

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY

Research Support in Operations Research/Systems
Analysis Applications to Army Needs and Objectives

DAAD05-74-C-0777

AN APPLICATION OF MULTIVARIATE STATISTICAL
METHODS IN DEVELOPING OPERATIONAL USAGE
PATTERNS OF U.S. ARMY VEHICLES

Leslie G. Callahan Jr.

Project Director

Randall Brannon Medlock

Principal Investigator

Georgia Institute of Technology

April 1975

This report is based on Captain Randall Brannon Medlock's thesis, submitted in partial fulfillment of the requirements of the Master of Science in Industrial Engineering degree at the School of Industrial and Systems Engineering, Georgia Institute of Technology. The thesis committee was:

Dr. Leslie G. Callahan Jr., Chairman

Dr. Russell G. Heikes

Dr. Douglas C. Montgomery

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY

Research Support in Operations Research/Systems
Analysis Applications to Army Needs and Objectives

DAAD05-74-C-0777

AN APPLICATION OF MULTIVARIATE STATISTICAL
METHODS IN DEVELOPING OPERATIONAL USAGE
PATTERNS OF U.S. ARMY VEHICLES

Leslie G. Callahan Jr.

Project Director

Randall Brannon Medlock

Principal Investigator

Georgia Institute of Technology

April 1975

ACKNOWLEDGMENTS

The author wishes to acknowledge the many individuals who have assisted in the preparation of this thesis.

Mr. Thomas Burnette is acknowledged for the original idea and initial conjectures that developed into the initial concept. In addition Mr. Burnette has been a continual friend and assistant in sounding out ideas and in preparation of computer programs. Mr. Frank Alt, instructor in the department, has my utmost appreciation for his helpful suggestions and the loan of his extensive library in multivariate analysis. Mrs. Mary Burkes of TRADOC was the main source of information and instruction in the preparation of Chapter IV and my thanks go to her for her time and effort.

Dr. L. G. Callahan, my thesis advisor, has been the main source of support and encouragement during the entire preparation. The remainder of the reading committee comprising of Dr. Montgomery and Dr. G. Thompson, have provided strong support and many constructive comments which have added considerably to the thesis. My special thanks go to Dr. Heikes who at the last moment agreed to sit as proxy for Dr. Montgomery at my oral defense. His directive comments were highly valued and added to the final product.

The author also graciously acknowledges the financial

and extensive assistance provided by AMSAA and Mr. Wormert.

And finally to my wife, Marya, and my two boys, Randy, Jr. and Robert, I owe the most for the sacrifices they made in order that this thesis could be completed. Marya, in addition, provided the typing expertise which produced the final rough drafts.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS.	ii
LIST OF TABLES	vi
LIST OF ILLUSTRATIONS.	vii
SUMMARY.	viii
Chapter	
I. INTRODUCTION.	1
Background Leading to this Research	
Purpose of this Thesis	
General Approach and Overview	
II. TECHNIQUES OF MULTIVARIATE ANALYSIS	7
General Discussion	
Explanation of Techniques	
Major Literature in Multivariate Analysis	
III. APPLICATION OF MV TECHNIQUES TO OBTAIN OPERATIONAL USAGE PATTERNS (METHODOLOGY).	44
Introduction	
Selection of a MV Technique	
Preparation of Data	
Selection of Distance Measure	
Application of the M.V. Technique	
IV. PROJECTED UTILIZATION WITHIN THE TOE SYSTEM	71
The Present System for the Development of TOE	
Method for Review and Change of TOE	
The Implementation of the Proposed Methodology Within the TOE System	
V. CONCLUSIONS AND RECOMMENDATIONS	83
Limitations of the Research	
Conclusions	
Recommendations	

	Page
APPENDIX A.	87
APPENDIX B.	89
APPENDIX C.	100
BIBLIOGRAPHY.	105

LIST OF TABLES

Table		Page
1.	Cluster Groupings.	63
2.	A Cluster Group.	64
3.	Distance Columns from Table 6.	65
4.	Splitting Cluster.	66
5.	Operational Usage Patterns	70
6.	Comparison of Results.	101
7.	Validation Results	103

LIST OF ILLUSTRATIONS

Figure	Page
1. Sums of Squares Matrix.	29
2. Partitioned Matrix.	30
3. Unit Ball	33
4. Connected Tree.	36
5. Cluster Shapes.	60
6. TOE Development Cycle	72
7. TOE Management Model.	73
8. Daily Trip Data Report.	87

SUMMARY

This research develops a methodology which establishes operational usage patterns of Army vehicles using field data. The field data for this research was material supplied by the U. S. Army Materiel Systems Analysis Agency. An examination of the data revealed that a number of correlated variables could be extracted. These variables described the manner in which the particular type vehicle was utilized. Since it would be ideal to examine these variables in their entirety, multivariate techniques were considered. These techniques included principal component analysis, factor analysis, discriminant analysis, canonical correlation analysis, and cluster analysis. Each of these techniques were examined in some detail to determine its suitability for producing operational usage patterns. The cluster analysis technique was chosen based on its simplicity, low cost and the ability to provide meaningful groupings of data units. A nonhierarchical clustering technique known as McQueen's convergent K-means method was selected as the most appropriate method for this case.

The data was subjected to outlier analysis techniques to eliminate multivariate outliers. In addition the data was centered by subtracting the means and standardized by dividing by the standard deviations.

It was hypothesized that the clustering should reveal between four and ten "natural" clusters. Consequently, the analysis was accomplished to produce partitionings that included five to nine clusters. The "optimum" or "best" partition was chosen based on two criteria, one of which compared cluster centroids between partitions and between different methods of selecting initial starting points (i.e. seed points) to determine at which partition clusters become most stable. The other criterion established an upper limit for the number of clusters and was based on nonoptimal splitting of stable clusters. If nonoptimal splitting occurred then it was an indication that the optimal partition existed at a smaller number of clusters. Once the "optimum" or "best" partition is determined then in fact the operational usage patterns have been established.

It is envisaged that this methodology can be effectively utilized in the area of assigning and reassigning vehicles within the Army TOE system. In particular the technique would establish a baseline of usage patterns which describes how a new vehicle is being utilized. This baseline can then be used to periodically identify misassigned vehicles. This is done by first establishing new usage patterns based on field data collected at different times during the vehicles inventory life. A comparison of these new usage patterns with the base would then identify possible outlier clusters which would indicate the possibility of

misassigned vehicles.

It is recommended that this methodology be implemented by the Army in a limited case to determine its feasibility in application.

CHAPTER I

INTRODUCTION

Background Leading to this Research

The performance characteristics and requirements, and the mission of U. S. Army Vehicles are initially determined and detailed in the Required Operational Capabilities (ROC). Although these documents categorize the expected and intended use of the vehicle, they in no way can project the actual use of the vehicles. Actual input to these documents has been determined by various studies. One of the earliest of these was the Motor Vehicle Requirements, Army in the Field, 1965-1970 (MOVER), October 1960 [65]. The requirements presented by this study were based on an analysis of the functions to be performed by wheeled vehicles, the operational environment and the concept of operations. An additional study Tactical Mobility of Land Forces, 1971-1980 (U), January, 1965 [87], analyzed vehicle characteristics and performance required and conclusions were drawn to initiate certain improvements. The performance requirements described in the study resulted from an analysis of the operational environment and the concept of operations. Other studies [73], [76], [87] were also conducted which in essence added to or supported the above studies. Analysis

of these studies provided a basis for the REVAL-WHEELS study [77]. This landmark study identified in detail the tasks which make up a combat action and described those tasks in terms of the environment. Tasks were categorized by functional area and were oriented more toward a narrative of what the vehicle did rather than toward a specific set of task parameters. In order to provide a consistent means of recording information required for task definition and analysis and for collecting information related to individual tasks, a worksheet was prepared which listed each task parameter. A worksheet was completed for the task performed by each vehicle and trailer authorized in the force structure. This information was the input that resulted in the REVAL-WHEELS data bank [78]. This study was one of the first which attempted to gain feedback on how vehicles are actually utilized. One of the major drawbacks to the method was that it was generally based on the opinions and experience of the commander and not on actual vehicle performance and usage. A final study of vehicle performance was conducted by the Family of Army Vehicles Study (FAVS) [49] completed in 1971 and was designed to support recommendations for a family of tactical vehicles for the future Army. This was a limited study which collected data on cargo dimensions and on relative time spent in each of three theater zones, i.e., Division, Corp/Army and COMMZ.

The REVAL-WHEELS and the FAVS have developed a type

of "mission profile" for Army vehicles, but are open to criticism in that they are based on opinion and belief rather than factual data of utilization. The Army Material System Analysis Agency (AMSAA) has recognized this need for factual "mission profiles" of the tactical truck in order to properly assess effectiveness [1].

In January, 1972, the Department of the Army's Project WHEELS Study Group was established and directed to conduct a comprehensive analysis of the Army's wheeled vehicle program to include the Army's need for, and managerial concepts and utilization of wheeled vehicles. AMSAA was in the process of establishing a mission profile for the 5 ton truck fleet organic to an Armored Division when, in March, 1972, Project WHEELS requested that AMSAA provide a "mission profile" for 5 ton cargo and 5 ton tractor and semitrailer trucks operating in logistic support of forward divisions. AMSAA developed a technique for generating mission analyses of logistics vehicles [29]. The procedure involves a map study wherein a semi-war game is played using a scenario of an actual situation. The play of the game is modified to concentrate on the movement of the truck type being evaluated. This technique has been used to generate mission profiles for the 5 ton truck, the 5 ton tractor, the Heavy Equipment Transporter (HET), and the 1-1/4 ton limited mobility truck. Emanuel [29] reports on AMSAA's mission analysis of 1-1/4 ton truck where the validity of the semi-war gaming technique

is evaluated by comparing results with FAVS and REVAL-WHEELS data banks. The report concluded that the validity was of an acceptable nature. The results of the AMSAA war game was not found to be in conflict with mission analyses derived from the previously mentioned data banks. A further conclusion was that mission analysis is generally "based solely on military judgment and experience." A final conclusion of the study was that "real life data from actual field operations is badly needed to validate these judgments"[29].

In response to the concluded need for vehicle utilization data from actual field operations, AMSAA developed a data collection form (Appendix A). This form was distributed during REFORGER operations in Germany during October, 1973, to collect data on the 1-1/4 ton and 3/4 ton trucks. In October, 1974, data was again collected during REFORGER operations but on this occasion the 1/4 ton truck was targeted. The 1-1/4 ton and 3/4 ton truck data was made available for this research.

Purpose of this Research

The purpose of this research is to develop a methodology which will establish operational usage patterns of U. S. Army vehicles. The term operational usage patterns is used instead of mission profile since actual operational data is available for the research and it is believed that

several usage patterns will result rather than a single mission profile.

General Approach and Overview

A review of the data indicated that six good variables could be extracted for evaluation. These included amount of travel on hard surface roads, on trails or unimproved roads, and cross country plus the load by weight, number of personnel, and total miles traveled. Additionally, the mission type and unit designation were identified. This is clearly identified as multivariate data when the variables are displayed in vector form (i.e. $X_n = (X_1, X_2, X_3, \dots, X_p)$ where each X_p is a random variable and X_n is a random vector). In each of the above mentioned studies, data of this nature is evaluated one variable at a time by univariate analysis. It is the intent of this research to evaluate the data in its complete form using multivariate statistical analysis. The ultimate result would be a classification or clustering scheme to identify operational usage patterns. A brief discussion of multivariate statistical analysis literature and techniques will comprise Chapter II.

The actual evaluation of multivariate data would be impossible without the modern day computer with its large memory capacity and the ability to make rapid calculations. As a consequence a large portion of this research utilizes the computer with its flexible capabilities. The data must

be processed by computer, and programs to accomplish this have been selected from many different sources. Chapter III will include a brief discussion of each program and the techniques each one uses to accomplish its specific purpose. The complete methodology which produces the final result is also detailed in Chapter III.

Chapter IV discusses the projected utilization for this research. This research will allow vehicles to be more effectively assigned and reassigned within the Army's Tables of Organization and Equipment (T.O.E.). A brief discussion of the present system of assigning and reassigning will be presented and then finally a comparison with the proposed utilization will establish improved effectiveness.

A discussion of results will be made in Chapter III emphasizing the need for validity. The final conclusions and recommendations will be presented in Chapter V.

CHAPTER II

TECHNIQUES OF MULTIVARIATE STATISTICAL ANALYSIS

General Discussion

Tatsuoka [90] describes multivariate statistical analysis, or multivariate analysis for short, "as that branch of statistics which is devoted to the study of multivariate (or multidimensional) distributions and samples from those distributions." This is how he believes the mathematical statistician would characterize this discipline. For the applied statistician and researcher who uses statistics as a tool this definition would not be adequate. Press [72] gives a more applied characterization by stating that multivariate analysis is "that branch of statistics that is devoted to the study of random variables which are correlated with one another." These random variables are studied to determine the interrelation, similarity, or association between them. If some type of correlation can be determined, then the behavior of one provides some knowledge about the behavior of the other. This can result in drawing inferences relevant to these variables concerning the populations from which the samples were selected. In order to accomplish the study of multivariate or multidimensional data, considerable study has produced several techniques and

models which simplify the interpretation of this data. Thus, the field involves a collection of tools, techniques, and methods of thinking which can be applied to the immense task of simultaneously handling and interpreting many related random variables.

Multivariate techniques start from a multivariable data matrix. This data matrix normally results from N observations (units, cases, entities) on n variables (attributes, characteristics, measurements) simultaneously. The techniques normally analyze the variables and less often they are used to analyze the observations. What operations are performed depend on the specific model that is being utilized and on the type results desired.

The orientation of the data matrix has been flexible depending on which field of study is describing it. For the purposes of this paper the variables will be represented by the columns of the matrix and the observations will be represented by the rows of the matrix.

The theoretical input to the multivariate field of knowledge is widely diverse and emerges from various dispersed sources. The models of multivariate analysis were developed as a result of real problems in many different disciplines. Historical accounts indicate that Galton developed the concept of "correlation" in the late 19th century. Probably the most crucial article was Karl Pearson's [71] 1901 paper which set forth "the method of principal

axes." This article provides a basis for two multivariate techniques, factor analysis and principal component analysis. Charles Spearman [84] is generally ascribed with the first development of factor analysis. Harold Hotelling [45,47] extended Pearson's original study to basically the technique of principal component analysis that is used today. Godfrey Thomson [91,92] and Cyril Burt [18] extended the development of the principles of factor analysis while Karl Holzinger [44] and Leon Thurstone [93] developed the more precise techniques. Thurstone is generally credited with the term "Factor Analysis."

Hotelling [46,48] provided another multivariate technique when he developed canonical correlation. This technique establishes relations between two sets of variables. The development of discriminant analysis is attributed to R. A. Fisher [32,33,34] with more recent work done by E. Fix and J. L. Hodges [35], and M. G. Kendall [55]. Finally, cluster analysis has developed as an outgrowth of techniques of classification used by biologists, zoologists, and botanist. This technique might best be catalogued as a data analysis technique as opposed to a statistical analysis technique. It could be said that cluster analysis is a descriptive technique such as the mean, variance, or range of a set of data.

The origin of cluster analysis is vague even though Tryon [100] takes credit for its development. Admittedly

Tryon is a pioneer in the area of clustering variables but few of the later works credit him with initiating essential contributions to clustering data units.

A closer look at each of these techniques will determine their applicability to the solution of the problem.

Explanation of Techniques

In this section an attempt will be made to keep the explanation in expositive terms and relying on mathematical terms only when necessary. Verbal descriptions of complex methods without the use of exact mathematical terms tend to be misleading if taken verbatim. In order to avoid this pitfall some discussion of the mathematics will be included. The reader is referred to the last section of this chapter for literature which gives complete and rigorous development of each technique.

We will begin our discussion by examining the principal component technique developed by Hotelling [45]. The discussion will be restricted to a multivariate sample approach since the procedures developed in this thesis are based on such a sample. In general, the technique involves a set of n random variables (X_1, \dots, X_n) which is transformed linearly and orthogonally into an equal number of new variables (Z_1, \dots, Z_n) which are uncorrelated (orthogonal). These are developed such that Z_1 has maximum variance and

Z_2 has maximum variance subject to being uncorrelated with Z_1 . This is completed for all variables insuring that each variable is uncorrelated with each other variable. The technique can utilize either the covariance matrix (S) or the product moment correlation matrix (R). The transformation is obtained by finding the eigenvalues (latent roots) λ_j and eigenvectors (latent vectors) α_j by solving the equation $(\tilde{S} - \lambda\tilde{I})\alpha=0$ or $(\tilde{R} - \lambda\tilde{I})\alpha=0$. A different set of latent roots and latent vectors will be obtained depending on the use of either the covariance or the correlation matrix.

The first principal component of the observations X is the linear combination

$$\begin{aligned} Z_1 &= \alpha_{11}X_1 + \dots + \alpha_{n1}X_n \\ &= \alpha_1 \hat{X} \end{aligned}$$

where α_1 are the elements of the latent vector associated with the largest latent root λ_1 . The latent root λ_1 is interpretable as the sample variance of Z_1 the linear combination

$$\begin{aligned} Z_2 &= \alpha_{12}X_1 + \dots + \alpha_{n2}X_n \\ &= \alpha_2 \hat{X} \end{aligned}$$

where α_2 is the latent vector corresponding to the second largest latent root and α_1 and α_2 are orthogonal (uncorrelated). This insures that the variance of Z_2 is a maximum and Z_2 is uncorrelated with Z_1 . This is accomplished for each of j principal components such that the sum of the latent roots equals the total sample variance (i.e. $\lambda_1 + \dots + \lambda_n = \text{trace } \underline{S}$).

This technique can allow the system to be described more parsimoniously by taking the first major components to describe the complete system (i.e. those which account for the majority of the sample variance). The technique, however, is scale dependent and therefore not invariant under changes in scales. Furthermore, there is no provision for variance that is attributable only to the unreliability or sampling variation of the observations. Finally, all the components are required to reproduce the correlations among the variables.

Factor analysis is a technique for the reduction of the number of dimensions of a body of data (parsimony) so that a maximum of the correlation is reproduced. The factor analysis thereby overcomes some of the shortcomings of the previous technique. In order to accomplish parsimony, factor analysis utilizes the correlation coefficients for a specific set of variables to determine some underlying pattern of relationships which might exist in the data. This is done such that the data may be rearranged or reduced

to a smaller set of factors that will account for the observed interrelations in the data. The general field of factor analysis provides many specific procedures but each normally includes four customary steps. Given a data matrix these steps include (1) the computation of the correlation matrix, (2) the extraction of the unrotated factors, (3) the rotation of the factors, and (4) the interpretation of the rotated factors. The differences in procedures stem from the various methods of extraction and rotation.

The principal (also referred to as common or classical) factor technique is obtained from the principal component by replacing the main diagonal of the correlation matrix with estimates of communality (h^2). Communality describes that proportion of a variable which shares something in common with the other variables. The remainder of the correlation diagonal ($1-h^2$) is that proportion that uniquely defines that particular variable. Common estimates of communality are (1) the squared multiple correlation between a variable and the rest of the variables and (2) the absolute value of the largest element in each column of the correlation matrix. This matrix is then referred to as the reduced correlation matrix. The reduced matrix is then rotated, as in the principal components case, to produce the principal (common) factors. The principal factors represent only that proportion of the variance which is

described by the communality. Consequently, we are assuming the existence of a unique factor not involved with the other variables. The principal factor model is similar to the principal component linear compound model and is as follows:

$$Z_i = a_{1i}F_1 + \dots + a_{mi}F_m + d_i U_i \quad (i = 1, 2, \dots, n)$$

where F_m are the common factors and U_i is the unique factor and a_{mi} are the factor weights or "loadings." The unique factor (U_i) can be decomposed into two separate entities such that

$$d_i U_i = b_i S_i + e_i E_i$$

where S_i is the uniqueness or specificity factor and E_i is the error factor. Since the specific and error factors are uncorrelated, the following relationship exists,

$$d_i^2 = b_i^2 + e_i^2$$

Therefore, the total variance and the communality are expressed in the following way,

$$S_i^2 = h_i^2 + b_i^2 + e_i^2 = 1$$

and

$$h_i^2 = 1 - d_i^2$$

In actual practice an iterative scheme is employed to obtain a best factoring based on an improving communality. First, the number of factors is estimated using principal components (i.e. normal correlation matrix). The main diagonal elements of the correlation matrix are then replaced with initial estimates of the communality. The factors are then extracted using this reduced matrix, and the variances accounted for by these factors become new communality estimates. The matrix diagonal elements are replaced by these new estimates. This process continues until the differences in two successive communality estimates are negligible.

Several variations of this "classical" technique have been developed. The centroid factor method was developed before the computer age in order to ease the strain of hand computation. As it is more of a historical method it is not discussed further. Those who are interested are referred to Harman [43] and Comrey [24].

One major refinement of the classical model is Lawley's [60] maximum likelihood factor method which provided a statistical basis for judging the adequacy of the model. In essence this method provides for the maximum likelihood estimation of the factor loadings based on an assumption of a given number (m) of common factors from a

sample of observations of n variables. In order to utilize this type estimation, it must be assumed that the unique factors are mutually independent and independent of the common factors. In addition all factors are assumed to be normally distributed with zero means. An additional restriction on the common factors is that they must have unit variances. The unique factors are permitted to be heteroscedastic (differing variances). These assumptions imply that the observed data must have a multivariate normal distribution.

