New York State Offender-Based Transaction System database, were used to study the behavior of the method as a whole for both the selection of predictor variable sequences and the definition of terminal subgroups.

The discussions which follow for each set of simulations are parallel in format and include remarks on the 1) objectives, 2) procedures, 3) observations, and 4) conclusions from each analysis.

2x2 Nodes

Objectives. This set of simulations, initiated during the early stages of the project, was designed to provide a 'feel' for the characteristics of the four statistical measures that were to be available in the computer program. Levels of association, predictability, and cell frequencies were varied for a series of 2x2 tables. Particular attention was given to the magnitude of the

TABLE 3.1
Summary of Simulation Analyses

TYPE OF DATA SYSTEM

Focus of Analysis	Defined Interrelationships			Low-Order Interrelations	High-Order Interrelations	
	2 x 2 Tables	Causal Factors	Symmetric Interrelation	Random Data	PROB 8Ø NYC	PROB 8Ø Suburb
NODES	*	*	*	*		
TREES		*	*	*	*	*
SUBGROUPS					*	*

instability of the statistics arising from random sampling error in tables having small cell sizes.

Procedure. The series of 2x2 tables was constructed by selecting some generic tables of interest and varying the ratios of row and column marginal frequencies as well as the absolute frequency counts of cell entries. Both artificial and published data were used. The tables were considered to be asymmetric (in the sense that one dimension of the table was regarded as a criterion to be predicted by the other dimension). Chi-square, Phi, Uncertainty, and Lambda coefficients were calculated for each table.

Observations. The relative magnitudes of the statistical coefficients were in general agreement with expectation. All of the statistical measures were unstable when small cell counts (especially <5) were varied -- more so where the marginal ratios were high and some cells had very low relative frequencies. Figure 3.2 presents examples. The movement of a small number of observations could change the value of the statistic (and hence the PAA-selected predictor variable) markedly.

The association measures (Chi-square, Phi) and the prediction measures (Lambda, Uncertainty) were not equally affected by changes in the 2x2 tables. It was instructive to observe the divergence of Lambda from the other coefficients for certain types of tables. An initial impression was that Lambda was a rather coarse measure when compared with the other coefficients. The Uncertainty coefficient, while asymmetric in nature, behaved empirically much like the Chi-Square measure.

Conclusions. An important by-product of looking at these simplified contingency tables and their associated statistical measures was an appreciation for the potential instability of the statistically-driven PAA predictor selection process for tables with small or unbalanced cell counts. This instability has implications for the specification of such PAA parameters as branch termination values, but even a conservative specification of these

FIGURE 3.2
Analysis of Some Selected 2x2 Tables

-a-

Some Abstract Ratios

	1 5 / 5 Ø	1 5 / 5 1	1 5 / 5 2	1 5 / 5 3	1 5 / 5 4	1 5 / 5 5
Chi-Square	7.64	5.33	3.90	2.94	2.27	1.78
Phi	.833	.667	.548	.458	.389	.333
Lambda	.800	.667	.500	.333	.167	0.
Uncertainty	.643	.350	.232	.164	.120	.090

	2 5 / 5 Ø	2 5 / 5 1	2 5 / 5 2	2 5 / 5 3	2 5 / 5 4	2 5 / 5 5
Chi-Square	6.12	3.90	2.57	1.73	1.17	.78
Phi	.714	.548	.429	.339	.270	.214
Lambda	.600	.500	.429	.286	.143	0.
Uncertainty	.486	.232	.137	.085	.054	.035

	Info. from neither category of X 25 25 / 25 25	Info. from one category of X 20 20 / 20 40	Info. from both categories of X 15 35 / 35 15	Info. from one category of X 20 20 / 20 200
Chi-Square	Ø.	2.78	16.ØØ	43.51
Phi	Ø.	.167	.4ØØ	.4Ø9
Lambda	Ø.	Ø.	.4ØØ	Ø.
Uncertainty	Ø.	.Ø21	.119	.151

-b-

Some Empirical Ratios

	Predicted Classification (Simon, p. 48) (six vars)	(seven vars)	Predicted Classification (Simon, p. 47) Clinical judgement	Behavioral Rating
	39 | 9 / 9 | 10	39 | 9 / 7 | 12	40 | 8 / 8 | 11	35 | 13 / 8 | 11
Chi-Square	7.69	12.47	11.39	5.62
Phi	.339	.431	.412	.290
Lambda	.053	.158	.158	Ø.
Uncertainty	.091	.150	.135	.069

	Marriage Status (Simon p. 188) Construct	Valid	Age (Simon p. 188) Construct	Valid
	114 | 147 / 6 | 3	109 | 151 / 2 | 7	59 | 59 / 61 | 91	49 | 61 / 62 | 97
Chi-Square	1.86	1.39	2.62	.83
Phi	.083	.072	.099	.055
Lambda	Ø.	Ø.	Ø.	Ø.
Uncertainty	.024	.019	.007	.002

-b-
(continued)

	Last Penalty (Simon)		Previous Files (Simon)	
	67 | 53 / 115 | 35	59 | 52 / 113 | 45	75 | 45 / 112 | 38	65 | 46 / 110 | 49
Chi-Square	13.17	9.54	4.64	3.24
Phi	.221	.188	.131	.109
Lambda	.150	.063	.058	Ø.
Uncertainty	.036	.026	.012	.009

	# Previous Convictions (Simon)		# Jobs (Simon)	
	37 | 83 / 61 | 89	29 | 78 / 82 | 76	23 | 97 / 54 | 96	30 | 81 / 44 | 114
Chi-Square	2.79	16.11	9.27	.02
Phi	.102	.247	.185	.009
Lambda	Ø.	.019	.008	Ø.
Uncertainty	.008	.046	.026	.00006

criteria would not necessarily protect against the mistaken identification of significant effects. Also of interest was the differing best-predictor decisions resulting from the use of an associative versus a predictive measure.

Defined Models

Objectives. One device for investigating the performance of a particular analytic method is using it to process a data system with known characteristics. The ability (or inability) of the method to accurately and reliably recover these known characteristics can provide useful insights into the general performance capabilities of the method. We used this approach to address some of our concerns about the validity aspects of PAA by analyzing data generated from specified models with particular effects structures.

Also of interest here were the replicability characteristics for the analysis of simplified data systems (as a preface to the more complex systems to be considered subsequently).

Procedure. Three sets of analyses were performed. One analysis had a theoretical emphasis; the other two were concerned primarily with replicability issues and secondarily with the theoretical considerations implied by their effects. These were (1) the study of main effect-interaction effect models discussed in Chapter II, and analyses of small data systems containing (2) a causal factor and (3) a symmetric interaction. The latter two analyses were based on data used by Magidson (1982) in a discussion of problematic categorical data systems.

The design and specifications of (1) have been previously discussed in the context of theoretical considerations of PAA. The analyses associated with this part of the defined-models simulations were not subjected to cross-validation because our interest was in the confounding of interpretation that could result from using PAA in the context of model specification.

The causal factor data set (Table 3.2) is a data system that has an underlying factor which can be used to make the observations (almost) perfectly predictable. For these data, the completeness of treatment determines the survival of the patient, regardless of type of treatment or location of facility. The conclusions drawn for analyses of subsets of these factors can be quite misleading, however: analysis of survival by type-of-treatment suggests that the standard treatment is preferable, but analysis of survival by type-of-treatment by location suggests that the new treatment is to be preferred. In reality, either treatment is equally effective when carried to completion.

The causal-factor data system was analyzed, with and without inclusion of the causal factor (completeness) by conducting 100 replications with samples generated by the bootstrap resampling method. Of interest was the empirical consistancy of the PAA results for this four variable system.

The data system containing a symmetric interaction effect (Table 3.3) contains an interaction that is approximately equal in strength - but opposite in direction - for each of two subgroups. In particular, neither the main effect for medication nor the main effect for sex is significant by conventional tests. Since no effects are found in either of the two-way contingency tables, a PAA would normally cease processing. The medication by outcome subtables conditional on sex, however, reveal a significant benefit for the use of aspirin for males, and an approximately equal benefit from the placebo for females. This information would not usually be uncovered by a standard PAA.

The symmetric-interaction data system was also analyzed by using the bootstrap method to observe the consistency of results. Of secondary interest was the extent to which this type of interaction (as represented in this particular data system) might pass the initial main-effects test.

Observations. The results of the model-recovery study (1) had implications imbedded in the theoretical aspects of the method and were therefore presented in Chapter II.

TABLE 3.2
'Causal Factor' Data from Magidson (1982)

City A

	Complete Treatment				Abbreviated Treatment			
	Standard		New		Standard		New	
Alive	5	(100%)	100	(100%)	0	(0%)	0	(0%)
Dead	0	(0%)	0	(0%)	95	(100%)	900	(100%)

City B

	Complete Treatment				Abbreviated Treatment			
	Standard		New		Standard		New	
Alive	500	(100%)	95	(100%)	0	(0%)	0	(0%)
Dead	0	(0%)	0	(0%)	500	(100%)	5	(100%)

TABLE 3.3
'Omitted Interaction' Data from Magidson (1982)

MALES

Medication	Outcome: Stricken	Not Stricken	% Stricken
Aspirin	29	171	14.5%
Placebo	46	160	22.3%

FEMALES

Medication	Outcome: Stricken	Not Stricken	% Stricken
Aspirin	17	73	18.9%
Placebo	12	77	13.5%

The analyses of the causal-factor data (2) and the symmetric-interaction data (3) demonstrated a consistent stability of results; for each of the analyses described there was no variability in the selection of predictor variables.

