LOAN DOCUMENT

PHOTOGRAPH THIS SHEET

DTIC ACCESSION NUMBER

LEVEL

INVENTORY 0

RCS-AMCSM-156

DOCUMENT IDENTIFICATION

DEC 1991

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

DISTRIBUTION STATEMENT

HANDLE WITH CARE

ACCESSION FOR
NTIS GRA&I
DTIC TRAC
UNANNOUNCED
JUSTIFICATION

BY

DISTRIBUTION/

AVAILABILITY CODES

DISTRIBUTION	AVAILABILITY AND/OR SPECIAL	
A-1		

DISTRIBUTION STAMP

DATE ACCESSIONED

DATE RETURNED

20000203 102

DATE RECEIVED IN DTIC

REGISTERED OR CERTIFIED NUMBER

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM 70A JUN 90

DOCUMENT PROCESSING SHEET

PREVIOUS EDITIONS MAY BE USED UNTIL STOCK IS EXHAUSTED.

LOAN DOCUMENT

DTIC QUALITY INSPECTED 4

RCS AMCSM-156

DEPARTMENT OF THE ARMY SAMPLE DATA COLLECTION

SAMPLE DATA COLLECTION LOGISTICS MANAGEMENT ANALYSIS REPORT

Findings of the U.S. Army Safety Center Accident Data: A Mathematical Analysis

December 1991

Prepared for:

U.S. Army Aviation Systems Command
4300 Goodfellow Boulevard
St. Louis, Missouri 63120-1798

Prepared By:

COBRO Corporation

FOR FURTHUR INFORMATION CONCERNING DISTRIBUTION CALL (703) 767-8040

PLEASE CHECK THE APPROPRIATE BLOCK BELOW:

AQ# _____

☐ _____ copies are being forwarded. Indicate whether Statement A. B. C. D. E, F. or X applies.

☑ DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

☐ DISTRIBUTION STATEMENT B:
DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES ONLY; (Indicate Reason and Date). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office).

☐ DISTRIBUTION STATEMENT C:
DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES AND THEIR CONTRACTORS; (Indicate Reason and Date). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office).

☐ DISTRIBUTION STATEMENT D:
DISTRIBUTION AUTHORIZED TO DoD AND U.S. DoD CONTRACTORS ONLY; (Indicate Reason and Date). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office).

☐ DISTRIBUTION STATEMENT E:
DISTRIBUTION AUTHORIZED TO DoD COMPONENTS ONLY; (Indicate Reason and Date). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office).

☐ DISTRIBUTION STATEMENT F:
FURTHER DISSEMINATION ONLY AS DIRECTED BY (Indicate Controlling DoD Office and Date) or HIGHER DoD AUTHORITY.

☐ DISTRIBUTION STATEMENT X:
DISTRIBUTION AUTHORIZED TO U.S GOVERNMENT AGENCIES AND PRIVATE INDIVIDUALS OR ENTERPRISES ELIGIBLE TO OBTAIN EXPORT-CONTROLLED TECHNICAL DATA IN ACCORDANCE WITH DoD DIRECTIVE 5230.25. WITHHOLDING OF UNCLASSIFIED TECHNICAL DAT \FROM PUBLIC DISCLOSURE. 6 Nov 1984 (indicate date of determination). CONTROLLING DoD OFFICE IS (Indicate Controlling DoD Office).

☐ This document was previously forwarded to DTIC on _____ (date) and the AD number is _____

☐ In accordance with provisions of DoD instructions. the document requested is not supplied because:

☐ It will be published at a later date. (Enter approximate date. if known).

☐ Other. (Give Reason)

DoD Directive 5230.24, "Distribution Statements on Technical Documents," 18 Mar 87, contains seven distribution statements, as described briefly above. Technical Documents must be assigned distribution statements.

Cynthia Gleisberg
Print or Type Name

334-255-7924
Telephone Number

Cynthia Gleisberg
Authorized Signature/Date

SR91-206

SAMPLE DATA COLLECTION LOGISTICS MANAGEMENT ANALYSIS REPORT

Findings of the U.S. Army Safety Center Accident Data: A Mathematical Analysis

December 1991

Prepared For:

U.S. Army Aviation Systems Command (AVSCOM)
4300 Goodfellow Boulevard
St. Louis, Missouri 63120-1798

Prepared By:

COBRO Corporation
4260 Shoreline Drive
Earth City, Missouri 63045-1226

TABLE OF CONTENTS

I. Introduction 1

Section 1: Mathematical Analysis 2
A. Definitions and Terminology 2
B. Definition of Scoring 6
C. Compare Factor Analysis with Modified Factor Analysis 9
D. Jaccard Coefficient Insertion 12
E. Results of Analysis 16

Section 2: Methodology - Modified Factor Analysis . . . 19
A. Comparison of Methods Used 19
B. Comparison of Procedures 23
C. Comparison of Output 33

II. Conclusion 40

III. Recommendations 41

BIBLIOGRAPHY A-1

I. Introduction

Annually, the U.S. Army Safety Center (USASC), Fort Rucker, Alabama, analyzes approximately 20,000 reports of accidents to identify cause factors, develop countermeasures and evaluate effectiveness of fielded countermeasures. Efficient analysis of such massive data requires use of methods of data reduction to reveal the essential targets. One of the methods used by the Safety Center is factor analysis. However, the nature of much of the accident data (binary) is not amenable to the Pearsonian Product Moment Correlation Coefficient that drives the factor analysis. A substitute coefficient (Jaccard Similarity Coefficient) has been employed by USASC but the theoretical foundation for using this procedure has not been established.

In Section 1 of this report we will examine the feasibility of using similarity or matching/associative coefficient as a substitute for the correlation/covariance matrix in the factor analysis procedure. Also, examined in the report is the definition of dichotomous scoring compared to the binary data made available to the USASC.

From Section 1, it was determined that the method of data reduction had many mathematical pitfalls. A more efficient method should be utilized to reduce the accident data. In Section 2 of this report, we develop a methodology using accident data supplied by the USASC (Night Study data). This data is considered highly parsimonious in both the physical interpretation and mathematical complexity. The procedure which was investigated is the VARCLUS procedure contained in the SAS Institute INC. statistical package (SAS). It is felt that the parsimony mentioned above is minimized by using this procedure.

The VARCLUS procedure was investigated because of its usefulness in interpreting large amounts of variables. VARCLUS is a variable-reduction method and it is also useful in determining if there is a relationship between variables.

It should be noted that the analysis performed in this report was done using a Clustering Technique (Centroid Method) to duplicate Factor Analysis Procedure established by the USASC. Basically while performing this analysis, we found ourselves working towards answers already achieved. This technique allowed a thorough evaluation of the mathematical procedures used by the USASC.

Section 1: Mathematical Analysis

A. Definitions and Terminology

While investigating literature about Factor Analysis and Cluster Analysis, it became apparent that one term may have many different names. The name given to a term depends primarily on the author's background.

One such example is the correlation coefficient. In mathematics, the term is usually referred to as R. Phi (Φ) is mathematically equivalent to R when both variables are dichotomous. There are many other names for the correlation coefficient some of the more common ones, with their associated mathematical equation and references, are listed in this section.

The following is a list of correlation coefficients which can be algebraically reduced to equivalent equations. The only differences between these equations are the names and the form of the equations. This is important when reading any literature about cluster analysis and factor analysis, because these terms are used interchangeably depending upon the background of the author.

Product Moment Correlation Coefficient Pg 85 Anderburg
Cophentic Correlation Coefficient Pg 26 Romesburg

$$r=\frac{\sum_{i=1}^{n} x_i y_i-\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left\{\left[\sum_{i=1}^{n} x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]\left[\sum_{i=1}^{n} y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]\right\}^{\frac{1}{2}}}$$

Coefficient of Correlation (R) Pg 441 Freund and Smith

$$R=\frac{n\sum_{i=1}^{n} x_i y_i-\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right)-\left(\sum_{i=1}^{n} y_i\right)^2}}$$

Pearsonian Correlation Coefficient Pg 42 Basilevskey

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2]^{\frac{1}{2}}}$$

Correlation Coefficient in Standard score form Pg 27 Comrey

$$r_{ij}=\frac{\sum_{i=1}^{n}z_{ik}z_{jk}}{N}$$

$$r_{xy}=\frac{1}{N}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{[\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{N}\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{N}]^{\frac{1}{2}}}$$

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2]^{\frac{1}{2}}}$$

Phi Coefficient pg 149 Romesburg

$$c_{jk}=\frac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$$

Another source of confusion, is attributed to the similarity coefficients. Some of the coefficients have as many as three names for the same term. The Sorenson Coefficient is also known as the Dice Coefficient and Czehanowski's Coefficient. There are many similarity coefficient some of the more common ones, with their

associated mathematical equation and references, are listed.

