PILOT ACCIDENT POTENTIAL AS RELATED TO TOTAL FLIGHT EXPERIENCE AND RECENCY AND FREQUENCY OF FLYING

Henry Benjamin Myers, jr.

Naval Postgraduate School
Monterey, California

September 1974

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

PILOT ACCIDENT POTENTIAL AS RELATED TO
TOTAL FLIGHT EXPERIENCE AND
RECENCY AND FREQUENCY OF FLYING

by

Henry Benjamin Myers, Jr.

September 1974

Thesis Advisor: R.R. Read

Approved for public release; distribution unlimited.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD/A001256

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Pilot Accident Potential as Related to Total Flight Experience and Recency and Frequency of Flying		5. TYPE OF REPORT & PERIOD COVERED Master's Thesis; September 1974 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Henry Benjamin Myers, Jr.		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey CA 93940		12. REPORT DATE September 1974 13. NUMBER OF PAGES 78
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Naval Postgraduate School Monterey CA 93940		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Pilot Accident Potential

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This thesis introduces exploratory data analysis methods into the question of categorizing pilots and relating these categories to accident potential. The usually recorded flight data deals with the pilots' total flight experience, recency, and frequency of flying. The purpose of categorizing is to determine if the recorded flight data could help discriminate between two original sample groups of fifty pilots each, those

DD FORM 1 JAN 73 1473 (Page 1)

EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED (78)
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(20. ABSTRACT Continued)

pilots with accidents during FY73 and those without.

The technique of linear discriminant analysis indicated that there is a significant difference in the mean vectors of flight data for the two groups. The computed discriminant function produced an empirical correct classification rate of 81%. Techniques of cluster analysis (with the aid of principal components analysis) are also employed to detect patterns or differences in the data. Curiously, the amount of time flown in the last 48 hours is associated with relatively low accident potential, whereas time flown in the last 24 hours seems to be correlated with a higher accident potential.

Pilot Accident Potential as Related to Total Flight Experience and Recency and Frequency of Flying

by

Henry Benjamin Myers, Jr.
Lieutenant, United States Navy
B.S., Illinois Institute of Technology, 1966

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1974

Author Henry Benjamin Myers, Jr.

Approved by: R. R. Read
Thesis Advisor

Second Reader

David A. Schrady
Chairman, Department of Operations Research and Administrative Sciences

Jack R. Borsting
Academic Dean

ABSTRACT

This thesis introduces exploratory data analysis methods into the question of categorizing pilots and relating these categories to accident potential. The usually recorded flight data deals with the pilots' total flight experience, recency, and frequency of flying. The purpose of categorizing is to determine if the recorded flight data could help discriminate between two original sample groups of fifty pilots each, those pilots with accidents during FY73 and those without.

The technique of linear discriminant analysis indicated that there is a significant difference in the mean vectors of flight data for the two groups. The computed discriminant function produced an empirical correct classification rate of 81%. Techniques of cluster analysis (with the aid of principal components analysis) are also employed to detect patterns or differences in the data. Curiously, the amount of time flown in the last 48 hours is associated with relatively low accident potential, whereas time flown in the last 24 hours seems to be correlated with a higher accident potential.

TABLE OF CONTENTS

I. INTRODUCTION AND OBJECTIVES

Aviation safety in the United States Navy has always received considerable attention. With the rapidly increasing costs of naval aircraft and the increasing costs of training naval aviators, it is imperative that every possible aspect of aviation safety be thoroughly investigated. It is important to search all paths which may yield any information at all having a bearing on aircraft accident causation or prevention.

Over the last five years, approximately fifty per cent of the major and minor aircraft accidents in the Navy have included pilot error as either the primary factor involved or as a contributing factor to the cause of the accident.

There have been many reasons purported as to the causes of pilot error accidents, ranging anywhere from plain lack of physical coordination to mental incompetence. A general term which relates to both physical and mental abilities is experience. That is, as flying experience increases, the learning process should increase both of these abilities. Another general term which affects these two abilities is proficiency. That is, recency and frequency of flying should also have a direct bearing on these abilities.

This thesis explores methods for classifying or categorizing pilots according to variables associated with their experience and proficiency. Accident records are used to

determine if there is any relation between the classifications and the occurrence of accidents.

One would like to know if by investigating a pilot's experience and proficiency data whether or not he shows a high or low accident potential. Of specific interest is the question of whether pilot error accidents are related to lack of total flying experience, lack of experience in type of aircraft, or lack of practice due to insufficient current flying. If an individual were classified as having a high degree of accident potential, then corrective action could be taken to reduce this potential.

Only the pilots of Navy fixed-wing aircraft are studied. Marine and/or helicopter pilots are not included. The study encompasses those accidents that occurred during fiscal year 1973. Unfortunately, the data base contains the records of only fifty aviators who have been involved in pilot error accidents. Fifty other pilots were selected as a control. Even with these small numbers, a result appeared that may be worth pursuing further. Recency of flying may be overdone. The amount of time flown in the last 48 hours is positively correlated with low accident potential, but a reversal seems to take place when looking at the time flown in the last 24 hours.

II. FACTORS INFLUENCING EXPERIENCE AND PROFICIENCY

There are many factors affecting experience and proficiency. Situations encountered, crises faced, types of missions flown, and many other qualitative factors have a definite bearing. However, the only factors considered here are quantitative variables which can be obtained from accident records and IFARS (Individual Flight Activity Reporting System) pilot records.

The Naval Safety Center at Norfolk, Virginia maintains records of all accidents in which Naval aircraft are involved. The recorded data items which reflect a pilot's total experience are the following:

Number of years designated a naval aviator
Total flying hours
Total flying hours in the model aircraft in which the accident occurred
Total day carrier landings
Total night carrier landings

The data items which reflect his proficiency (i.e. his recency and frequency of flying) are the following:

Time all series this aircraft in last 90 days
Time this model this aircraft in last 90 days
Elapsed time since last previous flight
Time flown in the last 24 hours
Time flown in the last 48 hours
Number of missions flown in the last 24 hours
Number of missions flown in the last 48 hours
Number day carrier landings in last 30 days

Number night carrier landings in last 30 days
Instrument trainer time in last 90 days
Weapons system trainer time in last 90 days

The Individual Flight Activity Reporting System (IFARS), a part of the Naval Safety Center, maintains flight records on all naval aviators by fiscal year. The only data items pertaining to pilot experience which are retrievable from computer access for all fiscal years are:

Number of years designated a naval aviator
Total flying hours

At present, these following additional experience items are retrievable by computer only from the beginning of fiscal year 1969 and thus cannot be used as comparison variables since many of the aviators in both sample groups began flying prior to 1969.

Total time by model
Day and night carrier landings by model
Other type landings by model
Instrument time by model

A new compilation is now in progress by the IFARS section at the Naval Safety Center to record all flights on computer files for all fiscal years for all pilots so that future studies can be more encompassing.

The proficiency indicator data items for those pilots in the accident group have a natural base point from which to be measured. That is, an item such as "time flown in the last 48 hours" means the last 48 hours directly prior to the accident in which the pilot was involved. However, for

the non-accident (control) group, there is no such reference point from which to measure. Thus, comparison of proficiency data items becomes rather nebulous.

One reasonable way to give significant meaning to the term proficiency is to artificially construct similar data items by an averaging procedure. For example, prior to each flight (for the period in question) compute the time flown in the preceding 48 hours. Do this for every flight during the fiscal year and then obtain an average time flown in the preceding 48 hours. The necessary data can be obtained from a detailed flight listing for the pilots in the control group for FY73. This procedure can be utilized for the following data items:

Time all series this aircraft last 90 days
Elapsed time since last previous flight
Time flown in the last 24 hours
Time flown in the last 48 hours
Number of missions flown in the last 24 hours
Number of missions flown in the last 48 hours
Number day carrier landings in last 30 days
Number night carrier landings in last 30 days

With these artificially constructed data items one can include proficiency in the comparison between the control group and the accident group. The appropriateness of doing this can be determined by comparing the results of statistical analyses performed with and without these added variables. If these added variables give a better delineation between groups, then it is appropriate to include them.

To recap, the variables which are common to both groups and which are used for the analysis are:

(X_1) Number of years designated a naval aviator
(X_2) Total flying hours
(X_3) Time all series this aircraft last 90 days
(X_4) Time since last previous flight
(X_5) Time flown in the last 24 hours
(X_6) Time flown in the last 48 hours
(X_7) Number of missions flown in the last 24 hours
(X_8) Number of missions flown in the last 48 hours
(X_9) Number of day carrier landings in last 30 days
(X_{10}) Number of night carrier landings in last 30 days

III. SELECTION OF GROUPS

The accident group was composed of all those pilots who were involved in pilot error accidents during fiscal year 1973. This group comprised 66 different pilots; no pilot had more than one accident attributable to pilot error. Due to incomplete data in two cases, this was reduced to 64 pilots.

The control group was more difficult to establish since there were several thousand aviators from which to choose. A subset of these pilots was obtained that satisfied two criteria: (1) it appeared to be a sample representative of all naval aviators, and (2) the data was relatively easy to obtain. The sample taken was the first 100 aviators on the IFARS files. Since the IFARS files are ordered by increasing social security number and the increments between successive numbers was very large, examination of the biographical data leads us to believe that social security numbers had no bearing upon age, length of time in aviation duties, or even length of time in the Naval Service. There was no obvious reason to think that the sample was unrepresentative.

From the 100 pilots initially assigned to the control group, 20 were helicopter pilots and 15 were Naval Flight Officers, thus leaving 65 subjects in the control group. Since the size of the two groups under study is arbitrary, a further reduction in the size of each group was made to

meet a computational constraint which was imposed by a computer program employed in the actual analysis. Because of the extensive computational effort required in the analytical techniques used, the use of a digital computer was mandatory. One of the computer programs used for the analysis had a limitation of 100 data units. Therefore, a random selection of 50 subjects was chosen for each of the two groups under study. (The random selection was accomplished in the manner of drawing numbers out of a hat.)

IV. INVESTIGATIVE APPROACHES

The data describing the subjects is composed of ten pieces of information for each subject. This constitutes a multivariate data set. Therefore, some sort of multivariate statistical technique is appropriate. Which statistical techniques to employ depends upon the information desired to be obtained from the analysis, and is the primary concern of this section.