For the data system with the imbedded causal factor, results were consistantly confounding, as previously described, unless the causal factor was included. For the data system with the symmetric interaction, this effect was consistantly undetected using conservative statistical tests for significance. It could be detected by an appropriate manipulation of statistical cutpoint criterion or force-variable options, but unguided use of these options has other consequences which are discussed later in this chapter.

Conclusions. The most important findings from these simulations were the model-identification deficiencies of PAA discussed in Chapter II.

Results of the other two sets of analyses point out that, for simple data systems containing reasonably well defined effects, a PAA can be expected to provide replicable findings. However, as is also seen from the simulations, these findings may be misleading. This, of course, can be the case with any predictive method -- there is always the potential for misleading or erroneous interpretations of incomplete analyses where either important measures in the data system are omitted or where processing is not correctly guided in the direction of certain types of effects. We strongly suggest the use of ancillary analysis, especially in the absence of strong theory to guide interpretation of results.

Random Data

Objectives. We were interested in how PAA would function in the 'worst-case' situation where there were almost no systematic relationships present in the data. While there exist a number of statistical procedures for examining the effects of supplemental error added to the data base analyzed by a particular technique, for our purposes[6] we chose to look at the results of

analyses of a set of randomly-generated numbers. Of special interest were the potential effectiveness of various statistical and non-statistical branch-termination criteria available in the computer program.

Procedure. A data set consisting of 5ØØ 'cases' of 11 'variables' was created by using a FORTRAN computer program to sample data points from a uniform random distribution of range [Ø..1] and then dichotomizing these data using Ø.5 as the cutpoint. The characteristics of these data were checked by SPSS descriptive and correlation analyses.

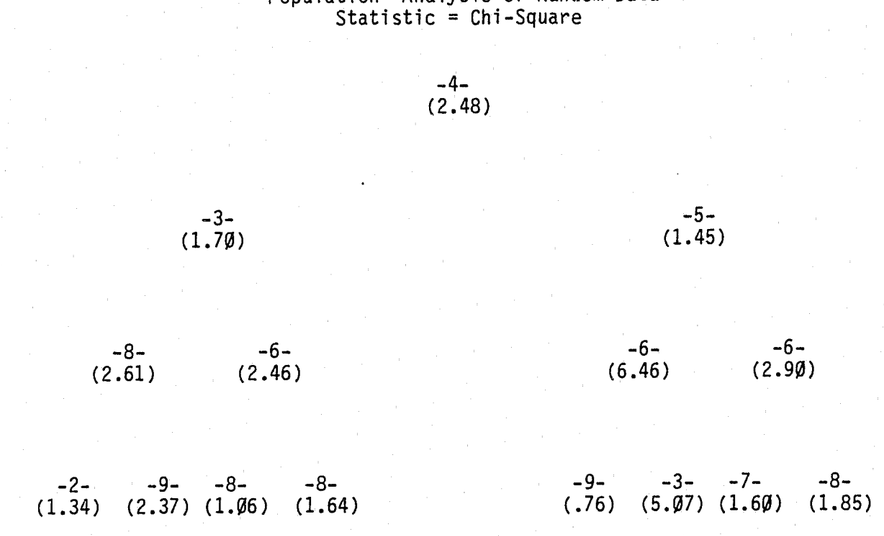
An analysis was run on the original data set using each of the four statistical coefficients; the program was allowed to proceed through four levels of PAA processing. These 'population' analyses provided the standard against which bootstrap sample replications (100 samples for each statistic) were compared.

Observations. For the population analyses, using conservative probabilistic criteria, the analysis of these data would not have proceeded beyond even the first predictor selection point. Values of the statistics for the strongest predictor variable were: Chi-square=2.484, Phi=Ø.Ø7Ø, Uncertainty=Ø.ØØ4, and Lambda=Ø.Ø64. These would not be considered significant at an alpha level of Ø.Ø5, for instance.

If the analysis is allowed to proceed, however, it is possible for the statistical measures at succeeding nodes to indeed be judged significant. At the sixth node of the third level, for example, Chi-square=5.Ø74; at the third node of the fourth level, Chi-square=5.788; at the fourth node and fourth level, Chi-square=4.455 (Figure 3.3).

For the 100 bootstrap replications that were run, results reflected the relative strengths of association found within the data: greater association generally led to better replicability. A chart of the predictor variables selected at each node is presented in Table 3.4 and illustrated in Figure 3.4.

FIGURE 3.3
'Population' Analysis of Random Data
Statistic = Chi-Square



Note: Index number of predictor variable is enclosed in dashes, value of the Chi-Square statistic is enclosed in parentheses.

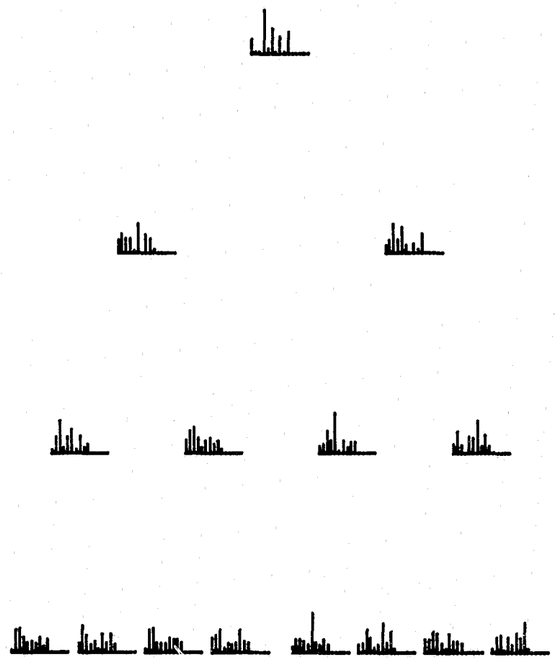
TABLE 3.4
Simulation Analysis for Random Data
Statistic = Chi-Square
100 Replications
(Entries are Frequencies that Each Predictor was Selected at Each Node)

		PREDICTOR VARIABLES									
		1	2	3	4	5	6	7	8	9	10
PAA Level = Ø	1	11	1	1	33	4	19	1	13	1	16
PAA Level = 1	2	10	15	11	11	2	22	1	14	11	3
	3	6	10	22	10	20	6	1	7	3	15
PAA Level = 2	4	3	12	24	4	12	18	3	13	4	7
	5	10	17	19	11	4	9	11	7	9	3
	6	5	7	17	10	30	2	9	4	8	8
	7	7	16	6	Ø	12	11	24	5	14	5
PAA Level = 3	8	2	17	18	12	8	6	7	12	5	10
	9	7	20	13	6	9	3	14	8	14	6
	10	3	17	19	8	7	7	11	10	10	8
	11	11	13	17	4	7	6	8	17	9	8
	12	5	10	10	9	6	30	8	6	10	6
	13	6	7	17	11	4	6	22	8	16	3
	14	10	10	16	15	7	3	14	9	9	7
	15	3	12	13	2	12	6	15	11	23	3

FIGURE 3.4

DISTRIBUTION OF PAA-SELECTED PREDICTOR VARIABLES

Random Data System, 100 Replications
Chi-Square Statistic



NOTES: 1) Each set of bars represents a node in the PAA tree.
2) Each vertical bar represents the relative frequency of selection for each predictor variable.

Notice that each variable was selected at least once as the best predictor of the criterion; the predictor variable with the largest 'population' Chi-square value was selected most frequently, but only 33% of the time.

Conclusions. Consideration of the results of the analysis of this random data system shows that it is possible (and indeed quite likely) for numerous 'chance' relationships to be selected by a PAA procedure. Theoretically, we expect on average that, for an alpha of Ø.Ø5, one in twenty contingency tables would be falsely identified as representing significant association/predictability in the population. For a PAA carried to five levels, there are 63 2x2 tables -- an average of three of which would be significant by chance alone. Simple methods for dealing with this problem are not available; one ad hoc approach would be to use the combination of both a probabilistic cutoff criterion for the statistic and a minimum size criterion for either the 2x2 cell or marginal frequencies.

Most likely the presence of this kind of background noise would not have pronounced effects at the early levels of a PAA of real data. Systematic relationships would be expected to exhibit strong enough association/prediction measures to be selected. But a few levels into the analysis, where strong effects have already been extracted and predictor-criterion tables begin to have similar statistical values, this effect becomes potentially a serious problem. The higher in the PAA tree this occurs, the more disruptive the net effect on the overall analysis.

The bootstrap replications conducted for this data system give an indication of just how tenuous the selection is for predictors in a noisy system. The strongest overall effect has borderline statistical significance (if we accept this sample as a 'population'), but it replicates over only one third of the samples. Proceeding farther down the PAA tree, the distribution of selected predictor variables (Figure 3.4) becomes even more uniform across samples.

OBTS Probation-Eligible Data

Objectives. Aside from the theoretical and simulation-based arguments presented, one might conjecture that there remains for some researchers the question of whether the PAA approach might still 'work' (in the very broadest sense of the word) for some empirical criminal justice data systems. We ask: Can useful information, which is trustworthy, be extracted from a Predictive Attribute Analysis of real criminal justice data?