Also, listed is the qualitative resemblance coefficients with the dissimilarity matches, the 0-0 or d cell. It was found that the coefficients reduce to the Jaccard Coefficient or the Dice Coefficient when the dissimilarities removed.

The list of coefficients below are all clustering coefficients. We are listing them here because they will be used throughout this text.

Jaccard Coefficient

$$J_{ij}=\frac{a}{a+b+c}$$

Dice Coefficient — Pg 89 Anderberg
Sorensen Coefficient — Pg 145 Romesburg
Czekanowski Coefficient — Pg 356 Seber

$$c_{ij}=\frac{2a}{2a+b+c}$$

Gower Coefficient — Pg 31 Aldenderfer & Blashfield

$$g_{ij}=\frac{\sum_{k=1}^{n} s_{ijk}}{\sum_{k=1}^{n} w_{ijk}},$$

where w_{ijk} has a value of one (1) if a comparison of variable k is considered and zero (0) if no comparison exists, and s_{ijk} has a value of one (1) if i and j are both similar. When using with dichotomous data the Gower Coefficient becomes the Jaccard Coefficient.

Many coefficients take the form of the Jaccard or Dice Coefficient when dissimilarities are not taken in consideration.

Simple Matching Coefficient:

$$c_{ij}=\frac{a+d}{a+b+c+d} \xrightarrow{d=0} \frac{a}{a+b+c}$$

Russell & Rao:

$$\frac{a}{a+b+c+d} \xrightarrow{d=0} \frac{a}{a+b+c}$$

Baroni-Urbani & Busser Pg 150 Romesburg

$$\frac{a+(ad)^{\frac{1}{2}}}{a+b+c+(ad)^{\frac{1}{2}}} \xrightarrow{d=0} \frac{a}{a+b+c}$$

Sokal & Sneath:

$$\frac{2(a+d)}{2(a+d)+b+c} \xrightarrow{d=0} \frac{2a}{2a+b+c}$$

B. Definition of Scoring

The Safety Center data, which we are observing contains, aircraft accidents. Each accident is looked at in terms of fifty-six (56) variables. Binary data is used in which one (1) indicates the presence of that particular variable and a zero (0) indicates the absence of that particular variable.

These 56 variables can be broken down into ten (10) categories.

A. Fatigue
B. Illumination
C. Aircraft
D. Mission
E. Aided
F. Problem Area
G. Pilot in Control
H. Task Error
I. Phase of Flight
J. Experience

The variables within each category are independent. For example, each individual accident can only involve one type of aircraft.

During the investigation of the Safety Center data, it was determined we are not dealing with truly dichotomous data. Dichotomous data is defined as data which is transformed to a binary set, 1 equalling the occurrence of an event and 0 being the complement of that event. Two sets of data are transformed and examined. The resultant data set is made up of four (4) independent outcomes.

VARIABLE/ EVENT 2	VARIABLE/EVENT 1	
	1	0
1	a	b
0	c	d

1 - indicates the presence of that variable/event
0 - indicates the absence of that variable/event

a - the # of times a variable/event 1 occurred and variable/event 2 occurred.
b - the # of times a variable/event 2 occurred and variable/event 2 did not occur.
c - the # of times a variable/event 1 did not occur and variable/event 2 occurred.

d - the # of times neither variable/event occurred.

The following example illustrates true dichotomous data. Variable 1 is defined as a turning error on a UH1 aircraft, and variable 2 is defined as a UH1 failure. The definitions for the a,b,c, and d terms are as follows:

a - the # of times a UH1 failed and there was a turning error.
b - the # of times a UH1 failed and there was not a turning error.
c - the # of times a UH1 did not fail but there was a turning error.
d - the # of times a UH1 did not fail and there was not a turning error.

The Safety Center utilizes binary scoring for their data. The difference between the Safety Center data and purely dichotomous data is the definition of the zero(0) terms for each variable.

The following is an example of the Safety Center data:

VARIABLE 1
1 - tuning error was made
0 - some other error was made

VARIABLE 2
1 - failure of a UH1
0 - some other aircraft failed

If the Safety Center was dealing with only two aircraft and two types of errors this scoring would be acceptable. However, the Safety Center is dealing with seven aircraft and a number of errors or other variables.

Since the Safety Center data is defined in such a way, any evaluation involving 0 - terms will be inaccurate. In the case of the c values, we are not comparing turning errors made on UH1s that did not have accidents. We are complicating the comparison away from a purely dichotomous relationship by broadening the relationship to comparing turning errors which contributed to accidents on all other aircraft. These values have no correlation with UH1s but are expected to be used in the calculation of a coefficient which will give a numerical value to the relationship between turning errors and UH1 aircraft.

When purely dichotomous data was used, it created another difficulty for this type of scoring. An example is the use of aided and unaided as two separate variables. When they were transformed into dichotomous data the definitions were as follows:

AID
1 - pilot was aided
0 - pilot was unaided

UNA
1 - pilot was unaided
0 - pilot was not unaided

The problem arises when these variables are compared. The AID - 0 is logically equivalent to the UNA - 1 , and the UNA -0 is equivalent to the AID -1. This would cause double weighing of these variables when they are compared with other variables.

C. Compare Factor Analysis with Modified Factor Analysis

After the data has been collected, the variables can then be investigated to determine which ones will be used in the analysis of the data. This procedure can be done by clustering and then some sort of analysis, or simply using factor analysis on the raw data.

If clustering is performed, the first step would be to group those variable which are closely associated with one another. There are many coefficients which achieve these correlations. One such coefficient commonly used by the Safety Center is the Jaccard Similarity Coefficient. This coefficient identifies similarities and disregards all dissimilarities. Once the clustering has been performed, analysis can be done by simply looking at the clusters or by some further statistical methods such as factor analysis.

Whatever the goals of an analysis, in most cases it will involve the following major steps: (a) selecting variables; (b) computing the matrix of correlation/covariance among the variables; (c) extracting the unrotated factors; (d) rotating factors; (e) interpreting the rotated factor matrix.

One common objective of the factor analysis is to provide a relatively small number of factor constructs that will serve as satisfactory substitutes for a much larger number of variables. The factor constructs themselves are variables that may prove to be more useful than the original variables from which they were derived. The first step would involve the decision of using a covariance matrix or a correlation matrix depending upon what the results are that the researcher is trying to obtain. Once this matrix is obtained, the number of factors should be found. This can be achieved by finding the rank of the correlation matrix. Another way to find the number of factors is to determine the eigenvalues for each variable and the number of eigenvalues which make up a predetermined percentage of the cumulative eigenvalues over the sum of all eigenvalues assigned to the variables.

Once the number of factors are specified, a linear combination of the variables can be determined for each factor. The first linear combination makes up the largest amount of variance and each combination after that is defined by a lesser amount of the variance. From these linear combinations, it can be determined which variables are responsible for the variance.

The Safety Center initially uses the Jaccard Similarity Coefficient to find variables within a specific group (ex. illumination) which occur simultaneously. A new variable is defined which consists of the two variables which exhibited a large Jaccard value.

The Safety Center then uses factor analysis to determine which

variables are important in aircraft accidents. One of the differences between the Safety Center factor analysis and what was described before is the use of the Jaccard coefficient in place of the correlation/covariance coefficient.

Jaccard Distribution Table

The Jaccard Similarity Coefficient is a ratio of the number of 1-1 matches (both variables occurring simultaneous) compared to the number of occurrences of every time at least one of the variables are present (a+b+c).