As stated in the introduction, one of the primary objectives is to establish a classification scheme and then to determine if this classification is related to the occurrence of accidents. One statistical procedure which treats this problem is that of discriminant analysis. Discriminant analysis is a multivariate statistical technique used for constructing decision rules by which data units (subjects, or pilots in the present context) can be classified as members of one group or another.[1] The goal is to assign subjects to the groups to which they have the greatest resemblance based upon a profile of their characteristics, while at the same time to minimize the effects of misclassification.[2]

---

[1] Anderberg, M.R., Cluster Analysis for Applications, p. 191, Academic Press, Inc., 1973

[2] Eisenbeis, R.A., and Avery, R.B., Discriminant Analysis and Classification Procedures, p. 3, Lexington Books, 1972

The procedure constructs a discriminant function based upon input data in which subjects are members of known groups. This discriminant function is usually linear but can be quadratic or have other forms. The data are used to make the function specific (determine the parameters). Typically, it is then used to reassign the original subjects to one of the two groups on the basis of their characteristics in order to make an empirical determination of the rate of misclassification. If all subjects are reassigned to the group from which they initially came, then there is zero percentage misclassification and perfect discrimination between groups. The discriminant function can also be used to categorize other observations (subjects), whose group membership is unknown, on the basis of their attributes.

If several (more than two) groups are present, then a set of discriminant functions is constructed to assign observations to the appropriate groups.

A linear discriminant function will be constructed for the two pilot groups on the basis of their experience and proficiency characteristics. If the function discriminates well, then one can determine what particular characteristics have the strongest influence on placing a subject in the accident group.[3] Also, by applying the discriminant function to subjects not in the original test groups one can determine their accident potential.

---

[3]Press, S.J., Applied Multivariate Analysis, p. 376-379, Holt, Rinehart and Winston, Inc., 1972

The assumptions upon which discriminant analysis is based and the actual mathematics will be covered in the next section.

If the discriminant function fails to separate the groups without a high rate of misclassification, the lack of success can be attributed to one of two causes. The first is that the variables characterizing the subjects do not distinguish between the groups to a strong enough degree or the groups overlap too much in the given measurement space. The second is that the groups cannot be separated by a function of the form chosen for the analysis. That is, maybe instead of a linear discriminant function we should have a quadratic or more complex one.

To illustrate the preceding concept, let the accident group be denoted by "A" and the control group by "C". Now, if one considers the groups in two dimensions only (instead of the actual ten) the groups might be clumped as in Figure (1).

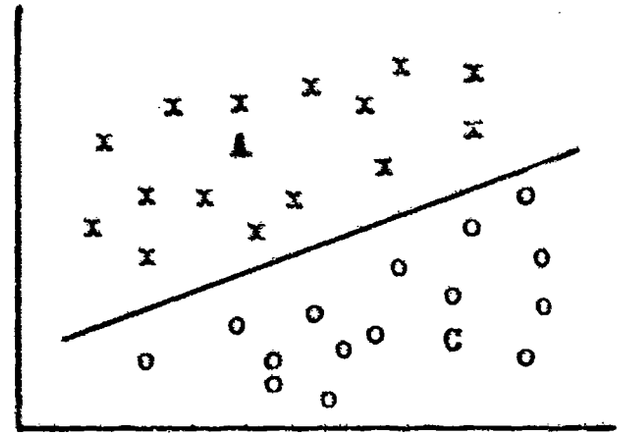


Figure (1)

In this case a linear discriminant function would serve to separate the groups well and it is not necessary to construct a quadratic function. If, however, the data ppeared as in Figure (2), then one can see that a linear discriminant function cannot discriminate among the groups without error. However, a quadratic form of discriminant function such as the curve depicted might very well have excellent discriminating capabilities.

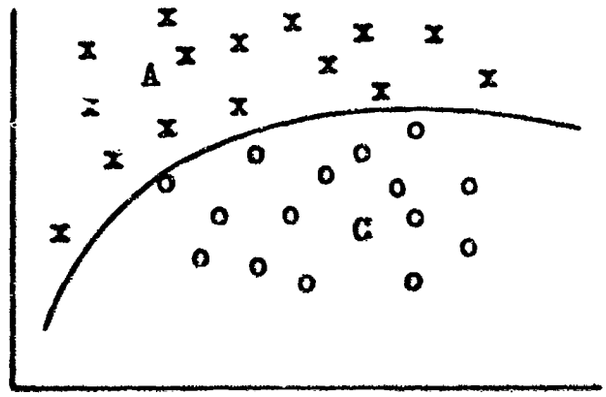


Figure (2)

The linear discriminant function is a tool that is immediately available in terms of computer programs. It is based upon the assumption that the data came from a multivariate normal population, and when this assumption is met, it works as well as any other discriminant function. Other discriminant functions are not readily available for use. Also, the linear discriminant function could do a good job even if the multivariate normal assumption is not met, i.e. when the natural separation of groups is so great that even a simple method would do the job.

For the problem at hand, the use of the linear discriminant function was encouraging, but since the assumption of multivariate normality is not appropriate (e.g. rotation policies split variables X_1 and X_2 so that their distributions are multimodal) it was decided to explore the nature of the data to see if a better job could be done.

Exploratory data analysis on 100 points in Euclidian 10-space is not easy. Some form of cluster analysis is called for, that is, cluster the subjects into groups. This leads to the question of how many groups we actually have and how the data are grouped.

Cluster analysis is actually a collection of techniques that are used to group multidimensional entities according to various criteria of their degrees of homogeneity or heterogeneity.[5] For example, in this problem grouping will be on the basis of the values of each variable which describes the pilot's flight experience and proficiency. Pilots with high total flight time might tend to cluster into one group while pilots with few carrier landings or with little time since last flight might tend to cluster into other groups. How close should the values of the variables be before subjects are grouped into the same cluster is the question of the degree of homogeneity desired, and how many

---

[5]Op. Cit., Press, S.J., p. 408-411

clusters there should be is the question of the degree of heterogeneity desired. This type of grouping is called grouping by subjects; that is, the entities are subjects. The entities can also be the variables themselves, in which case the clustering is said to be by attributes.

There are several pertinent questions to bear in mind when performing a cluster analysis. How many clusters are inherent in the data? Since attributes may be measured in different units, should the attributes be standardized before they are clustered? How large should the errors be before they are considered intolerable? There will be one type of error made by not assigning similar entities into the same group, and another type of error made by grouping dissimilar entities into the same cluster. Should all possible pairs of points (or attributes) be scrutinized for similarities?[6] Not all of these questions have definite answers, but they will be addressed in the next section.

In most other statistical techniques, such as analysis of variance, the variables usually possess some structure of belonging to particular populations a priori. Consequently, it is often possible to assume particular distributions for the populations and make associated inferences. In clustering problems, however, the principal concern is how to establish

---

[6]Ibid.

appropriate populations. Thus, clustering analysis logically precedes the application of most other multivariate procedures when the data do not possess structured form.[7]

There are two possible approaches to clustering. These are enumerative procedures and non-enumerative. Enumerative means simply to list all the possible groupings of subjects (attributes if the clustering is by this form of entities). The number of possible groupings is represented by a Stirling Number of the Second Kind. For example, in clustering twenty-five subjects into five groups there are between two and three quadrillion possibilities from which to choose the best grouping.[8] This is not feasible even with [illegible] ter, especially when the problem is much larger than [illegible]. Some feasible non-enumerative techniques are described in the next section.

If through the use of cluster analysis one can find a feasible set of groupings that have meaning to this problem then the groupings can be analyzed by a discriminant analysis to obtain the desired classification procedure.

Clustering by variables can also prove to be worthwhile in that it can help to determine if some of the variables are redundant and not providing any additional information.

---

[7] Ibid.

[8] Op. Cit., Anderberg, M.R., p. 3

If so, those redundant variables can be eliminated or combined, thus simplifying the required computations. A method of combining variables which was utilized was that of principal components analysis.

Multivariate analysis by the principal components model attempts to reduce the dimension of the problem while retaining as much information (i.e. variation) contained in the original data as possible. The method produces linear combinations of the original variables which maximize the variance of the resultant weighted sum. Thus attention is centered primarily on the variable with the greater variability by the appropriate assignment of the weights. This linear combination of the variables is called the first principal component and reduces our set of old variables to one variable. If it is desired to extract more variance from the data, one can construct a second principal component which is orthogonal to the first. The process can be repeated until there are as many components as original variables, and thus have extracted one-hundred percent of the total variance.[9]

The objective of principal components analysis is not merely to reduce the size and complexity of the problem, but also to glean information from the data which might not

---

[9]Op. Cit., Press, S.J., p. 283-285

otherwise be obvious. Specifically, in the problem under study here, the fifth, sixth, seventh and eighth variables listed on page eleven can be regarded as prime indicators of frequency of flying. However, when the data is analyzed (by cluster analysis) the exact effects of these variables might not readily be apparent. When all these variables are combined into one variable (i.e. the first principal component) the effect of frequency might be quite obvious. That is, it might be observed that frequency of flying has an inverse relationship with the occurrence of accidents.

For this analysis, of those variables listed in page eleven, the first and second (years designated naval aviator and total hours flown) were combined to get a "total experience" variable; and the fifth, sixth, seventh and eighth (time flown in the last 24 hours, time flown in the last 48 hours, number missions flown in the last 24 hours, and number missions flown in the last 48 hours) were combined to get a "frequency" variable.

V. ANALYTICAL TECHNIQUES

The first analytic technique applied to the two groups of pilots was that of discriminant analysis with the primary objective being to develop an accurate linear discriminant function. (Actually, the purposes of discriminant analysis are first to determine if there is a difference among population means or equivalently if there are any overlaps among the groups, and secondly, to construct classification schemes based upon the descriptive variables.)