To address this question, we conducted a series of analyses on data considered to represent a typical research application for the PAA method. We selected a 'Probation-Eligible' subset of the 1980 NYS Offender-Based-Transaction-Statistics database[7] ("PROB8Ø") for study. These data were considered to be representative of the kind generally used by criminal justice researchers; additionally, some of the relationships present in the data system were known from prior analytic work.[8]

We examined the results of a number of population and validation PAAs on these data, looking at (1) issues of replicability and the implications for statements of confidence appropriate to results, and (2) combinations of parameter specifications providing the most meaningful results (based on our understanding of the relationships existing within the data). In particular, we looked at both the sequence of selected predictor variables and the individuals comprising the terminal subgroups defined by the analysis. By looking at the membership of the terminal subgroups, we hoped to be able to assess whether, aside from the issues of model identification and variable-selection sequence, the individuals categorized as a result of the analysis were a meaningful and reliably-clustered group on which policy decisions might be appropriately based.

Procedure. A total of 7813 records for cases categorized as probation-eligible were drawn from the 1980 NYS OBTS database for two regions of New York

TABLE 3.5
Description of PROB8Ø Measures

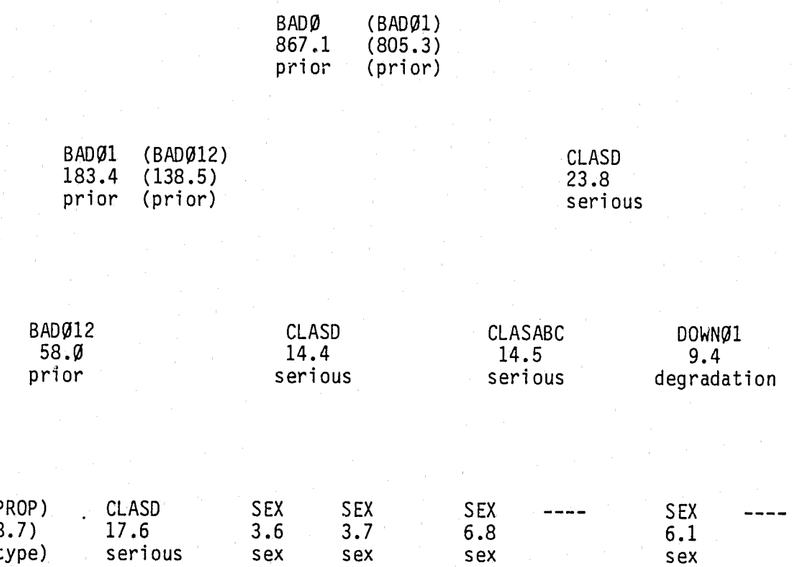
VAR #	VAR Name	VAR Type	Description
1	SEX	Sex	Sex of offender [Ø=F, 1=M]
2	BADØ	Prior	Prior criminal history: any [Ø=No, 1=Yes]
3	BADØ1	Prior	Prior criminal history: moderate [Ø=No, 1=Yes]
4	BADØ12	Prior	Prior criminal history: substantial [Ø=No, 1=Yes]
5	CLASD	Serious	Class D offense [Ø=No, 1=Yes]
6	CLASABC	Serious	Class A, B, C offenses [Ø=No, 1=Yes]
7	PERS	Type	Person crime [Ø=No, 1=Yes]
8	PROP	Type	Property crime [Ø=No, 1=Yes]
9	DRUG	Type	Drug crime [Ø=No, 1=Yes]
1Ø	DOWNØ	Degradation	Charge reduced more than one class [Ø=No, 1=Yes]
11	DOWNØ1	Degradation	Charge reduced one class [Ø=No, 1=Yes]
12	AGERISK	Age	Offender age between 2Ø and 3Ø [Ø=No, 1=Yes]
13	BLACK	Race	Offender was Black [Ø=No, 1=Yes]
14	INCARC	(Dep Measure)	Offender Incarceration [Ø=No, 1=Yes]

State: 6Ø78 for New York City (5 boroughs) and 1735 for NYC suburban counties. The variables chosen for analysis (Table 3.5) were among those used in prior studies of incarceration predictability. The measures included individual-level attributes of age, sex, race, and prior criminal history, as well as offense-level characteristics such as type and seriousness of crime. The criterion to be predicted was whether or not the individual was incarcerated. Variables not already coded as binary data were dichotomized, principally on the basis of the meaning attached to various ranges of the categorical and continuous measures but with some consideration given to the empirical distributions.

As with the previous data systems, a 'population' analysis was run on the original data for each of the four statistical coefficients; this was done separately for the New York City and the NYC Suburban regions. Selected PAA trees are presented in Figures 3.5 (NYC, Chi-square), Figure 3.6 (NYC, Lambda), Figure 3.7 (suburban, Chi-square), and Figure 3.8 (suburban, Lambda). This set of population analyses was again the standard to which bootstrap replications were compared. In general, the strategy was to conduct a set of 100 resamplings for each of the four statistical criteria for both regions; the processing was limited to a depth of four levels (16 terminal subgroups) where population-sample terminal subgroup comparisons were made. In addition to the results reported directly here (through Tables and Figures), a number of analyses that were either partial or extended versions of the ones reported were run to insure that conclusions being drawn were appropriate.

We wished here to have a more formal means of evaluating the information provided by the resampling replications with regard to the terminal subgroups delineated by the analysis. A supplementary computer program was written to compare the case membership of the terminal subgroups selected by the population analysis to the terminal subgroups selected by each of the bootstrap replications. Phi coefficients were calculated for all population-sample comparisons and averaged across the 100 replications to provide (1) an index of the goodness-of-fit for the samples to the population and (2) and indication of

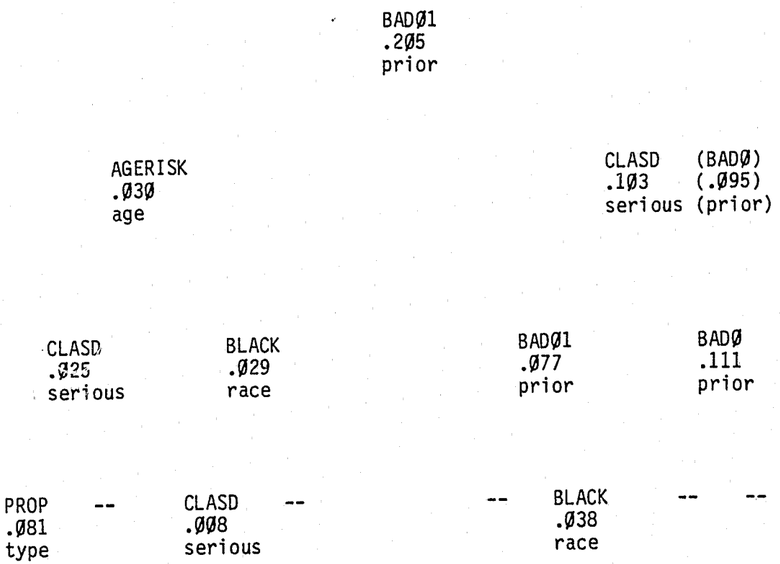
FIGURE 3.5
'Population' Analysis for PROB8Ø Data
NYC, Statistic = Chi-Square



Note: Information at each node is in the format:
VARIABLE (RUNNER-UP)
stat val (stat val)
info type (info type)

Runner-up predictors are given only when they have very similar stat values to the selected predictor.

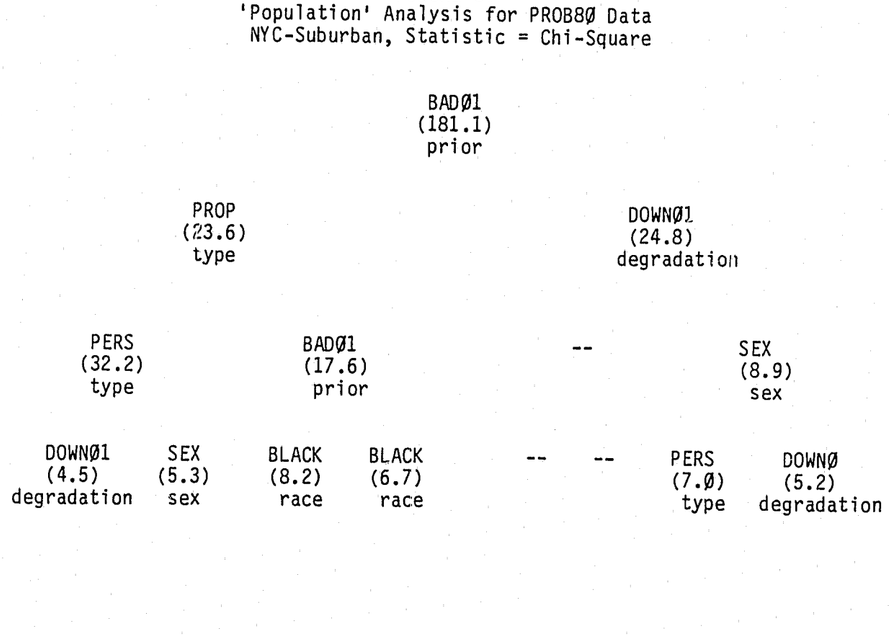
FIGURE 3.6
'Population' Analysis for PROB8Ø Data
NYC, Statistic = Lambda



Note: Information at each node is in the format:
VARIABLE (RUNNER-UP)
stat val (stat val)
info type (info type)

Runner-up predictors are given only when they have very similar stat values to the selected predictor.