VARIABLE/ EVENT 2	VARIABLE/EVENT 1	
	1	0
1	a	b
0	c	d

1 - indicates the presence of that variable/event
0 - indicates the absence of that variable/event

The Jaccard Similarity Coefficient is defined as:

$$J_{ij}=\frac{a}{a+b+c} \quad 0\leq J_{ij}\leq 1$$

Hence, when creating the Jaccard Similarity Coefficient Table it was found that the coefficients were uniformly distributed on the integer 1,2,...,n, n being the total sample size (a+b+c).

$$f(x)=\begin{cases}\frac{1}{n} & \text{for } n=1,2,\ldots,n,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore, the total sample size (a+b+c) can be multiplied by the needed confidence level to determine what the minimum number of 1-1 matches is needed to give the tolerable sample size.

The following chart is an example of the Confidence Levels of the Jaccard Coefficients.

Chart 1.

n Sample Size (a+b+c)	1%	2.5%	5%	10%	90%	95%	97.5%	99%
2	0	0	0	0	2	2	2	2
3	0	0	0	0	3	3	3	3
4	0	0	0	0	4	4	4	4
5	0	0	0	1	5	5	5	5
6	0	0	0	1	5	6	6	6
7	0	0	0	1	6	7	7	7
8	0	0	0	1	7	8	8	8
9	0	0	0	1	8	9	9	9
10	0	0	1	1	9	10	10	10
11	0	0	1	1	10	10	11	11
12	0	0	1	1	11	11	12	12
13	0	0	1	1	12	12	13	13
14	0	0	1	1	13	13	14	14
15	0	0	1	2	14	14	15	15

The matrix shown below is an of the Jaccard Similarity Coefficient Matrix all possible combinations calculated for n = 1,2,...,10.

	Number of "a" terms									
n (a+b+c)	10	9	8	7	6	5	4	3	2	1
10	1	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
9		1	0.89	0.78	0.67	0.56	0.44	0.33	0.22	0.11
8			1	0.88	0.75	0.63	0.50	0.38	0.25	0.13
7				1	0.86	0.71	0.57	0.43	0.29	0.14
6					1	0.83	0.67	0.50	0.33	0.17
5						1	0.80	0.60	0.40	0.20
4							1	0.75	0.50	0.25
3								1	0.67	0.33
2									1	0.50
1										1

D. Jaccard Coefficient Insertion

The operation being scrutinized was accomplished by the insertion of the Jaccard Coefficient Matrix into the Factor Analysis problem at the point where the correlation matrix or covariance matrix between the variables is computed. The Jaccard Coefficient is a similarity coefficient. A similarity coefficient measures the resemblance between two (2) objects based on either or both of two (2) logically distinct kinds of information pertaining to a set of variables.

The similarity coefficient provides information on the existence or absence of the variables being compared. The use of this coefficient is amicable when comparing attributes of an object. Coefficients of this type are dichotomous. The term dichotomous is reserved for characters that are either present or absent and whose absence in both of a pair of objects is not taken as a match. This approach would be appropriate if the variables were all nominal with two (2) states, the states simply being alternatives with equal weight.

If ρ is the population of variables, then we can define a similarity as a function that maps $\rho \times \rho$ into $\mathbb{R}^1$ and satisfies the following axioms:

(1) $0 \leq C_j(h,i) \leq 1$ for all h and i in ρ, were ρ is the population of objects.
(2a) $C_j(h,h) = 1$.
(2b) $C_j(h,i) = 1$ only if $i = h$.
(3) $C_j(h,i) = C_j(i,h)$.

The Jaccard Coefficient, $C_j(h,i)$, satisfies the above axioms.

With quantitative variables, one measure of similarity between $\mathbf{x}_h$ and $\mathbf{x}_i$, the observations on objects h and i is the correlation of the pairs (x_{hk},x_{jk}), $k = 1,2,\ldots,n$, namely, the Pearsonian Correlation Coefficient,

$$r_{hi} = \frac{\sum_{k=1}^{n} (x_{hk}-\bar{x}_{h.})(x_{ik}-\bar{x}_{i.})}{\left[\sum_{k=1}^{n} (x_{hk}-\bar{x}_{h.})^2 \sum_{k=1}^{n} (x_{ik}-\bar{x}_{i.})^2\right]^{\frac{1}{2}}},$$

and $-1 \leq r_{hi} \leq 1$. As can be seen the Pearsonian Correlation Coefficient does not satisfy Axiom 1 above. Other problems are observed which do not make the Pearsonian Correlation Coefficient and the Jaccard Similarity Coefficient mathematically consistent.

When using dichotomous data, the correlation coefficient can be reduced to a, b, c and d terms

VARIABLE/ EVENT 2	VARIABLE/EVENT 1		
	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	a+b+c+d

Variable 1 = x_i
Variable 2 = y_i

$$\sum_{i=1}^{n} x_i y_i = a$$

$$a+b=\sum_{i=1}^{n} x_i^2=\sum_{i=1}^{n} x_i$$

$$a+c=\sum_{i=1}^{n} y_i^2=\sum_{i=1}^{n} y_i$$

$$n=a+b+c+d$$

$$r=\frac{\sum_{i=1}^{n} x_i y_i-\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left\{\left[\sum_{i=1}^{n} x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]\left[\sum_{i=1}^{n} y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]\right\}^{\frac{1}{2}}}$$

Substituting the identities given above into Product Moment Correlation Coefficient:

$$r=\frac{a-\frac{(a+b)(a+c)}{n}}{\{[(a+b)-\frac{(a+b)^2}{n}][(a+c)-\frac{(a+c)^2}{n}]\}^{\frac{1}{2}}}$$

$$r=\frac{an-(a+b)(a+c)}{\{(a+b)[n-(a+b)^2](a+c)[n-(a+c)^2]\}^{\frac{1}{2}}}$$

$$r=\frac{ad-bc}{\{(a+b)(c+d)(a+c)(b+d)\}^{\frac{1}{2}}} \quad -1\leq r\leq 1$$

Hence, the Product Moment Correlation Coefficient equals the Phi Coefficient.

The Jaccard Similarity Coefficient is being used, hence all dissimilarities are removed from the data. The dichotomous data table is as follows:

VARIABLE/ EVENT 2	VARIABLE/EVENT 1		
	1	0	
1	a	b	a+b
0	c	ϕ	c
	a+c	b	a+b+c

Variable 1 = x_i
Variable 2 = y_i

The Jaccard Similarity Coefficient is defined as:

$$J_{ij}=\frac{a}{a+b+c} \quad 0\leq J_{ij}\leq 1$$

and the Correlation Coefficient for the Jaccard Coefficient is given by:

$$r_J = \frac{-bc}{\{(a+b)(a+c)(b)(c)\}^{\frac{1}{2}}} \quad -1 \leq r_J < 0$$

The range of J_{ij} is from 0 to 1, and is made up of a proportion of pure similarities (1-1) to the total data set of the dichotomous data being compared minus dissimilarities (0-0). The Jaccard Coefficient is completely void of magnitude of the raw data. The raw data has a possible range of $\pm \infty$. To further complicate the data, the Correlation Coefficient for the Jaccard Coefficient ranges from -1 to 0, and is strongly dependent on the 1-0 and 0-1 similarities. Note that the matching coefficient is discontinuous if b or c is zero (0). Opposed to the Pearsonian Correlation Coefficient which has range of $-1 \leq r_{xy} \leq +1$ and is dependent on a calculated Euclidean distance.

E. Results of Analysis

Factor Analysis is one procedure that is very useful in those situations in which one wants to reduce the number of variables under consideration while at the same time retain as much subject-to-subject variability as is possible. In most of the literature concerning Factor Analysis, it is suggested that binary data not be used. And, it is strongly recommended that data of this type not be used when manipulating large data sets. This is due to the fact that the correlation matrix is based on the Euclidean distance of the vectors in the plane and/or the covariance matrix which is based on the variance of the data being examined. Hence, the use of binary data (dichotomous) is not recommended.

Cluster Analysis, however, is the grouping of similar variables using data from the cases. It is part of the general scientific process of searching for patterns in data and then trying to construct laws that explain the pattern. Clustering can compare case to case situations and would be appropriate when working with dichotomous data. The data will form oblique-transformed data set within the cluster.