There are three basic underlying assumptions of discriminant analysis. They are (1) that the groups being investigated are discrete and identifiable, (2) that each observation (subject) in each group can be described by a set of measurements on m characteristics or variables, and (3) that these m variables are assumed to have a multivariate normal distribution in each population and equal covariance matrices among populations. The first two assumptions are seen to be satisfied as discussed in previous sections. The third assumption indicates the need for separate statistical tests to determine if the variables are multivariate normal and if the covariance matrices are equal. It has been mentioned that non-normal multivariate data does not necessarily bias the results of a discriminant analysis. Also, since no satisfactory tests exist for testing populations to be multivariate

normal, it is difficult to routinely test the normality assumption. Finally, the central limit theorem suggests that as the number of observations increases, the discriminant values for each group approaches a normal distribution.[10]

The assumption of equality of covariance matrices (i.e. equality of within group dispersions) appears to be more critical in biasing the results. Eisenbeis and Avery suggest that linear classification rules are not adequate when unequal covariance matrices exist and that quadratic classification rules should be employed.[11]

The within group dispersion matrices for the two groups of data were computed and are shown in Table VI in Appendix C. The pooled within-groups dispersion matrix is also shown. The group dispersion matrices were tested for equality by the procedure given in Appendix D.

After satisfying the assumptions preparatory to the actual analysis one can first test the equality of group means. The null hypothesis is:

$$H_o: \bar{u}_1 = \bar{u}_2$$

---

[10] Kirk, R.E., Experimental Design: Procedures for the Behavioral Sciences, p. 62, Brooks/Cole Publishing Co., 1968

[11] Op. Cit., Eisenbeis, R.A., and Avery, R.B., p. 16

where

$$\bar{\mu}_1 = (\bar{\mu}_{1,1}, \bar{\mu}_{1,2}, \ldots, \bar{\mu}_{1,10})$$

and

$$\bar{\mu}_2 = (\bar{\mu}_{2,1}, \bar{\mu}_{2,2}, \ldots, \bar{\mu}_{2,10}).$$

The following steps are used by the BIMED04M computer program to test for the equality of group means:

Step (1) -- the means for each group are computed

$$\bar{x}_i = (\bar{x}_{i,1}, \bar{x}_{i,2}, \ldots, \bar{x}_{i,10}) , \quad i = 1,2$$

Step (2) -- the differences in group means are computed.

$$\bar{x}_1 - \bar{x}_2 = (\bar{x}_{1,1} - \bar{x}_{2,1}, \ldots, \bar{x}_{1,10} - \bar{x}_{2,10})$$

Step (3) -- the matrices S^1 and S^2 are computed where an element of S^i is given by

$$s^i_{u,v} = \sum_{j=1}^{n_i} (x_{iju} - \bar{x}_{i,u})(x_{ijv} - \bar{x}_{i,v}) \quad \text{and}$$

$i = 1,2; \quad u = 1,2,\ldots, 10; \quad \text{and} \quad v = 1,2,\ldots, 10$

Step (4) -- the matrix A is computed

$$A = S^1 + S^2$$

where $(a^{j1}, a^{j2}, \ldots, a^{j,10})$ is the $j^{\underline{th}}$ row of A

Step (5) -- the Mahalanobis D^2 statistic is computed

$$D^2 = (n_1 + n_2 - 2) \sum_{i=1}^{m} \sum_{j=1}^{m} a^{ij}(\bar{x}_{1,i} - \bar{x}_{2,i})(\bar{x}_{1,j} - \bar{x}_{2,j})$$

Step (6) -- the F statistic is computed

$$\frac{n_1 n_2 (n_1 - n_2 - m - 1)}{m(n_1 - n_2)(n_1 - n_2 - 2)} \cdot D^2 \sim F^m_{n_1 - n_2 - m - 1}$$

where n_1 and n_2 are the respective sizes of the two groups and m is the number of variables.[12]

The null hypothesis can be rejected when the value of the test statistic is greater than the tabled value of F for the desired level of significance.

The construction of the discriminant function is predicated upon minimizing the effects of misclassification and assigning subjects to the group to which they have the greatest resemblance. The effects of misclassification

---

[12] BMD Manual, Biomedical Computer Programs, Health Sciences Computing Facility, UCLA, University of California Press, 1973, p. 211-220

depend upon the a priori knowledge of group membership and the costs or penalties of misclassification. The BIMED programs assume no special a priori probabilities of group membership, i.e. the probability of belonging to either group (in the two-group case) is one-half. They also assume the costs of misclassification to be equal, i.e. the cost of assigning an actual member of group number one to group number two is the same as assigning a member of group number two to group number one.

The measure of resemblance is determined by the m characteristics which describe each subject. By substituting the values of the characteristics into each group's probability density function it is determined how closely the subject resembles the group as compared with the rest of the population. The BIMED programs yield the coefficients and constants for the linear discriminant function for each group in the total population.[13]

In order to determine what effect the chosen variables had on proficiency and experience it was desirable to measure the association among the variables. The association measure employed was the product-moment correlation coefficient. The correlation computations and correlation matrix for the entire data set is given in Appendix F.

---

[13]Ibid.

The problem of how to group the variables given this association measure can be solved through the use of hierarchical clustering techniques. These techniques can also be useful to cluster by data units (subjects) which have a different association measure.

For the association measure among data units, most investigators use metric measures when the data units are described by interval variables. Metric measures must satisfy certain properties. If E is a given measurement space and X, Y, and Z are points in E, then an association function D is a metric measure if and only if it satisfies the following conditions: [14]

(1)	$D(X,Y) = 0$	if and only if $X = Y$
(2)	$D(X,Y) \geq 0$	for all X and Y in E
(3)	$D(X,Y) = D(Y,X)$	for all X and Y in E
(4)	$D(X,Y) \leq D(X,Z) + D(Y,Z)$	for all X, Y and Z in E

The most common metric measure is the Euclidian distance function, $D_2(X_j, X_k) = [\sum_{i=1}^{n} (x_{ij} - x_{ik})^2]^{\frac{1}{2}}$. This is a special case of the general class of metrics called Minkowski metrics which have the form $D_p(X_j, X_k) = [\sum_{i=1}^{n} |x_{ij} - x_{ik}|^p]^{1/p}$, where $p \geq 1$ and $X_j^T = (x_{1j}, x_{2j}, \ldots, x_{nj})$ is the vector

---

[14] Op. Cit., Anderberg, M.R., p. 98-102

of scores on the $j^{\underline{th}}$ data unit. In this analysis the Euclidian distance function was used to cluster the data units.[15]

The hierarchical methods are used to construct a tree (dendrogram) depicting the relationship among the entities. The entities are grouped into clusters in order of their association measures or similarities. The ordering provides a hierarchy, thus the name. The similarities can be of many forms of association measures; the general term applied to the matrix being a similarity matrix.

A breakdown of hierarchical methods yields agglomerative and non-agglomerative procedures. The agglomerative procedures start with the branches (each entity) and combine these entities until there is but one remaining cluster (the root). The alternative procedures work from the root backward. Only the former was used in this analysis.

There are many actual techniques and criteria of hierarchical clustering. Initially each entity is considered to be a cluster of one. The first method searches the similarity matrix for the pair of entities with the highest degree of association (e.g. largest correlation among the variables) and groups these two entities. It then searches all remaining clusters and groups those two clusters which are closest,

---

[15] Ibid.

i.e. the correlation among their closest members is highest. This step is repeated until there is but one cluster remaining. This method is called the "single linkage" method by Anderberg or the "connectedness" method by Johnson.[16] The names derive from the fact that each cluster is joined by the single shortest or strongest link (thus most strongly connected) between them.

The second procedure, called complete linkage, is the same as single linkage except that the association between groups is the association between their farthest members. Johnson calls this the diameter method because all entities in a cluster are linked to each other at some maximum distance (or diameter).

Hierarchical clustering is usually not too enlightening for the clustering of data units. The non-hierarchical methods are more appropriate for classifying the data units into a single classification of k clusters. The basic concept in most of the non-hierarchical methods is to begin with an initial partition of the data units and adjust the cluster members to obtain a "best" partition.

The simplest and most common non-hierarchical clustering procedure is that of centroid sorting. Beginning with the initial partition of k clusters (each usually consisting of

---

[16] Johnson, S.C., "Hierarchical Clustering Schemes", Psychometrika, Vol. 32, No. 3, p. 241-254, 1967

one data unit) a new data unit is assigned to the cluster with the nearest centroid by some sort of distance measure. Centroids are recomputed after a data unit is assigned and the procedure repeated for remaining data units. After all data units are assigned, the entire procedure can be reapplied to all data units over and over until there are no more changes in cluster memberships, i.e. until convergence.[17]

There are more complex methods than the centroid methods for clustering data units and these are based on multivariate statistical analysis techniques. The scatter of two variables is the inner product of two centered score vectors. The scatter matrix T is a square matrix that has the entry t_{ij} which is the scatter of variables i and j computed over all the data units. Each of the h clusters has its own scatter matrix W_k computed over the data units in the $k^{\underline{th}}$ cluster. The within groups scatter matrix is given by $W = \sum_{k=1}^{h} W_k$. The between groups scatter matrix is denoted by B. An element $b_{ij} = \sum_{k=1}^{h} m_k \bar{x}_{ik} \bar{x}_{jk}$ where m_k is the number of data units in the $k^{\underline{th}}$ cluster, $\bar{x}_{ik}$ is the mean (centered around the grand mean in the entire data set) of the $i^{\underline{th}}$ variable

---

[17] Op. Cit., Anderberg, M.R., p. 156-173

in the $k^{\underline{th}}$ cluster. The three scatter matrices can be shown to satisfy the relation $T = B + W$. [18]

An important element in many clustering criteria is the determinantal equation $|B - \lambda W| = 0$. The eigenvectors of the matrix $W^{-1}B$ provides the λ_1 solutions to this equation.