Other refinements are the Rao canonical factor, the alpha factor, and the image factor methods. They differ mainly in how each method estimates the communality. The canonical factoring method is an extension of the maximum likelihood method. Hence, the correlation matrix is based on a sample of cases and thereby permits test of significance to be applied. This method enables the common factors of the population to be determined such that they have maximum canonical correlation with the sample data, hence the name. In contrast the alpha factoring assumes a sample of variables from a population of total variables and produces the reduced correlation matrix based on the communality while canonical factoring is based on an estimate of the unique variance. These techniques are also iterative and continue until the communalities converge. The name alpha comes from the fact that the factors are defined in terms of maximum

"generalizability" which is a measure known as Cronbach's alpha. Only those factors which indicate some generalizability to other variables in the total population are retained.

The image factoring differs from the common method only in that it provides a different approximation of communality. This is based on the best estimate of the common part of a variable being given by the image of variable j , denoted by P_j where:

$$P_j = \sum_{k=1}^n a_{jk} Z_k \quad k = 1, 2, \dots, n-1$$

where a_{jk} are the standardized coefficients for predicting variable j from the rest of the variables. The best estimate of the unique portion of a variable is the anti-image (u_j) which is given by

$$u_j = z_j - P_j$$

A more complete explanation of this estimate of communality would be lengthy and the reader is referred to Rummel [79].

Other factor methods such as minres, and multiple factor are too involved to be discussed here and the reader is directed to Comrey [24] and Harman [43] for detailed discussions.

Until this point only orthogonal (rigid) rotations

have been mentioned due to the assumption that the common factors have unit variances and zero correlations. In some situations there should be no reason to assume that the factors are orthogonal. In these situations there would be some correlation between factors dictated by knowledge of the circumstances. Consequently, the best fit of factors would follow some oblique rotation. This would allow a relaxation of the assumption of noncorrelation.

The factor rotation presents an additional problem in that it could lead to multiplicities of solutions. By choosing different orthogonal transformations an infinity of factor loading matrices can be computed which would lead to the same covariance matrix. One method to obtain a unique solution is to add to our list of assumptions that the latent roots associated with the common factors are distinct. This technique could force a structure that would not give the best factor solution. Various other procedures have been proposed for eliminating the ambiguity due to rotation. Thurstone [94] proposed the concept of "simple structure" as a means of selecting the most meaningful loadings. The five criterion he set up places certain restrictions on the factor matrix. The ultimate goals of any rotation is to obtain some meaningful factors and if possible, the simplest factor structure. To gain the simplest structure we make as many row and column values of the factor matrix as close to zero as possible.

Several alternatives are available for specific rotational methods which strive for a simple type structure. For orthogonal rotations the quartimax and the varimax methods have been developed. The quartimax method rotates the initial factors such that a variable loads high on one factor, but almost zero on all others. This simply means that the cross-product of factor loadings are minimized for each variable, i.e.

$$\text{minimize } \sum_{p < q = 1}^m \sum_{i = 1}^n (a_{ip} a_{iq})^2$$

where $p < q$ and both are common factors. A problem exists, however, when one of the factor loadings is zero. This will result in the cross product being zero. In order to get around this problem, it is noted that the communalities remain constant under orthogonal rotation and, consequently, the amount of variance accounted for by the orthogonal solution will remain constant. Therefore, the square of the communalities will also remain constant.

$$\sum_{i=1}^n h_i^2 = \sum_{i=1}^n \sum_{p=1}^m a_{ip}^2 = \text{constant}$$

and

$$\left(\sum_{p=1}^m \sum_{j=i}^n a_{ip}^2 \right)^2 = \sum_{p=1}^m \sum_{i=1}^n a_{ip}^4 + 2 \sum_p \sum_{q=1}^m \sum_{i=1}^n a_{ip}^2 a_{iq}^2 = \text{constant}$$

If we want to minimize the cross product terms in this equation, we need only maximize $\sum_{p=1}^m \sum_{i=1}^n a_{ip}^4$ which is the quartimax method.

The varimax method developed by Kaiser [54] focuses on simplifying the columns of the factor matrix as opposed to the quartimax which simplifies the rows of the factor matrix. In addition, varimax defines the simple factor as one with only ones and zeros in the column. This simplification is equivalent to maximizing the variance of the squared factor loadings in each column. The objective function is,

$$\text{maximize } n \sum_{p=1}^m \sum_{i=1}^n \left(\frac{a_{ip}}{h_i} \right)^4 - \sum_{p=1}^m \left(\sum_{i=1}^n \frac{a_{ip}^2}{h_i^2} \right)^2.$$

In order to gain simplification under oblique rotation another method has evolved. Again the idea is to minimize the cross products of the factor loadings on some reference axes fitted from the oblique rotation. A more direct solution involves the minimization of the following criterion,

$$\sum_{p < q=1}^m \left(\sum_{i=1}^n a_{ip}^2 a_{iq}^2 - \frac{\delta}{n} \sum_{i=1}^n a_{ip}^2 \sum_{i=1}^n a_{iq}^2 \right)$$

where δ is an arbitrary value which controls the obliqueness

of the factor rotation.

Up to this point we have only discussed what is termed R-technique factor analysis (i.e. factor analysis of n variables over N observations). In this method correlation is computed by taking a pair of columns (variables) and determining cross-product terms. If instead of correlating two data variables over the sample of data units two data units are correlated over the sample of data variables, we then have Q-technique factor analysis. This technique is also termed "inverse" factor analysis and is the most commonly considered alternative to the R-technique. There are several problems with this technique but if the sample is large and is standardized, many of these problems are overcome. The reader is directed to Comrey [24] for details of the technique and to Fleiss and Zubin [36] for a detailed discussion of problems associated with the Q-technique.

If we observe the variables at specific intervals of time (occasions) and factor the variables accordingly, the technique is termed P-factor analysis. This correlates pairs of variables over the data occasions. The O-technique reverses the procedure and correlates pairs of occasions over the data variables.

In addition there are the T-technique and the S-technique. The T-technique correlates pairs of occasions over a sample of individuals for a given variable. The

S-technique correlates pairs of individuals over the occasions for a given variable. The interested reader is referred to Cattell [20] or Rummel [79] for a more complete discussion of all methods of factors analysis.

Discriminant Analysis is another multivariate technique which deals with the linear function. The discriminant function is similar to that derived for the principal components technique. The principal components search is for parsimony however the discriminant search is normally for classification of individuals.

The standard classification procedure for n variables assumes that the observations come from one of two multivariate normal populations. The two populations are denoted W_1 and W_2 and are assumed normal with mean $\mu_i^{n \times 1}$ and covariance matrix $\Sigma_i^{n \times n}$ ($i=1$ or 2). In addition it is assumed that the covariance matrices are equal.

In the development of the procedure it is assumed that the parameters are known. A discriminant function, as defined by Fisher [32] is given by

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n,$$

where the α_i 's are constants, and X is classified into W_1 if $Z \geq C$ and into W_2 if $Z < C$ where C is also a constant. The next step is to determine the α_j 's and C which minimize the probabilities of making a misclassification.

If the vector \tilde{X} is from W_1 , then Z is normal with mean

$$\xi_1 = \sum_{j=1}^n \alpha_j \mu_{1j}$$

and variance

$$\sigma_z^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \sigma_{ij} \alpha_j$$

Similarly, if \tilde{X} is from W_2 , then Z is normal with mean

$$\xi_2 = \sum_{j=1}^n \alpha_j \mu_{2j}$$

and the same variance σ_z^2 .

The α_j 's should then be chosen to maximize the distance between ξ_1 and ξ_2 relative to σ_z^2 . Mahalanobis [62] proposed a measure of distance between two populations which is used to define the above distance,

$$\Delta^2 = \frac{(\xi_1 - \xi_2)^2}{\sigma_z^2}$$

Therefore, the α_j 's are chosen which maximize Δ^2 . According to Fisher [32] the α_j 's are solutions to the equations

$$\alpha_1 \sigma_{11} + \alpha_2 \sigma_{12} + \dots + \alpha_n \sigma_{1n} = \mu_{11} - \mu_{21},$$

$$\begin{aligned} \alpha_1 \sigma_{21} + \alpha_2 \sigma_{22} + \dots + \alpha_n \sigma_{2n} &= \mu_{12} - \mu_{22}, \\ &\vdots \\ \alpha_1 \sigma_{n1} + \alpha_2 \sigma_{n2} + \dots + \alpha_n \sigma_{nn} &= \mu_{1n} - \mu_{2n}. \end{aligned}$$

The α_j 's are then used to establish the discriminant functions.

The constant C is determined as that value which minimized the sum of the probabilities of misclassifying. This value is achieved by choosing C halfway between the two means (average),

$$C = \frac{\xi_1 + \xi_2}{2}.$$

In summary we classify any observation vector \tilde{X} into W_1 if the value of the discriminant function (Z) evaluated for this vector, is greater than or equal to the constant C ; otherwise \tilde{X} is classified into W_2 .

A Bayes classification procedure is more theoretical and consists of classifying \tilde{X} into W_1 if

$$\Pr (W_1 | \tilde{X}) \geq \Pr (W_2 | \tilde{X})$$

and classifying \tilde{X} into W_2 if

$$\Pr (W_1|\tilde{X}) < \Pr (W_2|\tilde{X})$$

where $\Pr (W_1|\tilde{X})$ and $\Pr (W_2|\tilde{X})$ are the posterior probabilities of classification which are given by the Bayes theorem

$$\Pr(W_i|\mathbf{x}) = \frac{q_i f_i(\mathbf{x})}{q_1 f_1(\tilde{X}) + q_2 f_2(\tilde{X})}$$

where q_i are the prior probabilities of classification and $f_i(\mathbf{x})$ are the density functions.

If the Bayes theorem substitution is made the procedure reduces to classifying \tilde{X} into W_1 if

$$\frac{q_1 f_1(\tilde{X})}{q_2 f_2(\tilde{X})} \geq 1$$

and into W_2 if

$$\frac{q_1 f_1(\tilde{X})}{q_2 f_2(\tilde{X})} < 1.$$

By substituting in the density functions and taking logs it can be shown that this is equivalent to classify \tilde{X} into W_1 if

$$\sum_{i=1}^n \alpha_i x_i \geq \frac{\xi_1 + \xi_2}{2} + \ln\left(\frac{q_2}{q_1}\right)$$

and into W_2 if

$$\sum_{i=1}^n \alpha_i x_i < \frac{\xi_1 + \xi_2}{2} + \ln \left(\frac{q_2}{q_1} \right)$$

Rao [75] has shown that the solution to this minimizes the expected probability of misclassification,

$$q_1 \Pr(2|1) + q_2 \Pr(1|2).$$

So far only classification into two normal populations with known parameters has been considered. The above criteria can be generalized to classification into one of k arbitrary or normal populations. The assumption of known parameters is only a theoretical simplification and in most applications independent random samples are available from the populations from which estimates of parameters may be made. For a generalization of the Bayesian approach to classification into one of k populations, the reader is referred to Afifi and Azen [1].

The discriminant criterion that has been discussed is based on Mahalanobis distance Δ^2 (or D^2 if population samples are available). This criterion selects the α_j 's such that D^2 is a maximum. This is basically equivalent to minimizing the sample pooled covariance (dispersion) matrix S since $D^2 = \frac{(\bar{Z}_1 - \bar{Z}_2)^2}{S_z^2}$ where $S_z^2 = \sum_{j=1}^p \sum_{m=1}^p a_j s_{jm} a_m$.

A similar approach is based on the sums of squares (scatter) matrices, i.e. $\tilde{T} = \tilde{W} + \tilde{B}$ where \tilde{T} is the total sums of squares, \tilde{W} is the within group sums of squares, and \tilde{B} is the between group sums of squares for the discriminant functions. The analogy to the previous procedure should be obvious if the relation between the corrected sums of squares and the sample covariance is remembered, i.e. in the univariate case.

$$SS_{yy} = \sum_{j=1}^a \sum_{i=1}^n (y_{ij} - \bar{y}_i)^2$$

and

$$S_y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

where SS_{yy} is the within group sums of squares and

$$S_y = SS_{yy}/K$$

where $K = \text{constant}$.

Since the two sums differ by only a constant then minimizing SS is equivalent to minimizing S .

If K groups exist with N observations, the F ratio is given by

$$F = \frac{SS_b}{SS_w} \frac{N-K}{k-1}$$

where SS_b and SS_w are the sums of squares between and within groups, respectively. Since $(N-k)/(k-1)$ is a constant then SS_b/SS_w is the only essential quantity for measuring how widely a set of group means differ among themselves relative to the amount of variability within the groups. It can be shown that

$$\frac{SS_b(z)}{SS_w(z)} = \frac{\alpha' B \alpha}{\alpha' W \alpha} \equiv \lambda$$

where λ is defined as the discriminant criterion. Therefore this criterion would select the α_j 's such that λ is a maximum which would be the case if the within group sums of squares is minimized.

Canonical correlation analysis and discriminant analysis are closely related. In fact, Tatsuoka [89] has shown that the discriminant criterion and canonical correlation produce identical results. In canonical correlation the objective is to find a linear compound of X-variables that has maximum correlation with a linear compound of Y-variables. In mathematical terms we want to determine a set of weights $\underline{a}' = (a_1, a_2, \dots, a_p)$ for the X-variables and a set of weights $\underline{b}' = (b_1, b_2, \dots, b_q)$ for the Y-variables such that the correlation r_{zw} is maximized between

$$Z = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$$

and

$$W = b_1 Y_1 + b_2 Y_2 + \dots + b_q Y_q$$

In the development of the procedure, the sums of squares and cross-products matrix depicted in Figure 1 will be used.

$$\left[\begin{array}{cccc|cccc}
 \Sigma X_1^2, & \Sigma X_1 X_2, & \dots, & X_1 X_p & \Sigma X_1 Y_1, & \Sigma X_1 Y_2, & \dots, & \Sigma X_1 Y_q \\
 \Sigma X_2 X_1, & \Sigma X_2^2, & \dots, & \Sigma X_2 X_p & \Sigma X_2 Y_1, & \Sigma X_2 Y_2, & \dots, & \Sigma X_2 Y_q \\
 \vdots & & & & \vdots & & & \vdots \\
 \vdots & & & & \vdots & & & \vdots \\
 \Sigma X_p X_1, & \Sigma X_p X_2, & \dots, & \Sigma X_p^2 & \Sigma X_p Y_1, & \Sigma X_p Y_2, & \dots, & \Sigma X_p Y_q \\
 \hline
 \Sigma Y_1 X_1, & \Sigma Y_1 X_2, & \dots, & \Sigma Y_1 X_p & \Sigma Y_1^2, & \Sigma Y_1 Y_2, & \dots, & \Sigma Y_1 Y_q \\
 \Sigma Y_2 X_1, & \Sigma Y_2 X_2, & \dots, & \Sigma Y_2 X_p & \Sigma Y_2 Y_1, & \Sigma Y_2^2, & \dots, & \Sigma Y_2 Y_q \\
 \vdots & & & & \vdots & & & \vdots \\
 \vdots & & & & \vdots & & & \vdots \\
 \Sigma Y_q X_1, & \Sigma Y_q X_2, & \dots, & \Sigma Y_q X_p & \Sigma Y_q Y_1, & \Sigma Y_q Y_2, & \dots, & \Sigma Y_q^2
 \end{array} \right]$$

Figure 1. Sums of Squares Matrix

This matrix is partitioned in the form depicted in Figure 2.

$$\begin{bmatrix} S_{xx}(pxp) & S_{xy}(pxq) \\ \text{---} & \text{---} \\ S_{yx}(qxp) & S_{yy}(qxq) \end{bmatrix}$$

Figure 2. Partitioned Matrix

A matrix \tilde{A} is then formed by the following matrix multiplications.

$$\tilde{A} = \begin{bmatrix} S_{xx}^{-1} & S_{xy} \\ S_{yx} & S_{yy}^{-1} \end{bmatrix} \begin{bmatrix} S_{xy} \\ S_{yx} \end{bmatrix}$$

The eigenvalues μ_i^2 and the eigenvectors α_i of the matrix \tilde{A} are then computed. The largest eigenvalue μ_1^2 , is the square of the maximum correlation r_{zw} , where

$$r_{zw} = \frac{a' \tilde{S}_{xy} b}{(\tilde{a}' \tilde{S}_{xx} \tilde{a}) (\tilde{b}' \tilde{S}_{yy} \tilde{b})}$$

This is termed the maximum canonical correlation between the two sets of variables. The elements of this eigenvalue are then the weights or loadings which are used to form the linear compound of the X-variables. The \tilde{b}_i weights for the y-variables compound can be determined by the relationship

$$\tilde{b}_1 = \frac{1}{\mu_1} S_{yy}^{-1} S_{yx} \tilde{a}_1$$

which is a direct result of the solution to the partial derivatives of the Lagrange function used in maximizing r_{zw} . The reader is directed to Van de Geer [102] or Tatsouka [90] for details. It might be noted that the situation is exactly parallel to that of principal components analysis and discriminant analysis and consequently even factor analysis. In each case a set of combining weights is determined which will maximize a specified criterion for a resulting linear compound. In each situation the vector elements of the eigenvector associated with the largest eigenvalue of a particular matrix specify the weights or loadings of the compound.

Up until this point the discussion has circumvented the question of measures of association or similarity between variables and between data units. As usual there is considerably semantic difference between terms. This discussion will follow the terminology of Sneath and Sokal [82].

Variables are classified according to type and scale of measurement. The type variables include continuous, discrete, and binary. A continuous variable may assume an uncountably infinite number of values while a discrete variable may assume a finite (or at most countably infinite) number of values. A binary variable may assume only two

values. The various scales of measurement include nominal, ordinal, interval, and ratio. Nominal and ordinal scales are referred to as qualitative variables, whereas interval and ratio scales are referred to as quantitative variables.

In applied multivariate analysis the comparison of variables with different scales and types can present problems of interpretation. Consequently, techniques have been developed for conversion of variables from one type to another in order to provide some homogeneity of scale types. Anderberg [2] gives a comprehensive presentation of some of these techniques of conversion.

The general term in describing relationships between variables and between data units is "similarity coefficients" which includes distance, association, and correlation coefficients. It might be noted that distance and correlation coefficients are actually dissimilar in nature but shall be included in the similarity category.

Distance measures are the most popular and practical for use when describing similarity between data units. The Minkowski metric is by far the most popular metric for measuring distances. The Minkowski metric in general terms is

$$D_p(X_j, X_k) = \left[\sum_{i=1}^n |X_{ij} - X_{ik}|^p \right]^{1/p}$$

where $p \geq 1$.

By selecting various values of p many different metric distances are obtained. The most common being the Euclidian distance or L_2 metric for $p = 2$. If $p = 1$ the "city-block" or L_1 metric is obtained. The L -infinity metric is defined as:

$$D_{\infty}(X_j, X_k) = \max_{i=1, \dots, n} |X_{ij} - X_{ik}|$$

Figure 3 depicts the unit balls of each of the metrics described.

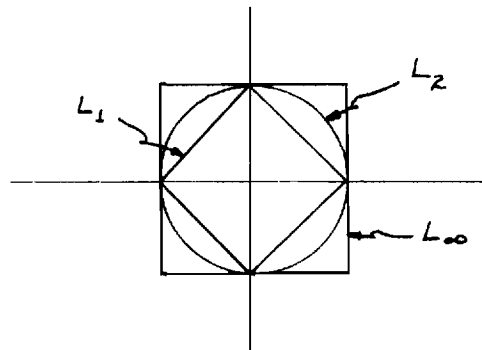


Figure 3. Unit Ball

The unit ball is the set of points for which $D_p(X_j, X_x) = 1$.

Another distance measure which takes into account the correlation between variables is the generalized Mahalanobis' distance (D^2 if derived from a sample). A general formulation is as follows:

$$D_y^2 = (\tilde{X}_i - \tilde{X}_j)' \tilde{W}^{-1} (\tilde{X}_i - \tilde{X}_j)$$

where \tilde{W}^{-1} is the inverse of the pooled within groups variance-covariance (dispersion) matrix or the pooled within groups sums of squares (scatter) matrix. It might be noted that if $W = I$ then the distance becomes the Euclidian metric.

Association coefficients are normally used to describe the similarity between pairs of variables over an array of two-state (binary) or multi-state characters. The most popular of these are the matching coefficients. In the final analysis this type of similarity measure will be of no use in the development of the methodology in this paper. Anderberg [2] has a most complete and comprehensive discussion of these coefficients.

Correlation coefficients are the most commonly used similarity coefficients when measuring the similarity between variables which are described on an interval (quantitative) scale. Pearson's product moment correlation coefficient is the most frequently employed of these. The following formulation of this correlation coefficient should be familiar to the reader:

$$r = r(X,Y) = \frac{\text{cov}(X,Y)}{[\text{var}(X)\text{var}(Y)]^{1/2}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \left[\sum_{i=1}^n (Y_i - \bar{Y})^2 \right] \right)^{1/2}}$$

There are various other measures such as the Chi-square based measures which are normally used to describe the relationship between nominal (qualitative) variables. Since this measure and others are of minimal assistance in the development of the methodology, their discussion will not be made. The reader will find a presentation of other measures in Anderberg's and Sneath and Sokal's books.