The essence of the clustering method consists of representing the factors by reference axes passing through the centriods of the respective groups of variables (clusters). Since the clusters of variables would not ordinarily be at right angles to one another, it can be assumed that the common factors within an individual cluster is not orthogonal. We must assume some criterion for establishing some Cluster Structure Matrix as a starting point, with communalities in the principal diagonal. The actual analysis begins with an appropriate clustering of n variables into m clusters on some a priori basis or purely arbitrary basis. For the "Night Study Data", we chose to establish the commonality on the eight (8) Problem Areas (PA1 through PA8).

One major problem that occurs when clustering techniques are being used is scaling. The effect of scaling can depend very much on the skewness of the data. Dichotomous data can reflect a highly skewed binary variable taking values 1 and 0 with relative frequencies p and 1-p in the n objects. The simple variance p(1-p) will be less than one (1) and division by $[p(1-p)]^{1/2}$ will inflate the importance of the variable.

To overcome not only the scaling problem, but also correlation effects among the variables, the Mahalanobis distance

$$\Delta(x_r, x_s) = [(x_r - x_s)' S^{-1} (x_r - x_s)]^{1/2}$$

where

$$S=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'}{(n-1)},$$

was implemented as a distance measure. This measure is invariant to affine transformation.

Suppose that the data consists of to similar, well-separated spherical clusters, each with n_k points and centriods $\underline{\mathbf{x}}_k$ (k = 1,2) as indicated in Figure 1. The grand mean

$$\bar{x}=\frac{(n_1\bar{x_1}+n_2\bar{x_2})}{(n_1+n_2)}$$

will lie on the line L joining the two centriods. The first principle component is $\mathbf{a}'\mathbf{x}$, where $\mathbf{a}$ the direction of the line through $\underline{\mathbf{x}}$ that minimizes the sum of the squares of the distances of all the points on the line. Clearly this line will be close to L and the second component will correspond approximately to the line M through $\underline{\mathbf{x}}$ perpendicular to L. Since the distances from M are much greater than from L, the first principle component (obtained by projecting orthogonally onto L) will have a large sample variance compared with the second component: These two variances are the eigenvalues of $\mathbf{S}$.

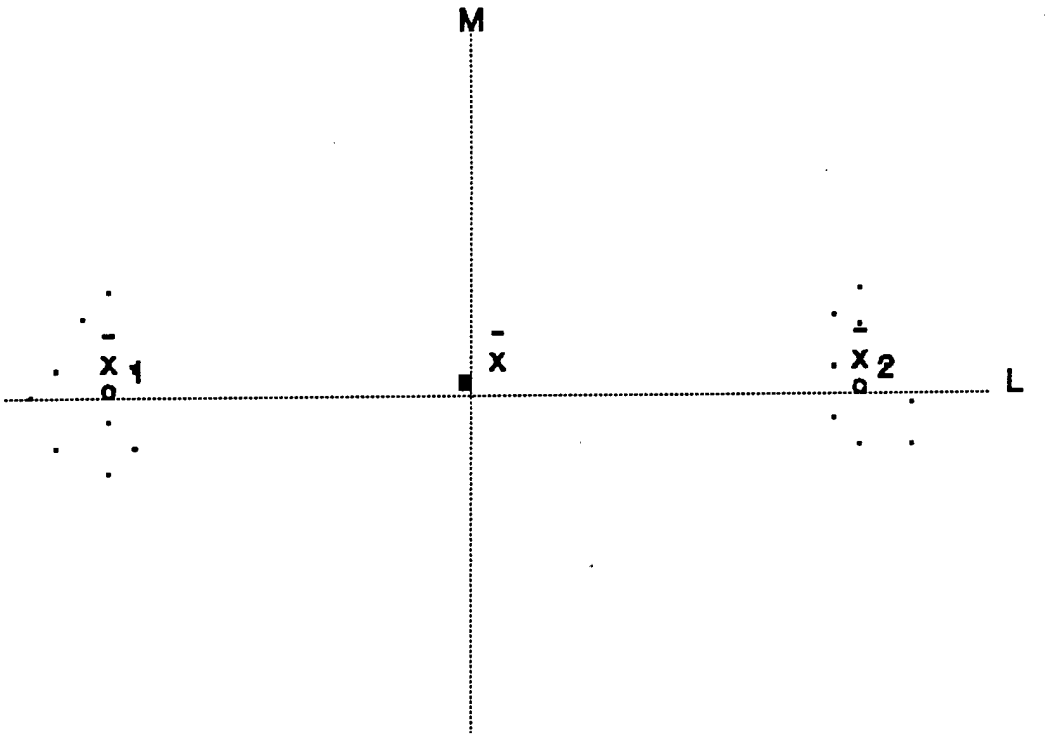

Figure 1.

Two similar well-separated spherical clusters.

Using the Mahalanobis distance instead of the of the Euclidean distance, we are effectively replacing the $\mathbf{x}_i$ by $\mathbf{S}^{-1/2}\mathbf{x}_i$, which have a sample covariance matrix $\mathbf{S}^{-1/2}\mathbf{S}\mathbf{S}^{-1/2} = \mathbf{I}_d$. This mean that the two principal components are standardized to have equal sample variances. With this new scaling, the within-cluster distances increase relative to the between-cluster distances and the cluster become less distinct. In our analysis, the point of interest is the within-cluster distances.

The VARCLUS Procedure was selected from SAS to perform the Night Study Analysis. The procedure meets all the mathematical criterion and overcomes the restrictions stated in this section. In the next section a comparison of procedures will be made and statistical evaluation of the output products of the previous procedure and the VARCLUS Procedure.

Section 2: Methodology - Modified Factor Analysis

A. Comparison of Methods Used

FACTOR ANALYSIS

1. The original data is reduced in the number of variables which will be used for final analysis.
   a. The maximum n is determined, where n is the number of times a variable is present.
   b. All of the variables are compared to one another to yields a matrix composed of the simultaneous occurrences of each paired set of variables.
   c. A Jaccard Coefficient Matrix is formed and those matches which have Jaccard values are either combined to form one variable or the variable may be discarded.

2. Manipulate the original data before putting it into the SAS program.
   a. Correlation Matrix
   b. Covariance Matrix
   Note: This is were USASC placed the Jaccard Similarity Matrix (Discussed in Section 1).

3. Once the data is in the proper form the SAS FACTOR procedure is implemented.
   a. The SAS Program used

```
DATA CORREL (TYPE = CORR);
TYPE_ = 'CORR';
INFILE SASONE;
INPUT _NAME_ $ FAT OBS ILL U60 UH1 A64 ADM PRO TAC AID
               UNA PA1 PA2 PA3 PA4 PA5 PA6 PA7 PA8 TE8
               T10 T11 PMS LND CRU HOV FIV EXP;
PROC PRINT;
PROC FACTOR METHOD = PRINCIPAL N=8 ROTATE = VARIMAX MSA
     OUTSTAT = FACT DATA =CORREL SCORE;
```

```
DATA RAW;
   INFILE SASTWO;
   INPUT _NAME_ SEQ FAT OBS ILL U60 UH1 A64 CH47 H6 ADM
                    PRO TAC AID UNA PA1 PA4 PA8 PA6 PA5
                    PA2 PA3 PA7 PCT PNC TE8 T10 T11 PMS
                    LND CRU HOV FIV EXP;
PROC PRINT;
PROC SCORE DATA = RAW SCORE = FACT OUT = SCORES;
PROC PRINT DATA = SCORES;
RUN;
```

b. The number of factors used can then be chosen in the following manners:
   i. variance explained
   ii. when specifically chosen variables are contained in separate factors
   iii. when factor pattern coefficients fit a predetermined condition
   iv. knowledge of data
   v. inter-factor correlations are at a minimum.
c. Once a number of factors are chosen, the Factor Pattern Matrix is multiplied with the Data Matrix. The resultant matrix is the Scoring Matrix.
d. When the Scoring Matrix is obtained it consists of a score for each case in relation to every factor.
e. Cases are then separated into factors depending upon which factor, a case, had the highest score.