D. J. McCrae has developed a FORTRAN IV computer program called K-MEANS which utilizes these concepts to cluster the data into k clusters. He provides for four possible criteria for determining when assignment of a data unit to a particular cluster results in the "best" partition of the data set. These criteria are: (1) minimize the trace of W; (2) maximize the largest eigenvalue of $W^{-1}B$; (3) maximize the trace of $W^{-1}B$: and (4) minimize the ratio of the determinants $|W|/|T|$. This last criterion is more commonly known as Wilk's Lambda statistic. Since T is the same for all partitions, this is equivalent to minimizing det W. The last procedure was the one used to cluster the data units in this particular analysis.[19]

McCrae's K-MEANS also allows three choices of diatnce measures between clusters. These are Euclidian distance, scaled Euclidian distance, and Mahalanobis distance. Assuming normal populations, $N(\theta_j, \Sigma_j)$, with equal covariance

---

[18] Op. Cit., Anderberg, M.R., p. 173-176

[19] Ibid.

matrices $\Sigma_1 = \Sigma_2 = \ldots = \Sigma$ so that the populations differ only in location, the Mahalanobis distance between the populations is given by $D^2 = (\theta_1 - \theta_j)^T \Sigma^{-1}(\theta_1 - \theta_j)$. This was the distance measure used in this cluster analysis.[20]

The question of how many clusters are present in the data was mentioned in the previous section. It can be shown that one prime indicator of the discriminability of variables in the data set is given by the log of the ratio det T/det W. When this quantity is plotted against the number of clusters one can gain insight as to the appropriate number of clusters within the data set. As the number of clusters is increased the ratio begins to reach a stabilizing value indicating that the discriminability of the data is decreasing. Thus, one can approximate the maximum number of natural clusters by observing when the curve levels off. It should be reemphasized that it is a primary objective of most cluster analysis problems to produce a set of clusters that are well differentiated from each other.

As stated before, when cluster analyses are performed on data with several variables actually measuring the same characteristic, it might be profitable to reduce the problem to one of only a few primary variables by the techniques of principal components analysis.

---

[20]Op. Cit., Press, S.J. p. 372-323

For this analysis, the computer program BIMED01M was utilized to extract the first principal component from the first two variables and the first principal component from the fifth, sixth, seventh and eighth variables. BIMED01M performs the following four basic steps: (1) the data are normed and centered; (2) the correlation matrix of the centered and normed data is computed; (3) the eigenvalues and corresponding eigenvectors of the correlation matrix are calculated; and (4) the centered and normed data are transformed into their orthogonal components.[21]

---

[21] BMD Manual, Biomedical Computer Programs, Health Sciences Computing Facility, UCLA, University of California Press, 1973, p. 193-201

VI. RESULTS OF ANALYSIS

The two data groups, control and accident, were first investigated by discriminant analysis with the use of the computer program BIMED04M.

The test for equality of group covariance matrices (or equivalently, group dispersion matrices) was performed according to the procedure developed by G. E. P. Box and illustrated in Appendix D. They were found to be equal at the .10 level of significance so it was appropriate to apply the discriminant analysis procedures.

Testing for the equality of group means, BIMED04M computed an F statistic of 8.94. For the $\alpha = .001$ level of significance, the tabled F value is $F_{n_1+n_2-m-1}(1-\alpha) = F^{10}_{50+50-10-1}(1-.001) = F^{10}_{89}(.999) = 3.39$ and one can conclude that there is definitely a difference in location of group means.

The computed discriminant function coefficients were (-0.00152, 0.00001, -0.00035, 0.00360, -0.00988, 0.00685, 0.00218, -0.00191, -0.00245, 0.00231). If after applying the coefficients to a data unit vector X_j, $-0.00151x_{j1} + 0.00001x_{j2} + \ldots + 0.00231x_{j,10} \leq 0$ then data unit j is assigned to group number two. Otherwise, the data unit is assigned to group number one.

Those subjects who had high values for the variables with positive coefficients and low values for the variables

with negative coefficients were classified as being in the control group, and those with opposite attributes were classified as belonging to the accident group.

The discriminant function was applied to the original data units to determine the performance of the function. Fifteen subjects of the fifty in the control group were classified as being in the accident group, while only four of the fifty in the accident group were classified as being in the control group. It is important to observe that although the overall misclassification rate is nineteen percent, the misclassification rate of the original accident group is only eight percent. This is encouraging. The question of identifying correctly those in the accident group is of greater concern than that of misclassifying those individuals in the control group.

To obtain the preceding results, it should be noted that the discriminant analysis was performed on the raw data as listed in Appendix A. An analysis was also performed on the standardized data, listed in Appendix B, but the results were much poorer. Using standardized data, the overall misclassification rate was fifty-five percent, quite a loss of discriminating power. It should be recognized that standardizing data has the drawback of providing answers to a problem different than the one originally posed.[22]

---

[22]Op. Cit., Press, S.J., p. 416.

In addition to learning the misclassification rates, it was also desired to determine which variables had the strongest effect on classifying the data units and in which direction the effect was observed. The discriminant function coefficients indicate whether each variable has a positive or negative effect, but because of the difference in magnitudes of the variables, the discriminant function coefficients alone do not tell how much of an effect. It is of interest, therefore, to compare how much a one standard deviation change in each variable will affect the discriminant function. Table I presents the standard deviations of each variable in the second column, the discriminant function coefficients in the third column, and in the last column the effect on the discriminant function of a one-sigma change in each variable.

TABLE I

Variable	Standard Deviation	Disc. Funct. Coefficient	Effect of a 1σ Change
1	6.03	-0.00152	-0.00916
2	1390.00	0.00001	0.01390
3	28.20	-0.00035	-0.00987
4	2.99	0.00360	0.01080
5	1.42	-0.00988	-0.01400
6	1.78	0.00685	0.01219
7	0.68	0.00218	0.00148
8	1.03	-0.00191	-0.00248
9	5.14	-0.00245	-0.01259
10	2.45	0.00231	0.00565

The results of Table I indicate that variable two (total hours) has the strongest positive effect in classifying a subject as not being in the accident group. A surprising result, however, is that variable five (time flown in the last twenty-four hours) has the strongest negative effect while variable six (time flown in the last forty-eight hours) has a strong positive effect. This would suggest that flying every other day is beneficial, but that too much flying (i.e. everyday) is detrimental. Similar interpretations can be made for the remaining variables although their effects are less pronounced.

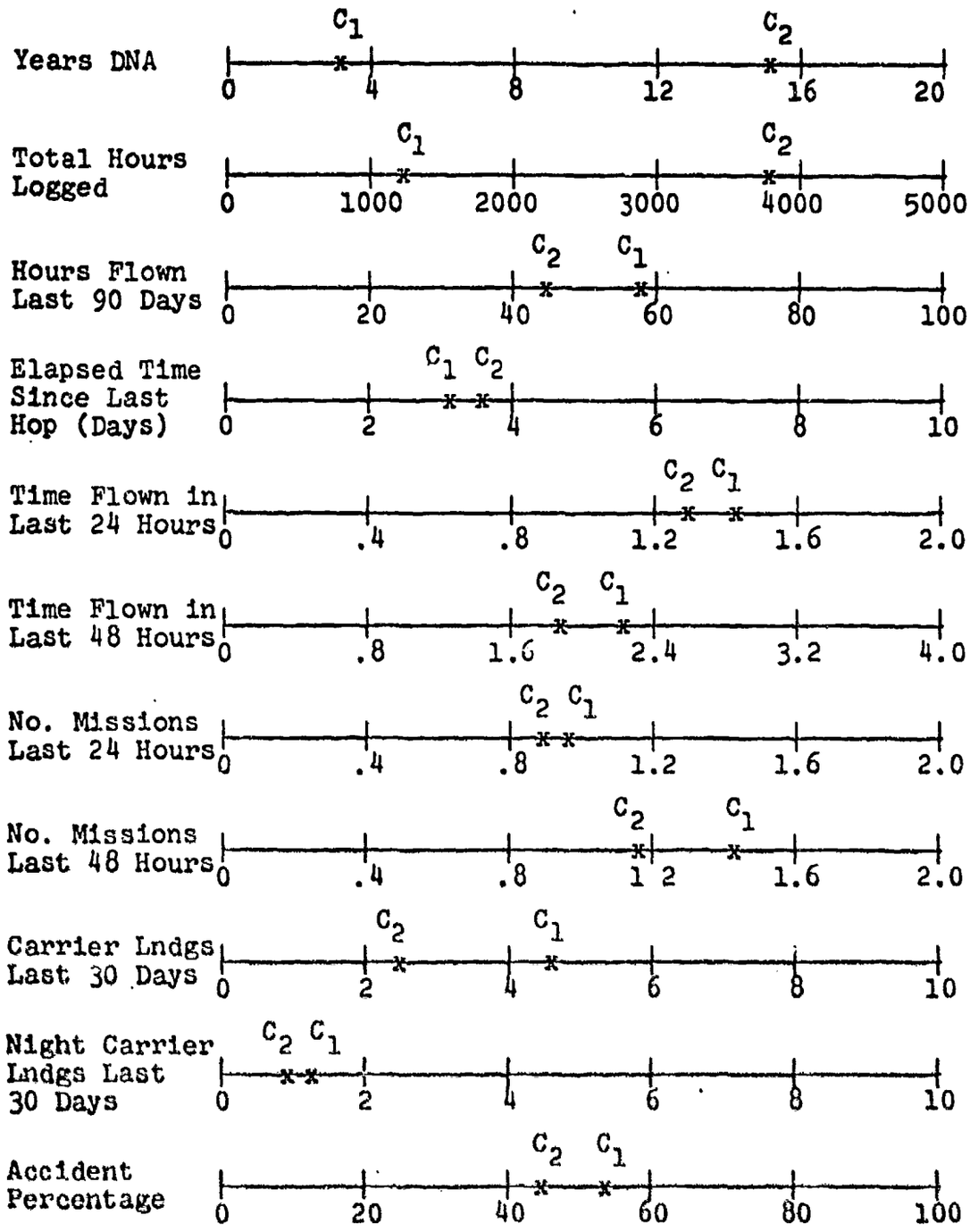
Although an overall misclassification rate of nineteen percent tends to indicate that there are meaningful differences between the two groups, the classification capabilities of the discriminant function are not as sharp as one would like. One cannot say with assurance how a pilot not initially a member of either group should be classified. It was desired to learn more about the variables' effects to be able to apply conclusions to subjects beyond the range of the data. To do this, a second method of analysis was employed; that of cluster analysis.