The final multivariate technique to be discussed is that of cluster analysis. The technique is by far the most diverse of all and would require many pages to properly discuss and describe the myriad of individual points of view and methods. The discussion will be limited to the more popular and useful of the methods. The basic methods are normally assigned to two general groups, the hierarchical methods and the non-hierarchical (optimization or partitioning) methods.

The hierarchical methods are further grouped into two general sub-groups, agglomerative and divisive. The hierarchical methods' objective is to produce a connected tree graph which in some way describe the relationships between data units. Such a connected tree is illustrated in Figure 4.

If the method begins with the data units as single entities and attempts to associate them in some manner until only one main grouping remains, then this method is termed agglomerative. On the other hand, the method is termed

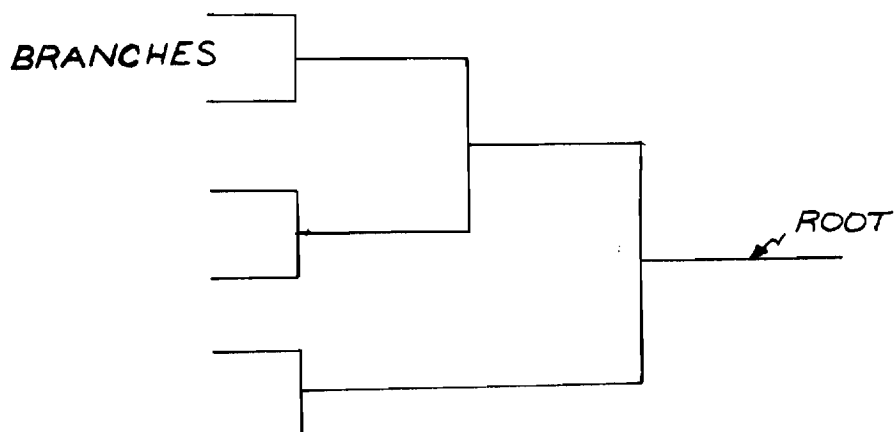


Figure 4. Connected Tree

divisive if the data units are taken as a single grouping and then partitioned in some manner until all data units are a group of one.

Agglomerative hierarchical methods are probably the most popular of all clustering methods. They include single linkage (nearest neighbor), complete linkage (furthest neighbor), centroid, and Ward's error sums of squares methods. The single linkage method groups data units according to the distance or correlation between their nearest neighbors. The cluster groups with the smallest distance or largest correlation are grouped together to form the new cluster groupings. The complete linkage method groups data units according to the distance or correlation between their furthest neighbors. The cluster groups with the largest distance or smallest correlation are grouped together to form the new cluster groupings. The

interpretation of the clusters is different in each case. The single linkage method is interpreted in terms of the relationships between cluster groupings. In contrast, the complete linkage method is interpreted in terms of the relationships within cluster groupings.

The centroid method clusters groups which have the most similar mean vectors or centroids. Groups are combined which have the smallest distance between their centroids. The only interpretable distance for the centroid method is the squared Euclidian distance and is therefore the normal measure of similarity between groups.

The Ward [103], [104] Method was developed to cluster groups which maximize a particular objective function. The procedure is known as the error sum of squares, since Ward illustrated his technique with this type objective function. At each step of clustering, all possible pairs of groupings are considered and that one which increases the total within groups error sums of squares the least is selected. Wishart [108] has shown that this minimum increase in error sums of squares is proportional to the squared Euclidean distance between the centroids of the combined groups. This result is different from the centroid method since it weights the distance between centroids.

The divisive methods are normally termed monothetic or polythetic. The monothetic is based on having a single attribute while polythetic is based on all the attributes.

The monothetic approach is normally used in the case where data units are described by binary variables. Lance and Williams [56] present this approach as "association analysis." They divide the data in terms of one attribute variable with one subset being those data units which have the attribute and one subset being those data units which do not possess the attribute. The attribute is chosen which will maximize the distance or minimize some similarity measure between the two groups. The polythetic methods have few applications and will not be discussed here.

The most popular and versatile of the nonhierachical methods are the partitioning or optimization methods. Hereafter, the terms nonhierachical and partitioning (optimization) will refer to the same group of techniques. The nonhierachical methods normally are suitable for clustering data units and not variables. As opposed to hierachical methods, these methods cluster data units into a single grouping of K clusters which is specified prior to initiating the algorithm. Some techniques will change the number of clusters during execution of the particular algorithm. These techniques normally attempt to partition the data units so as to optimize some criterion. The user normally selects some initial partition or some initial k seed points for k initial partitions. The algorithm then reassigns or reallocates data units to improve the initial partitioning. The reallocations are based on optimizing the specific

criterion of the algorithm. These techniques are sometimes referred to as hill-climbing and are analogous to the steepest descent algorithms of nonlinear programming. There are several methods of determining initial partitions and seeds points and Anderberg [2] covers the subject quite well.

The nonhierarchical methods most commonly used are Forgy's [37] method, Jancey's [51] variation of Forgy's method, and McQueen's [66] K-means method. All of these techniques are based on distance measures and assign data units according to their distance from the centroids of initial partitions. If seed points are used, initial partitions are formed using these seed points by assigning each data unit to the nearest one. In Forgy's method the centroids are computed for these partitions and then data units are reallocated to the closest centroid. All data units are reassigned before centroids are recomputed.

Jancey's method differs from Forgy's method only in the manner of determining the centroid. Actually Jancey determines a new seed point by reflecting the old seed point through the new centroid. This is done to speed convergence and bypass possible local minimums. Both Jancey's and Forgy's methods implicitly minimize a within cluster error function.

McQueen's K-means method differs from Forgy's method in that the K-means method recomputes the gaining cluster centroid after each data unit is relocated. In the simple K-means technique the entire data set is reassigned only

once. McQueen has also developed a variate known as McQueen's convergent K-means which continues to cycle through the data units until the data set fails to cause any changes in cluster membership. McQueen [66] has established the convergence properties of this technique. Anderberg gives a brief explanation of the convergence properties of all three of these methods.

These methods however do not allow the number of clusters to vary during the execution. Several methods exist which include this ability within a specific algorithm. McQueen's [66] method with coarsening and refining parameters and Wishart's [109] variant on K-means in his CLUSTAN IA computer package are two of these methods. Ball and Hall's [14] ISODATA computer package also provides this ability with interaction with graphic display devices.

Other criteria have been suggested by Friedman and Rubin [38] which have been derived for multivariate analysis. In each case the object is to select that partition or cluster which best meets the selected criterion. The criteria are based on the previously mentioned matrix equation for the sums of squares, $\underline{T} = \underline{W} + \underline{B}$. The first of the criterion is minimize trace \underline{W} or minimize the total within group sums of squares of the partition. This is the criterion implicitly used in Forgy's, Jancey's, and McQueen's methods. The second criterion is to minimize the ratio of determinants $|\underline{W}|/|\underline{T}|$ which can be shown to be equivalent to

minimizing $|\tilde{W}|$ or maximizing $|\tilde{T}|/|\tilde{W}|$ which additionally can be shown to be equivalent to maximizing $|\mathbf{I} + \tilde{W}^{-1} \tilde{B}|$. This leads to the third criterion which is to maximize the trace of $\tilde{W}^{-1} \tilde{B}$. A fourth criterion which is closely related has been proposed by McRae [67]. This criterion is S. N. Roy's largest root criterion, i.e. maximize the largest eigenvalue of $\tilde{W}^{-1} \tilde{B}$.

Several computer programs have been compiled which utilize several or all of these criteria. McRae's [67] computer program MIKCA combines McQueen convergent K-means with any one of the four criteria. Demiremen [25] combines Forgy's method with the first three criteria into his computer program. There is very little reported on how to determine which criterion to choose under a particular situation. In addition there has been very little work done in showing a comparison of any of these clustering techniques. Friedman and Rubin [38] have provided some comparison but with inconclusive results.

It has already been mentioned that cluster analysis as applied to data units is not necessarily a statistical analysis technique but rather a data analysis technique. The author will not attempt to support either contention but shall use cluster analysis in its descriptive form. This will entail utilizing the principal output of cluster analysis, i.e. relative homogeneous groupings of the data. The advantage of homogeneity should be obvious to the

statistician whose analysis is normally based on the assumption of homogeneity between data units. Cluster analysis provides an opportunity to investigate empirically the degree of success achieved in attempts to fulfill assumptions of homogeneity.

Major Literature in Multivariate Analysis

This chapter will be concluded with a brief discussion of the major literature in the multivariate analysis field. Several books under the general heading of multivariate statistical analysis have been published. The more advanced works are headed by the classic by T. W. Anderson [8]. Other advanced works include books by M. G. Kendall [55], Morrison [64], A. P. Dempster [26] and C. R. Rao [75]. A relatively readable and complete work is S. J. Press' [72] excellent book. M. M. Tatsuoka's [90] and J. P. Van de Geer's [102] books are complete works with respect to the various multivariate techniques. Introductory works from a computer application viewpoint are books by Afifi and Azen [1] and Bolch and Huang [16].

Specific multivariate techniques enjoy a prolific production of books especially in the factor analysis field. The classic works in factor analysis are H. H. Harman's [43] and R. B. Cattell's [20] books. R. J. Rummel's [79] book on applied factor analysis is very readable and complete. Introductory works include A. L. Comrey's [24]

and D. N. Lawley and A. E. Maxwell's [60] books. Tatsuoka's [90] and Van de Geer's [102] previously mentioned books are excellent discussions on the principal components, discriminant, and canonical correlation techniques of multivariate analysis.

The classic work in cluster analysis is P. N. A. Sneath and R. R. Sokal's [82] book which has recently been updated in a new edition. This book tends to be difficult to read due to the interrelationships that are drawn between the technique and biological taxonomy. Jardine and Sibson [52] is another book based on biological taxonomy. This book, however, is more advanced and requires the mathematical sophistication necessary for Rao's or Anderson's books which have been previously mentioned. For the researcher more interested in application, there are two excellent and complete works. The first is M. R. Anderberg's [2] book which contains several computer programs and detailed discussions on how to apply and interpret cluster analysis. The second work is Brian Everitt's [30] brief but concise book which when coupled with Anderberg's produces a complete and thorough examination of the field of cluster analysis.

This is not an exhaustive list of the books that have been published in these fields. However, the list provides an enumeration of some of the better works for both theory and application in all areas of multivariate analysis.

CHAPTER III

APPLICATION OF MV TECHNIQUES TO OBTAIN OPERATIONAL USAGE PATTERNS (METHODOLOGY)

Introduction

This chapter will evaluate each of the multivariate techniques discussed in Chapter II and determine which of them will adequately provide operational usage patterns. The various advantages and limitations of these techniques will be presented and one technique will be selected based on these advantages and limitations coupled with certain criteria and assumptions set forth in this section.

The major criterion is simplicity; the methodology should be relatively easy to implement and relatively simple to interpret. In addition, it is felt that expense is a factor and therefore, the methodology should be relatively inexpensive to execute. The relativity refers to the adequacy of the methodology to consistently provide valid usage patterns. The expense and simplicity base begins with the establishment of an adequate technique. The term validity shall refer to the ability of the technique to provide essentially the same results when applied to the sample data using various alternative methods. Cross validation techniques, which validate by applying the

methodology to a different sample, will not be possible due to the nonavailability of another sample.

In addition to technique selection, this chapter will involve a discussion of the data preparation, the selection of an appropriate distance or similarity measure, and the actual application of the technique to produce the usage patterns. The establishment of validity will also be discussed and will conclude this chapter.

Selection of a MV Technique

The main selection criterion is whether or not the MV technique will provide meaningful groupings (usage patterns) of data units. However, the methodology should also take into account methods of providing for parsimony. The assumption that the data has a joint multivariate normal distribution will also play a major role in determining which technique is selected.

The techniques which will provide parsimony are principle component or factor analysis. In the first case there is no need for assumptions of normality. However, the technique is seriously limited in the amount of parsimony allowed depending on the amount of total variance required. Factor analysis will provide considerable parsimony but is limited by a normality assumption and by the number of data units and variables it can cope with on today's computers. The computer storage requirements and time demands sky

rocket for data sets containing more than 250 data units. This problem will surface when there is a requirement for a similarity matrix. This is the case with all the techniques except nonhierarchical clustering. There are methods to circumvent this problem by utilizing special computer procedures but at large additional costs. In addition there are tradeoffs between the number of variables and the number of data units handled by the computer and although sampling theory allows for small samples of the total population to be used, the more data units sampled the more accurate the results. And finally, there is the traditional difficulty in interpreting the factors which are generated. Some interpretation should be made in order to determine the validity of the resulting factors. One should not take the factors as good results unless some, at least intuitive, credibility can be associated with them. It would be dangerous to assume with little basis that any resulting factors are correct and meaningful.

If there is no requirement to limit or reduce the number of variables, the problem becomes one of mere grouping of data units. Under factoring techniques the Q-factor is considered. This technique suffers from some of the short comings of its parent, in that there is a restriction on the number of data units and that normality should be assumed. In addition there is the question as to how to interpret correlations between data units. Fleiss

and Zubin [36] object to the Q-technique because of this question and consider it an idle exercise when factor analysis is performed on data units when there is no reason to believe there is an underlying linear model. And finally, if there are only five variables, there can be at most four distinct clusters, i.e. one less than the number of variables. This last restriction severely limits this procedure.

Discriminant analysis and canonical correlation analysis are exceptionally good techniques when applied to the class of problems for which they are designed to analyze. They normally cannot separate out groupings of data units. However, once meaningful groupings have been discovered, then evaluation using these techniques can be very enlightening. For example, discriminant analysis can be used to determine the discriminant power of the grouping. In other words how well does each group of the partition differ from each other? In addition other groups could be assigned or compared to this partition according to the discriminant function of each group. The question answered would be, is this specific group a member of one of the groups belonging to the partition? Canonical correlation on the other hand would be useful in comparing two different partitions or groupings. This would occur when there existed a grouping from each of two populations and a comparison of the two is desired. In order to implement these techniques

it is normally necessary to assume joint normality.

Since joint normality is considered desirable in the above techniques it would be advantageous if a discussion of this topic be inserted at this point. It is well known that the existence of marginal normality of each variable does not insure nor imply joint normality. Consequently, a test for marginal normality would be fruitless. Tests for joint multivariate normality are few and sparsely documented. Malkovich and Afifi [63] present four such tests based on extension and generalization of univariate test of normality. These include, generalization of (1) measures of skewness and Kurtosis, (2) the Shapiro-Wilk test criterion, (3) the Cramer-Von Mises test criterion, and (4) the Kolmogorov-Smirnov test criterion. A Monte Carlo study was conducted to examine the power of each test with each having advantages in different situations. Even if this type examination does not reveal normality, nonlinear transformations have been developed which might induce normality. Tukey's [101] basic article on transformations presents a family of transformations which will induce normality under a variety of situations. Dolby [28] applies an approximation technique to Tukey's family of transformations to provide a quick method of choosing a transformation. Box and Cox [17] proposed a set of data transformations of multivariate observations to enhance the normality of the distribution and improve homoscedasticity. Andrews, et al. [6] and

Andrews [4] have further expounded and refined these concepts. These techniques of testing and inducing normality tend to be difficult and time consuming to apply. As a result researchers such as Kolchi Ito [50] have presented arguments showing the robustness of the assumption of normality and homoscedasticity. Under certain circumstances the two assumptions will not greatly affect the results. One conclusion drawn from a review of literature is that it is difficult to define exactly what multivariate normality really is. The question still remains, what is meant by normality in the multivariate case? In any case the basic assumption of normality and even homoscedasticity reduces the enormity of the multivariate analysis.

It is interesting to note that the above conventional statistical data analysis techniques have been shown to be inadequate in determining known partitions in certain situations. For example, Ball [13] has shown where three sets of data which are quite different but have identical covariance matrices. The question is, does the presence of identical covariance matrices imply the populations are identical? In order to produce partitions of data units without resorting to normality and homoscedasticity assumptions the technique to use is that of cluster analysis.

At first glance cluster analysis appears to be a ready answer to the desired criteria of simplicity. This could be misleading for cluster analysis does not provide a

simple neat packaged solution. Jardine and Sibson [52] object to several agglomerative methods based on mathematical arguments. They specify certain conditions which clustering methods should meet in order to be mathematically acceptable, for example, continuity and minimum distortion. The single linkage method is the only method which satisfies all their conditions. In the same vein Fisher and Van Ness [31] have introduced nine admissibility criteria for all clustering techniques. Their criteria include image admissibility, convex hull admissibility, connectedness admissibility, etc. The single linkage method surfaces as the "best" method by satisfying all criteria except the convex hull. The complete linkage fails on two criteria, the convex hull and connectedness. The latter criterion is the same criterion for which Sibson and Jardine object to the complete linkage method. According to Everitt this criticism has caused some people to object to these admissibility criteria since in several cases the application of single linkage has produced less satisfactory solutions than those produced by other agglomerative methods. Such results have led Lance and Williams [58] to consider single linkage an obsolete technique. However this is probably a bit harsh.

Nonhierarchical or optimization methods which seek to optimize some criterion frequently find suboptimal solutions. This problem of local optima also exists in the case of nonlinear hill-climbing algorithms. It is impossible

to check each possible location in order to obtain the global optimal solution. Techniques have therefore been developed to increase the likelihood of finding the global optimum. In addition optimization methods meet only two of seven of Fisher and Van Ness's relevant admissibility criteria. In Chapter II it was noted that McQueen's, Forgy's, and Jancey's methods implicity minimize the trace of \tilde{W} . Everett [30] has shown that minimizing trace \tilde{W} produces or attempts to provide spherical, homogeneous clusters. In addition this criterion is not invariate under linear transformations. On the other hand the criteria of minimizing the determinant of \tilde{W} attempts to provide clusters of the same shape but is invariate under linear transformations. In addition $\det \tilde{W}$ criterion requires more computer time than the trace \tilde{W} technique. The trade-offs are obvious in this situation. The other two criteria mentioned before (i.e. Roy's largest root and Hotelling's trace criterion) would also tend to attempt to locate the same size clusters and in addition would require additional computer time to invert \tilde{W} . Anderberg [2] comments that "it is difficult to identify any clear-cut advantages stemming from the use of" the last three criteria as opposed to the use of the trace \tilde{W} criterion. On the other hand Everitt opts for the determinate \tilde{W} criterion. Friedman and Rubin [38] observe that there are no guidelines for making choices among the criteria. This appears to be essentially true

except for the latest work of Everitt which has been discussed above. Except for this, there remains a lack of research into which of the criteria perform best under which set of circumstances.

In order to apply any of the above methods, fairly sophisticated computer programs are necessary for implementation. The writing of such programs could be very time consuming. Consequently, the researcher is basically bound to the computer and to the availability of existing computer programs. The University of California presents programs for all the above mentioned techniques, except cluster analysis, in its Biomedical Computer Programs book [27]. The Principal Component analysis program is limited to 25 variables and 400 data units. A general factor analysis program provides several alternatives but is limited to 198 variables with 99 factors rotated. The number of data units is then restricted to at least 200 and not much more. A program for discriminant analysis for two groups and a program for multiple group stepwise analysis is also available with practical limits set at 300 data units and 25 variables in the first case and 80 variables and 80 groups in the latter case. With 80 variables and 100 data units the computer time required would be approximately 12 minutes. A canonical correlation analysis program is also provided with limits of 99 variables and about 100 data units, with a tradeoff of computer time between the two. In addition

to the BMD programs there is the Statistical Package for the Social Sciences (SPSS) by Nie, et al. [68] which provides sophisticated programs for general factor analysis techniques and the stepwise discriminate analysis technique. These programs demand even more computer time than the BMD programs for equal input.

There are several programs available for cluster analysis techniques. They range in complexity from the relatively simple to the highly sophisticated. Anderberg [2] provides a thorough discussion of almost all of them and several of his own. He provides hierarchical programs using different applications of minimizing the error sums of squares (basically Ward's method). In addition, he provides a centroid method program. All of these can be utilized using a stored similarity matrix approach or a stored data approach. Anderberg also provides nonhierarchical programs allowing selection of Jancey, Forgy, and McQueen's methods allowing user selection of distance measures, and different methods of selecting seed points, and initial partitions. Other nonhierarchical programs include Ball and Hall's [14] ISODATA method which is an extension of Forgy's basic method and allowing splitting and lumping of clusters based on different input parameters. Ball and Hall [15] and Sammon [80] have included the ISODATA method in the PROMENADE and OLPARS systems, respectively. These systems provide an on-line data analysis package utilizing

interactive graphics to better determine appropriate clusters. Demiremen [25] provides a computer program which is also an extension of the Forgy method but in addition allows the partition to be improved using the trace \tilde{W} criterion, the determinant \tilde{W} criterion, and Hotelling's trace criterion. A linear discriminant analysis and a multivariate analysis of variance are also performed as an aid in evaluating each partition.

McRae [67] extends McQueen's convergent K-means method by allowing improvement of the K-means final partition using any of the four sums of squares criteria.