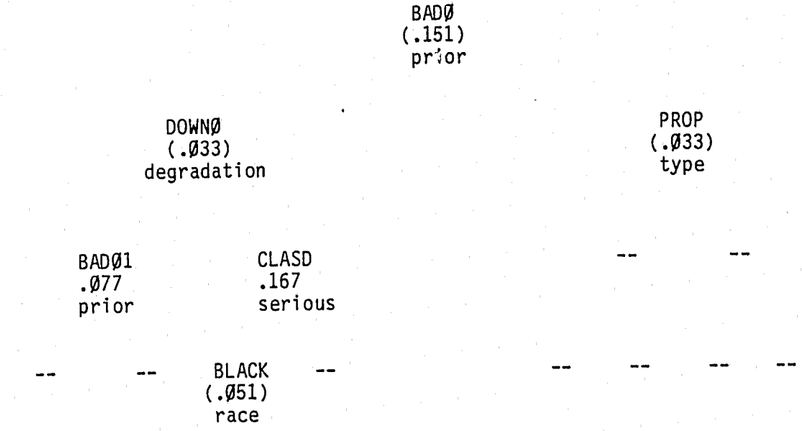
FIGURE 3.7
'Population' Analysis for PROB8Ø Data
NYC-Suburban, Statistic = Chi-Square



Note: Information at each node is in the format:
VARIABLE (RUNNER-UP)
stat val (stat val)
info type (info type)

Runner-up predictors are given only when they have very similar stat values to the selected predictor.

FIGURE 3.8
'Population' Analysis for PROB8Ø Data
NYC-Suburban, Statistic = Lambda



Note: Information at each node is in the format:
VARIABLE (RUNNER-UP)
stat val (stat val)
info type (info type)

Runner-up predictors are given only when they have very similar stat values to the selected predictor.

variability across the set of sample analyses. Figure 3.9 illustrates this process.

Observations. We discuss here only the analyses for the New York City data; specific results for the two regions were generally different (as anticipated), but the conclusions drawn regarding the PAA method itself were equivalent. We begin with the series of population analyses.

For the analysis using the Chi-Square statistic as the predictor-selection criterion, a probability-based termination criterion (alpha = 0.05) was used. Figure 3.5 displays the PAA tree for the analysis, conducted to a maximum of four levels for each branch. At nodes where alternative measures were close runners-up to the selected variable, they are noted in parentheses.

The initial split was based on the BADØ measure (a measure of prior criminal history) with the BADØ1 measure being a relatively close runner-up (Chi-square of 867. versus 805.). Of interest is the far left branch of the tree: [BADØ->BADØ1->BADØ12->BLACK]. Although these four measures are highly correlated, they are successively selected in a branch that might be thought of as a progressively-more-incriminating characterization.

For the analysis using Phi as the predictor-selection criterion, results were the same as the Chi-square analysis except where branch processing was allowed to continue beyond the point where the Chi-square minimum terminated the analysis (a less-restrictive minimum statistic was used here).

For the analysis using the Uncertainty coefficient, results were the same as for the Chi-square/Phi analyses through the third level of processing. The far left branch [BADØ->BADØ1->BADØ12->PROP] did not precisely replicate the Chi-square results described previously; it would, however, be speculation to

FIGURE 3.9
Procedure for Comparing Population and Sample Subgroup Membersip

Comparison Table for Each Terminal Group of the PAA:

SAMPLE CASES	POPULATION CASES: Absent from Group	POPULATION CASES: Present in Group
Absent From Group	A	B
Present in Group	C	D

Where: A and D indicate corresponding classifications
B and C indicate differing classifications

attempt substantive explanations for the points of divergence found at the nodes in the fourth level.

For the analysis using the Lambda coefficient, results were entirely different from the above three analyses (see Figure 3.6). Recall, of course, that Lambda measures improvement in predictability -- not simply predictability. The first split was based on BADØ1 rather than BADØ, and the left-hand branch proceeds [BADØ1->AGERISK->CLASD->PROP] rather than [BADØ->BADØ1->BADØ12->BLACK]. Sex, which appeared quite frequently in the third level of the previous three analyses, did not appear at all here. Race appeared much more frequently for the Lambda-based PAA than the other analyses.

For the bootstrap validation series of analyses, 100 replications of the population PAA for each statistical criterion were run. Results are discussed in terms of (1) replication of predictor variable sequences in the PAA tree, and (2) replication of subgroup membership.

In brief, the PAA tree structures did not replicate well beyond the first level. For NYC data, Tables 3.6 through 3.8 present the distribution of predictors selected at each node of the analysis for Chi-Square, Uncertainty, and Lambda coefficients. Figures 3.1Ø, 3.11, and 3.12 correspond to these tables and illustrate graphically how the results at successive nodes of a particular branch become less and less reliable. (A perfectly-replicated PAA would, for example, have only one bar for each node.) These representations require careful study to determine the source of dispersion at each node, however, since the variables selected at each node are conditional on the variables selected at the previous decision point, and hence the counts at lower levels for any particular variable can be the by-product of a number of different paths. The end result, however, seems to be a tendency toward a more uniform distribution of predictors as one proceeds down the PAA tree.

With regard to the stability of subgroup membership across samples, there were three general questions: (1) what statements can be made about the global

TABLE 3.6
Simulation Analysis for PROB8Ø Data
Region = NYC Statistic = Chi-Square
1ØØ Replications
(Entries are Frequencies that Each Predictor was Selected at Each Node)

		Variable Number												
Node		1	2	3	4	5	6	7	8	9	10	11	12	13
Level=Ø	1			100										
Level=1	2				100									
	3		100											
Level=2	4	10				2			57	3	4			24
	5	2				89			1			8		
	6	5				37					58			
	7	59				11	2		19		5	1	3	
Level=3	8	11				52		2	11	2	2	4		16
	9	23				1			9	1	19	39	5	3
	10	5				9			70	1			15	
	11	18				2			4	52		24		
	12	52				22	1	2	2	1	13	4		3
	13	8				6	2	11	1	49	13		6	4
	14	5				1	30	27	15	2	1	7	6	6
	15	29				36	1	3	3	8	1	4	15	

TABLE 3.7
Simulation Analysis for PROB8Ø Data
Region = NYC Statistic = Uncertainty
1ØØ Replications
(Entries are Frequencies that Each Predictor was Selected at Each Node)

		VARIABLE NUMBER												
NODE		1	2	3	4	5	6	7	8	9	10	11	12	13
Level=Ø	1			100										
Level=1	2				100									
	3		100											
Level=2	4	3				2			65		5			25
	5	1				90			1				8	
	6	5				37					58			
	7	64				10	2		18		2	1	3	
Level=3	8	6				59		4	13	2	1	5	1	9
	9	19				2			6	1	32	29	11	3
	10	5				9			71				15	
	11	14				1			6	49		30		
	12	53				21	1	2	2	1	12	5		3
	13	6				5	2	11	1	50	15		7	3
	14	4				1	17	34	24	2	1	7	5	5
	15	28				37	1	2	3	7	3	4	15	

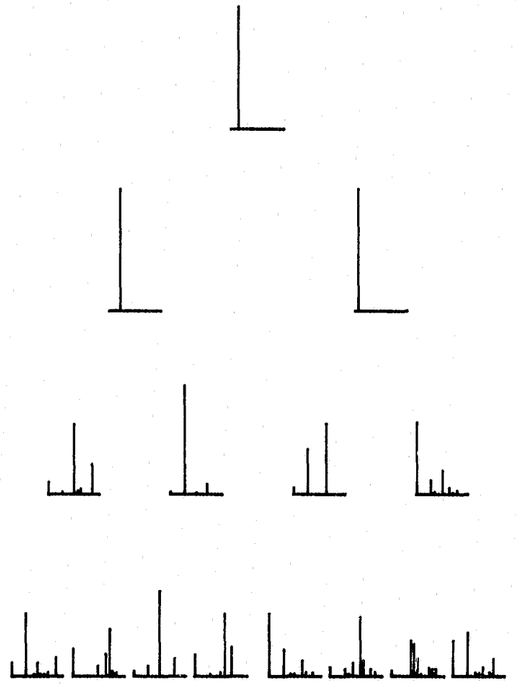
TABLE 3.8
Simulation Analysis for PROB8Ø Data
Region = NYC Statistic = Lambda
1ØØ Replications
(Entries are Frequencies that Each Predictor was Selected at Each Node)

		VARIABLE NUMBER												
NODES		1	2	3	4	5	6	7	8	9	10	11	12	13
Level=Ø	1			100										
Level=1	2					51		23						26
	3		3			95								2
Level=2	4					34			51					9
	5							1					29	
	6		12			2		11						47
	7		17											1
Level=3	8				3	11			26				10	50
	9				64			11			10			15
	10				44									
	11					5		53					3	31
	12		70					16					3	9
	13		66			3								30
	14					3								
	15													95

FIGURE 3.10

DISTRIBUTION OF PAA-SELECTED PREDICTOR VARIABLES

PROB80 Data System, 100 Replications
Chi-Square Statistic

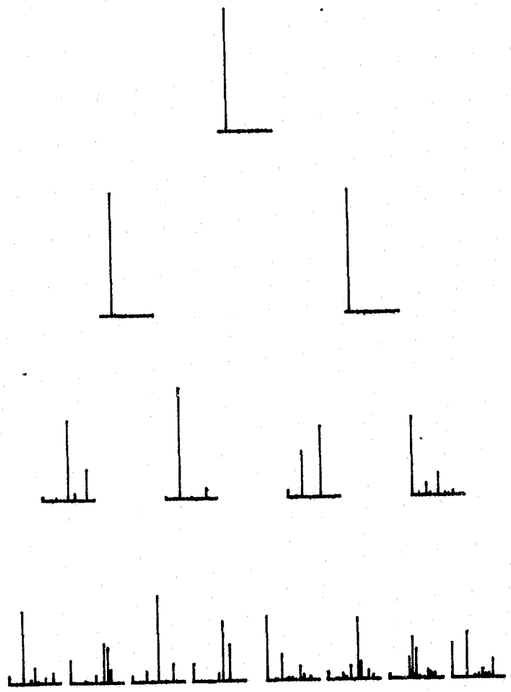


NOTES: 1) Each set of bars represents a node in the PAA tree.
2) Each vertical bar represents the relative frequency of selection for each predictor variable.