The first type of cluster analysis used was hierarchical clustering by data units. The computer program HI-CLUST was used with Euclidean distance measure between data units as the indicator of association. The results from both the single linkage and complete linkage methods were not at all satisfactory. When clustered into the final two groups,

one cluster consisted of ninety-nine units and the other of a single unit. The cluster of ninety-nine units was composed of clusters of ninety-three units and six units; again shedding no light on relation to accidents. Therefore, the clustering on data units was reworked using the non-hierarchical techniques of the computer program K-MEANS.

Initially, the one hundred data units were clustered into two groups to ascertain if there was any association directly with the two original groups, control and accident. Unfortunately, there did not appear to be any association, as cluster number one contained thirty-three subjects from the control group and thirty-seven from the accident group while cluster number two had seventeen and thirteen, respectively.

Figure (3) graphically depicts the cluster means of the two-group cluster results, and the number of subjects in the clusters. It is interesting to note that fifty-three percent of cluster number one was composed of subjects from the accident group while only forty-three percent of cluster number two was from the accident group. By inspecting the cluster means of variables one and two, one can see that the cluster compositions are inversely related to total experience, i.e. cluster number one has higher accident composition and fewer years designated naval avaiator and fewer total hours. The same kind of relation is seen to apply to the recency and frequency variables (variables four through ten) but the separation is not as great. Cluster number one which has the



Cluster Means

	Cluster One	Cluster Two	Total
Control Group	33	17	50
Accident Group	37	13	50
Total	70	30	

Number in Clusters

Figure (3)

higher accident composition has cluster means which indicate more recent and frequent flying than the subjects of cluster number two. Again, the results here are in basic agreement with those of discriminant analysis in that they indicate less frequent flying is beneficial. But of course, the support is very thin and the results are far from conclusive, especially since the cluster means are seen to be relatively close for all of variables four through ten.

It was stated in Section V that it is possible to get a rough idea of the number of natural clusters present in the data by plotting log(det T / det W) versus the number of clusters. Figure (4) is a plot of this information for the data under study. As the number of groups is increased the curve begins to level off. It appears that beyond nine groups there is not much additional information to be gained by grouping further.

The primary interest lies in the analysis of two groups, since there were two groups initially, and in the analysis of the natural number of groups. Between two and nine groups the results are believed to be less useful.

Figure (5) graphically portrays the cluster means of the nine cluster results, and the number of data units in each cluster. The relationships among clusters here are not apparent and there is no one-to-one correspondence such as an inverse relation between the cluster means of total hours flown and composition of clusters by accident percentages.

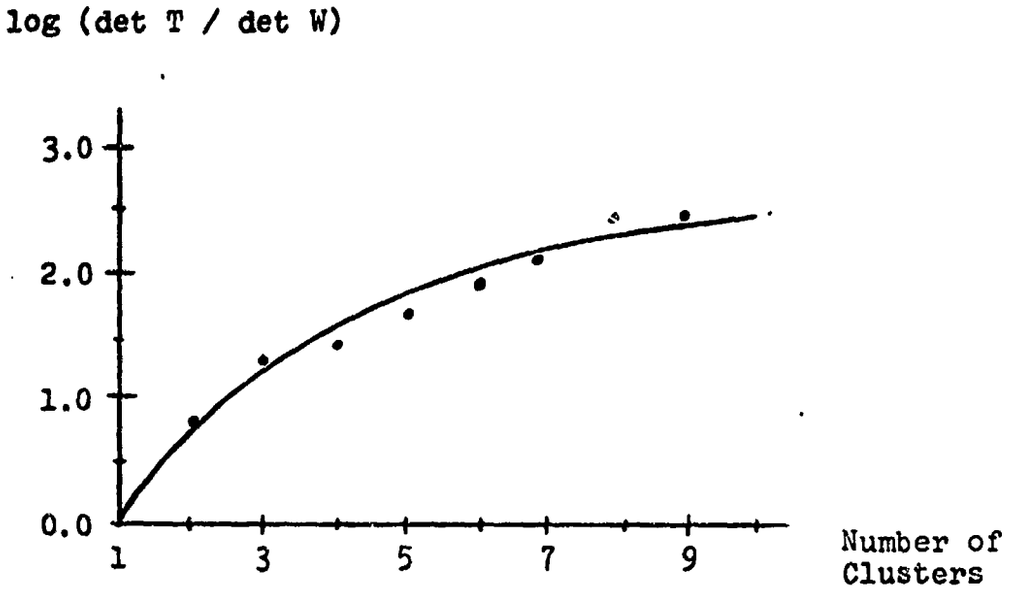
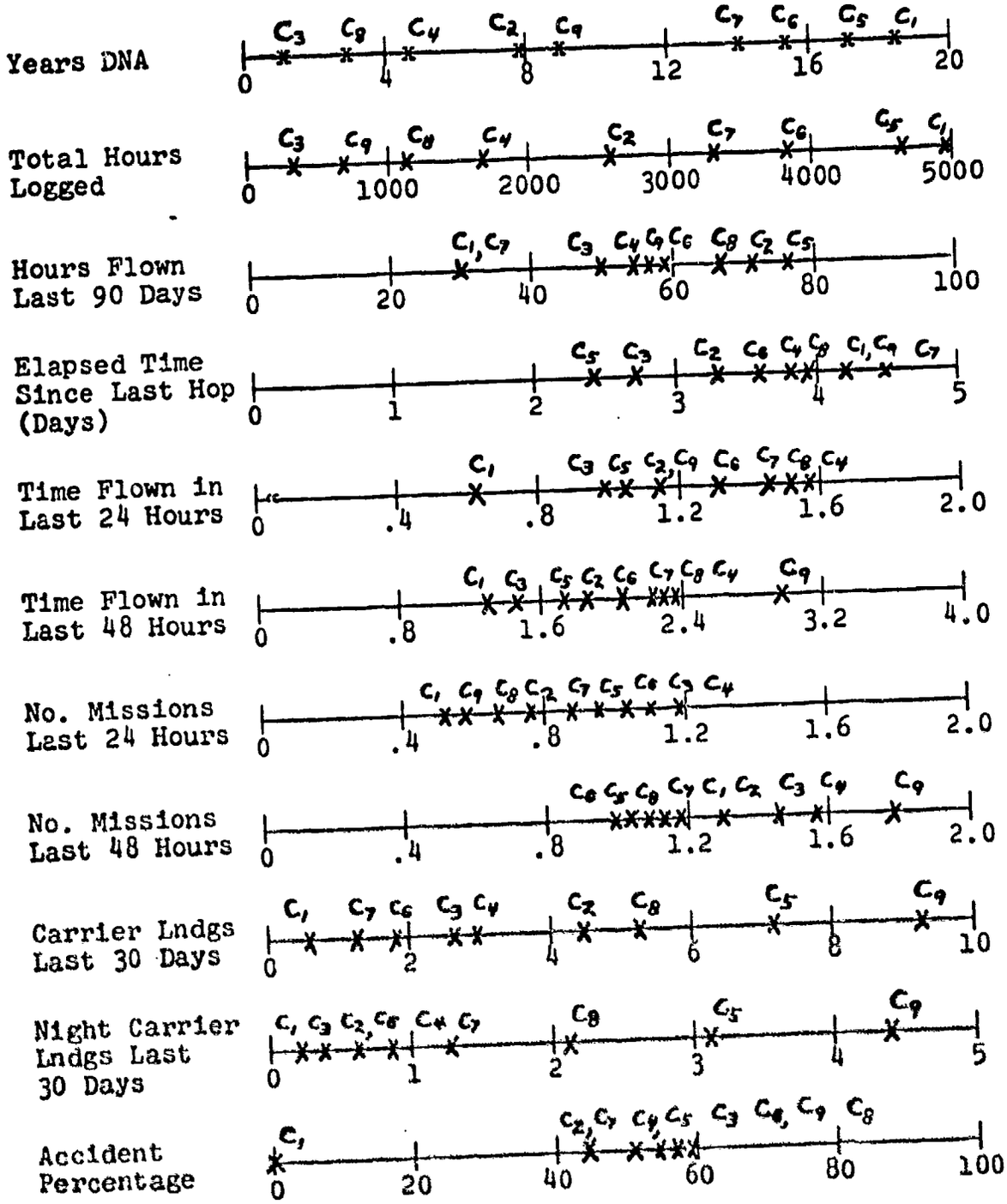


Figure (4)

It is desirable, therefore, to plot the proportion of each cluster from the accident group versus the cluster means for each variable. By so doing, trends might appear and factors influencing the accident proportions might become more readily observable. These plots are depicted in Appendix E as Figures (10) through (19). Figures (10) through (19) are similar in that none of them reveal any prominent relationships that their respective variables have with the proportion of the clusters composed of accident subjects. Intuitively, one might have hypothesized that as the cluster means increased (as in Fig. (11) for instance) that the proportion of the clusters composed of accident units would decrease. Since this kind of relationship did not appear for the total hours variable, nor did similarly anticipated



Cluster Means

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	Totals
Control Group	3	4	9	10	2	3	8	8	3	50
Accident Group	0	2	10	10	3	4	6	11	4	50
Totals	3	6	19	20	5	7	14	19	7	

Number in Clusters

Figure (5)