Wishart's [109] CLUSTAN 1A package is a complete package of both hierarchical and nonhierarchical methods. This package is highly sophisticated and provides for the changing of the required number of clusters during execution of a convergent K-means method.

Wolfe's [110] NORMIX program should be included in this discussion even though it requires the assumption of normality. The NORMIX program makes explicit use of likelihood methods in that it seeks that partition which maximizes the likelihood function.

Each multivariate technique has now been examined in some detail and the time for final selection has arrived. Since the number of variables that is of concern is not great, the matter of parsimony of variables is of little import. However the distinguishing of usage patterns from

among the data is the main concern. As a consequence, this narrows the field of techniques to those of Q-factor and clustering. The various shortcomings of the Q-factor technique make it unsuitable for this purpose or at least less desirable than cluster analysis.

The selection of one of the clustering methods could be difficult. The field can be narrowed further by the elimination of the hierarchical methods which are best suited to the biological fields of science. The nonhierarchical methods are better suited to the purpose for which they are intended, i.e. discovering natural clusters within a data set. The choice has now been narrowed to the nonhierarchical methods but the final selection from these is no less difficult. However a final selection is made of MacQueen's convergent K-means method. This method is selected because of its greater intuitive appeal but more important because of its convergence properties so painstakingly and expertly shown by MacQueen [66]. This does not exclude completely the other methods especially those which include the different sums of squares criteria. But in the spirit of simplicity and ease of implementation and interpretation, MacQueen's convergent K-means method is the final selection.

Preparation of Data

The preparation of data into a multivariate format

is in itself a huge task. The first step is to determine what variables should be measured and how the data is collected. The data for this thesis was collected on the collection sheet shown in Appendix A. Based on this form and the resulting entries only six distinct variables can be isolated. These variables are, (1) miles traveled on paved roads (HB), (2) miles traveled on unimproved roads (SB), (3) miles traveled cross country (CC), (4) total miles traveled (MILE), (5) estimated load or cargo by weight (LOAD), and (6) number of personnel carried (PERS). Two other descriptive variables were also isolated but are not used in the actual analysis. These are (1) military unit controlling the vehicle and (2) mission of a particular vehicle. In addition, the total miles traveled category is not used in conjunction with any other mileage variable in order to insure full rank sums of squares matrices. Consequently the initial analysis clusters data units based on only five variables.

Once the data units (vectors) have been formed the next step is to determine outliers. Outliers are data units that obviously do not belong to the sample population. Cluster analysis is sensitive to outliers and can influence the optimal partition. However, cluster techniques can and will isolate outlier groups. Hierarchical methods will normally isolate individual outliers based on the fact that they will join the main body of points near the final level

of clustering. However individual outliers will not necessarily be isolated by nonhierarchical techniques. Therefore, it is necessary to identify and eliminate these type of outliers. One technique is to produce histograms of each variable. This normally identifies outliers resulting from errors in data construction. They are evidenced by zero values and large unreasonable values surfacing where they should not. Gnanadesikan and Kettenring [39] explain and demonstrate several techniques for the identification of multivariate outliers. These include two and three dimensional plots of the first two or three principal components and Mahalanobis' generalized distance for uncovering observation which lie far afield from the general scatter of points. They further discuss discriminant and canonical correlation techniques for discovering outliers in two or more groups of observations or two or more sets of variables. The last two techniques become expensive in computer time for large data sets. The BMD program file includes a program which identifies multivariate outliers based on Mahalanobis' distance and the first and second principal components.

With multivariate data sets it is difficult to define the difference between units for each variable. For example, how do you interpret the difference between miles and pounds? A cluster of these two units would emphasize pounds more than miles. As a consequence the data should

be transformed in some manner to take into account these differences. Techniques of data standardization have been developed such as dividing each data variable by its mean, or its standard deviation, or even its range. Each of these techniques reduces some aspects of the differences between variable units. These standardization techniques are normally applied without consciousness of the reasons for or the result of the particular technique used. Anderberg [2] cautions the researcher in unilaterally applying these techniques in all situations. Another way to circumvent the problem is to require the final clustering to be invariant under changes of unit for each variable. These methods have been previously discussed under the terms of sums of squares criteria such as the determinant \tilde{W} or Hotelling's trace criterion. The reader should remember that the trace \tilde{W} criterion is not invariant under linear transformations. It was also noted that MacQueen's convergent K-means method implicitly minimizes the trace \tilde{W} criterion. Consequently, clustering of raw data using this technique alone will produce a different partitioning than the one produced by standardized data. If final clustering is done according to MacRae's algorithm, standardization would not be necessary since one of the invariant criteria can be applied to the resulting convergent K-means clustering. However, in this case, the demonstration of the methodology is by the convergent K-means method only.

In any case the large difference in variable units is considered the overriding criterion and therefore the data is standardized. This is accomplished by dividing the centered variable elements by the standard deviation for each variable. Centering was accomplished by subtracting the mean of each variable from each variable element.

It is at this point that a determination must be made of whether or not there is a need for determining multivariate normality. If such a need arises this step would be to test for normality and if normality does not exist then to transform the data in some manner to induce normal conditions. However, in the majority of clustering techniques there is little need to show normality.

Selection of Distance Measure

Prior to implementing the specific cluster technique, it is necessary to identify which distance measure is to be used. There are three possible alternatives which lend themselves to use in clustering data units, (1) the squared Euclidian distance, (2) the weighted squared Euclidian distance, and (3) Mahalanobis' generalized distance. In determining which to use, it is best to remember the actual distance which is being measured. In the case of Euclidian distance, a spheroidal distance results and in the case of Mahalanobis' distance, an ellipsoidal distance results. If the clusters are of the two-dimensional form in Figure 5a

then the Euclidian measure will suffice. The Mahalanobis measure would be best if the clusters are of the form in Figure 5b but if the situation was as in Figure 5c the Mahalanobis measure would tend to place observation X in cluster two even though it is closer to cluster one. This results since the Mahalanobis measure tends to emphasize the vertical axis of a cluster. Of course, if a prior knowledge of cluster shapes were known then a general weighted distance measure would be best. Therefore, prior knowledge or experience could identify the relative importance of variables. This relative importance can then be reduced to weightings for each variable. The Euclidian measure thus transformed by these weights could provide more acceptable optimal partitions. In this case such knowledge or experience is either not available or not reliable. Therefore a simple squared Euclidian distance measure will be used in the analysis, even though there is no evidence to assume that the clusters would be spherical.

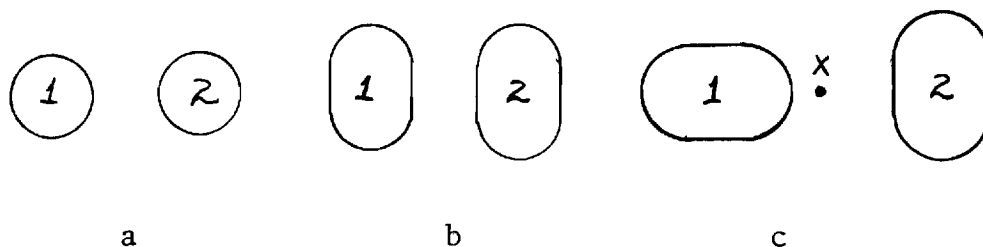


Figure 5. Cluster Shapes

Application of the M.V. Technique

Once the data has been prepared and a distance measure selected, the technique is then applied to the data through the use of a computer. Anderberg's program using MacQueen's convergent K-means was modified for use on the UNIVAC 1108. A listing of the program appears in Appendix B. The program provides for three methods for inputting or selecting initial seed points or an initial partition, (1) seed points can be read directly into the program on data cards, (2) seed points can be selected as individual data units (selection can be random or otherwise), and (3) the data units can be grouped in any manner determined by an input sequence of numbers (e.g. 100, 150, 50... where the first initial cluster is composed of the first 100 data units, the second composed of the next 150 data units, etc.). For the purposes of this application seed points were initially chosen in a semi-random manner by selecting them equidistant from each other and spanning the data set. Partitions were selected based on equal partitions spanning the data set. The centroid results of these two methods were compared and the most identical centroids by pairs were extracted, averaged and then used as inputted seed points for the third method. The final three results were then compared to determine their similarity and whether or not they return essentially the same partition. All three methods were employed to decrease the likelihood of being

trapped into a local optimum. Similar techniques are used in nonlinear programming to avoid the same entrapment. In addition if the same partition is returned by each method then this increases the acceptance of the overall technique's validity.

The determination of the optimal partition or the "best" number of clusters is not a straight forward choice. Some techniques have been developed whereby when a certain criterion is met then the optimal number has been reached. These techniques usually involve a plot of the number of groups against the value of the criterion used in clustering (e.g. minimize trace W). The optimal partition is the one where there is a sharp increase or decrease in the plot. Everitt [30] suggests that in general these procedures are unsatisfactory. Normally, heuristic techniques are utilized such as MacQueen's coarsening and refining parameters or Wishart's various control parameters within the CLUSTAN 1A package or Ball and Hall's control parameters within the ISODATA method. Rather than go to these expensive techniques a simple heuristic technique is utilized. This technique can easily be applied to data sets of less than ten variables where there is some prior knowledge or belief as to the possible optimum number of clusters in the partition. In this case it is determined that there are probably more than four basic usage patterns and probably less than ten. The convergent K-means program is then run to seek clusters

numbering five to nine. After obtaining the partitions for each "K" number of clusters, each partition is compared to group those clusters which most resemble each other. This is based on the number of data units, and the sign and value of the variables in the centroid. Table 1 is supplied to aid understanding of this procedure. It should be noted that two centroids are shown for the 6 and 7 cluster partition.

Table 1. Cluster Groupings

K	NUM	HB	SB	CC	LOAD	PERS
7	231	-.28	-.21	-.21	-.80	-.51
6	265	-.32	-.21	-.21	-.78	-.42
7	199	-.32	-.18	-.11	1.05	-.21
6	220	-.28	-.17	-.06	1.07	-.21

The results in Appendix C have been arranged in this manner to form "g" groups. Table 2 shows a particular grouping (i.e. group 1) from the results of Table 6.

The next step is to determine which is the "best" partition. Two criteria are selected upon which to determine that partition which is the "best." The first of these is stability of the cluster to remain as a cluster in all three methods of choosing initial starting points and

Table 2. A Cluster Group

g	K	NUM	HB	SB	CC	LOAD	PERS
	9	106	-.37	-.21	-.21	-.59	-.56
	9	207	-.29	-.19	-.20	-.83	-.56
1	8	179	-.56	-.21	-.20	-.80	-.50
	7	231	-.28	-.21	-.21	-.80	-.51
	6	265	-.32	-.21	-.21	-.78	-.42
	5	312	-.33	-.21	-.22	-.74	-.23

during the changes in the numbers of clusters required. The second criterion focuses on the splitting of relatively stable clusters. If and when a relatively stable cluster splits to form two separate clusters and their centroids do not differ significantly, then they should not be split but should remain as a single cluster. With both criteria the comparison of centroids is the determining factor. The comparison is made on the basis of a distance measure. In this case, the squared Euclidian distance will be used to make comparisons for both criteria. In the first criterion the distance is computed between the centroids using all variables. If the distance between the centroids of any two k-partitions for any g-group is less than or equal to .05, the cluster is said to be highly stable between those two k-partitions. In addition if the distance between any two centroids for each method of a particular k-partition is

less than or equal to .05, then the cluster is said to be highly stable and valid for that k-partition. Table 3 is extracted from Table 6 to aid understanding. It should be noted that the squared Euclidian distance is recorded in the distance between cluster's column for both methods (e.g. the distance between 6 and 7 cluster for method 1 is .013 and also .013 for method 2). In addition a center column denotes the squared Euclidian distance between the two methods for each k-cluster (e.g. the distance between methods for the 6 cluster is .0002 and 0 for the 7 cluster).

Table 3. Distance Columns

K	NUM	HB	SB	CC	LOAD	PERS	DIST BETWN CLUST	DIST BETWN MTHDS	DIST BETWN CLUST	NUM
7	231	-.28	-.21	-.21	-.80	-.51	.013	0	.013	231
6	265	-.32	-.21	-.21	-.78	-.42		.0002		265

In the second criterion the distance measure is computed between centroids of the splitting cluster. In computing the distance only p-1 variables are used. The reasoning being that if the cluster splits when it really should be one cluster then splitting probably occurred based on only one specific variable. Therefore the variable which differs the most between the two clusters is not used

when computing the centroid. If the computed distance is less than or equal to .05, then the cluster should not have been split and it can readily be assumed that the best k-partition will be found at a smaller k. Table 4 is also extracted from Table 6 and depicts a splitting cluster in group 1. The cluster appears to have split on one variable LOAD (note values of -.59 and -.83). Consequently, this variable is not included in the distance computation. The squared Euclidian distance between the two clusters is computed based on the other four variables. This distance is recorded in the distance between clusters' column and designated by an asterisk (e.g. the distance equals .007* in this case).

Table 4. Splitting Cluster

K	NUM	HB	SB	CC	LOAD	PERS	DIST BETWN CLUST
9	106	-.37	-.21	-.21	-.59	-.56	.007*
9	207	-.29	-.19	-.20	-.83	-.56	

When this type splitting occurs then the number of clusters should be fewer. At a particular partition where clusters are highly stable and unnecessary splitting occurs at a larger number of clusters, then it can safely be assumed

that the best partition has been found. The results of this technique are outlined in table format in Appendix C which depicts the results of two starting methods for all partitions between four and ten clusters. In addition distance figures are shown which compare clusters centroids between partitions for each K. Distance figures are also shown between the first two starting methods. These two results determine the stability of the clusters. Significant nonoptimal stability is shown for partitions with less than six clusters and more than seven clusters. Significant nonoptimal splitting is shown for partitions with more than eight clusters (see group 1, 9--partition in Table 6). This leaves partitions with six, seven, and eight clusters. Nonoptimal stability was exhibited at the six and eight cluster levels (see e.g. group 1, 8 partition and group 3, 6 partition in Table 6). The seven cluster partition was chosen as the "best" because it resulted in the most stable clusters at that level and between starting methods.

A lumping criteria could be established but it is felt that it would be redundant since the stability and splitting criteria are adequate in determining the optimal partition. If there are more than ten variables this procedure would probably become untenable and the analyzer would have to resort to the more sophisticated techniques of ISODATA and CLUSTAN 1A.

Once the final partition is obtained there is a need

to insure its validity. One method has already been described which entailed the use of three different methods of selecting starting points. If results agree significantly between each method then this increases the acceptance of validity. An additional validation technique is to divide the data set into two equal subsets by taking every other data unit. The same clustering method using three different starting points is employed and the result of the two partitions is recorded in Appendix D. It should be noted that membership assignment and centroid values are essentially the same. It might also be noted that an arithmetic average of the two half data sets centroids closely reproduces the total data sets centroid for any k-cluster. This shows more evidence of cluster stability in addition to added validation of the partition. A third method of validation is to omit or replace some variables to determine any effect or change this might have on the partition. In this case the three variables describing distances traveled for three categories are replaced by the total miles traveled. The resulting partition considerably changes the membership list from the original full variable partition. In addition it is hypothesized that the elimination of variable three (CC) should have little effect on the partition relative to the elimination of variable four (LOAD). An analysis is conducted for each case and results conclusively support the hypothesis. From these three

validation techniques the validity of the partition can be accepted.

An interesting situation occurred when it was noted that the smallest cluster (i.e. 19 data units) was also the most stable of all the clusters which indicated a strong possibility of it being an outlier group. A check of the original data indicated that this was not the case. All units belonged to the Division Signal Battalion and carried long distance communication gear. These individual units provided the most distinct cluster since they are utilized in a distinctly different manner for this particular vehicle.

The final selection of the best partition produces the operational usage patterns which are the objects of this research. In this case the seven cluster centroids are the operational usage patterns. Of course, these values are still in standardized form and can be readjusted to give more meaningful values. The standardized and readjusted forms of the usage patterns are shown in Table 5.

Table 5. Operational Usage Patterns

Group	Standardized Centroid Results						Operational Usage Patterns					
	HB	SB	CC	LOAD	PERS		HB	SB	CC	LOAD	PERS	
1	-.28	-.21	-.21	-.80	-.51		23	2	0	290	2	
2	-.27	-.20	-.11	1.11	-.29		23	2	0	1193	2	
3	-.56	-.15	-.19	-.32	.60		15	3	0	517	4	
4	2.11	-.06	-.26	-.32	-.16		90	4	0	517	2	
5	.34	.24	3.99	.37	-.26		40	7	6	844	2	
6	1.49	5.41	-.32	.87	.22		73	70	0	1080	3	
7	-.27	-.02	-.11	-.10	3.10		23	4	0	621	8	

CHAPTER IV

PROJECTED UTILIZATION WITHIN THE TOE SYSTEM

The Present System for the Development of TOE

The development of the Tables of Organization and Equipment (TOE) is very intricate and requires input and partial development at all major levels of Army Organization. AR 310-31, "Management System for Tables of Organization and Equipment" [10] establishes the TOE system and prescribes the policies, concepts, and procedures concerning the development, preparation, processing, review, approval, and publication of TOE documents. The TOE system provides the method by which the personnel and equipment requirements of the Army are structured and documented. The TOE documents ultimately produced by the system prescribe the normal mission, organization structure, and personnel and equipment requirements for specific military units.

A detailed explanation of the system would be lengthy, therefore, a brief explanation supplemented by illustrations will suffice. As an aid to understanding, two illustrations are included in this chapter. Figure 6 depicts the TOE management model showing the levels of command and the action network between them. Figure 7 depicts the TOE development cycle showing the flow of documents and information.