4. After the cases are assigned to there perspective factors the cases can then be analyzed to determine if there is any sort of pattern within a factor.

The VARCLUS procedure has many options which can be programmed depending upon the desired results the researcher is hoping to obtain. The procedure used for this report did not use any special options. The procedure was allowed to run until the clusters contained only one or two variables. The researcher can specify the maximum or minimum number of clusters desired. Clusters can also be separated based on several different methods. The method used here was the centroid method because it allowed for no interaction between the clusters. All of the options available are listed in the SAS User's Guide: Statistics, Edition 5 Chapter 40 page 801.

VARCLUS

1. The original data is reduced in the number of variables which will be used for final analysis.
   a. The maximum n is determined, where n is the number of times a variable is present.
   b. Then all of the variables are compared to one another to yield a matrix composed of the simultaneous occurrences of each paired set of variables.
   c. Then a Jaccard Coefficient Matrix is formed and those matches which have Jaccard values are either combined to form one variable or the variable may be discarded.

2. Once the data is reduced the SAS VARCLUS procedure is implemented.
   a. SAS Program used

```
DATA SAFETY;
     SET NGT.EVT;
RUN;
PROC VARCLUS DATA=SAFETY SUMMARY OUTSTAT=CLSTR;
RUN;
PROC TREE;
RUN;
```

   b. The number of clusters used can then be chosen in the following manners:
      i. variance explained
      ii. when specifically chosen variables are contained in separate clusters
      iii. when cluster structure coefficients fit a predetermined condition
      iv. knowledge of data
      v. inter-cluster correlations are at a minimum.
   c. Once a number of clusters are chosen, the Cluster Structure Matrix is multiplied with the Data Matrix. The resultant matrix is the Scoring Matrix (Refer to next page for Sorting Criterion).
   d. Once a Scoring Matrix is obtained it consists of a score for each case in relation to every cluster.
   e. Cases are then separated into clusters depending upon which cluster a case had the highest score.

3. After the cases are assigned to there perspective clusters, the variables can then be analyzed to determine if there is any sort of pattern within a cluster.

Sorting Criterion

The Sorting Procedure is composed of a series of matrix manipulations. By definition:

The Cluster Structure Matrix is made up of linear combination of the variables contained in the cluster being examined (Cluster Coefficients x Variables).
The Data Matrix is composed of dichotomous data identifying the variables occurring in each case (Variables x Cases).
The Scoring Matrix is created by the multiplication of the Cluster Structure Matrix and the Data Matrix. The resultant matrix is composed of Cluster Coefficients (row vectors) and Cases (column vectors). Giving a Scoring Matrix defined as Cluster Coefficient associated to a variable identified for a particular case.

Example:

$$\begin{matrix} 0.43 & 0.25 & 0.65 \\ 0.79 & -0.32 & 0.09 \\ -.072 & 0.89 & -0.45 \\ 0.65 & 0.02 & -0.89 \end{matrix} \times \begin{matrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{matrix} = \begin{matrix} 043 & 0.25 & 0.65 \\ 0.79 & -0.32 & 0.09 \\ -0.72 & 0.89 & -0.45 \\ 0.65 & 0.02 & -0.89 \end{matrix}$$

The Scoring Matrix in this study was created by the use of LOTUS Software by selecting the defined Cluster Structure Matrix and the defined Data Matrix and multiplying the two matrixes. It should be noted that a SCORE (Sorting) Procedure is available in SAS (SAS User's Guide: Statistics, Version 5 Edition Chapter 34 page 735).

B. Comparison of Procedures

One noted difference between the VARCLUS procedure and the factor analysis is the treatment of the original data. In factor analysis the first step is to form a correlation matrix which compares variable to variable until every combination has been made. In the VARCLUS procedure the first step is to plot each case as a point on a multidimensional hyper-space. Each variable is treated as though it were one plane in the hyper-space.

The advantage of the VARCLUS procedure is that it does not compare the variables themselves. Instead this procedure compares the cases and then determines which variables are driving those cases. When the data is compared case to case, it is now truly dichotomous data.

A second noted difference is the use of the correlation coefficient. As indicated in Section 1 of this report, factor analysis uses the correlation coefficient matrix as the foundation for all of the calculations, which was proven mathematically incorrect due to lack of dichotomy in the data. In the VARCLUS procedure the correlation coefficient is used to separate variables into clusters, however it is calculated between the centroid of a cluster and the variable in question. If a variable has a relatively small coefficient when compared to the cluster it is in, that variable then becomes part of a new cluster. This eliminates the original problem of dichotomous data because when the cases are compared the variables are then truly dichotomous.

The third difference involves the Factor Pattern Matrix and the Cluster Structure Matrix. Both matrixes are a linear representation of each variable with respect to the individual factors or clusters. However, the factor pattern for each factor can be squared and summed which will yield the amount of variance explained for that factor. This can also be done with the Cluster Structure Matrix except that not all of the variables are squared and summed. Only those variables which are present in that cluster are used to determine the amount of variance explained. Not all of the variables are used to determine variance explained because they are not all present in the particular cluster being examined. The following eight (8) tables contain the linear equations for each cluster and each factor which is obtained from the Cluster Structure Matrix and the Factor Pattern Matrix.

Table 1.

VARIABLE	CLUSTER 1	CLUSTER 1 SQUARED	FACTOR 5	FACTOR 5 SQUARED
FAT	0.1336		0.1402	0.0197
OBS	-0.0132		0.0013	0.0000
ILL	0.0570		0.6641	0.4410
UH60	-0.0195		0.0485	0.0024
UH1	0.0184		-0.0393	0.0015
AH64	-0.1181		-0.0369	0.0014
ADM	0.1389		0.1062	0.0113
PRO	-0.0782		-0.0626	0.0039
TAC	-0.0323		0.0723	0.0052
AID	-0.2158		0.0051	0.0000
UNA	0.2158		0.1247	0.0155
PA1	-0.2009		-0.1065	0.0113
PA2	-0.0869		-0.0634	0.0040
PA3	0.0426		0.1921	0.0369
PA4	-0.0889		-0.00008	0.0000
PA5	0.9052	0.8193	0.8171	0.6677
PA6	0.1212		-0.0536	0.0029
PA7	-0.0830		0.0009	0.0000
PA8	-0.1304		-0.0960	0.0092
TE8	0.2177		0.1251	0.0156
TE10	-0.0550		0.0200	0.0004
TE11	-0.0277		0.0140	0.0002
PMS	0.9052	0.8194	0.8378	0.7019
LND	-0.1954		-0.1723	0.0297
CRU	-0.1549		-0.0378	0.0014
HOV	-0.1009		0.0108	0.0001
FIV	-0.0170		-0.0052	0.0000
EXP	0.0115		0.1093	0.0119
VARIANCE EXPLAINED		1.6389		1.9953

Table 2.

VARIABLE	CLUSTER 2	CLUSTER 2 SQUARED	FACTOR 3	FACTOR 3 SQUARED
OBS	0.1200		0.1971	0.0389
ILL	-0.0065		0.0499	0.0025
UH60	0.1308		0.1922	0.0369
UH1	-0.0335		0.0626	0.0039
AH64	0.0153		0.0395	0.0016
ADM	-0.0495		0.0332	0.0011
PRO	0.0545		0.1815	0.0329
TAC	-0.0058		0.0057	0.0000
AID	0.8534		0.0586	0.0034
UNA	-0.0853		0.0729	0.0053
PA1	-0.2907		-0.1595	0.0254
PA2	-0.1257		-0.0572	0.0033
PA3	-0.1257		-0.0340	0.0012
PA4	0.9330	0.8704	0.8488	0.7204
PA5	-0.1366		-0.0409	0.0017
PA6	-0.1754		-0.1036	0.0107
PA7	-0.1200		-0.0025	0.0000
PA8	-0.1887		-0.0448	0.0020
TE8	-0.1801		-0.0094	0.0001
TE10	-0.0796		-0.0026	0.0000
TE11	0.9330	0.8704	0.8630	0.7448
PMS	0.0235		0.0721	0.0052
LND	-0.0842		0.0370	0.0014
CRU	-0.0776		-0.0220	0.0005
HOV	-0.0232		0.0188	0.0004
FIV	-0.0480		0.0039	0.0000
EXP	0.0764		0.1813	0.0329
VARIANCE EXPLAINED		1.7409		1.6849

Table 3.