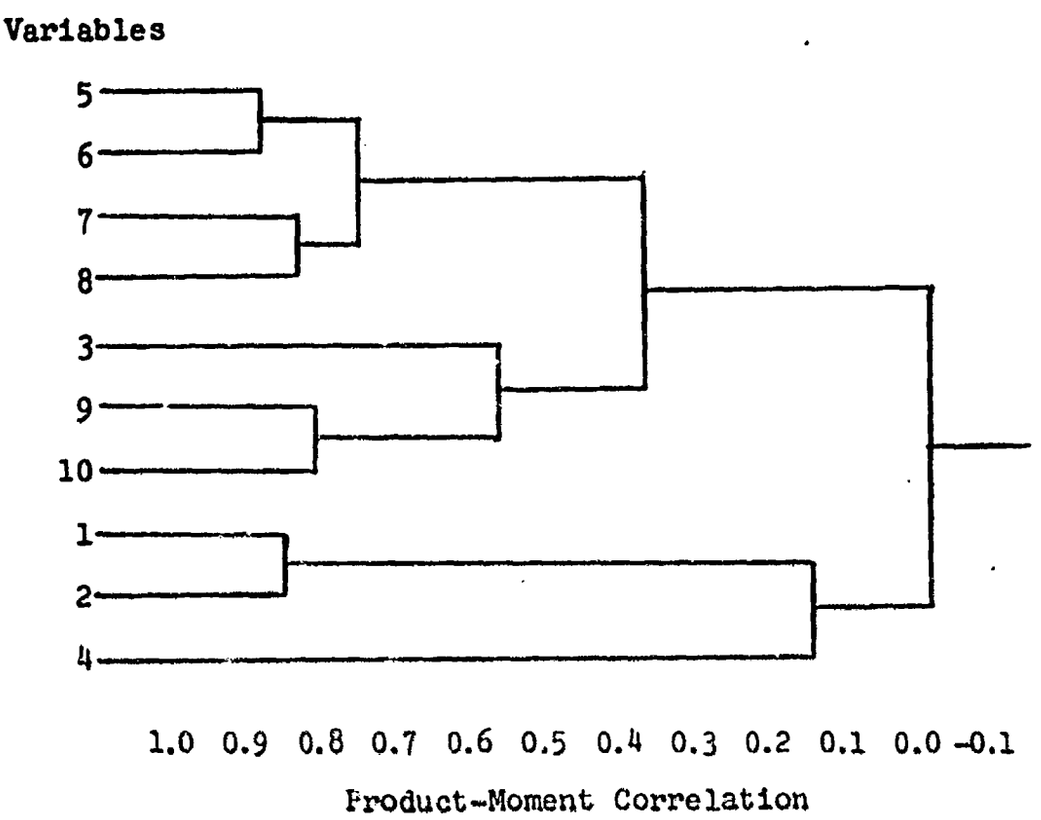
relations hold for the other variables, a final type of analysis was performed on the data.

It seemed plausible that although the variables individually did not reflect the contributions they had upon accidents, certain variables collectively might demonstrate such an effect. To determine which variables to combine, a hierarchical clustering analysis was performed by the computer program HI-CLUST. Product-moment correlation was used as the association measure between variables. Both the methods of single linkage and complete linkage clustering as discussed in Section V were employed. The results are shown as hierarchical trees (dendrograms) in Figures (6) and (7).

The results of both hierarchical methods are similar. Variables number one and two are highly correlated and variables five, six, seven and eight are highly correlated. Therefore, it was decided to combine those respective variables, calling the first the experience variable and the second the frequency variable. In order to eliminate all unnecessary or distracting influences it was also considered prudent to eliminate variables nine and ten since very few accidents involved carrier landings and many subjects in both groups were not involved in carrier operations during the period investigated.

As discussed in Sections IV and V, BIMED01M was used to extract the first principal components from those combinations of variables listed above to obtain the total

Hierarchical Tree for Variables
by Connections or Single Linkage Method



VARIABLE DEFINITIONS

1 - Years DNA
2 - Total Hours Logged
3 - Hours Flown in Last 90 Days
4 - Days Since Last Flight
5 - Time Flown in Last 24 Hours
6 - Time Flown in Last 48 Hours
7 - Missions in Last 24 Hours
8 - Missions in Last 48 Hours
9 - Car. Lndgs. Last 30 Days
10 - Night Car. Lndgs. in Last 30 Days

Figure (6)

Hierarchical Tree for Variables
by Diameter or Complete Linkage Method

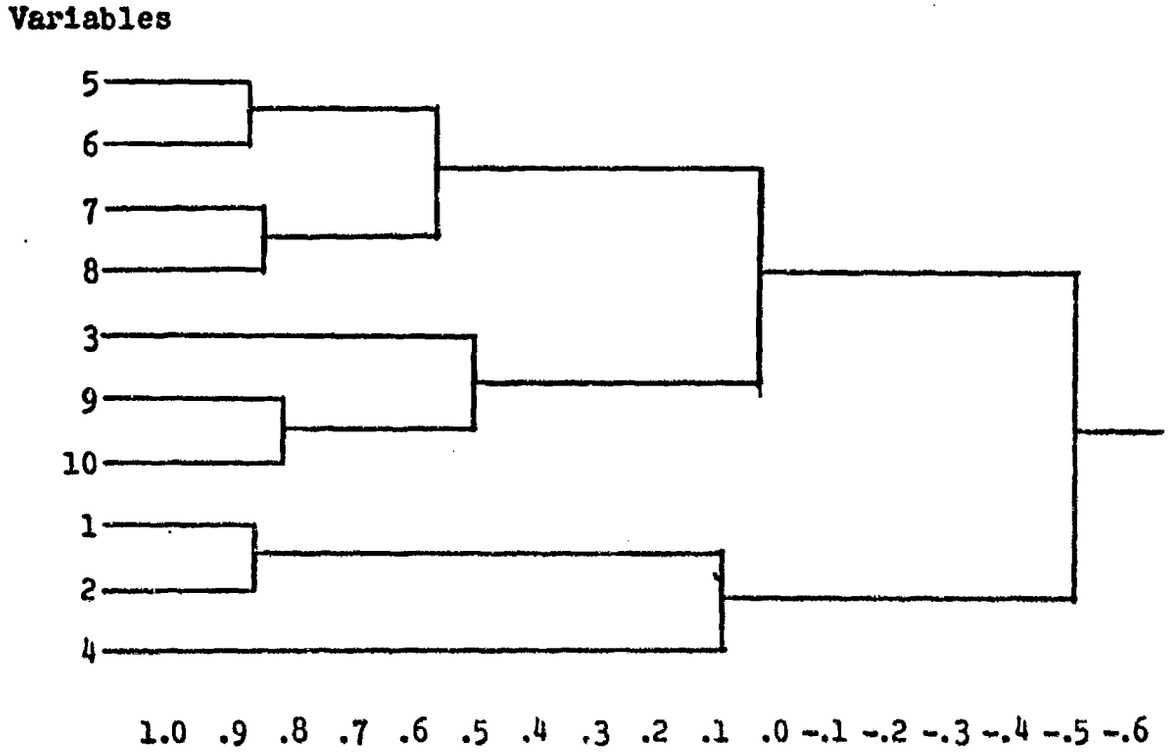
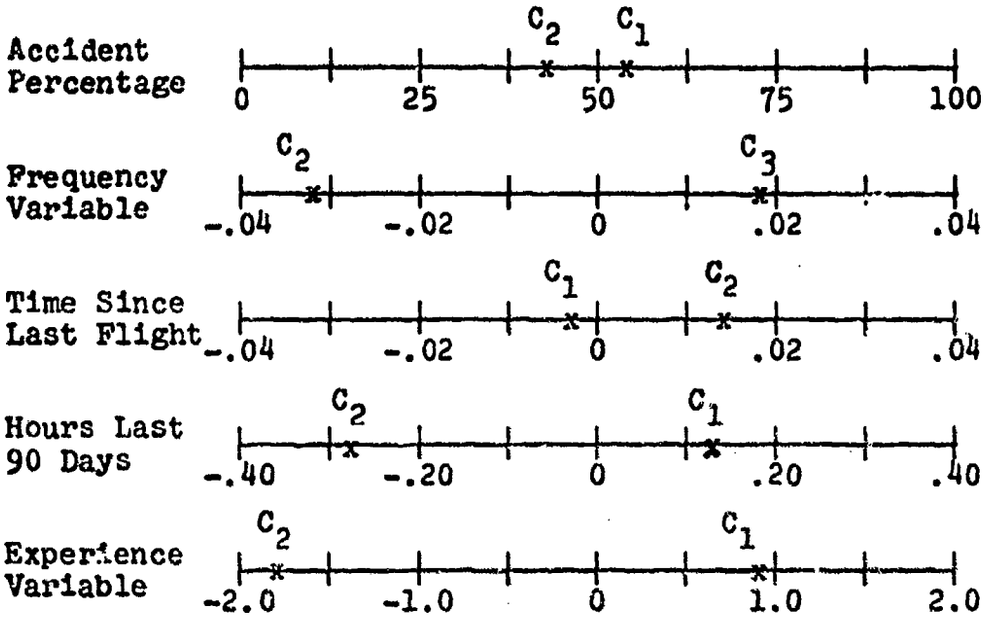


Figure (7)

experience variable and the frequency variable. The principal components (exhibited in Appendix G) were extracted from the standardized data (exhibited in Appendix B) as required by BIMED01M.

With the data reduced to four main variables, the cluster analysis program K-MEANS was again used to investigate the data. As was done with the data in ten variables (or regular space) the data was first investigated by clustering into just two groups. The cluster means are depicted in Figure (8). As was true in the regular space analysis, there does not appear to be any association between clustering of data units and membership in the accident group. Again, there were thirty-three subjects from the control group and thirty-seven from the accident group in cluster number one, and seventeen and thirteen respectively in cluster number two. Thus, fifty-three percent of cluster number one was from the accident group and forty-three percent of cluster number two was from the accident group.

The graph of log (det T / det W) versus number of clusters was plotted for the reduced space analysis in Figure (9) and was also found to indicate that beyond nine clusters, minimal information is gained. Therefore, a plot of the proportion of clusters from the accident group versus the cluster means was constructed for each of the four variables in the reduced space with nine cluster groupings. Figures (20) through (23) in Appendix H are graphs of the results.