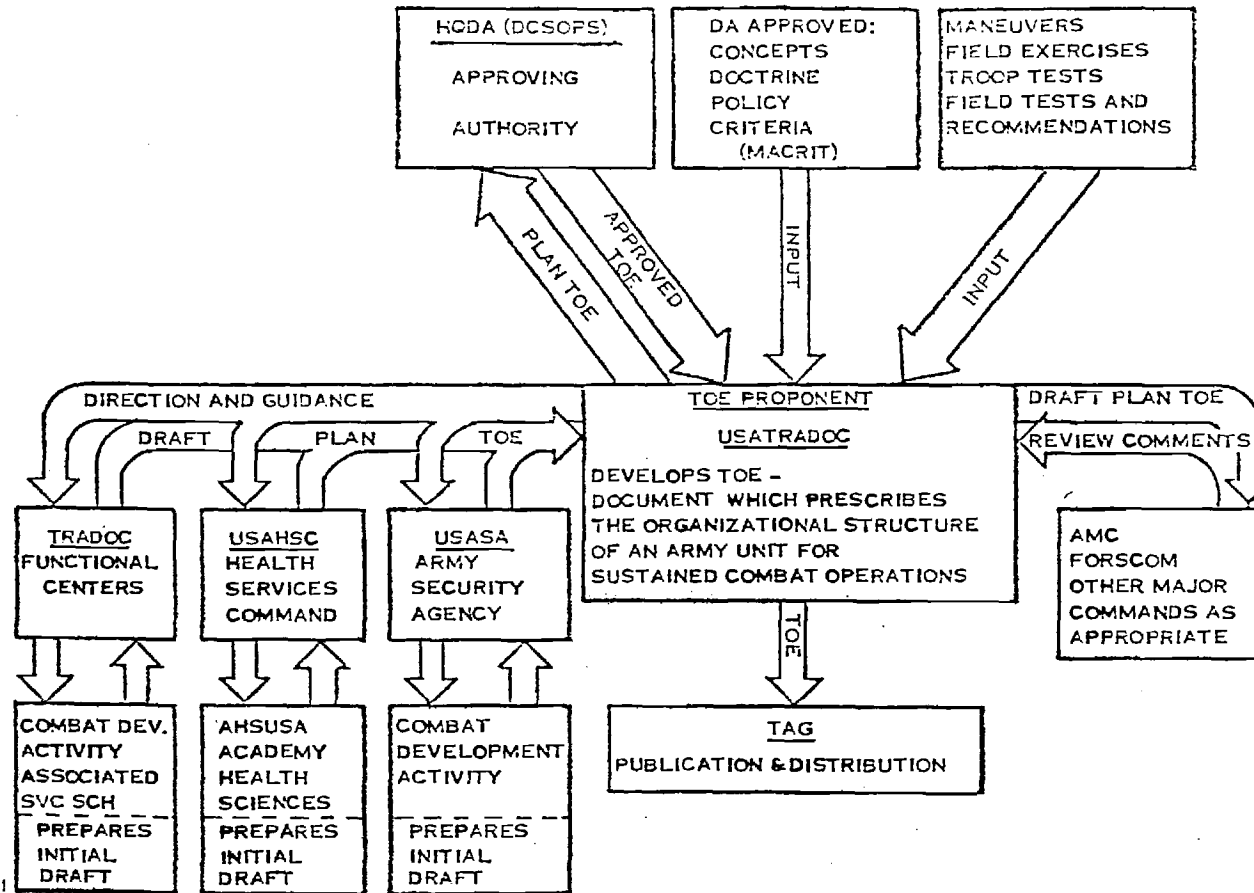


Figure 6. TOE Development Cycle

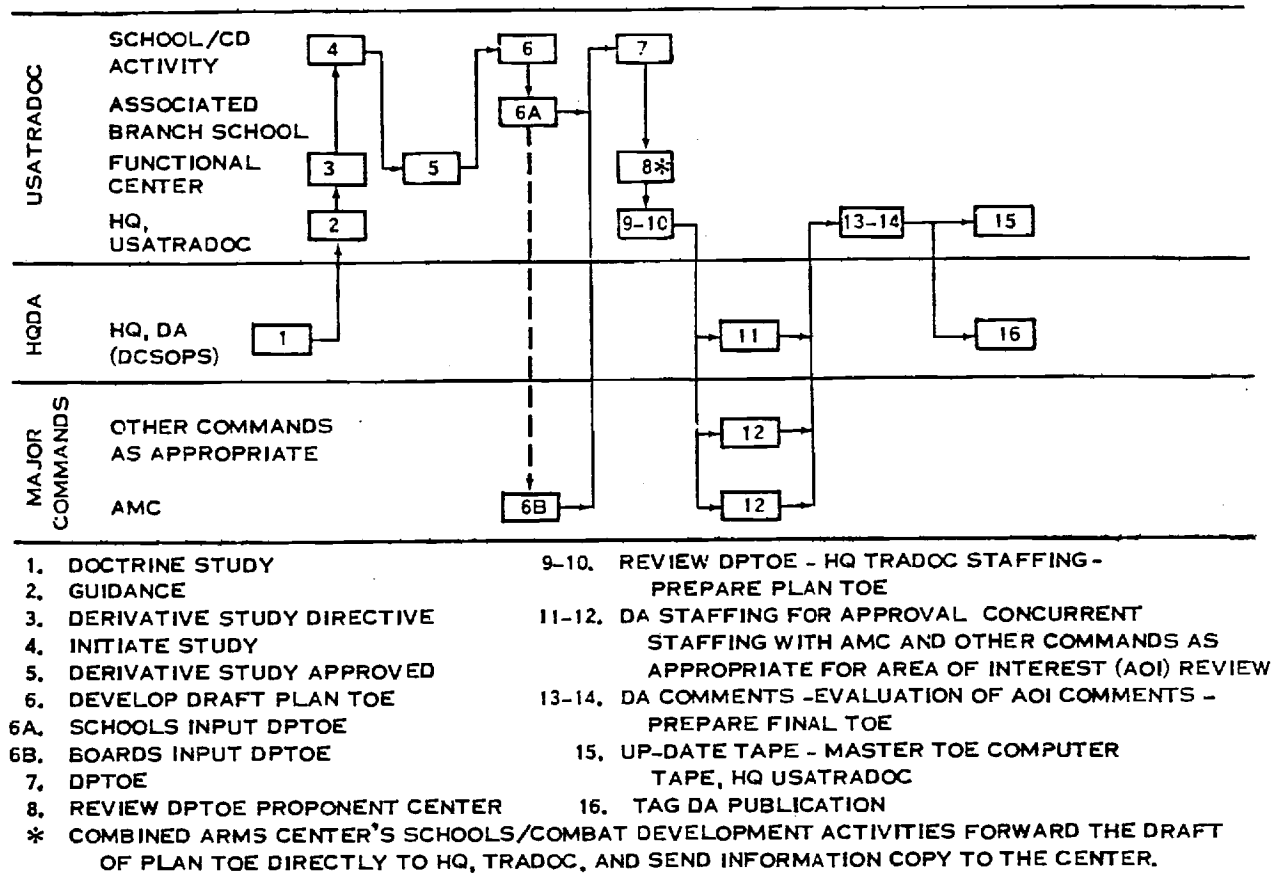


Figure 7. TOE Management Model

Within the system, a TOE normally goes through two major phases prior to publication. These phases are referred to as the draft plan TOE and the plan TOE. There are basically four levels of Army Organization which are directly responsible for the development of TOE. The Department of the Army (DA) heads the list; the Training and Doctrine Command (TRADOC) has the principle responsibility for development of TOE; the functional centers, of which there are three, are responsible in the Combat, Combat Support, and Combat Service Support groupings (e.g. Combat center is located at Fort Leavenworth, Kansas); and finally there are the combat development activity associated with each service school (e.g. Infantry School, Fort Benning, Georgia). These four levels interact to develop the TOE.

The combat development activity is responsible for developing the draft plan TOE based on (1) directives from their functional center, (2) DA approved concepts, doctrine, policy (AR 310-31), and criteria (AR 310-34), (3) direction and guidance from TRADOC staff, and (4) input from the school and board of each respective branch. Upon completion of the draft plan TOE, the combat development activities forward it directly to HQ, TRADOC, and send an information copy to their functional center. At TRADOC the draft plan goes through three weeks of extensive and detailed staffing to correct errors and to make it a better document. At the end of this period the document is presented to the TOE

Review Board. The TOE Review Board insures that the TOE is justified by approved "doctrine and complies with specific DA guidance. The board also provides a listing reflecting all changes approved by the board. The plan TOE is the resulting document of the TRADOC staffing. This document is then referred back to the functional centers and the service schools for their review and possible reclama. Upon completion of all TRADOC and proponent staffing the plan TOE is submitted to Army Material Command (AMC), Forces Command (FORSCOM), and other commands for area of interest review. In addition and most important it is submitted to the Deputy Chief Staff for Operations (DCSOPS) for concurrent HQ DA review. Significant changes arising from area of interest reviews are forwarded to HQ DA for resolution. DCSOPS will review, coordinate with HQ DA staff agencies, resolve conflicting recommendations, and approve the plan TOE. The approved plan TOE is then returned to TRADOC, with listed modifications, for preparation of final TOE. Another review board is conducted at HQ TRADOC, normally composed of the HQ TRADOC members. This board reviews and evaluates all comments received, and again accepts those that correct errors or make the TOE a better product. The completed final TOE is then sent to the Adjutant General (TAG) for publication and distribution.

At this point it is necessary to interject a brief discussion of the Basis of Issue Plan (BOIP) in order to

fully understand the TOE process and its relationship to the materiel acquisition cycle. The purpose of the BOIP is to project early in the material acquisition cycle for planning purposes, quantitative requirements for a new item of equipment in the TOE. In addition the BOIP projects other equipment and personnel changes that may be necessary in TOE to accommodate the new item of equipment. The BOIP is used to forecast new equipment densities for procurement programming purposes and is the main driving force behind the development of TOE. This document is therefore essential in revising of TOE by TRADOC and its proponents.

Procedures for developing BOIP closely parallels the procedure of the TOE. The development of the BOIP includes the same levels as the TOE development cycle and follows the same basic channels of production and review as the TOE. There is a BOIP Review Board at TRADOC which essentially is the same as the TOE Review Board. There should be complete interface between the TOE and BOIP development cycles to insure up to date and essential equipment in the Army Inventory. The policy and procedures for development of BOIP are contained in AR 71-2, "Force Development Basis of Issue Plan" [9] and TRADOC supplement 1 to AR 71-2 [98]. TRADOC Memorandum No. 15-1 [95] and 15-5 [96] discuss the policy and procedures of the TOE and BOIP Review Boards, respectively.

Method for Review and Change of TOE

Changes to present TOE are usually precipitated as a result of the following: (1) derivative study directed by the functional centers; (2) changes in policies, objectives, operational concepts, and doctrine at HQ DA; (3) proposed introduction of new equipment into the Army inventory; (4) special studies directed at all Army levels especially HQ DA (e.g. WHEELS study group). A closer look at the WHEELS study will give a specific account of how changes in TOE are made as a result of a special study directed by HQ DA.

The WHEELS study group in addition to their tasks mentioned in Chapter I, also evaluated the factors that are used as guidelines by those who structure the Army's Tables of the Organization and Equipment (TOE), Tables of Distribution and Allowances (TDA), and other factors that generate vehicle requirements. Their objective was to formulate recommendations that, when implemented, would reduce vehicle authorizations to minimum essential levels. The study group also made recommendations of vehicle adjustment (REVA) which resulted in tactical vehicle savings through the application of the recommended reductions [85].

The factors used as guidelines by those who structure TOE's are found in AR 310-34 "Equipment Authorization Policies and Criteria, and Common Tables of Allowances" [11]. They are obviously based on the experiences, opinions, and

desires of those who formulate them. As an example paragraph 4-62a (2) of AR 310-34 is quoted:

4-62. Functional Requirements for Vehicles in TOE units.

The following criteria will be used as a guide in determining requirements for vehicles in TOE/MTOE units--

(a) Category I TOE units. The following vehicles, limited to minimum quantities required in support of unit's missions, may be included in Category I TOE units:....

(2) One 2-1/2 ton truck with 1-1/2 ton cargo trailer for each company or battery supply function whose aggregate personnel strength does not exceed 220. For units with strength in excess of 220 but less than 300, one additional 2-1/2 ton truck may be authorized. For units in excess of 300, an additional 1-1/2 ton cargo trailer may be authorized [11].

It is the last part which the REVA targeted for change and with which their review was most specific.

To evaluate the changes, the WHEELS study group implemented a data collection which gathered data on all vehicles in the inventory. One of the major drawbacks to the method was that the data collected was based on the opinions and experience of commanders rather than on actual vehicle usage and performance. It is in these areas of TOE change and review where the methodology developed in Chapter IV would be applicable.

The Implementation of the Proposed Methodology

Within the TOE System

The U. S. Army is presently in an "austerity program" reflecting the present economic atmosphere. The Army has

solicited the field for suggestions which would reduce cost. The emphasis of austerity has also entered the TOE development phase. Col. Hicks, Deputy Chief of Staff for Combat Development (DCSCD), HQ TRADOC has stated that TOE or BOI Plans "which are approved in times of relative prosperity sometimes contain allowances which can only be viewed with suspicion in periods of austerity" [70]. AR 310-31 [10] states that TOE should contain requirements for "minimum essential equipment" only. Certain TRADOC publications [69] have contained austerity suggestions such as "survey using field units when developing new/updated/revised TOE." This suggestion supports the assertion that TOE in the past have been developed and revised based on opinion and experience only and not on actual operational usage data obtained in the field.

The methodology presented in Chapter III is an attempt to take operational usage data and evaluate it to gain some insight into how certain vehicles are utilized. It is contended that this methodology can be utilized in support of the Army's austerity program by identifying analytically utilization patterns of vehicles. This can then be used to assist Army elements at each and every level of TOE review by identifying potential vehicles which are misassigned (i.e. under or over utilized). In order to be more specific, two implementations will be presented for clarity. The two situations for implementation will be, (1) when a new vehicle

enters the inventory, and (2) when a vehicle has been in the inventory long enough for usage patterns to have been determined. In the first case historical data is not available on which to accurately determine which cluster set developed by the methodology is the most representative of the population. As a consequence, the experience and opinions of those who develop the vehicle are essential in establishing a baseline profile. Such a profile can be developed much in the same way as has been done in the past for mission profiles, i.e. (war games, questionnaires, etc.). This baseline profile can then be utilized to assist in determining the most accurate of the cluster sets which are developed from field collected operational usage data. The cluster sets are initially identified by Chapter III methodology. After the "best" cluster set has been determined, then grossly misassigned vehicles can be identified. These vehicles can then be evaluated for possible exclusion from the TOE or replacement by a type vehicle that has the prerequisite usage pattern. The cluster set can then be identified as representative operational usage patterns for that particular vehicle. These patterns then revert to historical data available for future evaluations. This situation would also apply to vehicles which have been in the inventory for some time but for which operational usage patterns or historical data are not available.

The second case follows directly from the first,

since historical data is now assumed available. This data of usage patterns can be utilized as a new baseline to assist in determining new cluster set from more recent field collected operational data. This procedure should identify misassigned vehicles more precisely. For example, if the baseline partition or cluster set exhibits six clusters, and the new partition reveals seven clusters then this extra cluster (outlier) could possibly represent vehicles which are not being used correctly or are misassigned. Of course as mentioned in Chapter III this extra cluster could be an outlier cluster due to other reasons, such as poorly recorded data. These particular vehicles can be evaluated for possible exclusion from the TOE or replacement by a more suitable vehicle type. A vehicle can be determined more suitable by comparing the recently obtained usage pattern with the complete file of usage patterns to establish a more accurate match. This identified vehicle then becomes a candidate to replace the misassigned vehicle.

The new cluster set becomes the new operational usage patterns which then become the new historical data. It is intended that this new data is "better" or more representative of that particular vehicle population. It is hypothesized that after two or three updates of the operational usage patterns that they will have obtained a high degree of accuracy. Consequently, that particular vehicle population need only be sampled periodically (i.e.

5-10 years) to determine if the usage patterns are still representative.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Limitations of the Research

The research which has been accomplished has been limited by the available data. The data was supplied by AMSAA as previously mentioned. The author had no control over the preparation of the collection form nor over the actual collection of the data. As a consequence, the data has been taken as is with no attempts made to discover nor discuss the validity of this data.

The research has also been limited to readily available computer programs. The author lacks the time and the expertise to develop and test sophisticated computer programs. However, the author has a complete understanding of the techniques underlying those computer programs discussed and utilized in this research.

The author has further limited himself to techniques which do not require the assumption of multivariate normality. The testing for multivariate normality and techniques of data transformation necessary to induce multivariate normality are available. However, usage of such techniques is beyond the scope of this research. The author feels that an assumption of multivariate normality is meaningless in this

case. Therefore, multivariate normality should be shown to exist if such is needed.

Conclusions

It is concluded from researching the multivariate field that cluster analysis provides the best technique for discovering operational usage patterns from a data set. It is further concluded that MacQueen's convergent K-means method is the optimal method to use based on the criteria of simplicity and low cost. In addition, this method does produce the optimal partition when constrained by the criteria set forth by the author to delineate the "best" partition. And finally, the resultant methodology created by this research produces valid results in the form of an optimal partition which when interpreted provides the required operational usage patterns.

Recommendations

Several recommendations can be made for further research using additional cluster analysis techniques which could provide additional confirmation of the partition established by this research. In light of this it is recommended that the sums of squares criteria be applied to the partition to determine if any significant change would result. This could be accomplished using McRae's MIKCA program. In addition it is recommended that test of multivariate normality be made on the data to determine if

in fact such condition exists. If such does not exist then transformation should be applied to induce normality. This would be done in order that certain statistical analysis might be accomplished to establish additional evidence that the clusters are in fact representative of the sample population. The partition that has been determined by this methodology should be cross validated. As a consequence, it is recommended that additional data be collected on the same vehicle in order that the methodology might be applied to a different sample population to provide cross validation.

It is recommended that this methodology be implemented by the Army in a limited case to determine its feasibility in application.

APPENDICES

INSTRUCTIONS FOR FILLING OUT REVERSE SIDE OF THIS FORM -

Item 1. If more than one driver uses the vehicle, only the first driver of the day needs to be recorded; if the vehicle does not have an assigned driver, the individual using the vehicle the most or the immediate supervisor or section chief may be recorded.

Item 2. Any vehicle in the 3/4 - 1-1/4 ton class must be recorded; or any vehicle used as a substitute vehicle must be recorded. Use model number if known, in any event provide enough information to adequately describe the vehicle. If a commercial type vehicle from a transportation motor pool is used, describe the vehicle, for example: "1/2 ton commercial pick-up."

Item 3. It is important that the company and battalion or separate unit designation be recorded here, as well as the parent unit such as Infantry, Artillery, Transportation, for example: "Co., 12nd Inf., 1st Bn., Detach., 58th Trans Bn.," "Hq & Hq Co., 175th Armored Cav Sqdn."

Item 4. Describe the section of your company or battalion, for example: "wire section," "S-4 section," "maintenance section," "postal section," etc.

Item 5. Put down beginning mileage for this date from the odometer (mileage gauge) of the vehicle.

Item 6. Circle "yes" or "no" to describe whether or not vehicle has winch.

Items 7 thru 16. Under destination, briefly state where you are going, such as: "post office," "Brigade Headquarters," "main post," "ration break down," "hospital," etc. Under purpose, state why you are making the trip, such as: "to pick up mail," "to pick up laundry," "message run," "transport troops," "administrative," etc. Under cargo, describe briefly, such as: "60 pounds mail," "4 troops," "600 pounds commo equip," "carry tool box," "none," etc. Under estimated mileage, put down the mileage by type of road traveled as best you can determine. For example, your first trip might look like this: Paved road 12, Unpaved road 2, Trail 0, Cross Country 0. A description of each type of road is as follows:

a. Paved road - any paved 4-lane, 2-lane, or single-lane road.

b. Unpaved road - any road normally used by vehicles which is not paved.

c. Trail - a vehicle passage-way, not considered to be A or B above; could be old logging road, wagon track in field or pasture, a farm tractor path, or passage-way recently made to accommodate vehicles. It is recognizable as a vehicle passage-way.

d. Cross-country - no recognizable road or passage-way; not considered A or B or C above.

Item 17. If the vehicle was used to tow anything, circle "yes" and describe the object which was towed, such as: 3/4 ton trailer, "disabled 1/4 ton truck," "water trailer," etc. Also, estimate the miles traveled with towed load.

Item 18. Check "a" or "b" or "c" or "d" or "e"; if you check "e" then explain briefly why the vehicle was not used on this date, such as: "used as display," "used as supply vehicle," "used as an office in field," etc.

Items 19, 20, and 21: self explanatory; be sure to circle either the "Yes" or the "No."

Item 22. If vehicle has four-wheeled drive, circle the "Yes," if not, circle "No."

Item 23. If the four wheeled-drive was used circle "Yes," if not circle "No." (Circle "Yes" if four-wheeled drive was used for any purpose.)

Item 24. If at any time during the daily 24-hour report period the vehicle was driven with black-out (BO) lights only, estimate the miles driven in the blank space provided.

Item 25. Record the reading from the odometer (mileage gauge) after finishing with the vehicle for the day.

Other Information -

(1) Record the date by day, month, and year; for example: 28 Sep 73.

(2) Record the Vehicle USA # as it is on the vehicle; if this number is not available, the bumper number may be used. In any event, use the same number throughout the period of the exercise.

(3) If items 7 thru 16 are completely used up and the vehicle is still in use for the day, additional sheets may be used and stapled together.

(4) Any unusual circumstances not provided for on the front of this form may be reported under remarks below.

REMARKS

APPENDIX B

This Appendix includes a computer program for implementing MacQueen's convergent K-means nearest centroid sorting method.

Program DRIVER is the dummy main program which sets initial dimensions and calls subroutine EXEC. Subroutine EXEC includes input specifications, computes storage allocations, and calls other program segments. Subroutine KMEAN performs the actual clustering and calls function DIST to compute distances between a data unit and a seed point. Subroutine RESULT prints the cluster membership lists and the mean vector for each cluster in the final partition. The user supplies three of the program segments: (1) program DRIVER, (2) subroutine DIST, and (3) subroutine USER, which reads the scores on all variables for one data unit.

In addition program POSTDU is provided to assist in the analysis of a given partition. This program reorders the original data and computes summary statistics. This program takes as input the original data and a sequence list. This list can be punched on cards or saved on tape from subroutine RESULT. The program is limited to a maximum of 50 clusters and 10 variables.

SUBROUTINE EXEC(X,LIMIT)

C
C THIS SUBROUTINE READS PARAMETERS, COMPUTES STORAGE AND CALLS MAJOR
C PROGRAM SEGMENTS NEEDED FOR A NON-HIERARCHICAL CLUSTERING JOB
C USING SUBROUTINE *K-MEAN*.

C
C EVERY JOB REQUIRES THREE USER SUPPLIED DECK SEGMENTS.

- C
C 1. PROGRAM *DRIVER* PERFORMS THE FOLLOWING TASKS.
C A. ESTABLISHES THE DIMENSION OF THE *X* ARRAY AND SETS THIS
C DIMENSION TO *LIMIT*
C B. CALLS SUBROUTINE *EXEC*.

C
C THE FOLLOWING EXAMPLE IS USED IN THIS CASE.

C
C PROGRAM DRIVER

C
C DIMENSION X(5000)
C LIMIT=5000
C CALL EXEC(X,LIMIT)
C END

- C
C 2. SUBROUTINE *USER* IS EMPLOYED TO READ THE COMPLETE SET OF SCORES
C ON THE VARIABLES FOR ONE DATA UNIT. THE FOLLOWING EXAMPLE
C IS USED IN THIS CASE. IT IS POSSIBLE TO MERGE FILES AND
C TRANSFORM VARIABLES IN THIS SUBROUTINE.

C
C SUBROUTINE USER(X)
C DIMENSION X(5)
C READ(5,200,END=999) (X(I),I=1,5)
C RETURN
C 200 FORMAT(5F10.2,30X)
C 999 END

- C
C 3. FUNCTION *DIST* COMPUTES THE DISTANCE BETWEEN TWO DATA UNITS OR
C BETWEEN A DATA UNIT AND A CLUSTER CENTROID. THE USER CAN SPECIFY
C ANY DESIRED DISTANCE FUNCTION AND WEIGHT THE VARIABLES IN ANY
C MANNER. THE FOLLOWING EXAMPLE IS USED IN THIS CASE AND
C ILLUSTRATES THE SQUARED EUCLIDIAN DISTANCE BETWEEN TWO DATA UNITS
C DENOTED AS X AND Y.