FIGURE 3.11

DISTRIBUTION OF PAA-SELECTED PREDICTOR VARIABLES

PROB80 Data System, 100 Replications
Uncertainty Statistic

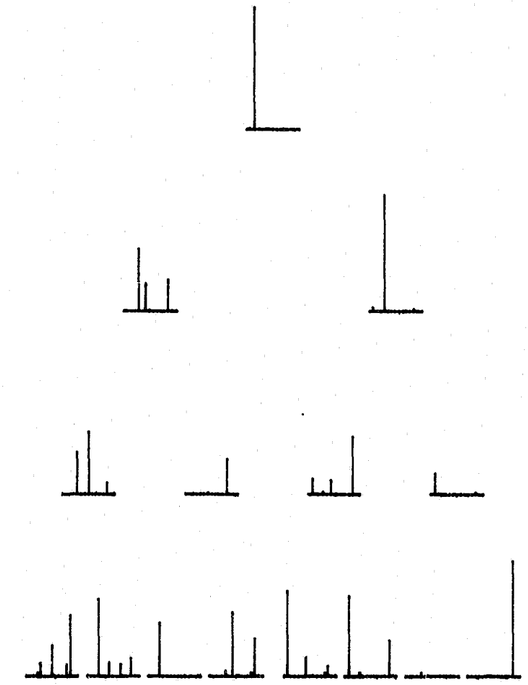


NOTES: 1) Each set of bars represents a node in the PAA tree.
2) Each vertical bar represents the relative frequency of selection for each predictor variable.

FIGURE 3.12

DISTRIBUTION OF PAA-SELECTED PREDICTOR VARIABLES

PROB80 Data System, 100 Replications
Lambda Statistic



NOTES: 1) Each set of bars represents a node in the PAA tree.
2) Each vertical bar represents the relative frequency of selection for each predictor variable.

level of sample-population agreement, (2) what is the variability around this measure of agreement, and (3) are there any meaningful differences in stability for different predictor-selection statistical criteria? We found that (1) there was generally a poor correspondence between sample and population subgroup classification - typically less than 50 percent 'correct' classification, (2) there was considerable variability around this correspondence, generally becomming more variable (and unreliable) as one progresses to deeper levels of the analysis, and (3) these results held true regardless of the statistical criterion used.

Table 3.9 and Figures 3.13-3.15 present the goodness-of-fit coefficients and respective frequency distributions for each statistical criterion, averaged across the 100 replications. Looking at the distribution for the Chi-square criterion (Figure 3.13), one can see that only 8 percent of the sample terminal subgroups (16 subgroups times 100 replications) contained exactly the same individuals as the corresponding population terminal subgroups. The median goodness-of-fit coefficient, which is equivalent here to a correlation coefficient, is .44; that is, the expected correlation between the membership of the population and any corresponding sample terminal subgroup is only .44. The variablity of these measures is high for each of the statistical criteria, as evidenced by the spread of the distributions and the resulting quartile statistics (Table 3.9).

Conclusions. Consideration of the series of analyses of the PROB80 data leads us to conclude that Predictive Attribute Analysis does not have particular utility when used with 'real' criminal justice data. PAA tree structures did not replicate well beyond the initial levels of analysis. Even considering only terminal subgroup membership, a single PAA cannot be expected to provide a reliable representation of the groups of individuals who cluster together.

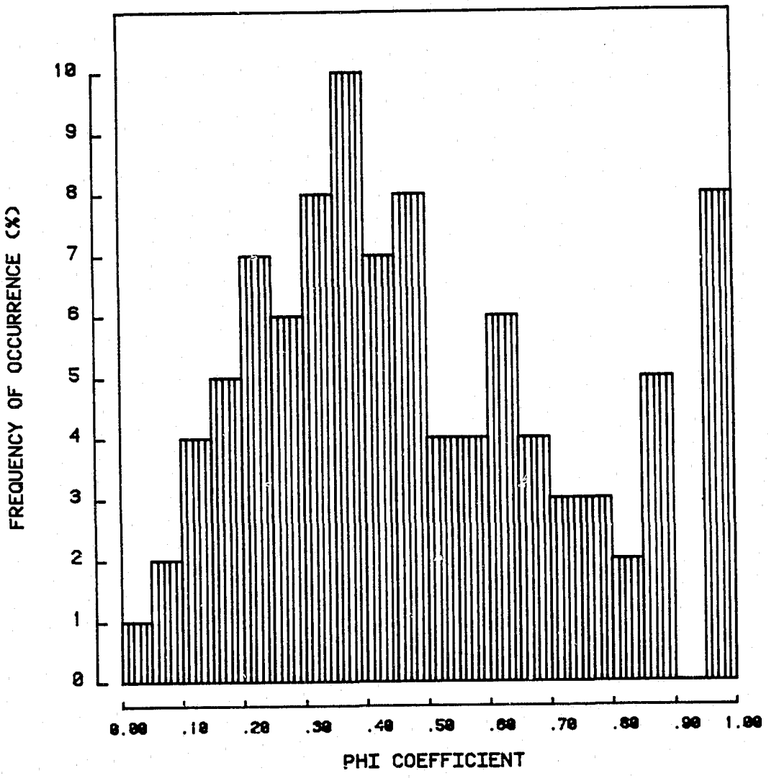
TABLE 3.9
Terminal Subgroup Analysis for PROB80 Data
Comparison of PHI Goodness-of-Fit Coefficients
(Population-Sample) For 100 Samples

		NYC Region	Suburban Region
Chi-Square	U Quartile	.617	.671
	Median	.408	.440
	L Quartile	.285	.306
Uncertainty	U Quartile	.610	.739
	Median	.402	.499
	L Quartile	.284	.377
Lambda	U Quartile	.764	.671
	Median	.640	.362
	L Quartile	.453	.232

FIGURE 3.13

PROB80 SUBGROUP ANALYSIS: CHI-SQUARE STATISTIC

Distribution of Population-Sample Goodness-of-Fit Coefficients



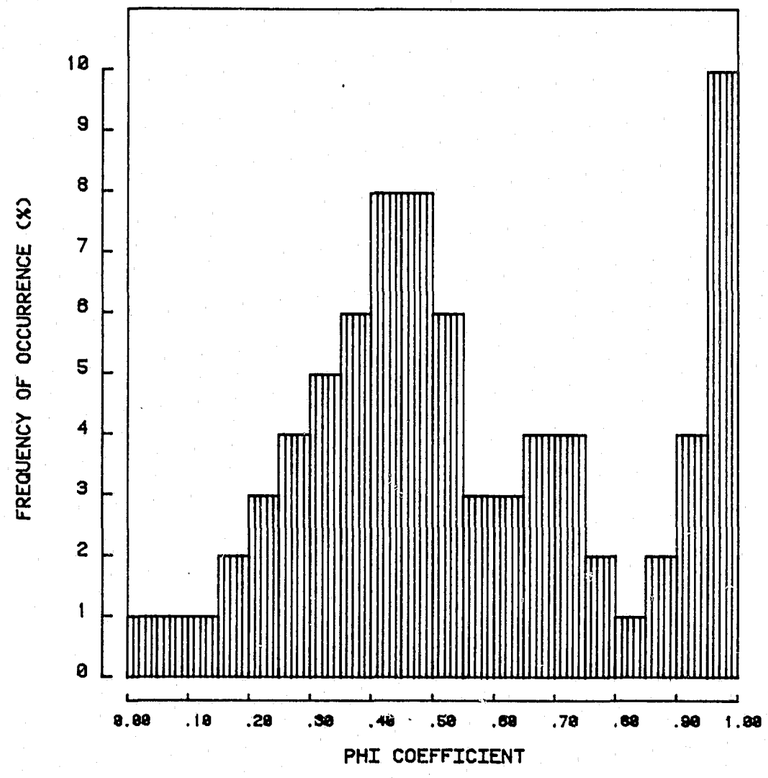
NOTES

Values are from the analysis of NYC-suburban data,
100 replications,
statistical criterion = chi-square.

FIGURE 3.14

PROB80 SUBGROUP ANALYSIS: UNCERTAINTY STATISTIC

Distribution of Population-Sample Goodness-of-Fit Coefficients



NOTES

Values are from the analysis of NYC-suburban data,
100 replications,
statistical criterion = uncertainty.