VARIABLE	CLUSTER 3	CLUSTER 3 SQUARED	FACTOR 6	FACTOR 6 SQUARED
FAT	-0.0152		-0.0256	0.0007
OBS	0.0213		0.0701	0.0049
ILL	-0.1277		0.0229	0.0005
UH60	-0.1322		-0.0327	0.0011
UH1	0.0037		0.0444	0.0020
AH64	-0.1147		-0.0136	0.0002
ADM	-0.1021		-0.0882	0.0078
PRO	0.0331		0.1582	0.0250
TAC	0.0524		0.0593	0.0035
AID	-0.1019		0.0162	0.0003
UNA	0.1019		0.1270	0.0161
PA1	-0.1092		-0.0200	0.0004
PA2	-0.0726		-0.0056	0.0000
PA3	-0.1001		0.0020	0.0000
PA4	-0.1262		-0.0020	0.0000
PA5	-0.0575		0.0078	0.0001
PA6	-0.0348		-0.0049	0.0000
PA7	0.7512	0.5643	0.7822	0.6118
PA8	-0.0270		0.0165	0.0003
TE8	-0.1241		0.0040	0.0000
TE10	0.8200	0.6725	0.7785	0.6061
TE11	-0.0816		-0.0041	0.0000
PMS	-0.0687		0.0115	0.0001
LND	0.0766		0.1122	0.0126
CRU	-0.0237		0.0566	0.0032
HOV	-0.0615		0.0405	0.0016
FIV	0.6521	0.4252	0.3813	0.1454
EXP	-0.1152		-0.0128	0.0002
VARIANCE EXPLAINED		1.6621		1.4438

Table 4.

VARIABLE	CLUSTER 4	CLUSTER 4 SQUARED	FACTOR 4	FACTOR 4 SQUARED
FAT	0.0034		0.2392	0.0572
OBS	-0.1023		-0.0136	0.0002
ILL	0.0099		0.1425	0.0203
UH60	-0.0526		-0.0271	0.0007
UH1	-0.1505		-0.0464	0.0021
AH64	0.6595	0.4349	0.7659	0.5866
ADM	-0.2627		-0.0067	0.0000
PRO	0.5020	0.2520	0.2306	0.0532
TAC	-0.2395		0.1079	0.0117
AID	0.1689		0.2058	0.0424
UNA	-0.1689		-0.0419	0.0018
PA1	0.5627	0.3166	0.2305	0.0531
PA2	-0.1157		0.0815	0.0066
PA3	-0.1533		-0.0832	0.0069
PA4	-0.0108		0.1493	0.0223
PA5	-0.1772		0.0292	0.0009
PA6	-0.1288		0.1271	0.0161
PA7	-0.0532		0.0299	0.0009
PA8	-0.2282		-0.0211	0.0004
TE8	-0.2152		0.1562	0.0244
TE10	-0.0680		0.0414	0.0017
TE11	-0.1819		-0.0309	0.0010
PMS	-0.2042		-0.0055	0.0000
LND	-0.1743		0.0086	0.0001
CRU	-0.1660		-0.1505	0.0226
HOV	0.6409	0.4108	0.6889	0.4746
FIV	-0.1160		-0.0692	0.0048
EXP	0.0108		0.3324	0.1105
VARIANCE EXPLAINED		1.4143		1.5232

Table 5.

VARIABLE	CLUSTER 5	CLUSTER 5 SQUARED	FACTOR 1	FACTOR 1 SQUARED
FAT	-0.2524		-0.0113	0.0001
OBS	-0.1021		0.4614	0.2128
ILL	0.4527	0.2050	0.7206	0.5193
UH60	0.4936	0.2436	0.5367	0.2881
UH1	-0.4728		0.0494	0.0024
AH64	0.0307		0.0211	0.0004
ADM	-0.3883		-0.0114	0.0001
PRO	-0.2783		0.2770	0.0767
TAC	0.5325	0.2835	0.7099	0.5040
AID	0.5947	0.3537	0.7877	0.6205
UNA	-0.5947		0.0998	0.0100
PA1	-0.0536		0.4363	0.1903
PA2	0.0005		0.0538	0.0029
PA3	0.4213	0.1775	0.1785	0.0319
PA4	0.0597		0.1770	0.0313
PA5	-0.1094		0.0385	0.0015
PA6	-0.0049		0.1700	0.0289
PA7	-0.1832		0.0023	0.0000
PA8	-0.0888		0.1208	0.0146
TE8	-0.1449		0.3700	0.1369
TE10	-0.0786		-0.0343	0.0012
TE11	0.0398		0.0948	0.0090
PMS	0.0116		0.0525	0.0028
LND	-0.1781		0.1953	0.0381
CRU	0.0371		0.5011	0.2511
HOV	0.0925		0.1177	0.0139
FIV	-0.1409		0.2164	0.0468
EXP	0.4028	0.1623	0.1516	0.0230
VARIANCE EXPLAINED		1.4255		3.0587

Table 6.

VARIABLE	CLUSTER 6	CLUSTER 6 SQUARED	FACTOR 2	FACTOR 2 SQUARED
OBS	0.4851	0.2354	0.3531	0.1247
ILL	-0.1864		0.1753	0.0307
UH60	-0.2071		0.0071	0.0000
UH1	0.6240	0.3894	0.6645	0.4416
AH64	-0.2013		0.0214	0.0005
ADM	0.6792	0.4614	0.6799	0.4622
PRO	-0.1188		0.2474	0.0612
TAC	-0.4075		-0.0581	0.0034
AID	-0.6550		-0.0172	0.0003
UNA	0.6550	0.4290	0.6691	0.4476
PA1	-0.1140		0.3055	0.0933
PA2	-0.1177		0.0896	0.0080
PA3	-0.1330		-0.0420	0.0018
PA4	-0.1580		0.0358	0.0013
PA5	0.0946		0.1237	0.0153
PA6	-0.0260		0.1373	0.0189
PA7	-0.0501		0.0235	0.0005
PA8	0.5015	0.2515	0.2306	0.0532
TE8	0.2949		0.2517	0.0634
TE10	-0.0319		-0.0577	0.0033
TE11	-0.0448		0.0715	0.0051
PMS	0.0964		0.1182	0.0140
LND	0.2211		0.4232	0.1791
CRU	-0.1178		0.0623	0.0039
HOV	-0.1907		-0.0122	0.0001
FIV	0.0708		0.2383	0.0568
EXP	-0.0101		0.0665	0.0044
VARIANCE EXPLAINED		2.0130		2.3449

Table 7.

VARIABLE	CLUSTER 7	CLUSTER 7 SQUARED	FACTOR 7	FACTOR 7 SQUARED
OBS	0.0439		0.1978	0.0391
ILL	0.0324		0.0608	0.0037
UH60	-0.0085		-0.0408	0.0017
UH1	0.0404		0.1196	0.0143
AH64	-0.0829		0.0213	0.0005
ADM	0.0049		0.0623	0.0039
PRO	-0.0306		-0.1546	0.0239
TAC	0.0328		0.1732	0.0300
AID	-0.0007		0.0658	0.0043
UNA	0.0007		0.0638	0.0041
PA1	-0.3343		-0.5416	0.2934
PA2	-0.1972		0.0330	0.0011
PA3	-0.1475		-0.0408	0.0017
PA4	-0.1426		0.0377	0.0014
PA5	0.0405		0.0268	0.0007
PA6	0.7796	0.6078	0.3712	0.1378
PA7	-0.0899		-0.0118	0.0001
PA8	0.1435		0.5735	0.3289
TE8	0.7156	0.5121	0.6222	0.3871
TE10	-0.1405		0.0509	0.0026
TE11	-0.2436		-0.0436	0.0019
PMS	-0.0910		-0.0180	0.0003
LND	-0.1317		0.1924	0.0370
CRU	0.5976	0.3571	0.0646	0.0042
HOV	-0.1286		-0.0231	0.0005
FIV	0.0363		-0.0222	0.0005
EXP	-0.0643		0.2493	0.0621
VARIANCE EXPLAINED		1.4771		1.3869

Table 8.