C
C FUNCTION DIST(X,Y)
C DIMENSION X(1),Y(1)
C DIST=0.
C DO 10 I=1,5
C DIST=DIST+(X(I)-Y(I))**2
C RETURN
C END

C
C NOTE THAT SCALING AND TRANSFORMATION CAN BE ACCOMPLISHED
C EITHER IN SUBROUTINE *USER* OR IN FUNCTION *DIST*.

C
C -----

C INPUT SPECIFICATIONS

C

C CARD 1 TITLE

C CARD 2. PARAMETER CARD

C CØLS 1- 5 NE=NUMBER ØF ENTITIES (DATA UNITS)

C CØLS 6-10 NV=NUMBER ØF VARIABLES

C CØLS 11-15 NC=NUMBER ØF CLUSTERS

C CØLS 16-20 NTIN=INPUT UNIT FØR DATA SET

C NTIN=5, CARD READER

C NTIN.NE.5, TAPE ØR DISK FILE

C CØLS 21-25 NTØUT=OUTPUT UNIT FØR SAVING CLUSTER MEMBERSHIP LISTS

C NTØUT=1, CARD PUNCH

C NTØUT=8, OUTPUT TØ TAPE ØR DISK FILE

C CØLS 26-30 MINREL=TERMINATION PARAMETER. CLUSTERING ENDS WHEN A

C CYCLE THROUGH THE DATA SET RESULTS IN *MINREL*

C ØR FEWER CHANGES IN CLUSTER MEMBERSHIPS

C MINREL.LE.0, ITERATE TØ COMPLETE CONVERGENCE

C CØLS 31-35 IPART=1, SEED PØINTS ARE SELECTED FØM THE DATA UNITS.

C READ THE SEQUENCE NUMBERS FØR THE CHØSEN DATA

C UNITS FØM CARD(S) 3 IN 2Ø14 FØRMAT. IF THE

C DATA SET IS NOT STØRED IN CORE, THE LIST ØF

C SEQUENCE NUMBERS MUST BE IN ASCENDING ØRDER

C IPART=2, THE DATA UNITS ARE GRØUPED INTO AN INITIAL

C PARTITION IN THE INPUT SEQUENCE WITH THE

C FIRST *NUMBR(1)* IN CLUSTER 1, THE NEXT

C *NUMBR(2)* IN CLUSTER 2, ECT. READ THE

C *NUMBR* ARRAY FØM CARD(S) 3 IN 2Ø14 FØRMAT.

C IPART=3, THE SCORE VECTØRS FØR THE SEED PØINTS ARE

C READ FØM CARD(S) 4 IN FØRMAT *FMT* WHICH IS

C READ FØM CARD 3.

C CØLS 36-40 METHØD=PARAMETER FØR CHØØSING THE ALGORITHM IN ØNE

C VERSION ØF SUBRØUTINE *K-MEAN*

C METHØD=0, MACQUEEN ALGORITHM

C METHØD=1, JANCEY ALGORITHM

C METHØD.NE.1, FØRGY ALGORITHM

C

C NOTE THAT ØNLY THE MACQUEEN ALGORITHM HAS BEEN USED IN THIS CASE.

C

C ***CARDS 3 AND 4 ARE READ IN SUBRØUTINE *K-MEAN* ACCØRDING TØ THE

C ***PRØCEDURE SPECIFIED BY THE CHØSEN VALUE ØF *IPART*.

C

C

C STORAGE ALLØCATIONS IN THE *X* ARRAY

C X(N1) TØ X(N2-1) NC*NØ WORDS--STØRAGE IN THE CENTR ARRAY

C X(N2) TØ X(N3-1) NC WORDS--STØRAGE ØF THE NUMBR ARRAY

C X(N3) TØ X(N4-1) NE WORDS--STØRAGE ØF THE MEMBR ARRAY

C X(N4) TØ X(N5-1) NC*NØ WORDS--STØRAGE ØF THE TØTAL ARRAY

C X(N5) TØ X(N6) NV ØR NV*NE WORDS--STØRAGE ØF THE DATA ARRAY

C X(N6) TØ X(N7) NE WORDS--STØRAGE ØF THE LIST ARRAY IN *RESULT*

C

```

DIMENSION X(1),TITLE(20)
READ(5,1000) TITLE
READ(5,1100) NE,NV,NC,NTIN,NTOUT,MINREL,IPART,METHOD
WRITE(6,2000) TITLE
WRITE(6,2100) NE,NV,NC,NTIN,NTOUT,MINREL,IPART,METHOD
N1=1
N2=N1+NC+NV
N3=N2+NC
N4=N3+NE
N5=N4+NC+NV
C *N6* MAY BE INCREASED IN *K-MEAN*.
N6=N5+NV-1
N7=N4+NE-1
MAX=N6
IF(N7.GT.MAX) MAX=N7
WRITE(6,2200) MAX,LIMIT
IF(MAX.GT.LIMIT) STOP
CALL KMEAN(X(N1),X(N2),X(N3),X(N4),X(N5),N5,NE,NV,NC,NTIN,MINREL,
IPART,METHOD,LIMIT)
CALL RESULT(X(N1),X(N2),X(N3),X(N4),TITLE,NE,NV,NC,NTOUT)
RETURN
1000 FORMAT(20A4)
1100 FORMAT(8I5)
2000 FORMAT(1H1,20A4)
2100 FORMAT(5H NE =,18,/,5H NV =,18,/,5H NC =,18,/,7H NTIN =,16,/,
18H NTOUT =,15,/,9H MINREL =,14,/,8H IPART =,15,/,9H METHOD =,14)
2200 FORMAT(19H REQUIRED STORAGE =,15,6H WORDS,/,
1 19H ALLOTTED STORAGE =,15,6H WORDS)
END

```

```

      SUBROUTINE RESULT(CENTR,NUMBR,MEMBR,LIST,TITLE,NE,NV,NC,NTOUT)
C   THIS SUBROUTINE PRINTS THE RESULTS FROM A CLUSTERING JOB BASED
C   ON SUBROUTINE *K-MEAN*
C
      DIMENSION CENTR(1),NUMBR(1),MEMBR(1),LIST(1),TITLE(20)
C
C   AS A CONTINGENCY PRECAUTION WRITE OUT THE RAW MEMBERSHIP LIST.
      WRITE(6,2000) TITLE
      WRITE(6,2100) (MEMBR(K),K=1,NE)
      WRITE(6,2200) (NUMBR(J),J=1,NC)
C
C   INVERT THE *MEMBR* ARRAY AND PUT THE RESULT IN THE *LIST* ARRAY.
C   FIRST REVISE THE *NUMBR* ARRAY TO CONTAIN START POINTS IN THE
C   *LIST* ARRAY FOR EACH CLUSTER.
      NUMBR(NC)=NE-NUMBR(NC)+1
      JJ=NC
      JJ1=JJ-1
      DO 10 J=2,NC
      NUMBR(JJ1)=NUMBR(JJ)-NUMBR(JJ1)
      JJ=JJ1
10    JJ1=JJ-1
C   BUILD *LIST* ARRAY
      DO 20 K=1,NE
      MEMBRK=MEMBR(K)
      NJ=NUMBR(MEMBRK)
      LIST(NJ)=K
      NUMBR(MEMBRK)=NUMBR(MEMBRK)+1
20    CONTINUE
C   SAVE THE SORTED MEMBERSHIP LIST IF DESIRED
      IF(NTOUT.LE.0) GO TO 30
      WRITE(NTOUT,3000) TITLE
      WRITE(NTOUT,3100) (LIST(K),K=1,NE)
C   RESTORE THE *NUMBR* ARRAY
30    JJ=NC
      DO 40 J=2,NC
      NUMBR(JJ)=NUMBR(JJ)-NUMBR(JJ-1)
      JJ=JJ-1
40    NUMBR(1)=NUMBR(1)-1
C   PRINT RESULTS FOR EACH CLUSTER
      WRITE(6,2000) TITLE
      K1=1
      DO 50 J=1,NC
      WRITE(6,2300) J,NUMBR(J)
      J1=(J-1)*NV
      WRITE(6,2400) (CENTR(J1+I),I=1,NV)
      K2=K1+NUMBR(J)-1
      WRITE(6,2500) (LIST(K),K=K1,K2)
      K1=K2+1
50    CONTINUE
      RETURN
2000  FORMAT(1H1,20A4)
2100  FORMAT(20HORAW MEMBERSHIP LIST,/, (1X,25I5))
2200  FORMAT(14HOCCLUSTER SIZES,/, (1X,25I5))
2300  FORMAT(8HOCCLUSTER,13,9H CONTAINS,15,11H DATA UNITS)
2400  FORMAT(21HOCENTROID COORDINATES,/, (1X,10E12.4))
2500  FORMAT(16HOMEMBERSHIP LIST,/, (1X,25I5))
3000  FORMAT(20A4)
3100  FORMAT(20I4)
      END

```

SUBROUTINE KMEAN(CENTR,NUMBR,MEMBR,TOTAL,DATA,NS,NE,NV,NC,NTIN,
IMINREL,IPART,METHOD,LIMIT)

```

C
C THIS SUBROUTINE ITERATIVELY SORTS *NE* DATA UNITS INTO *NC* CLUSTERS
C USING THE CONVERGENT K-MEANS METHOD
C
C CENTR(NV*(J-1)+1)=SCORE ON I-TH VARIABLE FOR J-TH CLUSTER CENTROID
C TOTAL(NV*(J-1)+1)=TOTAL SCORE ON I-TH VARIABLE FOR DATA UNITS THUS
C FAR ALLOCATED TO THE J-TH CLUSTER
C NUMBR(J)=NUMBER OF DATA UNITS THUS FAR ALLOCATED TO THE J-TH CLUSTER
C MEMBR(K)=CLUSTER TO WHICH THE K-TH DATA UNIT CURRENTLY BELONGS
C DATA(NV*(K-1)+1)=SCORE ON I-TH VARIABLE FOR K-TH DATA UNIT
C
    DIMENSION CENTR(1),TOTAL(1),NUMBR(1),MEMBR(1),DATA(1),FMT(20)
    WRITE(6,2000)
C CHECK FOR SUFFICIENT STORAGE
    N6=NS+NE+NV-1
    WRITE(6,2100) N6,LIMIT
    IF(N6.GT.LIMIT) STOP
C ESTABLISH INITIAL PARTITION
    IF(IPART.NE.3) GO TO 20
C SEED POINTS ARE READ DIRECTLY FROM CARDS
    READ (5,1000) FMT
    WRITE(6,2200) FMT
    WRITE(6,2300)
    J1=0
    DO 10 J=1,NC
    READ(5,FMT) (CENTR(J1+1),I=1,NV)
    WRITE(6,2400) (CENTR(J1+1),I=1,NV)
10  J1=J1+NV
    GO TO 30
C IPART=1 OR 2
20  WRITE(6,2500) IPART
    READ(5,1100) (NUMBR(J),J=1,NC)
    WRITE(6,2600) (NUMBR(J),J=1,NC)
C READ THE DATA SET INTO CENTRAL MEMORY
30  KI=1
    DO 40 K=1,NE
    CALL USER (DATA(KI))
40  KI=KI+NV
    IF(IPART.EQ.3) GO TO 51
C IF *IPART* IS 1 OR 2 SET UP THE SEED POINTS
    IF(IPART.EQ.2) GO TO 60
C IPART=1. THE DATA UNIT WITH SEQUENCE NUMBER *NUMBR(J)* IS USED AS
C THE J-TH SEED POINT
    DO 50 J=1,NC
    NJ=(NUMBR(J)-1)*NV
    J1=(J-1)*NV
    DO 50 I=1,NV
    CENTR(J1+1)=DATA(NJ+1)
50  CONTINUE
C THE INITIAL CONFIGURATION IS GIVEN IN TERMS OF SEED POINTS.
C CONSTRUCT AN INITIAL PARTITION BY ASSIGNING EACH DATA UNIT TO THE
C NEAREST SEED POINT. SEED POINTS REMAIN FIXED THROUGHOUT ASSIGNMENT
C OF THE FULL DATA SET.

```

```

51  DO 52 K=1,NE
52  MEMBR(K)=0
    J1=0
    DO 53 J=1,NC
      NUMBR(J)=0
      DO 53 I=1,NV
        J1=J1+1
53  TOTAL(J1)=0.
C   ALLOCATE EACH DATA UNIT TO THE NEAREST SEED POINT
    K1=0
    DO 55 K=1,NE
      K2=K1+1
      J2=1
C   COMPUTE DISTANCE TO THE FIRST SEED POINT
      DREF=DIS(TDATA(K2),CENTR(J2))
      JREF=1
C   TEST DISTANCES TO THE REMAINING SEED POINTS
      DO 54 J=2,NC
        J2=J2+NV
        DTEST=DIS(TDATA(K2),CENTR(J2))
        IF(DTEST.GE.DREF) GO TO 54
        DREF=DTEST
        JREF=J
54  CONTINUE
C   ALLOCATE DATA UNIT *K* TO CLUSTER *JREF*
      NUMBR(JREF)=NUMBR(JREF)+1
      MEMBR(K)=JREF
      J1=(JREF-1)*NV
      DO 55 I=1,NV
        J1=J1+1
        K1=K1+1
        TOTAL(J1)=TOTAL(J1)+DATA(K1)
55  CONTINUE
      GO TO 85
C   IPART=2. THE DATA UNITS ARE GROUPED INTO CLUSTERS WITH THE J-TH
C   CLUSTER HAVING *NUMBR(J)* MEMBERS
60  K=0
    J1=-NV
C   ACCUMULATE THE TOTAL SCORE ON EACH VARIABLE FOR EACH CLUSTER
    DO 80 J=1,NC
      NJ=NUMBR(J)
      J1=J1+NV
      DO 70 I=1,NV
60  TOTAL(J1+I)=0.
      DO 80 KJ=1,NJ
        K=K+1
        MEMBR(K)=J
        K1=(K-1)*NV
        DO 80 I=1,NV
          J2=J1+I
          TOTAL(J2)=TOTAL(J2)+DATA(K1+I)
80  CONTINUE
C   COMPUTE THE CENTRIDS
85  J1=0
    DO 90 J=1,NC
      DO 90 I=1,NV
        J1=J1+1
        CENTR(J1)=TOTAL(J1)/NUMBR(J)
90  CONTINUE

```

```

C INITIALIZE ARRAYS
100 NPASS=1
C BEGINNING OF MAIN LOOP
120 MOVES=0
    TDIST=0
C ALLOCATE EACH DATA UNIT TO THE NEAREST CLUSTER CENTROID
    K1=0
    DO 160 K=1,NE
        K2=K1+1
        J2=1
C COMPUTE DISTANCE TO THE FIRST CLUSTER CENTROID
        DREF=DIST(DATA(K2),CENTR(J2))
        JREF=1
C COMPUTE DISTANCES TO THE REMAINING CLUSTER CENTRIODS
        DO 140 J=2,NC
            J2=J2+NV
            DTEST=DIST(DATA(K2),CENTR(J2))
            IF(DTEST.GE.DREF) GO TO 140
            DREF=DTEST
            JREF=J
140 CONTINUE
        TDIST=TDIST+DREF
        IF(JREF.NE.MEMBR(K)) GO TO 155
        K1=K1+NV
        GO TO 160
C REALLOCATE DATA UNIT *K* FROM *MEMBR(K)* TO CLUSTER *JREF*
155 MOVES=MOVES+1
        J2=MEMBR(K)
        NUMBR(J2)=NUMBR(J2)-1
        NUMBR(JREF)=NUMBR(JREF)+1
        MEMBR(K)=JREF
        J1=(JREF-1)*NV
        J3=(J2-1)*NV
        DO 150 I=1,NV
            J1=J1+1
            J3=J3+1
            K1=K1+1
            TOTAL(J1)=TOTAL(J1)+DATA(K1)
            CENTR(J1)=TOTAL(J1)/NUMBR(JREF)
            TOTAL(J3)=TOTAL(J3)-DATA(K1)
            CENTR(J3)=TOTAL(J3)/NUMBR(J2)
150 CONTINUE
160 CONTINUE
C ALL DATA UNITS ARE ALLOCATED. TEST FOR CONVERGENCE
    WRITE(6,2700) MOVES,NPASS,TDIST
    NPASS=NPASS+1
    IF(MOVES.LE.MINREL) RETURN
    GO TO 120
1000 FORMAT(20A4)
1100 FORMAT(20I4)
2000 FORMAT(46HOCONVERGENT K-MEANS METHOD OF CLUSTER ANALYSIS,/,
1 24H DATA SET STORED IN CORE)
2100 FORMAT(19HOREQUIRED STORAGE =,15,6H WORDS,/,
1 19HALLOTTED STORAGE =,15,6H WORDS)
2200 FORMAT(7HOFORMAT,20A4)
2300 FORMAT(43HINITIAL CLUSTER CENTERS READ IN AS FOLLOWS///)
2400 FORMAT(1X,10E12.4)
2500 FORMAT(9HI IPART =,12,30H, NUMBR ARRAY READ AS FOLLOWS///)
2600 FORMAT(1X,10I7)
2700 FORMAT(1H0,15,37H DATA UNITS MOVED ON ITERATION NUMBER,13,/,
138H SUMMED DEVIATIONS ABOUT SEED POINTS =,E16.8)
END

```

```

C      PROGRAM POSTDU(INPUT,OUTPUT,TAPES=INPUT,TAPE6=OUTPUT,TAPE1)
C
C THIS PROGRAM IS DESIGNED TO ASSIST IN THE INTERPRETATION OF
C CLUSTERED DATA UNITS. ORIGINAL DATA IS PERMUTED TO THE SEQUENCE
C APPEARING IN THE HIERARCHICAL TREE (OR ANY OTHER SEQUENCE THE
C USER WISHES TO SPECIFY). CLUSTERS ARE IDENTIFIED BY SIMPLY STATING
C THE NUMBER OF DATA UNITS IN EACH CLUSTER, SAY N1,N2, ETC. THEN
C THE FIRST N1 UNITS IN THE SEQUENCE LIST ARE IN THE FIRST CLUSTER,
C THE NEXT N2 UNITS IN THE SECOND CLUSTER AND SO FORTH. EACH CLUSTER
C IS DESCRIBED BY A LISTING OF ITS DATA UNITS, THEIR SCORES ON
C SELECTED VARIABLES AND SUMMARY STATISTICS. THE PRINTED OUTPUT IS
C LIMITED TO 10 VARIABLES EACH RUN. IF MORE THAN 10 VARIABLES ARE
C OF INTEREST, SIMPLY PARTITION THE VARIABLES INTO SUBSETS AND RUN THE
C PROGRAM FOR EACH SUBSET.
C-----
C INPUT SPECIFICATIONS
C
C CARD 1 TITLE CARD
C
C CARD 2 PARAMETER CARD
C   COLS 1- 4 NE=NUMBER OF ENTITIES (DATA UNITS)
C   COLS 5- 6 NV=NUMBER OF VARIABLES (MAX 10)
C   COLS 7- 8 NC=NUMBER OF CLUSTERS (MAX 50)
C   COLS 9-10 NTIN=INPUT UNIT FOR DATA
C
C CARD 3 LABEL CARD FOR VARIABLES. A 4 CHARACTER LABEL IS REQUIRED
C FOR EACH VARIABLE (10A4 FORMAT)
C
C CARD(S) 4 LABEL CARDS FOR DATA UNITS. THERE ARE TWO OPTIONS
C 1. INCLUDE 1 CARD WITH THE 4 CHARACTERS *N0LB* IN COLUMNS 1-4.
C    UNDER THIS OPTION LABELS ARE NOT PRINTED ON THE TREE OUTPUT.
C
C 2. INCLUDE *NE* CARDS, COLUMNS 1 TO 20 CONTAINING A LABEL FOR ONE
C    DATA UNIT.
C
C CARD(S) 5 SEQUENCE LIST FOR DATA UNITS (2014 FORMAT). USE AS MANY
C CARDS AS NECESSARY TO LIST *NE* DATA UNITS. THIS LIST MAY BE
C PUNCHED IN SUBROUTINE *TREE* AS PART OF A HIERARCHICAL CLUSTERING JOB
C OR IN SUBROUTINE *RESULT* AS PART OF A NON-HIERARCHICAL CLUSTERING
C JOB.
C
C CARD(S) 6 NUMBER OF DATA UNITS IN EACH CLUSTER (2014 FORMAT). USE
C AS MANY CARDS AS NECESSARY TO LIST THE SIZE OF THE *NC* CLUSTERS
C WHOSE MEMBERS ARE ORDERED IN THE SEQUENCE LIST OF CARD 6.
C
C CARD 7 FORMAT FOR PRINTING DATA ON OUTPUT. GIVE FORMAT FOR *NV*
C FIELDS OF 10 CHARACTERS EACH. USE ANY COMBINATION OF E, F AND G
C FIELDS. THE FORMAT IS LEFT VARIABLE SO THE NUMBER OF SIGNIFICANT
C DIGITS CAN BE CONTROLLED FOR EACH VARIABLE. BEGIN THE FORMAT
C IN COLUMN 1 WITH A LEFT PARENTHESIS AND END WITH A RIGHT PARENTHESIS.
C
C CARD 8 FORMAT FOR READING DATA
C
C CARD(S) 9 ORIGINAL DATA (IF ON CARDS)
C-----