Cluster Means (on standardized data)

	Cluster One	Cluster Two	Totals
Control Group	33	17	50
Accident Group	37	13	50
Totals	70	30	

Number in Clusters

Figure (8)

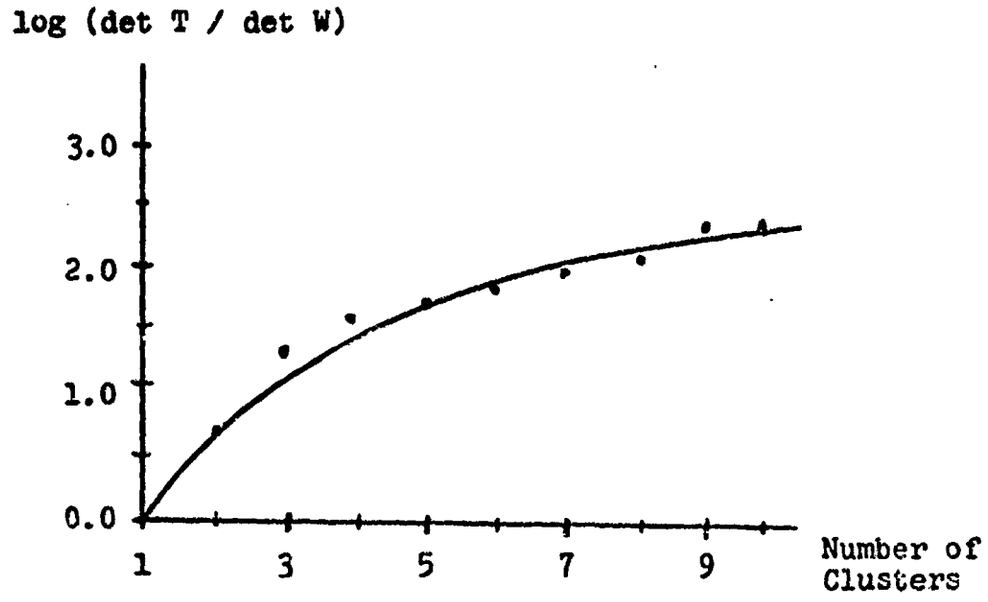


Figure (9)

The resultant plots in Appendix H are not too informative. Three of the four "new" variables do not appear to reveal any structure; but the first, the experience variable, may have some interest. A parabolic fit has been drawn in freehand and the accident rate seems to bottom out for experience in the interval (-1,0). This is misleading however. The interval (-1,0) of the experience variable corresponds to values of X_1 and X_2 which are between modes of their respective distributions. Only seven of the one-hundred aviators are in this range.

VII. CONCLUSIONS

The three analytical methodologies employed in this investigation were primarily utilized as exploratory tools to determine if there were significant differences in the various flight time statistics recorded for sample groups of pilots with and without accidents. The discriminant analysis techniques provided the best indication that there were differences which could be used to categorize the pilots according to the probability of belonging to the accident group.

It should be recognized that failure to distinguish among pilots according to their flight statistic attributes is not necessarily a fault of the analytical procedures, but inherent inability of the data as currently conceived to discriminate among subjects. This does not suggest, however, that this approach to accident analysis has no merit. It does point out the need to expand the investigation to include more quantitative aspects of flying. Many other variables such as instrument time, synthetic trainer time, number of instrument approaches, average time spent briefing flights, and subjective attributes such as training command flight grades and NATOPS quiz grades could be included. Breaking the investigation down into many more restrictive areas such as including only accidents in a particular phase of flight, or including only accidents by a particular type

of aircraft such as attack or patrol might also prove to be more relevant. It should be informative to expand the time span of the data base to include five or ten years so as to have a larger sample size on which to base results. Also, enlarging the size of the control group would help to eliminate the effects of non-randomness which could bias the data.

Despite the fact that the data investigated in this analysis did not contain those characteristics which could identify the underlying accident generating mechanism, it is still considered worthwhile to pursue the basic ideas developed here in future accident analysis.

APPENDIX A

TABLE II – Control Group Raw Data

	Variables									
Obs.	1	2	3	4	5	6	7	8	9	10
1	1	0137	10	10	0.5	0.7	0.4	0.6	00	00
2	15	3398	12	01	3.3	5.7	1.5	2.5	01	01
3	16	0589	36	05	0.9	1.8	0.5	0.9	00	00
4	12	4940	35	04	0.5	0.9	0.4	0.8	00	00
5	16	4436	18	04	0.6	1.7	1.6	2.0	03	00
6	14	3562	18	05	0.3	0.5	0.3	0.5	00	00
7	5	1681	32	05	0.9	1.6	0.5	0.7	00	00
8	8	1621	14	09	0.9	1.3	0.5	0.6	00	00
9	4	1539	47	04	0.8	1.9	0.7	1.0	00	00
10	2	0968	30	06	0.5	1.2	0.3	0.5	00	00
11	1	0320	41	03	0.7	1.4	0.8	1.4	00	00
12	3	1532	04	17	0.3	0.3	0.3	0.6	00	00
13	3	0803	42	03	1.1	1.8	0.9	1.3	03	01
14	2	1963	132	02	1.9	5.2	1.3	2.3	00	00
15	1	0325	36	03	0.6	1.6	1.1	1.6	00	00
16	1	0285	43	02	0.9	2.3	1.0	1.8	00	00
17	1	0187	32	03	0.7	1.3	1.0	1.3	00	00
18	2	0280	44	02	0.8	1.5	1.1	1.5	00	00
19	18	4134	64	03	0.8	2.2	0.6	0.9	00	00
20	16	3901	34	03	0.6	1.2	0.7	1.1	00	00
21	14	3549	13	13	0.2	0.3	0.2	0.2	00	00
22	10	3426	24	05	0.5	0.9	0.4	0.6	01	01
23	26	5592	44	03	0.7	1.8	0.7	1.8	00	00
24	4	2028	29	06	0.8	1.7	0.5	0.9	00	00
25	1	0507	46	03	0.9	1.6	1.1	1.8	03	01
26	4	1878	46	05	0.4	0.7	0.2	0.3	00	00
27	7	2031	18	10	0.8	1.7	0.4	0.8	00	00
28	4	1168	36	04	0.7	1.9	1.0	2.1	00	00
29	4	2523	52	04	0.8	1.3	0.5	0.7	00	00
30	3	1274	46	04	0.4	0.6	0.2	0.3	00	00
31	2	1107	55	04	0.6	1.3	0.3	0.5	00	00
32	1	0332	37	03	0.7	1.1	0.9	1.3	00	00
33	2	1511	69	02	0.8	1.4	0.9	1.3	00	00
34	5	2346	23	08	0.7	1.9	0.3	0.5	00	00
35	13	3310	39	04	1.4	2.7	0.9	1.5	01	00
36	12	2799	30	04	0.8	0.4	0.4	0.5	00	00
37	14	3335	14	03	0.8	0.6	0.5	0.7	00	00
38	13	3127	09	17	0.4	0.8	0.1	0.3	00	00
39	12	0301	31	06	1.5	3.4	1.0	1.4	00	00
40	10	3565	24	05	0.9	1.7	0.5	0.9	00	00

TABLE II (Continued)

Obs.	1	2	3	4	5	6	7	8	9	10
41	3	2673	76	02	2.3	4.0	1.3	1.9	10	03
42	5	1199	14	15	0.0	0.0	0.1	0.1	00	00
43	3	1143	38	05	0.3	0.4	0.1	0.1	00	00
44	3	0703	10	10	0.5	0.9	0.3	0.5	00	00
45	2	1000	46	05	0.5	0.8	0.7	0.9	00	00
46	3	1769	61	02	1.2	1.9	0.6	0.9	02	03
47	15	4172	31	04	1.5	3.1	0.9	1.4	00	00
48	16	0528	13	13	1.2	2.3	0.6	1.0	00	00
49	16	4843	82	02	1.6	2.9	0.8	1.1	06	03
50	17	5123	23	05	0.8	1.6	0.6	0.9	01	01

TABLE III – Accident Group Raw Data

	Variables									
Obs.	1	2	3	4	5	6	7	8	9	10
1	1	0405	78	01	0.3	1.4	1.0	1.0	09	01
2	6	2232	130	01	2.0	2.0	1.0	1.0	04	01
3	5	1760	90	05	0.9	0.9	1.0	1.0	00	00
4	3	0918	88	01	0.0	5.2	0.0	3.0	01	00
5	1	0421	64	01	3.0	6.0	2.0	4.0	00	00
6	1	0150	78	01	0.7	0.7	1.0	1.0	00	00
7	12	2860	45	00	3.5	4.8	2.0	3.0	06	03
8	3	1121	76	01	0.0	0.5	0.0	1.0	11	02
9	22	3487	17	00	1.1	3.5	1.0	2.0	00	00
10	7	2760	104	04	0.0	0.0	0.0	0.0	18	00
11	5	1724	92	00	3.7	5.5	2.0	4.0	17	07
12	5	0654	107	01	2.7	3.1	1.0	3.0	37	16
13	7	1462	79	01	3.4	5.4	2.0	3.0	15	00
14	1	0405	54	00	1.5	3.0	1.0	2.0	19	08
15	12	3204	133	03	4.7	4.7	1.0	1.0	04	06
16	1	0919	120	00	4.3	6.8	2.0	3.0	12	05
17	3	1142	106	07	0.0	0.0	0.0	0.0	16	12
18	1	0357	64	01	1.0	1.0	2.0	3.0	06	00
19	3	1085	132	01	1.9	1.9	1.0	1.0	07	04
20	2	0679	89	03	0.0	1.3	0.0	1.0	06	02
21	18	4001	77	03	1.3	1.3	1.0	1.0	06	00
22	4	1144	41	03	9.3	9.3	1.0	1.0	00	00
23	9	2658	40	01	1.7	3.2	2.0	4.0	00	00
24	3	1153	86	01	3.4	5.6	2.0	3.0	10	05
25	6	1762	57	02	0.0	2.0	0.0	1.0	05	00
26	1	0457	93	01	3.5	4.0	2.0	3.0	10	02
27	4	1241	69	02	1.8	2.9	1.0	2.0	11	00
28	3	1129	111	03	0.0	0.0	0.0	0.0	09	06
29	12	3933	27	03	0.0	0.0	0.0	0.0	00	00
30	18	4041	73	08	0.0	0.0	0.0	0.0	00	00
31	5	1640	26	00	2.1	3.7	2.0	3.0	00	00
32	12	3285	20	01	1.3	1.3	1.0	1.0	00	00
33	18	4461	86	03	0.0	0.0	0.0	0.0	09	04
34	5	1595	159	01	3.3	6.7	3.0	6.0	00	00
35	0	0329	26	01	1.0	1.0	1.0	1.0	00	00
36	3	0577	89	03	0.0	0.0	0.0	0.0	05	01
37	5	0866	43	01	4.2	4.2	2.0	2.0	01	00
38	12	3467	82	01	3.3	3.3	1.0	1.0	12	04
39	12	3772	109	01	6.6	6.6	3.0	3.0	07	04
40	3	0957	49	06	0.0	0.0	0.0	0.0	13	06
41	3	1393	32	01	1.0	1.0	1.0	1.0	06	01
42	5	1457	98	01	4.2	4.2	2.0	2.0	10	01
43	18	4737	77	01	1.7	1.7	1.0	1.0	11	06
44	5	1500	45	01	2.5	2.5	1.0	1.0	00	00
45	2	0594	80	02	0.0	2.7	0.0	1.0	10	06
46	5	1454	21	00	3.3	3.3	3.0	3.0	00	00
47	18	2939	26	01	0.6	0.6	1.0	1.0	00	00
48	1	0513	56	03	0.0	0.0	0.0	0.0	11	02
49	3	1049	47	00	2.0	4.5	1.0	2.0	00	00
50	1	0241	75	00	2.5	3.9	2.0	3.0	00	00