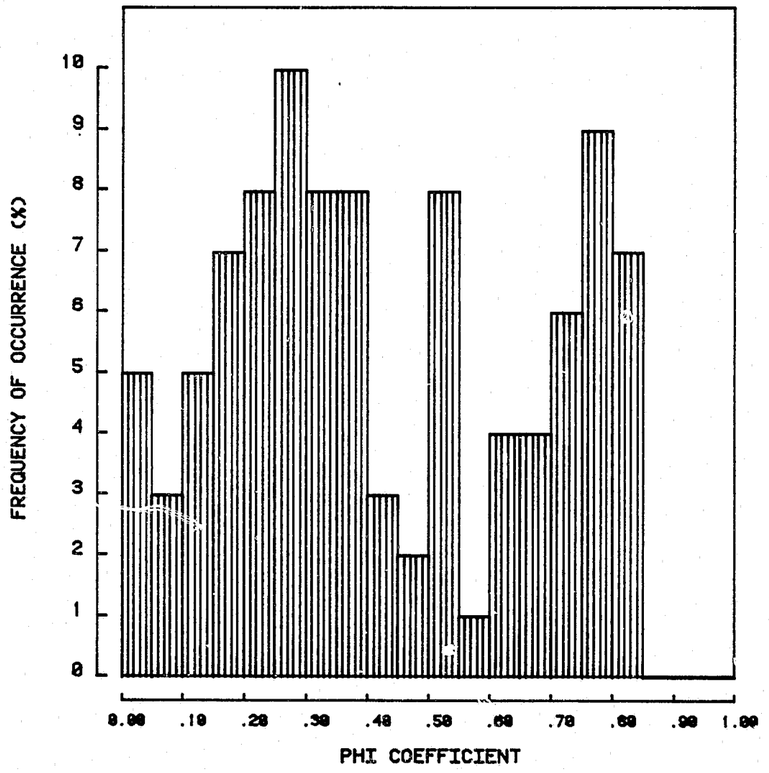
CONTINUED

1 OF 2

FIGURE 3.15

PROB80 SUBGROUP ANALYSIS: LAMBDA STATISTIC

Distribution of Population-Sample Goodness-of-Fit Coefficients



NOTES

Values are from the analysis of NYC-suburban data,
100 replications,
statistical criterion = lambda

Footnotes

[1]We use this acronym primarily to allow us to easily distinguish the computer program from the analytical technique which it implements, secondarily to emphasize the evaluation capabilities of the program.

[2]Even at one page per node, a 4-level analysis with 100 replications can generate 3200+ pages of output; with four statistical criteria, 12,800+ pages.

[3]These are documented in the User's Guide and the internal program documentation.

[4]The program has also been compiled and run on the SPERRY 1100/83 using ASCII FORTRAN.

[5]Program dimensions are easy to modify. See the User's Guide for discussion.

[6]Initially, we expected to re-analyze some of the data systems discussed in this report contaminated with various percentages of randomly-distributed error (perhaps 5%, 10%, 20%). Because of our findings of poor-reproducability from the original analyses, however, we doubted the utility of much further effort toward documenting subtleties of PAA performance.

[7]For a more complete description of this database than will be presented here, refer to New York State Criminal Justice Processing: Felony Offenders Disposed in 1980, 3 Vol., Harig, Thomas, Division of Criminal Justice Services, OPARSS, (1983).

[8]Frederick, Bruce C. and Sherwood E. Zimmerman. Discrimination and the Decision to Incarcerate, (Albany, NY) Division of Criminal Justice Services, OPARSS, (1983).

IV
CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS AND RECOMMENDATIONS

This chapter presents a brief recapitulation of the findings detailed in the other sections of this report. Recommendations are based on the discussions contained in this Technical Report as well as our own experiences in using the PAA method. We attempt to provide generalizations which are germaine to practical applications of PAA in criminal justice research problems.

Appropriate Applications for PAA

The proper design of a PAA requires a prior knowledge of some of the characteristics of the data system under study. The proper interpretation of the results of a PAA requires efforts toward the validation of initial findings. Therefore, a PAA should not be the first step in the analysis of a complex data system, nor should it be the last.

Predictive Attribute Analysis is not appropriate for model-development applications. This is contrary to popular perception and usage. It is the case regardless of whether the application is exploratory or confirmatory in nature. Because a PAA cannot, in principle, distinguish between a main effect and an interaction effect, it is incorrect to presume that one has uncovered a complex (yet significant) interaction effect defined by the sequence of variables selected at the nodes along a particular PAA branch. The PAA algorithm simply detects the series of strongest main effects (which may or may not be components of higher-order interaction terms) that are conditional on the defining characteristics of the subgroups found at the particular nodes of the analysis. Thus, in either exploratory or confirmatory types of analysis, one cannot be certain as to the particular model that is most appropriate to the data.

Predictive Attribute Analysis is perhaps best used as a somewhat serendipitous pre-formal-analysis data description technique. As such, the analysis is used as an exploratory tool to suggest effects and relationships which might not have previously occurred to the analyst. Any hypotheses resulting from this exploratory activity, however, require validation by independent means.

Choice of a PAA Design

Although we have suggested that the method may be appropriately used in an exploratory manner, that is not to say that we have recommended its use in a haphazard manner. If the use of a Predictive Attribute Analysis has been deemed appropriate, there are important considerations in the selection of the parameters to guide the processing within a PAA. The interpretability of the information resulting from any analysis, even exploratory, is a function of the precision of the questions asked beforehand.

As a general approach, we suggest the accumulation of as much prior evidence as possible when addressing particular research questions; complex social science models suggested by post hoc deduction seldom replicate convincingly. The development of a useful prediction model should be expected to require iteration, and a hypothesis-based approach provides a good sequential analysis strategy for validating true effects and for rejecting competing alternative explanations.

Some prior theory, however complete, should guide the specification of parameters supplied to a Predictive Attribute Analysis. Especially important among these parameter-specification considerations are the statistical criterion used to measure predictive power, the stopping criteria used, and the depth at which the results of the analysis are to be considered acceptably valid. The conclusions drawn from particular PAAs were shown differ markedly depending on whether a measure of association (or a measure of predictability) was used.

In assembling the set of predictor variables to be used in an analysis, it is important to remember that the PAA method is not a multivariate procedure where interrelationships among the predictor variables are explicitly controlled. The pattern of predictor variable selection as well as the final subgroup definitions are a function of the interrelationships of the prediction measures. This characteristic must be kept in mind, lest one particular 'factor,' by virtue of a relative over-representation in the set of predictors, appear inordinately influential. Conversely, the exclusion of a demonstrably good predictor can seemingly elevate the importance of less effective predictors. We suggest that at least one of the ancillary statistical techniques used in conjunction with the PAA should focus on interrelationships among the variables in the data system.

Interpretation of Results

The PAA method has observed instability with respect to replication across samples drawn from a particular population data system. This is true for both the branching pattern of predictor variables and for the particular individuals comprising the membership of the terminal subgroups.

Analysts should, therefore, exercise caution when determining the depth of the analysis at which results are to be reported. While appropriately conservative stopping criteria may be of some use in this regard, they typically do not provide either definitive or statistically-interpretable bounds. The bootstrap resampling capabilities provided by the computer program provide valuable information for such decisions, and their use is strongly encouraged.

Since we classify PAA as an exploratory type of method, we caution analysts to conscienciously examine alternative hypotheses when using the method. An everpresent consideration is whether or not unrepresented effects are a significant underlying influence behind the observations that are the focus of analysis. A PAA should be regarded as only one component of a comprehensive analytic strategy, typically incorporating several statistical techniques, which attempts to converge on a proper interpretation.

Summary

A Predictive Attribute Analysis, as a cautiously conducted and properly interpreted component of a well-planned and thorough research design, may provide useful information to the criminal justice data analyst. Its inherent limitations, however, argue against casual use and informal interpretation. Research conclusions which are based on the results of a Predictive Attribute Analysis should always be accompanied by additional supporting evidence.

REFERENCES

Anderson, T.W. (1959).
An Introduction to Multivariate Statistical Analysis.
New York: John Wiley & Sons.

Bishop, Y.M.M.; Feinberg, S.E.; Holland, P.W. (1975).
Discrete Multivariate Analysis: Theory and Practice
Cambridge: MIT Press.

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

Cox, D.R. (1970).
The Analysis of Binary Data
Methuen: London.

Davis, J.A. (1974).
"Hierarchial Models for Significance Tests in Multivariate Contingency Tables: An Exegesis of Goodman's Recent Papers," in Sociological Methodology: 1973-1974.
San Francisco: Jossey-Bass.

Dixon, W.J. (Ed.) (1981).
BMDP Statistical Software.
Berkeley: University of California Press.

Dolmatch, T.B. (Ed.) (1981).
Information Please Almanac: 1982.
New York: Simon and Schuster.

Efron, B. (1979).
Bootstrap Methods: Another Look at the Jackknife.
Annals of Statistics, 7, 1-26.

Efron, B. (1982).
The Jackknife, the bootstrap and other resampling plans.
SIAM, Monograph # 38, CBMS-NSF.

Efron, B. and Gong, G. (1983).
A Leisurely look at the bootstrap, the jackknife, and cross-validation.
American Statistician, Feb. 1983, 37 #1, p. 36-48.

Everitt, B.S. (1977).
The Analysis of Contingency Tables.
New York: Halstal Press.

Feinberg, S.E. (1977).
The Analysis of Cross-Classified Categorical Data.
Cambridge: MIT Press, 1977.

Finney, D.J. (1952).
Statistical Methods in Biological Assay.
London: Griffin.

Fleiss, Joseph L. (1981).
Statistical Methods for Rates and Proportions (2nd Ed.)
New York: John Wiley & Sons.

Goodman, L.A. (1972).
A General Model for the Analysis of Surveys.
American Journal of Sociology, 77, 1035-1086.