```

```

C  VARIABLES IN THE PROGRAM
C  TITLE=IDENTIFYING TITLE FOR RUN
C  LABELV(I)=4 CHARACTER LABEL FOR I-TH VARIABLE
C  LABELD(I,J)=J-TH OF 5 WORDS (4 CHARACTERS EACH) LABELLING I-TH DATA
C  UNIT
C  LIST(I)=I-TH DATA UNIT IN THE SEQUENCE LIST
C  NUMBR(I)=NUMBER OF DATA UNITS IN THE I-TH CLUSTER
C  DATA(I,J)=VALUE OF J-TH VARIABLE FOR I-TH DATA UNIT
C  GTOT(I)=TOTAL FOR I-TH VARIABLE OVER ENTIRE DATA SET
C  CTOT(I)=TOTAL FOR I-TH VARIABLE OVER CURRENT CLUSTER
C  GSS(I)=SUM OF SQUARES FOR I-TH VARIABLE OVER ENTIRE DATA SET
C  CSS(I)=SUM OF SQUARES FOR I-TH VARIABLE OVER CURRENT CLUSTER
C  DIMENSION TITLE(20),FMT(24),NUMBR(50),FMTD(20)
C  DIMENSION LABELV(10),GTOT(10),GSS(10),CTOT(10),CSS(10)
C  THE FOLLOWING DIMENSION STATEMENT IS SET TO HANDLE 1000 DATA UNITS
C  DIMENSION LABELD(5,1000),LIST(1000),DATA(10,1000)
C  DATA FMT1,FMT2A,FMT2B,FMT3,FMT4A,FMT4B/
C  A4H(1X,,4H5A4,,4H20X,,4H15.2,4HX, ,4H(28X/
C  INTEGER FIRST
C  FMT(1)=FMT1
C  FMT(2)=FMT2B
C  FMT(3)=FMT3
C  FMT(4)=FMT4A
C  READ(5,1000) TITLE
C  WRITE(6,2000) TITLE
C  READ(5,1100) NE,NV,NC,NTIN
C  WRITE(6,2100) NE,NV,NC,NTIN
C  READ(5,1000) (LABELV(I),I=1,NV)
C  READ(5,1000) (LABELD(I,1),I=1,5)
C  IF(LABELD(1,1).EQ.4HNO LB) GO TO 20
C  READ REMAINING LABELS
C  DO 10 J=2,NE
10  READ(5,1000) (LABELD(I,J),I=1,5)
C  FMT(2)=FMT2A
20  READ(5,1200) (LIST(I),I=1,NE)
C  READ(5,1200) (NUMBR(I),I=1,NC)
C  WRITE(6,2200) (I,NUMBR(I),I=1,NC)
C  READ(5,1300) (FMT(I),I=5,24)
C  WRITE(6,2300) (FMT(I),I=1,24)
C  READ(5,1000) (FMTD(I),I=1,20)
C  WRITE(6,2300) (FMTD(I),I=1,20)
C  READ DATA SET
C  DO 25 J=1,NE
25  READ(NTIN,FMTD) (DATA(I,J),I=1,NV)
C  INITIALIZE GRAND STATISTICS FOR THE ENTIRE DATA SET
C  DO 30 I=1,NV
30  GTOT(I)=0.
C  GSS(I)=0.
C  COMPUTE STATISTICS FOR EACH CLUSTER AND PRINT RESULTS
C  LAST=0
C  DO 90 IC=1,NC
C  FIRST=LAST+1
C  NEC=NUMBR(IC)
C  LAST=LAST+NEC
C  DO 40 I=1,NV

```

```

CTOT(I)=0.
40  CSS(I)=0.
    WRITE(6,2000) TITLE
    WRITE(6,2400) IC,NEC
    WRITE(6,2500)
    WRITE(6,2600) (LABELV(I),I=1,NV)
    DO 70 J=FIRST, LAST
    JE=LIST(J)
    DO 50 I=1, NV
    CTOT(I)=CTOT(I)+DATA(I,JE)
50  CSS(I)=CSS(I)+DATA(I,JE)**2
    IF(FMT(2).EQ.FLB) GO TO 60
C   NO LABELS
    WRITE(6,FMT) JE,(DATA(I,JE),I=1,NV)
    GO TO 70
C   WITH LABELS
60  WRITE(6,FMT) (LABELD(I,JE),I=1,5),JE,(DATA(I,JE),I=1,NV)
70  CONTINUE
C   UPDATE GRAND STATISTICS AND PRINT CLUSTER STATISTICS
    DO 80 I=1, NV
    GTOT(I)=GTOT(I)+CTOT(I)
    GSS(I)=GSS(I)+CSS(I)
    CTOT(I)=CTOT(I)/NEC
80  CSS(I)=CSS(I)/NEC-CTOT(I)**2
    WRITE(6,2700) (CTOT(I),I=1,NV)
    WRITE(6,2800) (CSS(I),I=1,NV)
    FMT(4)=FMT4A
90  CONTINUE
C   PRINT GRAND STATISTICS
    WRITE(6,2000) TITLE
    WRITE(6,2600) (LABELV(I),I=1,NV)
    DO 100 I=1, NV
    GTOT(I)=GTOT(I)/NE
100  GSS(I)=GSS(I)/NE-GTOT(I)**2
    WRITE(6,2700) (GTOT(I),I=1,NV)
    WRITE(6,2800) (GSS(I),I=1,NV)
1000 FFORMAT(20A4)
1100 FFORMAT(14,3I2)
1200 FFORMAT(20I4)
1300 FFORMAT(1X,20A4)
2000 FFORMAT(1H1,20A4)
2100 FFORMAT(5H ONE =,18,/,5H NV =,18,/,5H NC =,18,/,7H NTIN =,16)
2200 FFORMAT(21H SIZE OF EACH CLUSTER,/, (1X,2I10))
2300 FFORMAT(7H OFFORMAT,24A4)
2400 FFORMAT(8H CLUSTER,13,11H CONTAINING,14,12H DATA UNITS.)
2500 FFORMAT(11H DATA UNITS,13X,2HID,2X,20H SCORES ON VARIABLES)
2600 FFORMAT(28X,10(6X,A4))
2700 FFORMAT(6H MEANS,22X,10(E10.3))
2800 FFORMAT(10H VARIANCES,18X,10(E10.3))
2900 FFORMAT(31H STATISTICS FOR ENTIRE DATA SET)
    END

```

APPENDIX C

Table 6. Comparison of Results

g	Method 1. Random Partition							DIST BETWN CLUST	DIST BETWN MTHDS	DIST BETWN CLUST	Method 2. Random Seed Points					
	K	NUM	HB	SB	CC	LOAD	PERS				NUM	HB	SB	CC	LOAD	PERS
1	9	106	-.37	-.21	-.21	-.59	-.56	.007*	**	.012*	140	-.72	-.18	-.18	-.78	-.42
	9	207	-.29	-.19	-.20	-.83	-.56	***	**	***	117	.24	-.23	-.26	-.83	-.39
	8	179	-.56	-.21	-.20	-.80	-.50	.080	.092	.002	221	-.26	-.20	-.21	-.84	-.51
	7	231	-.28	-.21	-.21	-.80	-.51	.013	0	.013	231	-.28	-.21	-.21	-.80	-.51
	6	265	-.32	-.21	-.21	-.78	-.42	.013	.0002	.013	264	-.33	-.21	-.21	-.78	-.41
	5	312	-.33	-.21	-.22	-.74	-.23	.040	.018	.099	330	-.35	-.21	-.20	-.73	-.10
2	9	58	-.49	-.21	-.01	1.83	-.23	.012*								
	9	149	-.45	-.21	-.11	.59	-.25	***	.019	.001	143	-.38	-.23	-.09	.56	-.33
	8	195	-.29	-.21	-.10	1.12	-.29	.0007	.306	.301	138	-.39	-.24	-.10	.58	-.35
	7	199	-.27	-.20	-.11	1.11	-.29	.013	0	.013	199	-.27	-.20	-.11	1.11	-.29
	6	220	-.32	-.18	-.11	1.05	-.21	.013	0	.013	220	-.32	-.18	-.11	1.05	-.21
	5	230	-.28	-.17	-.06	1.07	-.21	.005	.009	.004	226	-.32	-.19	-.11	1.04	-.15
3	9								**	.290	89	-.50	-.06	-.16	-.26	1.15
	8	94	-.56	-.13	-.17	-.22	.65	.013	.004	.013	103	-.54	-.15	-.20	-.35	.63
	7	108	-.56	-.15	-.19	-.32	.60	.013	0	1.58	108	-.56	-.15	-.19	-.32	.60
	6	94	-.38	-.09	-.15	-.32	1.84	1.58	0		94	-.38	-.09	-.15	-.32	1.84
4	9	46	2.41	-.13	-.32	-.85	-.21	.293	.242	.011	63	2.22	-.10	-.27	-.40	-.21
	8	46	2.58	-.02	-.29	-.36	-.24	.231	.222	.007	69	2.12	-.09	-.27	-.40	-.18
	7	72	2.11	-.06	-.26	-.32	-.16	0	.0006	.0005	72	2.09	-.06	-.27	-.33	-.16
	6	72	2.11	-.06	-.26	-.32	-.16	0	.001	0	73	2.09	-.06	-.27	-.34	-.18
	5	92	1.66	-.04	-1.03	-.27	-.23	.803	.770	0	73	2.09	-.06	-.27	-.34	-.18

Table 6 (concluded)

g	Method 1. Random Partition							DIST BETWN CLUST	DIST BETWN MTHDS	DIST BETWN CLUST	Method 2. Random Seed Points					
	K	NUM	HB	SB	CC	LOAD	PERS				NUM	HB	SB	CC	LOAD	PERS
5	9	29	.38	.26	4.04	.32	-.26	.007	0	0	30	.34	.24	3.99	.37	.26
	8	30	.34	.24	3.99	.37	-.26		0	0	30	.34	.24	3.99	.37	.26
	7	30	.34	.24	3.99	.37	-.26		0	0	30	.34	.24	3.99	.37	.26
	6	30	.34	.24	3.99	.37	-.26		0	0	30	.34	.24	3.99	.37	.26
	5									**		30	.34	.24	3.99	.37
6	9	19	1.49	5.41	-.32	.87	.22	0	0	0	19	1.49	5.41	-.32	.87	.22
	8	19	1.49	5.41	-.32	.87	.22		0	0	19	1.49	5.41	-.32	.87	.22
	7	19	1.49	5.41	-.32	.87	.22		0	0	19	1.49	5.41	-.32	.87	.22
	6	19	1.49	5.41	-.32	.87	.22		0	0	19	1.49	5.41	-.32	.87	.22
	5	19	1.49	5.41	-.32	.87	.22		0	**						
7	9	41	-.27	-.02	-.11	-.10	3.10	.053	**							
	8	41	-.27	-.02	-.11	-.10	3.10		0	0	41	-.27	-.02	-.11	-.10	3.10
	7	41	-.27	-.02	-.11	-.10	3.10		0	0	41	-.27	-.02	-.11	-.10	3.10
	5	47	-.30	-.05	-.08	-.14	2.88		***		41	.62	2.27	-.21	.45	2.17
8	9									79	-.08	-.12	-.12	1.9	-.18	
	8	96	.66	-.21	-.24	-.63	-.26		**	0	79	-.08	-.12	-.12	1.9	-.18
9	9	45	1.24	.13	-.23	1.13	-.14		**		20	-.04	-.11	-.19	-.09	1.42

* This distance is a measure between a splitting cluster and is based on p-1 variables.

** The distance between methods is not meaningful in this case.

*** The distance between clusters is not meaningful in this case.

Table 7. Validation Results

g	K	NUM DATA UNITS	Model 1. Random Partition					DISTANCE BETWEEN CLUSTERS	Model 2. Random Seed Points					
			NUM	HB	SB	CC	LOAD		PERS	NUM	HB	SB	CC	LOAD
1	6	350	151	-.36	-.21	-.22	-.80	-.22	154	-.33	-.19	-.21	-.70	-.24
		700	265	-.32	-.21	-.21	-.78	-.42	264	-.33	-.21	-.21	-.78	-.41
		350	154	-.33	-.21	-.21	-.71	-.24	130	-.34	-.22	-.20	-.81	-.43
	7	350	127	-.34	-.22	-.24	-.82	-.43	114	-.30	-.22	-.21	-.85	-.52
		700	231	-.28	-.21	-.21	-.80	-.51	231	-.28	-.21	-.21	-.80	-.51
		350	107	-.18	-.22	-.22	-.80	-.51	121	-.28	-.21	-.21	-.76	-.47
2	6	350	77	-.48	-.23	-.07	.58	-.26	112	-.24	-.12	-.14	1.07	-.17
		700	220	-.32	-.18	-.11	1.05	-.21	220	-.32	-.18	-.11	1.05	-.21
		350	112	-.24	-.12	-.14	1.07	-.17	110	-.37	-.23	-.09	1.04	-.24
	7	350	111	-.38	-.23	-.09	1.03	-.25	103	-.34	-.24	-.07	1.06	-.32
		700	199	-.27	-.20	-.11	1.11	-.29	199	-.27	-.20	-.11	1.11	-.29
		350	93	-.14	-.14	-.14	1.20	-.24	99	-.18	-.15	-.13	1.15	-.25
3	6	350	combined with GPS and split different					25	-.36	-.09	-.14	-.16	2.77	
		700	94	-.38	-.09	-.15	-.32	1.84	94	-.38	-.09	-.15	-.32	1.84
		350	25	-.36	-.09	-.14	-.16	2.77	48	-.34	-.18	-.14	-.32	1.81
	7	350	48	-.34	-.18	-.14	-.33	1.18	58	-.57	-.20	-.21	-.41	.78
		700	108	-.56	-.15	-.19	-.32	.60	108	-.56	-.15	-.19	-.32	.60
		350	70	-.65	-.11	-.15	-.24	.36	49	-.57	-.08	-.17	-.23	.63
4	6	350	34	2.15	-.05	-.28	-.35	-.12	35	2.16	-.01	-.28	-.27	-.13
		700	72	2.11	-.06	-.26	-.32	-.16	73	2.09	-.06	-.27	-.34	-.18
		350	38	2.02	-.14	-.18	-.45	-.23	37	2.02	-.14	-.25	-.48	-.23

Table 7 (concluded)

g	K	NUM DATA UNITS	Model 1. Random Partition					DISTANCE BETWEEN CLUSTERS	Model 2. Random Seed Points					
			NUM	HB	SB	CC	LOAD		PERS	NUM	HB	SB	CC	LOAD
4		350	35	2.16	-.01	-.28	-.27	-.13	35	2.16	-.02	-.28	-.27	-.13
	7	700	72	2.11	-.06	-.26	-.32	-.16	72	2.09	-.06	-.27	-.33	-.16
		350	37	2.05	-.13	-.18	-.43	-.21	37	2.02	-.14	-.25	-.48	-.23
6		350	combined with GP3 and split different					16	.31	.26	3.63	.38	-.20	
	7	700	30	.34	.24	3.99	.37	-.26	30	.34	.24	3.99	.37	-.26
		350	13	.26	.21	4.55	.32	-.33	14	.39	.21	4.39	.35	-.35
5		350	for some reason split cluster					16	.31	.26	3.63	.38	-.20	
	7	700	30	.34	.24	3.99	.37	-.26	30	.34	.24	3.99	.37	-.26
		350	13	.25	.21	4.55	.32	-.33	14	.39	.21	4.39	.35	-.35
6		350	11	1.35	5.32	-.32	.89	.30	11	1.35	5.32	-.32	.89	.30
	7	700	19	1.49	5.41	-.32	.87	.22	19	1.49	5.41	-.32	.87	.22
		350	8	1.69	5.55	-.32	.83	.11	8	1.69	5.55	-.32	.83	.11
7		350	11	1.35	5.32	-.32	.89	.30	11	1.35	5.32	-.32	.89	.30
	7	700	19	1.49	5.41	-.32	.87	.22	19	1.49	5.41	-.32	.87	.22
		350	8	1.69	5.55	-.32	.83	.11	8	1.69	5.55	-.32	.83	.11
8		350	no match due to group 5 split					13	.02	-.15	-.04	.26	3.83	
	7	700	41	-.27	-.02	-.11	-.10	3.10	41	-.27	-.02	-.11	-.10	3.10
		350	22	-.36	.15	-.17	-.14	2.95	22	-.36	-.15	-.17	-.14	2.95
8*	6	350	43	-.08	-.16	-.16	1.86	-.20						
9*	6	350	34	-.01	-.04	1.38	-.11	1.83						

*These groups produced when groups 3 and 5 were combined and split differently.

BIBLIOGRAPHY

1. Afifi, A. A. and Azen, S. P., Statistical Analysis, A Computer Oriented Approach, Academic Press, New York, 1972.
2. Anderberg, M. R., Cluster Analysis for Applications, Academic Press, Inc., New York, 1973.
3. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York, 1958.
4. Andrews, D. F., "A Note of the Selection of Data Transformations," Biometrika, Vol. 58, No. 2, pp. 249-254.
5. Andrews, D. F., "Plots of High-Dimensional Data," Biometrics, Vol. 28, pp. 125-136.
6. Andrews, D. F., Gnanadesikan, R. and Warner, J. L., "Transformations of Multivariate Data," Biometrics, Vol. 27, pp. 825-840.
7. Andrews, D. F., Gnanadesika, R. and Warner, J. L., "Methods for Assessing Multivariate Normality," Multivariate Analysis III, Paruchuri R. Krishnaian, editor, Academic Press, Inc., New York, 1973.
8. Andrews, F. M. and Messenger, R. C., Multivariate Nominal Scale Analysis, The University of Michigan, Ann Arbor, 1973.
9. AR 71-2, dated 1 June 1974, Force Development Basis of Issue Plan, Headquarters, Department of the Army, Washington, D. C.
10. AR 310-31, dated 25 May 1970, Management System for Tables of Organization and Equipment (The TOE System), Headquarters, Department of the Army, Washington, D. C.
11. AR 310-34, dated 1 October 1970, Equipment Authorization Policies and Criteria, and Common Tables of Allowances, Headquarters, Department of the Army, Washington, D. C.

12. AR 611-201, dated 1 October 1973, Personnel Selection and Classification. Enlisted Career Management Fields and Military Occupational Specialty, Headquarters, Department of the Army, Washington, D. C.
13. Ball, G. H., "Data Analysis in the Social Sciences," Proceedings of Fall Joint Computer Conference, 1965, pp. 533-559.
14. Ball, G. H. and Hall, D. J., "ISODATA, A Novel Method of Data Analysis and Pattern Classification," AD 699616, Stanford Research Institute, Menlo Park, California, 1965.
15. Ball, G. H. and Hall, D. J., "PROMENADE--An On-Line Pattern Recognition System," AD 822174, Stanford Research Institute, Menlo Park, California, 1967.
16. Bolch, B. W. and Huang, C. J. Multivariate Statistical Methods for Business and Economics, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.
17. Box, G. E. P. and Cox, D. R., "An Analysis of Transformations," J. R. Statist. Soc. B, Vol. 26, pp. 211-252, 1964.
18. Burt, C., "General and Specific Factors Underlying the Primary Emotions," Report of the British Association for the Advancement of Science, Vol. 85, 1915, pp. 694-696.
19. Cacoullos, T., "Some Characterizations of Normality," SANKHYA, A, Vol. 29, pp. 399-404.
20. Cattell, R. B., Factor Analysis, Harper, New York, 1952.
21. Chen, Chi-hau, Statistical Pattern Recognition, Hayden Book Co., Inc., Rochelle Park, New Jersey, 1973.
22. Chernoff, H., "Metric Considerations in Cluster Analysis," In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, pp. 621-629, 1970, University of California Press, Berkeley, 1970.
23. Collins, LTC N. H., Wheeler, J. P., Niemeyer, W. A. and Woomert, D. E., "Mission Analysis of Army Trucks," Interim Note No. 22, U. S. Army Materiel Systems Analysis Agency, Aberdeen Proving Ground, Maryland, June, 1973.

24. Comrey, A. L., A First Course in Factor Analysis, Academic Press, New York, 1973.
25. Demiremen, F., "Multivariate Procedures and FORTRAN IV Program for Evaluation and Improvement of Classification," Computer Contribution, State Geological Survey, University of Kansas, Lawrence, Kansas, 1969.
26. Dempster, A. P., Elements of Continuous Multivariate Analysis, Addison-Wesley Publishing Co., Reading, Mass., 1969.
27. Dixon, W. J., BMD Biomedical Computer Programs, 3rd Edition, University of California Press, Berkeley, 1974.
28. Dolby, J. L., "On a Quick Method of Choosing a Transformation," Technometrics, Vol. 5, pp. 317-325.
29. Emanuel, J. C. "Mission Analysis of 1-1/4 Ton Limited Mobility Truck," Technical Report No. 96, U. S. Army Materiel Systems Analysis Agency, Aberdeen Proving Ground, Maryland, April, 1974.
30. Everitt, B., Cluster Analysis, John Wiley and Sons, Inc., 1974.
31. Fisher, L. and Van Ness, J. W., "Admissible Clustering Procedures," Biometrika, Vol. 58, No. 1, pp. 91-104.
32. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, Vol. VII, Part II, pp. 179-188, 1936.
33. Fisher, R. A., "The Statistical Utilization of Multiple Measurements," Annals of Eugenics, Vol. VIII, Part IV, pp. 376-386, 1938.
34. Fisher, R. A. "The Precision of Discriminant Functions," Annals of Eugenics, Vol. X, Part IV, pp. 122-429, 1940.
35. Fix, E. and Hodges, J. L., "Discriminatory Analysis," Project Report 12-49-004, Numbers 4 and 11, U. S. Air Force School of Aviation Medicine, Randolph Field, San Antonio, Texas, 1951.
36. Fleiss, J. L. and Zubin, J., "On the Methods and Theory of Clustering," Multivariate Behavior Research, Vol. 4, pp. 235-250, 1969.