VARIABLE	CLUSTER 8	CLUSTER 8 SQUARED	FACTOR 8	FACTOR 8 SQUARED
FAT	0.0168		-0.0670	0.0045
OBS	-0.0696		0.0194	0.0004
ILL	-0.0962		0.0571	0.0033
UH60	-0.0794		0.0279	0.0008
UH1	0.0768		-0.0349	0.0012
AH64	-0.0086		0.0389	0.0015
ADM	0.0710		0.0592	0.0035
PRO	-0.0703		-0.2383	0.0568
TAC	0.0146		0.1997	0.0399
AID	-0.0905		0.0711	0.0051
UNA	0.0905		0.0540	0.0029
PA1	-0.1702		-0.0236	0.0006
PA2	0.8044	0.6471	0.6120	0.3746
PA3	-0.0786		0.2479	0.0614
PA4	-0.1581		-0.0445	0.0020
PA5	-0.1531		-0.0520	0.0027
PA6	-0.0306		-0.4366	0.1906
PA7	-0.0047		0.0013	0.0000
PA8	0.0354		0.1138	0.0130
TE8	-0.1350		-0.2175	0.0473
TE10	0.0055		0.0680	0.0046
TE11	-0.0853		0.0136	0.0002
PMS	-0.1646		0.0397	0.0016
LND	0.8044	0.6471	0.3948	0.1559
CRU	-0.2623		-0.2975	0.0885
HOV	-0.1771		-0.0343	0.0012
FIV	0.0049		-0.1096	0.0120
EXP	0.0104		0.2930	0.0858
Variance Explained		1.2942		1.1617

The Factor Pattern Matrix and the Cluster Structure Matrix can both be used to assign cases, which are aircraft accidents in our investigation, to specific factors or clusters. If the Factor Pattern Matrix is multiplied with the original data matrix, the result is a score for each case in each factor. Cases are assigned to the factors based on the sum of the multiplication of the rotated Factor Pattern Matrix and the original data matrix. Once the case has been assigned to a specific factor, the groups of cases can then be further analyzed to determine any trends the variables may show. This same procedure can be accomplished using the Cluster Structure Matrix.

The final difference involves the transformation used in each of the procedures. In the VARCLUS procedure, after a group of clusters are formed, a oblique transformation is used to calculate the distance of the variable within the particular cluster. In the factor analysis, a VARIMAX rotation is done after the initial oblique transformation is accomplished, to maximize the variance of the squared loadings for each factor.

c. Comparison of Output

The first section of the SAS output from the VARCLUS procedure contains the breakdown of variables into clusters. Each cluster contains a group of variables whose R-squared is closest to that cluster than any other. Another output of the VARCLUS procedure is the Cluster Structure Matrix explained in the previous section.

The steps of this procedure are repeated until each cluster contains only one or two variables or unless otherwise specified. (In the example that follows, the night study data from the USASC was used and compared to the factor analysis which was performed on the same data.) In the VARCLUS procedure eight (8) clusters were used because at that point the variables representing problem areas were split into separate clusters. The following is a list of which variables were contained in which cluster:

Table 9.

CLUSTER	VARIABLE
Cluster 1	PA5 PMS
Cluster 2	PA4 TE11
Cluster 3	PA7 TE10 FIV
Cluster 4	AH64 PA1 PRO HOV
Cluster 5	ILL TAC UH60 AID PA3 EXP
Cluster 6	FAT OBS UH1 UNA PA8 ADM
Cluster 7	PA6 TE8 CRU
Cluster 8	PA2 LND

The next step is to take the original data matrix and multiply it by the cluster structure matrix. The result of this multiplication is a scoring matrix which determines which cluster the cases belong. Once the cases are put into their respective cluster those cases are further analyzed.

To determine if the clusters were an accurate description of the factors which have been previously used, a comparison between VARCLUS clusters and factor analysis factors was made. A cluster

was matched to a factor based on variance explained and variables contained in each cluster. An example is the match between Cluster 1 and Factor 5 (Refer to Table 1). In Cluster 1, the variables PA5 and PMS contributed to the variance explained. In Factor 5, these same two variables contributed the most to the variance explained.

The following tables represent the distribution of variables from the cases in those clusters and factors. A test of proportions was used to determine if the proportions in the clusters were equivalent to the proportions in the factors. The test statistic for this test was the z-test.

$$Z = \frac{\frac{X_1}{N_1} - \frac{X_2}{N_2}}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{N_1}+\frac{1}{N_1}\right)}}$$

where

$$\hat{p} = \frac{X_1+X_2}{N_1+N_2}$$

The tests are based upon a 95% confidence interval which would give a critical z value of ± 1.96. Therefore any z value which is larger than 1.96 or less than -1.96 does not support the hypothesis that the proportions are equal. The shaded values in the following tables are those values which do not support the hypothesis. (STATISTICS: A First Course, Freund and Smith, page 368.)

$$H_o : P_1 = P_2$$

$$H_a : P_1 \neq P_2$$

Table 10.

	CLUSTER 1		FACTOR 5			
VAR	%	#	%	#	P-HAT	Z-VALUE
OBS	42.9%	3	50.0%	4	0.47	-0.27
ILL	100.0%	7	87.5%	7	0.93	0.97
OIL	42.9%	3	37.5%	3	0.40	0.21
UH60	14.3%	1	25.0%	2	0.20	-0.52
TAC	57.1%	4	50.0%	4	0.53	0.27
UNA	57.1%	4	50.0%	4	0.53	0.27
PA5	87.5%	6	87.5%	7	0.88	0.00
PCT	71.4%	5	87.5%	7	0.80	-0.78
TE8	87.5%	6	75.0%	6	0.81	0.61
CASES	3.8%	7	4.3%	8		

Table 11.

	CLUSTER 2		FACTOR 3			
VAR	%	#	%	#	P-HAT	Z-VALUE
OBS	62.5%	15	55.6%	10	0.60	0.45
ILL	75.0%	18	66.7%	12	0.71	0.59
OIL	45.8%	11	33.3%	6	0.40	0.82
UH60	41.7%	10	33.3%	6	0.38	0.55
PRO	50.0%	12	50.0%	9	0.50	0.00
AID	79.2%	19	72.2%	13	0.76	0.53
PA4	100.0%	24	100.0%	18	1.00	0.00
PCT	70.8%	17	77.8%	14	0.74	-0.51
TE11	75.0%	18	100.0%	18	0.86	-2.29
CASES	13.0%	24	9.8%	18		

Table 12.

	CLUSTER 3		FACTOR 6			
VAR	%	#	%	#	P-HAT	Z-VALUE
OBS	60.0%	6	33.3%	3	0.47	1.16
ILL	60.0%	6	55.6%	5	0.58	0.19
UH1	0.0%	0	22.2%	2	0.11	-1.58
TAC	70.0%	7	33.3%	3	0.53	1.60
PRO	30.0%	3	66.7%	6	0.47	-1.60
UNA	60.0%	6	55.6%	5	0.58	0.19
PA7	60.0%	6	100.0%	9	0.79	-2.14
PCT	50.0%	5	77.8%	7	0.63	-1.25
TE10	50.0%	5	44.4%	4	0.47	0.24
LND	30.0%	3	44.4%	4	0.37	-0.65
FIV	70.0%	7	33.3%	3	0.53	1.60
CASES	5.4%	10	4.9%	9		

Table 13.

	CLUSTER 4		FACTOR 4			
VAR	%	#	%	#	P-HAT	Z-VALUE
ILL	76.5%	13	75.0%	12	0.76	0.10
AH64	52.9%	9	81.3%	13	0.67	-1.73
PRO	76.5%	13	37.5%	6	0.58	2.27
TAC	23.5%	4	62.5%	10	0.42	-2.27
AID	88.2%	15	87.5%	14	0.88	0.06
PA1	88.2%	15	43.8%	7	0.67	2.70
PCT	88.2%	15	87.5%	14	0.88	0.06
TE8	11.8%	2	50.0%	8	0.30	-2.39
HOV	58.8%	10	62.5%	10	0.61	-0.22
EXP	23.5%	4	43.8%	7	0.33	-1.24
CASES	9.2%	17	8.7%	16		

Table 14.