Goodman, L.A. (1976).
The relationship between modified an usual multiple regression approaches to the analysis of dichotomous variables. In David R. Herse (Ed),
Sociological Methodology: 1976.
San Francisco: Jossey-Bass Inc.

Goodman, L.A. (1978).
Analyzing Qualitative Categorical Data.
Cambridge, MA; Abt Associates.

Goodman, L.A. and Kruskal, W.H. (1979).
Measures of Association for Cross Classifcations.
New York: Springer-Verlag

Hanushek, E.A. & Jackson, J.E. (1977).
Statistical Methods for Social Scientists.
New York: Academic Press.

Huber, P.J. (1981).
Robust Statistics.
New York: John Wiley & Sons.

Lewis, B.N. (1962).
On the Analysis of Interaction in Multi-dimensional Contingency Tables.
Journal of the Royal Statistical Society, A, 125-1, p. 88-117.

MacNaughton-Smith, P. (1963).
The classification of individuals by the possession of attributes associated with accriterion. Biometrics, 1963, 19(2), 364-366

MacNaughton-Smith, P. (1965).
Some Statistical and Other Numerical Techniques for Classifying Individuals.
London: HMSO.

Magidson, J. (1982).
"Some Common Pitfalls in Causal Analysis of Categorical Data."
Journal of Marketing Research.

Morgan, J.N. and Messenger, R.C. (1973).
THAID: A sequential Search Program for the Analysis of Nominal Scale Dependent Variables.
Ann Arbor: ISR, University of Michigan,

Mosteller, F., and Tukey, J.W. (1977).
Data Analysis and Regression.
Reading, MA: Addison-Wesley.

Nerlove, M. and Press, J. (1973).
Univariate and multivariate log-linear and logistic models. Prepared for the Economic Development Administration and the National Institute of Health, R-1306-EDA/NIH, December, 1973.

Nie, N.H. (1975).
Statistical Package for the Social Sciences.
New York: McGraw-Hill.

Perrault, W.D. Jr. & Barksdale, H.C. Jr. (1980).
A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research.
Journal of Marketing Research, Vol. 17, Nov. 1980, p 503-15.

Pruzek, R.M. & Walker, N.F. (1982).
"Improving Experimental Design Efficiency Through the Use of Relative Prior Means."
Paper presented at the Annual Conference of the American Educational Research Association, New York City.

Simon, F.H. (1971).
Prediction Methods in Criminology.
London: Her Majesty's Stationery Office.

Soloman, H. (1976).
Parole outcome: A multidimensional contingency table analysis. Journal of Research in Crime and Delinquency, 1976, 13, 107-119.

REFERENCES

Sonquist, J.A. & Morgan, J.N. (1964).
The Detection of Interaction Effects.
Ann Arbor: The University of Michigan.

Sonquist, J.A., Baker, E.L., and Morgan, J.L. (1971).
Searching for Structure.
Ann Arbor: ISR, The University of Michigan.

Stone, M. (1974).
Cross-validatory choice and assessment of statistical predictions.
J. Royal Statistical Society, Ser. B., 36, pp. 111-147.

Sutton, L.P. (1978).
Variations in federal criminal sentences: A statistical assessment at the national level (Utilization of Criminal Justice Statistics Project, Analytic Report 17, SD-AR-17 1978).
National Criminal Justice Information and Statistical Service, Law Enforcement assistance Administration, U.S. Department of Justice.

Theil, H. (1967).
Economics and Information Theory.
Chicago: Rand McNally.

Wilkins, L.T. and MacNaughton-Smith, P. (1964).
New Prediction and classification methods in criminology.
Journal of Research on Crime and Delinquency, 1964, 19, 19-32.

APPENDIX

PREDICTIVE ATTRIBUTE ANALYSIS

PROB80 : POPULATION ANALYSIS
REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 0 GROUP= 1

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		35.856	1	129.	1606.	7.4	92.6
2		181.067	2	1087.	648.	62.7	37.3
3		112.619	3	682.	1053.	39.3	60.7
4		59.107	4	277.	1458.	16.0	84.0
5		16.169	5	1050.	685.	60.5	39.5
6		0.029	6	1641.	94.	94.6	5.4
7		7.084	7	1347.	388.	77.6	22.4
8		9.944	8	691.	1044.	39.8	60.2
9		7.393	9	1624.	111.	93.6	6.4
10		8.707	10	1044.	691.	60.2	39.8
11		8.437	11	235.	1500.	13.5	86.5
12		17.327	12	1093.	642.	63.0	37.0
13		24.723	13	1155.	580.	66.6	33.4

CONTINGENCY TABLE ANALYSIS LEVEL= 0 GROUP= 1

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 2	-0-	276.	811.	1087.
	-1-	374.	274.	648.
		650.	1085.	1735.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 2	-0-	15.9	46.7	62.7
	-1-	21.6	15.8	37.3
		37.5	62.5	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 1 GROUP= 1

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -X0-	COMPOSITION -X1-
1		10.232	1	52.	1035.	4.8	95.2
2		0.000	2	1087.	0.	100.0	0.0
3		10.207	3	682.	405.	62.7	37.3
4		13.923	4	277.	810.	25.5	74.5
5		13.357	5	651.	436.	59.9	40.1
6		0.130	6	1055.	32.	97.1	2.9
7		2.600	7	882.	205.	81.1	18.9
8		23.590	8	392.	695.	36.1	63.9
9		7.022	9	1039.	48.	95.6	4.4
10		11.959	10	609.	478.	56.0	44.0
11		0.352	11	129.	958.	11.9	88.1
12		2.891	12	618.	469.	56.9	43.1
13		19.265	13	692.	395.	63.7	36.3

CONTINGENCY TABLE ANALYSIS LEVEL= 1 GROUP= 1

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 8	-0-	133.	259.	392.
	-1-	143.	552.	695.
		276.	811.	1087.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 8	-0-	12.2	23.8	36.1
	-1-	13.2	50.8	63.9
		25.4	74.6	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 1 GROUP= 2

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		9.525	1	77.	571.	11.9	88.1
2		0.000	2	0.	648.	0.0	100.0
3		0.000	3	0.	648.	0.0	100.0
4		0.000	4	0.	648.	0.0	100.0
5		3.667	5	399.	249.	61.6	38.4
6		5.639	6	586.	62	90.4	9.6
7		20.482	7	465.	183.	71.8	28.2
8		6.169	8	299.	349.	46.1	53.9
9		0.133	9	585.	63.	90.3	9.7
10		10.416	10	435.	213.	67.1	32.9
11		24.831	11	106.	542.	16.4	83.6
12		1.516	12	475.	173.	73.3	26.7
13		1.423	13	463.	185.	71.5	28.5

CONTINGENCY TABLE ANALYSIS LEVEL= 1 GROUP= 2

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 11	-0-	38.	68.	106.
	-1-	336.	206.	542.
		374.	274.	648.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 11	-0-	5.9	10.5	16.4
	-1-	51.9	31.8	83.6
		57.7	42.3	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 1

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		2.866	1	17.	375.	4.3	95.7
2		0.000	2	392.	0.	100.0	0.0
3		0.001	3	242.	150.	61.7	38.3
4		2.644	4	80.	312.	20.4	79.6
5		18.118	5	250.	142.	63.8	36.2
6		0.386	6	364.	28.	92.9	7.1
7		32.163	7	187.	205.	47.7	52.3
8		0.000	8	392.	0.	100.0	0.0
9		1.461	9	344.	48.	87.8	12.2
10		21.081	10	199.	193.	50.8	49.2
11		17.476	11	97.	295.	24.7	75.3
12		2.737	12	234.	158.	59.7	40.3
13		4.735	13	245.	147.	62.5	37.5

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 1

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 7	-0-	90.	97.	187.
	-1-	43.	162.	205.
		133.	259.	392.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 7	-0-	23.0	24.7	47.7
	-1-	11.0	41.3	52.3
		33.9	66.1	100.0

PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 2

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		8.509
2		0.000
3		17.574
4		9.157
5		0.728
6		0.048
7		0.000
8		0.000
9		0.000
10		0.005
11		11.024
12		0.318
13		16.962

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -X0-	COMPOSITION -X1-
1	35.	660.	5.0	95.0
2	695.	0.	100.0	0.0
3	440.	255.	63.3	36.7
4	197.	498.	28.3	71.7
5	401.	294.	57.7	42.3
6	691.	4.	99.4	0.6
7	695.	0.	100.0	0.0
8	0.	695.	0.0	100.0
9	695.	0.	100.0	0.0
10	410.	285.	59.0	41.0
11	32.	663.	4.6	95.4
12	384.	311.	55.3	44.7
13	447.	248.	64.3	35.7

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 2

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 3	-0-	69.	371.	440.
	-1-	74.	181.	255.
		143.	552.	695.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 3	-0-	9.9	53.4	63.3
	-1-	10.6	26.0	36.7
		20.6	79.4	100.0

PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 3

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		0.010
2		0.000
3		0.000
4		0.000
5		0.020
6		2.323
7		1.608
8		0.185
9		1.888
10		0.000
11		0.000
12		0.058
13		0.444

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -X0-	COMPOSITION -X1-
1	8.	98.	7.5	92.5
2	0.	106.	0.0	100.0
3	0.	106.	0.0	100.0
4	0.	106.	0.0	100.0
5	40.	66.	37.7	62.3
6	102.	4.	96.2	3.8
7	39.	67.	36.8	63.2
8	86.	20.	81.1	18.9
9	88.	18.	83.0	17.0
10	106.	0.	100.0	0.0
11	106.	0.	100.0	0.0
12	63.	43.	59.4	40.6
13	71.	35.	67.0	33.0