37. Forgy, E. W., "Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications," Biometric Soc. Meetings, Riverside, California, (Abstract in Biometrics, Vol. 21, No. 3, p. 768).
38. Friedman, H. P. and Rubin, J., "On Some Invariant Criteria for Grouping Data," J. Amer. Statist. Assoc., Vol. 62, pp. 1159-1178.
39. Gnanadesikan, R. and Kettenring, J. R., "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," Biometrics, Vol. 28, pp. 81-124.
40. Gnanadesikan, R., Srivastava, J. N., Fowlkes, E. B., and Lee, E. T., Analysis and Design of Certain Quantitative Multiresponse Experiments, Pergamon Press, Inc., New York, 1971.
41. Gnanadesikan, R. and Wilk, M. B., "Data Analytic Methods in Multivariate Statistical Analysis," Multivariate Analysis II, P. R. Krishnaian, editor, pp. 593-638, Academic Press, New York, 1969.
42. Hall, D. J., Duda, R. O., and Huffman, D. A., "Development of New Pattern-Recognition Methods," F33615-71-C-1894, Stanford Research Inst., Menlo Park, California, November, 1973.
43. Harman, H. H., Modern Factor Analysis, 2nd Edition, revised, The University of Chicago Press, Chicago, 1967.
44. Holzinger, K., Statistical Resume of the Spearman Two-Factor Theory, The University of Chicago Press, Chicago, 1930.
45. Hotelling, H., "Analysis of a Complex of Statistical Variables into Principle Components," Journal of Educational Psychology, 1933, pp. 417-441, 498-520.
46. Hotelling, H., "The Most Predictable Criterion," Journal of Educational Psychology, Vol. 26, pp. 139-142.
47. Hotelling, H., "Simplified Calculation of Principle Components," Psychometrika, Vol. 1, pp. 27-35.
48. Hotelling, H., "Relations Between Two Sets of Variates," Biometrika, Vol. 28, pp. 321-377.

49. Hull, M. H. and Malave-Garcia, S., "Family of Army Vehicles Study (FAVS), Vol. II," Headquarters, U. S. Army Combat Development Command, Fort Eustis, Virginia, 1971.
50. Koichi, I., "On the Effect of Heteroscedasticity and Nonnormality Upon Some Multivariate Test Procedures," Multivariate Analysis--II, P. R. Krishnaian, editor, Academic Press, Inc., New York, 1969.
51. Jancey, R. C., "Multidimensional Group Analysis," Australian Journal of Botany, Vol. 14, No. 1, pp. 127-130.
52. Jardine, N. and Sibson, R., Mathematical Taxonomy, John Wiley and Sons, New York, 1971.
53. Kabe, D. G. and Gupta, R. P., Multivariate Statistical Inference, North-Holland Publishing Co., Amsterdam, 1973.
54. Kaiser, H. F., "The Varimax Criterion for Analytic Rotation in Factor Analysis," Psychometrika, Vol. 23, 187-200.
55. Kendall, M. G., A Course in Multivariate Analysis, Charles Griffin Co., Ltd., London, 1963.
56. Lance, G. N. and Williams, W. T., "Computer Program for Monothetic Classification ('Association Analysis')," Computer Journal, Vol. 8, No. 3, pp. 246-249.
57. Lance, G. N. and Williams, W. T., "Computer Program for Hierarchical Polythetic Classification ('Similarity Analysis')," Computer Journal, Vol. 9, No. 1, pp. 60-64.
58. Lance, G. N. and Williams, W. T., "A General Theory of Classification Sorting Strategies, Hierarchical Systems," Computer Journal, Vol. 9, No. 4, pp. 373-380.
59. Lance, G. N. and Williams, W. T., "A General Theory of Classificatory Sorting Strategies, Clustering Systems," Computer Journal, Vol. 10, No. 3, pp. 271-276.
60. Lawley, D. N. and Maxwell, A. E., Factor Analysis as a Statistical Method, Butterworth and Co., Ltd., London, 1971.

61. Ling, Robert F., Cluster Analysis, Unpublished Dissertation, Yale University, Ph.D., 1971.
62. Mahalanobis, P. C., "On the Generalized Distance in Statistics," Proceedings of the National Institute of Science, India, Vol. 12, pp. 49-55.
63. Malkovich, J. F. and Afifi, A. A., "On Tests for Multivariate Normality," Journal of the American Statistical Association, Vol. 68, No. 341, pp. 176-179.
64. Morrison, D. F., Multivariate Statistical Methods, McGraw-Hill Book Co., New York, 1967.
65. Motor Vehicle Requirements, Army in the Field, 1965-1970 (MOVER), Headquarters, U. S. Continental Army Command, October, 1960.
66. McQueen, J. B., "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Symp. Math. Statist. and Probability, 5th, Berkeley, Vol. 1, pp. 281-297, University of California Press, Berkeley.
67. McRae, D. J., "MIKCA: A FORTRAN IV Iterative K-Means Cluster Analysis Program," Behavioral Science, Vol. 16, No. 4, pp. 423-424.
68. Nie, N., Bent, D. H., and Hull, C. H., Statistical Package for the Social Sciences, McGraw-Hill Book Company, New York, 1970.
69. Organization Directorate Bulletin No. 11-74, dated 1 November 1974, Headquarters, TRADOC, Fort Monroe, Virginia.
70. Organization Directorate Bulletin No. 1-75, 1 January 1975, Headquarters, TRADOC, Fort Monroe, Virginia.
71. Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space," Philosophical Magazine, Ser. 6, Vol. 2, pp. 559-572.
72. Press, S. J., Applied Multivariate Analysis, Holt, Rinehart and Winston, Inc., New York, 1972.
73. Projected Army Requirements for General Purpose Vehicles, Headquarters, U. S. Army Combat Development Command, December, 1966, Fort Eustis, Virginia.

74. Rao, C. R. "Some Characterizations of the Multivariate Normal Distribution," Multivariate Analysis II, P. R. Krishnaian, editor, Academic Press, Inc., New York, 1969.
75. Rao, C. R., Linear Statistical Inference and its Applications, 2nd Edition, John Wiley and Sons, New York, 1973.
76. Re-examination of Operational Concepts which Influence Unit Mobility (Unit Mobility Criteria)(U), Headquarters, U. S. Army Combat Developments Command, October, 1967, Fort Eustis, Virginia.
77. Reval-Wheels, Office Assistant Chief of Staff for Force Development, Washington, D. C., 1965, (confidential).
78. Reval-Wheels Data Bank, Updated to Reflect 'Z' Force Structure, Planning Research Corporation, McLean, Virginia, 1970, (secret).
79. Rummel, R. J., Applied Factor Analysis, Northwestern University Press, 1970.
80. Sammon, Jr., J. W., "On-Line Pattern Analysis and Recognition System (OLPARS)," Report No. RADC-TR-68-263, AD 675212, Rome Air Development Center, Griffiss Air Force Base, New York.
81. Sclove, S. L. "Population Mixture Models and Clustering Algorithms," AD-758 654, Stanford University, February, 1973.
82. Sneath, P. H. A. and Sokal, R. R., Numerical Taxonomy, W. H. Freeman and Company, San Francisco, 1973.
83. Sokal, R. R. and Rohlf, F. J., Biometry, W. H. Freeman and Co., San Francisco, 1969.
84. Spearman, C., "General Intelligence Objectively Determined and Measured," American Journal of Psychology, Vol. 15, pp. 201-293.
85. Special Analysis of Wheeled Vehicles (WHEELS) (U), Phase II Report, Part A, Vol. II, (confidential), Office, Chief of Staff, U. S. Army, 1 August 1972.
86. Stoloff, P. H., "User's Guide for Generalized Factor Analysis Program," Center for Naval Analyses, Arlington, Virginia, February, 1973.

87. Tactical Mobility of Land Forces, 1971-1980 (U), Headquarters, U. S. Army Combat Developments Command, January, 1965, Fort Eustis, Virginia.
88. Tactical Vehicles (Revised), Headquarters, U. S. Army Combat Developments Command and Headquarters, U. S. Army Materiel Command, 1 April 1965.
89. Tatsuoka, M. M., "The Relationship Between Canonical Correlation and Discriminant Analysis," Educational Research Corp., Cambridge, Mass., 1953.
90. Tatsuoka, M. M., Multivariate Analysis, John Wiley and Sons, Inc., New York, 1971.
91. Thomson, G. H., "A Hierarchy Without a General Factor," British Journal of Psychology, Vol. 8, pp. 271-281.
92. Thomson, G. H., The Factorial Analysis of Human Ability, 5th Edition, Houghton Mifflin Co., Boston, 1951.
93. Thurstone, L. L., "Multiple Factor Analysis," Psychology Review, Vol. 38, pp. 406-427.
94. Thurstone, L. L., Multiple Factor Analysis, University of Chicago Press, Chicago, 1947.
95. TRADOC Memo 15-1, dated 26 July 1974, Basis of Issue Plan (BOIP), Review Board, Headquarters, TRADOC, Fort Monroe, Virginia.
96. TRADOC Memo 15-5, dated 1 July 1973, Tables of Organization and Equipment Review Board, Headquarters, TRADOC, Fort Monroe, Virginia.
97. TRADOC Regulation 71-17, dated 1 July 1973, Unit Reference Sheets, Headquarters, TRADOC, Fort Monroe, Virginia.
98. TRADOC Supplement 1 to AR 71-2, dated 10 October 1974, Force Development Basis of Issue Plan, Headquarters, TRADOC, Fort Monroe, Virginia.
99. Trivedi, S. J. and Bargmann, R., "Configuration and Classification of Clusters in N-Dimensions," AD 735 129, Georgia Univ., Athens, December, 1971.
100. Tryon, Robert C. and Bailey, D. E., Cluster Analysis, McGraw-Hill, Inc., New York, 1970.

101. Tukey, J. W., "On the Comparative Anatomy of Transformations," Ann. Math. Statist., Vol. 28, pp. 602-632.
102. Van de Geer, J. P., Introduction to Multivariate Analysis, W. H. Freeman and Co., San Francisco, 1971.
103. Ward, Jr., J. H., "Hierarchical Grouping to Optimise and Objective Function," Journal Amer. Statistician Association, Vol. 58, No. 301, pp. 236-244.
104. Ward, Jr., J. H. and Hook, M. E., "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles," Education and Psychological Measurement, Vol. 23, No. 1, pp. 69-82.
105. Wheeler, J. P. and Collins, Major N. H., "Mission Analysis of Army Trucks," Interim Note No. 10, ASAMSAA, APG, Maryland, June, 1972.
106. Wheeler, Jr., J. P. and Collins, Lt. N. H., "Mission Analysis of Army Trucks," Interim Note No. 12, AMSAA, APG, Maryland, September, 1972.
107. Wheeler, Jr., J. P., "Mission Analysis of Army Trucks," Interim Note No. 13, AMSAA, APG, Maryland, November, 1972.
108. Wishart, D., "An Algorithm for Hierarchical Classifications," Biometrics, Vol. 22, No. 1, pp. 165-170.
109. Wishart, D., "Fortran II Programs for 8 Methods of Cluster Analysis (CLUSTAN I)," Computer Contributions No. 38, State Geological Survey, Univ. of Kansas, Lawrence, 1969.
110. Wolfe, J. H., "Normix 360 Computer Program," Rep. No. SRM-72-4, AD 731037, Personnel and Training Res. Lab., San Diego, California, 1971.

E-24-626

Final Summary Report
for the
U.S. Army Material Systems Analysis Agency
Aberdeen Proving Ground, Maryland 21005
under contract
"Research Support in Operations Research/Systems
Analysis Applications to Army Needs and Objectives"

DAAD05-74-C-0777

conducted by
The School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Leslie G. Callahan Jr. Principal Investigator

August 1975

I. Nature of the Research Program

A. Background: The School of Industrial and Systems Engineering of the Georgia Institute of Technology began to offer Operations Research/Systems Analysis courses at the graduate level in the 1950's. A small number of officers and civilians from the Department of Defense who were pursuing graduate degrees in established areas enrolled in these courses. In 1969 the U.S. Army developed a core curriculum for a formal graduate in OR/SA, and selected as one of the two civilian institutions for concentrated use in meeting Army graduate educational needs in this area. In 1972, the School was authorized to award a graduate degree in operations research MSOR. A number of joint reviews have been made in improving the Army OR/SA program requirement with the latest in April, 1974. Sixteen Army personnel entered the program in 1969, and by 1973, 35 students were in residence with approximately 20 graduating a year. At present 15 are in residence with a forecasted level of 20 in residence and an output of 10 a year.

B. The Theses Problem: For almost all Master's degree candidates, the identification and definition of a Thesis topic of interest both to the student and to his research advisor requires a disproportionate amount of time when compared with the course requirements or thesis research. One of the important objectives to be realized in this program is the development of readily available research topics relevant to Army needs and objectives and potentially interesting to Army personnel, and of competent, involved research advisors. These availabilities are critical if the Army personnel are to complete an acceptable thesis within the time constraint of their tenure in the program. A review of theses by Army officers prior to 1974

indicated a small percentage related to Army needs and problem areas. This situation was highlighted by Dr. Wilbur Payne, Deputy Under Sec. of the Army in Oct. 1973 in a letter to Georgia Tech commenting on the revised curriculum programs when he stated:

"I was very interested in the comments you received from the officer students in response to your Proposal Review memorandum. Of particular interest were their remarks concerning the lack of adequate communication between the Army and students, and the resulting scarcity of appropriate military related thesis topics. This has for some time also been a concern of mine. I believe that something can be done to improve this situation, and would be delighted to work with the Institute toward that goal."

C. Theses Support Program: During the fall of 1973 and spring of 1974 a number of conferences and seminars were held between the Georgia Tech faculty and Army agencies to improve the relevancy of these theses research. In June 1974 the Army Material Systems Analysis Agency contracted to support three officers and in the fall of 1974 the U.S. Army Operational Test and Evaluation Agency agreed to sponsor seven officers under two separate contracts. These contracts support the officer students by providing office space, leased computer terminals, and other logistic support at Tech. The contracts have also covered approximately 1/4 time salaries, overhead and limited travel for three faculty members for efforts beyond what would otherwise be required for their faculty duties. In addition to contract support, the sponsoring agency provides travel support and data sources for officer student. Actual thesis topics are developed between the student, the faculty and the sponsor to assure Army relevance, academic quality and within the individual officer's capabilities.

D. General Method of Approach: Literature search and problem definition in the two areas above began in the summer of 1974. The three faculty members met frequently with individual students and began to collect background material from OTEA, USAMSAA, Command and General Staff College, the Army

Logistic Management Agency, and other Army agencies as well as from the Georgia Tech Library. Frequent seminars and conferences between all the students and faculty were held to promote development of individual thesis topics.

E. Scope of Report: This report provides a final summary for work done for the U.S. Army Material Systems Analysis Agency under contract DAAD05-74-C-0777 subject "Research Support in Operations Research/Systems Analysis Applications to Army Needs and Objectives" awarded for 14 months of theses support during the period 3 June 1974 to 2 August 1975.

II. Results

A. Under the provisions of subparagraph A and B of Section F.3 of the contract two reports were submitted to the sponsor in April 1975 (Incls).

B. Work under subparagraph C of Section F.3 of the contract has not been completed. Capt. Everett D. Lucas, Arty, who proposed this task during contract negotiations in May 1974 is not scheduled to graduate until December 1975. As specified in the last paragraph of Section F of the contract, it was anticipated that work under this task might not be completed until after expiration of the contract period. Thus this summary will only reflect work accomplished up to the end of the contract period on 2 August 1975.

(a) During the summer and fall of 1974, Lucas frequently met individually with the principal investigator on developing his research areas at the same time carrying a full course load of 15-18 hours. He participated on a limited basis in the conferences and seminars cited in ID alone because of academic difficulties in meeting grade requirements. In February, 1975 he intensified the literature search phase and established contact with

the Assistant Director of USAMSAA, and with personnel at Fort Sill. He visited USAMSAA in June 1975 for data collection and sponsor guidance.

(b) At the end of the contract period he had completed the literature search phase and narrowed the problem area down to a feasible size for a master's thesis. His methodological approach involves the adaptation of linear programming assignment procedures to the modeling and evaluation of the effectiveness of various artillery system target configurations. He has selected four cognizant faculty members to serve on his thesis advisory committee which includes the principal investigator on this contract. It is anticipated that he will complete his oral defense and first draft by 1 December 1975.

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY

Research Support in Operations Research/Systems
Analysis Applications to Army Needs and Objectives

DAAD05-74-C-0777

AN APPLICATION OF MULTIVARIATE STATISTICAL
METHODS IN DEVELOPING OPERATIONAL USAGE
PATTERNS OF U.S. ARMY VEHICLES

Leslie G. Callahan Jr.

Project Director

Randall Brannon Medlock, *Capt, Inf*

Principal Investigator

Georgia Institute of Technology

April 1975

SUMMARY

This research develops a methodology which establishes operational usage patterns of Army vehicles using field data. The field data for this research was material supplied by the U. S. Army Materiel Systems Analysis Agency. An examination of the data revealed that a number of correlated variables could be extracted. These variables described the manner in which the particular type vehicle was utilized. Since it would be ideal to examine these variables in their entirety, multivariate techniques were considered. These techniques included principal component analysis, factor analysis, discriminant analysis, canonical correlation analysis, and cluster analysis. Each of these techniques were examined in some detail to determine its suitability for producing operational usage patterns. The cluster analysis technique was chosen based on its simplicity, low cost and the ability to provide meaningful groupings of data units. A nonhierarchical clustering technique known as McQueen's convergent K-means method was selected as the most appropriate method for this case.

The data was subjected to outlier analysis techniques to eliminate multivariate outliers. In addition the data was centered by subtracting the means and standardized by dividing by the standard deviations.

It was hypothesized that the clustering should reveal between four and ten "natural" clusters. Consequently, the analysis was accomplished to produce partitionings that included five to nine clusters. The "optimum" or "best" partition was chosen based on two criteria, one of which compared cluster centroids between partitions and between different methods of selecting initial starting points (i.e. seed points) to determine at which partition clusters become most stable. The other criterion established an upper limit for the number of clusters and was based on nonoptimal splitting of stable clusters. If nonoptimal splitting occurred then it was an indication that the optimal partition existed at a smaller number of clusters. Once the "optimum" or "best" partition is determined then in fact the operational usage patterns have been established.

It is envisaged that this methodology can be effectively utilized in the area of assigning and reassigning vehicles within the Army TOE system. In particular the technique would establish a baseline of usage patterns which describes how a new vehicle is being utilized. This baseline can then be used to periodically identify misassigned vehicles. This is done by first establishing new usage patterns based on field data collected at different times during the vehicles inventory life. A comparison of these new usage patterns with the base would then identify possible outlier clusters which would indicate the possibility of

misassigned vehicles.

It is recommended that this methodology be implemented by the Army in a limited case to determine its feasibility in application.

U.S. ARMY MATERIEL SYSTEMS ANALYSIS AGENCY

RESEARCH SUPPORT IN OPERATIONS RESEARCH/

SYSTEMS ANALYSIS APPLICATIONS TO ARMY

NEEDS AND OBJECTIVES

DAAP05-74-C-0777

EVALUATION OF COMPUTERIZED LAYOUT

ALGORITHMS FOR USE IN DESIGN OF

CONTROL PANEL LAYOUTS

Leslie G. Callahan Jr.

Project Director

Samuel D. Wyman III, *Capt, Armor*

Principal Investigator

Georgia Institute of Technology

April 1975

SUMMARY

This report examines the applicability of computer-aided design to the configuration of an army helicopter instrument panel. Two facility allocation algorithms, CRAFT and PLANET, were adapted to this purpose. Rotary wing aircraft instrumentation has not kept pace with the use of the helicopter and advancements in helicopter performance. This lack of instrumentation improvement does not allow the full utilization of the unique flight characteristics of the helicopter. The instrumentation problem that is studied in this report is the arrangement of instruments on the panel by use of computer-aided design. Since the current techniques are artisan in nature, computer-aided design offers an approach that can extend the engineer's problem-solving arm. The computer algorithms which have been used are readily available from their normal use in plant layout problems and require only minor modification to be used in laying out instrument panels. The design criteria used for the panel layouts was the minimization of pilot eye movement. This criteria readily fits into the context of the computer algorithms. Results from this research indicate that facility allocation algorithms with the eye movement optimization criteria offer a powerful tool for the computer-aided design of helicopter instrument panels.