	CLUSTER 5		FACTOR 1			
VAR	%	#	%	#	P-HAT	Z-VALUE
OBS	36.5%	23	48.2%	40	0.43	-1.41
ILL	87.3%	55	89.2%	74	0.88	-0.35
OIL	34.9%	22	45.8%	38	0.41	-1.33
UH60	44.4%	28	43.4%	36	0.44	0.12
TAC	88.9%	56	71.1%	59	0.79	2.61
AID	95.2%	60	90.4%	75	0.92	1.09
PA1	38.1%	24	42.2%	35	0.40	-0.50
PCT	77.8%	49	72.3%	60	0.75	0.76
TE8	30.2%	19	39.8%	33	0.36	-1.20
CRU	28.6%	18	42.2%	35	0.36	-1.69
CASES	34.2%	63	45.1%	83		

Table 15.

	CLUSTER 6		FACTOR 2			
VAR	%	#	%	#	P-HAT	Z-VALUE
FAT	38.2%	8	46.2%	12	0.43	-0.55
OBS	70.6%	15	65.4%	17	0.68	0.38
ILL	52.9%	11	61.5%	16	0.58	-0.59
UH1	73.5%	15	84.6%	22	0.80	-0.94
ADM	61.8%	13	76.9%	20	0.70	-1.12
UNA	82.4%	17	88.5%	23	0.86	-0.60
PA1	23.5%	5	30.8%	8	0.28	-0.56
PA8	35.3%	7	26.9%	7	0.31	0.62
PCT	94.1%	20	92.3%	24	0.93	0.24
TE8	100.0%	21	61.5%	16	0.79	3.21
CRU	57.1%	12	7.7%	2	0.30	3.68
LND	28.6%	6	53.8%	14	0.43	-1.74
CASES	11.4%	21	14.1%	26		

Table 16.

	CLUSTER 7		FACTOR 7			
VAR	%	#	%	#	P-HAT	Z-VALUE
OBS	28.6%	6	50.0%	6	0.36	-1.23
ILL	81.0%	17	41.7%	5	0.67	2.30
TAC	61.9%	13	58.3%	7	0.61	0.20
AID	85.7%	18	75.0%	9	0.82	0.77
PA6	71.4%	15	16.7%	2	0.52	3.02
PA8	19.0%	4	83.3%	10	0.42	-3.60
PCT	90.5%	19	100.0%	12	0.94	-1.10
TE8	100.0%	21	100.0%	12	1.00	0.00
CRU	75.1%	16	25.0%	3	0.57	2.80
CASES	11.4%	21	6.5%	12		

Table 17.

	CLUSTER 8		FACTOR 8			
VAR	%	#	%	#	P-HAT	Z-VALUE
ILL	62.5%	5	66.7%	8	0.65	-0.19
UH1	25.0%	2	8.3%	1	0.15	1.03
TAC	75.0%	6	75.0%	9	0.75	0.00
AID	62.5%	5	66.7%	8	0.65	-0.19
PA2	87.5%	7	83.3%	10	0.85	0.26
PCT	87.5%	7	91.7%	11	0.90	-0.31
LND	100.0%	8	75.0%	9	0.85	1.53
CASES	4.3%	8	6.5%	12		

According to the tables, the only cluster to factor pairing which may not be considered a good match is the Cluster 4 to Factor 4 (Table 13). This would indicate that the PROC VARCLUS was able to reproduce the same results as the factor analysis procedure. However the mathematical difficulties have been overcome with the use of the VARCLUS procedure.

II. Conclusion

Many studies conducted by the U.S. Army Safety Center involve the measurement of a large number of variables on different cases. Factor Analysis is one procedure that is very useful in those situations in which one wants to reduce the number of variables under consideration while at the same time retain as much subject-to-subject variability as is possible. In most of the literature concerning Factor Analysis, it is suggested the binary data not be used, and it is strongly recommended that data of this type not be used when manipulating large data sets. This is due to the fact that the correlation matrix is based on the Euclidean distance of the vectors in the plane and/or the covariance matrix which is based on the variance of the data being examined. Hence, the use of binary data (dichotomous) is not recommended.

Cluster Analysis, however, is the grouping of similar variables using data from the cases. It is part of the general scientific process of searching for patterns in data and then trying to construct laws that explain the pattern. Clustering can compare case to case situations and would be amicable to dichotomous data. The data will form oblique-transformed data set within the cluster.

The VARCLUS procedure has many options which can be programmed depending upon the desired results the researcher is hoping to obtain. The procedure used for this report did not use any special options. The procedure was allowed to run until the clusters contained only one or two variables. The researcher can specify the maximum or minimum number of clusters desired. Clusters can also be separated based on several different methods. The method used here was the centroid method because it allowed for no interaction between variables. All of the options available are listed in the SAS User's Guide: Statistics, chapter 40 page 801.

III. Recommendations

To reiterate, the analysis performed in this report was done using a Clustering Technique (Centroid Method) to duplicate Factor Analysis Procedure established by the USASC. Basically while performing this analysis, we found ourselves working towards answers already achieved. This technique allowed a thorough evaluation of the mathematical procedures used by the USASC. Listed below are some recommendations for further analysis:

1.) Re-analyze the Crew Co-ordination Study including Operation Desert Storm/Desert Shield Data using Clustering Techniques.

2.) Re-analyze the Night Study including Operation Desert Storm/Desert Shield Data using Clustering Techniques.

3.) Investigating the various options given with the VARCLUS Procedure.
   a. INITIAL = SEED assigns to cluster variables named in the SEED statement. The other variables are not specifically assigned until the clustering technique is performed.
   b. HIERARCHY requires the clusters at different levels to maintain a hierarchical structure.

4.) Investigate strategies to establish a criterion for recognizing a demand characteristic for clustering.

5.) Investigate different methods to calculate the correlation/covariance matrix with respect to dichotomous data used in the Safety Center Studies.

BIBLIOGRAPHY

Afifi, A. A. and Virginia Clark. Computer - Aided Multivariate Analysis. Belmont, CA.: Lifetime Learning Publications, 1984.

Aldenderfer, Mark S. and Roger F. Blasfield. Cluster Analysis. Beverly Hills, CA.: Sage Publications, 1986.

Anderberg, Michael R. Cluster Analysis for Application. New York, NY.: Academic Press, 1973.

Basilevskey, Alexander. Applied Matrix Algebra in the Statistical Sciences. New York, NY.: Elseview Science Publishing Co. Inc., 1983.

Comrey, Andrew L. A First Course in Factor Analysis. New York, NY.: Academic Press, 1973.

Cureton, Edward E. and Ralph B. D'Agostino. Factor Analysis: An Applied Approach. Hillsdale, NJ.: Lawrence Erlbaum Associates, Publishers, 1983.

Daniel, Wayne W. Applied Nonparametric Statistics. Boston, Mass.: PWS-KENT Publishing Company, 1990.

Fruend, John E. and Richard Manning Smith. Statistics: A First Course. Englewook Cliffs, NJ.: Prentice-Hall Inc, 1986.

Gorman, Bernard S. and Louis H. Primavera. "The Complementary use of Cluster and Factor Analysis." Journal of Experimental Education, vol 51 (Summer 1983), pp. 165 - 168.

Harman, Harry H. Modern Factor Analysis. Chicago, Il.: The University of Chicago Press, 1960.

Kim, Jae-on and Charles W. Mueller. Introduction to Factor Analysis - What it is and How to do It. Beverly Hills, CA.: Sage Publications, 1978.

Kim, Jae-on and Charles W. Mueller. Factor Analysis - Statistical Methods and Practical Issues. Beverly Hills, CA.: Sage Publications, 1978.

Romesburg, H. Charles. Cluster Analysis for Researchers. Belmont, CA.: Lifetime Learning Publications, 1984.

Serber, G. A. F. Multivariate Observations. New York, NY.: John Wiley and Sons, Inc., 1984.

SAS User's Guide: Statistics, Version 5 Edition Chapter 40 page 801.