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 3

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 6	-0-	38.	64.	102.
	-1-	0.	4.	4.
		38.	68.	106.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 6	-0-	35.8	60.4	96.2
	-1-	0.0	3.8	3.8
		35.8	64.2	100.0

CURRENT PAA BRANCH TERMINATES

THE FOLLOWING TEST CONDITION(S) HAVE NOT BEEN MET:

* THE OBSERVED CHI-SQ STATISTIC OF 2.323
WAS LESS THAN THE SPECIFIED MINIMUM OF 3.841

PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 4

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		8.881
2		0.000
3		0.000
4		0.000
5		1.039
6		9.572
7		7.760
8		0.835
9		0.370
10		3.254
11		0.000
12		0.554
13		0.622

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1	69.	473.	12.7	87.3
2	0.	542.	0.0	100.0
3	0.	542.	0.0	100.0
4	0.	542.	0.0	100.0
5	359.	183.	66.2	33.8
6	484.	58.	89.3	10.7
7	426.	116.	78.6	21.4
8	213.	329.	39.3	60.7
9	497.	45.	91.7	8.3
10	329.	213.	60.7	39.3
11	0.	542.	0.0	100.0
12	412.	130.	76.0	24.0
13	392.	150.	72.3	27.7

CONTINGENCY TABLE ANALYSIS LEVEL= 2 GROUP= 4

F R E Q U E N C I E S

PREDICTOR # 1	CRITERION #14 -0-	-1-	
-0-	54:	15:	69:
-1-	282:	191:	473:
	336.	206.	542.

P E R C E N T A G E S

PREDICTOR # 1	CRITERION #14 -0-	-1-	
-0-	10.0	2.8	12.7
-1-	52.0	35.2	87.3
	62.0	38.0	100.0

PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 1

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		0.237
2		0.000
3		0.076
4		2.667
5		2.209
6		0.237
7		0.000
8		0.000
9		1.080
10		2.543
11		4.533
12		1.286
13		1.191

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1	7.	180.	3.7	96.3
2	187.	0.	100.0	0.0
3	127.	60.	67.9	32.1
4	43.	144.	23.0	77.0
5	161.	26.	86.1	13.9
6	180.	7.	96.3	3.7
7	187.	0.	100.0	0.0
8	187.	0.	100.0	0.0
9	139.	48.	74.3	25.7
10	45.	142.	24.1	75.9
11	17.	170.	9.1	90.9
12	119.	68.	63.6	36.4
13	130.	57.	69.5	30.5

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 1

F R E Q U E N C I E S

PREDICTOR # 11	CRITERION #14 -0-	-1-	
-0-	4:	13:	17:
-1-	86.	84.	170.
	90.	97.	187.

P E R C E N T A G E S

PREDICTOR # 11	CRITERION #14 -0-	-1-	
-0-	2.1	7.0	9.1
-1-	46.0	44.9	90.9
	48.1	51.9	100.0

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 2

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		5.343
2		0.000
3		2.030
4		1.517
5		2.248
6		0.052
7		0.000
8		0.000
9		0.000
10		1.717
11		2.826
12		0.421
13		0.990

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -X0-	COMPOSITION -X1-
1	10.	195.	4.9	95.1
2	205.	0.	100.0	0.0
3	115.	90.	56.1	43.9
4	37.	168.	18.0	82.0
5	89.	116.	43.4	56.6
6	184.	21.	89.8	10.2
7	0.	205.	0.0	100.0
8	205.	0.	100.0	0.0
9	205.	0.	100.0	0.0
10	154.	51.	75.1	24.9
11	80.	125.	39.0	61.0
12	115.	90.	56.1	43.9
13	115.	90.	56.1	43.9

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 2

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 1	-0-	5.	5.	10.
	-1-	38.	157.	195.
		43.	162.	205.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 1	-0-	2.4	2.4	4.9
	-1-	18.5	76.6	95.1
		21.0	79.0	100.0

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 3

INDEP VAR #	STATUS	CHI-SQ STATISTIC
1		3.995
2		0.000
3		0.000
4		1.664
5		0.293
6		0.265
7		0.000
8		0.000
9		0.000
10		0.892
11		4.389
12		0.006
13		8.237

INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -X0-	COMPOSITION -X1-
1	23.	417.	5.2	94.8
2	440.	0.	100.0	0.0
3	440.	0.	100.0	0.0
4	197.	243.	44.8	55.2
5	249.	191.	56.6	43.4
6	436.	4.	99.1	0.9
7	440.	0.	100.0	0.0
8	0.	440.	0.0	100.0
9	440.	0.	100.0	0.0
10	246.	194.	55.9	44.1
11	14.	426.	3.2	96.8
12	215.	225.	48.9	51.1
13	270.	170.	61.4	38.6

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 3

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 13	-0-	53.	217.	270.
	-1-	16.	154.	170.
		69.	371.	440.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 13	-0-	12.0	49.3	61.4
	-1-	3.6	35.0	38.6
		15.7	84.3	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

PROB80 : POPULATION ANALYSIS

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 4

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		5.253	1	12.	243.	4.7	95.3
2		0.000	2	255.	0.	100.0	0.0
3		0.000	3	0.	255.	0.0	100.0
4		0.000	4	0.	255.	0.0	100.0
5		2.763	5	152.	103.	59.6	40.4
6		0.000	6	255.	0.	100.0	0.0
7		0.000	7	255.	0.	100.0	0.0
8		0.000	8	0.	255.	0.0	100.0
9		0.000	9	255.	0.	100.0	0.0
10		0.164	10	164.	91.	64.3	35.7
11		4.139	11	18.	237.	7.1	92.9
12		0.093	12	169.	86.	66.3	33.7
13		6.686	13	177.	78.	69.4	30.6

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 4

FREQUENCIES

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 13	-0-	60.	117.	177.
	-1-	14.	64.	78.
		74.	181.	255.

PERCENTAGES

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 13	-0-	23.5	45.9	69.4
	-1-	5.5	25.1	30.6
		29.0	71.0	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

PROB80 : POPULATION ANALYSIS

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 7

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		0.000	1	69.	0.	100.0	0.0
2		0.000	2	0.	69.	0.0	100.0
3		0.000	3	0.	69.	0.0	100.0
4		0.000	4	0.	69.	0.0	100.0
5		1.617	5	51.	18.	73.9	26.1
6		4.638	6	64.	5.	92.8	7.2
7		6.957	7	60.	9.	87.0	13.0
8		3.041	8	16.	53.	23.2	76.8
9		0.254	9	62.	7.	89.9	10.1
10		0.176	10	49.	20.	71.0	29.0
11		0.000	11	0.	69.	0.0	100.0
12		0.828	12	48.	21.	69.6	30.4
13		4.066	13	43.	26.	62.3	37.7

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 7

FREQUENCIES

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 7	-0-	50.	10.	60.
	-1-	4.	5.	9.
		54.	15.	69.

PERCENTAGES

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 7	-0-	72.5	14.5	87.0
	-1-	5.8	7.2	13.0
		78.3	21.7	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 8

INDEP VAR #	STATUS	CHI-SQ STATISTIC	INDEP VAR #	SUBGROUP -0-	SUBGROUP -1-	COMPOSITION -%0-	COMPOSITION -%1-
1		0.000	1	0.	473.	0.0	100.0
2		0.000	2	0.	473.	0.0	100.0
3		0.000	3	0.	473.	0.0	100.0
4		0.000	4	0.	473.	0.0	100.0
5		1.570	5	308.	165.	65.1	34.9
6		3.843	6	420.	53.	88.8	11.2
7		3.879	7	366.	107.	77.4	22.6
8		0.007	8	197.	276.	41.6	58.4
9		0.838	9	435.	38.	92.0	8.0
10		5.178	10	280.	193.	59.2	40.8
11		0.000	11	0.	473.	0.0	100.0
12		0.441	12	364.	109.	77.0	23.0
13		0.169	13	349.	124.	73.8	26.2

CONTINGENCY TABLE ANALYSIS LEVEL= 3 GROUP= 8

F R E Q U E N C I E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 10	-0-	155.	125.	280.
	-1-	127.	66.	193.
		282.	191.	473.

P E R C E N T A G E S

		CRITERION #14 -0-	CRITERION #14 -1-	
PREDICTOR # 10	-0-	32.8	26.4	59.2
	-1-	26.8	14.0	40.8
		59.6	40.4	100.0

<PREDICTIVE ATTRIBUTE ANALYSIS> <VERS:84.02><NYS*DCJS>

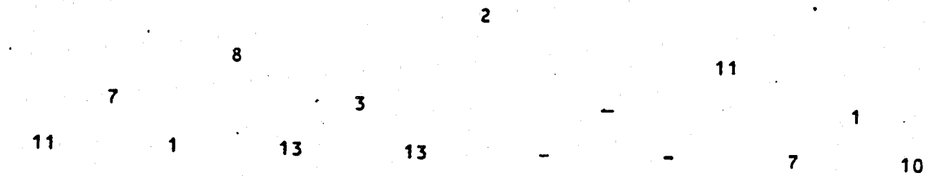
****

P R O B 8 0 : P O P U L A T I O N A N A L Y S I S

REGION=NYC-SUBS STATISTIC=CHISQ RUNDATE=1/29/84

****

P A A P R O C E S S I N G T R E E F O R P R E D I C T O R S



END