APPENDIX B

TABLE IV – Control Group Standardized Data

	Variables				
Obs.	1	2	3	4	5
1	-1.0761	-1.3095	-1.1829	1.1957	-0.5698
2	1.1200	0.8506	-1.0929	-1.1437	4.3852
3	1.2769	-1.0101	-0.0126	-0.1040	0.1380
4	0.6494	1.8721	-0.0576	-0.3639	-0.5698
5	1.2769	1.5581	-0.0576	-0.3639	-0.3929
6	0.9631	0.9593	-0.8228	-0.1040	-0.9238
7	-0.4486	-0.2867	-0.1926	-0.1040	0.1380
8	0.0220	-0.3265	-1.0028	0.9358	0.1380
9	-0.6055	-0.3808	0.4825	-0.3639	0.1380
10	-0.9192	-0.7591	-0.2827	0.1560	-0.5698
11	-1.0761	-1.1883	0.2124	-0.6239	-0.2159
12	-0.7624	-0.3855	-1.4529	3.0153	-0.9238
13	-0.7624	-0.8684	0.2575	-0.6239	0.4920
14	-0.9192	-0.0999	4.3804	-0.8838	1.9077
15	-1.0761	-1.1850	-0.0126	-0.6239	-0.3929
16	-1.0761	-1.2115	0.3025	-0.8838	0.1380
17	-1.0761	-1.2764	-0.1926	-0.6239	-0.2159
18	-0.9192	-1.2148	0.3375	-0.8838	-0.0389
19	1.5906	1.3382	1.2477	-0.6239	-0.0389
20	1.2769	1.1838	-0.1026	-0.6239	-0.3929
21	0.9631	0.9507	-1.0478	1.9755	-1.1007
22	0.3357	0.8692	-0.5527	-0.1040	-0.5698
23	2.8455	2.3040	0.3475	-0.6239	-0.2159
24	-0.6055	-0.0569	-0.3277	0.1560	-0.0389
25	-1.0761	-1.0644	0.4375	-0.6239	0.1380
26	-0.6055	-0.1563	0.4375	-0.1040	-0.7468
27	-0.1349	-0.0549	-0.8228	1.1957	-0.0389
28	-0.6055	-0.6266	-0.0126	-0.3639	-0.2159
29	-0.6055	0.2710	0.7076	-0.3639	-0.0389
30	-0.7624	-0.5564	0.4375	-0.3639	-0.7468
31	-0.9192	-0.6670	0.3875	-0.3639	-0.3929
32	-1.0761	-1.1804	0.0324	-0.6239	-0.2159
33	-0.9192	-0.3994	1.4727	-0.8838	-0.0389
34	-0.4486	0.1538	-0.4677	0.6758	-0.2159
35	1.5906	0.7223	0.1124	-0.3639	1.0229
36	0.6494	0.4538	-0.7328	-0.3639	-0.9238
37	0.9631	0.7433	-1.0028	-0.6239	-0.9238
38	0.8063	0.6711	-1.0279	3.0153	-1.0777
39	0.6494	-1.0032	-0.2377	0.1560	1.1998
40	0.3357	0.9613	-0.5527	-0.1040	0.1380

TABLE IV (Continued)

	Variables				
Obs.	6	7	8	9	10
1	-0.8123	-0.6847	-0.7165	-0.3559	-0.3736
2	3.6807	2.3281	2.5094	0.2181	0.9608
3	0.1761	-0.4108	-0.2071	-0.3559	-0.3736
4	-0.6326	-0.6847	-0.3769	-0.3559	-0.3736
5	0.0863	2.6020	1.6605	1.3660	-0.3736
6	-0.9921	-0.9586	-0.8862	-0.3559	-0.3736
7	-0.0036	-0.4108	-0.5467	-0.3559	-0.3736
8	-0.2732	-0.4108	-0.7165	-0.3559	-0.3736
9	0.2660	0.1370	-0.0373	-0.3559	-0.3736
10	-0.3630	-0.9586	-0.8862	-0.3559	-0.3736
11	-0.1833	0.4109	0.6418	-0.3559	-0.3736
12	-1.1718	-0.9586	-0.7165	-0.3559	-0.3736
13	0.1761	0.6848	0.4720	1.3660	0.9608
14	3.2314	1.7803	2.3396	-0.3559	-0.3736
15	-0.0036	1.2325	0.9813	-0.3559	-0.3736
16	0.6254	0.9586	1.3209	-0.3559	-0.3736
17	-0.2732	0.9586	0.4720	-0.3559	-0.3736
18	-0.0934	1.2325	0.8116	-0.3559	-0.3736
19	0.5356	-0.1369	-0.2071	-0.3559	-0.3736
20	-0.3630	0.1370	0.1324	-0.3559	-0.3736
21	-1.1718	-1.2325	-1.3956	-0.3559	-0.3736
22	-0.6326	-0.6847	-0.7165	0.2181	0.9608
23	0.1761	0.1370	1.3209	-0.3559	-0.3736
24	0.0863	-0.4108	-0.2071	-0.3559	-0.3736
25	-0.0036	1.2325	1.3209	1.3660	0.9608
26	-0.8123	-1.2325	-1.2258	-0.3559	-0.3736
27	0.0863	-0.6847	-0.3769	-0.3559	-0.3736
28	0.2660	0.9586	1.8302	-0.3559	-0.3736
29	-0.2732	-0.4108	-0.5467	-0.3559	-0.3736
30	-0.9022	-1.2325	-1.2258	-0.3559	-0.3736
31	-0.2732	-0.9586	-0.8862	-0.3559	-0.3736
32	-0.4529	0.6848	0.4720	-0.3559	-0.3736
33	-0.1833	0.6848	0.3022	-0.3559	-0.3736
34	0.2660	-0.9586	-0.8862	-0.3559	-0.3736
35	0.9849	0.6848	0.9813	0.2181	-0.3736
36	-1.0819	-0.6847	-0.8862	-0.3559	-0.3736
37	-0.9022	-0.4108	-0.5467	-0.3559	-0.3736
38	-1.2616	-1.5064	-1.3956	-0.3559	-0.3736
39	0.7153	0.9586	0.6418	-0.3559	-0.3736
40	0.0863	-0.4108	-0.2071	-0.3559	-0.3736

TABLE IV (Continued)

	Variables				
Obs.	1	2	3	4	5
41	-0.2918	0.3704	1.7878	-0.8838	2.6156
42	-0.4486	-0.6060	-1.0028	2.4954	-1.4547
43	-0.7624	-0.6431	0.0774	-0.1040	-0.9238
44	-0.7624	-0.9346	-1.1829	1.1957	-0.5698
45	-0.9192	-0.7379	0.4375	-0.1040	-0.5698
46	-0.7624	-0.2285	1.1127	-0.8838	0.6689
47	1.1200	1.3633	-0.2377	-0.3639	1.1998
48	1.2769	-1.0505	-1.0478	1.9755	0.6689
49	1.2769	1.8078	2.0579	-0.8838	1.3768
50	1.4337	1.0933	-0.5977	-0.1040	-0.0389

TABLE IV (Continued)

	Variables				
Obs.	6	7	8	9	10
41	2.1531	1.7803	1.4907	5.3837	3.6296
42	-1.4414	-1.5064	-1.5654	-0.3559	-0.3736
43	-1.0819	-1.5064	-1.5654	-0.3559	-0.3746
44	-0.6326	-0.9586	-0.8862	-0.3559	-0.3736
45	-0.7225	0.1370	-0.2071	-0.3559	-0.3736
46	0.2660	-0.1369	-0.2071	0.7921	3.6296
47	1.3443	0.6848	0.6418	-0.3559	-0.3736
48	0.6254	-0.1369	-0.0373	-0.3559	-0.3736
49	1.1646	0.4109	0.1324	3.0879	3.6296
50	-0.0036	-0.1369	-0.2701	0.2181	0.9608

TABLE V – Accident Group Standardized Data

	Variables				
Obs.	1	2	3	4	5
1	-0.9374	-1.0456	0.1413	-0.4195	-0.8172
2	-0.0531	0.4047	1.7109	-0.4195	0.0621
3	-0.2299	0.0300	0.5035	1.8479	-0.5069
4	-0.5836	-0.6384	0.4431	-0.4195	-0.9724
5	-0.9374	-1.0329	-0.2813	-0.4195	-0.5793
6	-0.9374	-1.2480	0.1413	-0.4195	-0.6103
7	1.0081	0.9032	-0.8548	-0.9863	0.8379
8	-0.5836	-0.4772	0.0809	-0.4195	-0.9724
9	2.7767	1.4009	-1.7000	-0.9863	-0.4034
10	0.1238	0.8238	0.9261	1.2810	-0.9724
11	-0.2299	0.0014	0.5639	-0.9863	0.9414
12	-0.2299	-0.8479	1.0166	-0.4195	0.4241
13	0.1238	-0.2065	0.1715	-0.4195	0.7862
14	-0.9374	-1.0456	-0.5832	-0.9863	-0.1965
15	1.0081	1.1762	1.8014	0.7142	1.4586
16	-0.9374	-0.6376	1.4090	-0.9863	1.2517
17	-0.5836	-0.4606	0.9864	2.9815	-0.9724
18	-0.9374	-1.0837	-0.2813	-0.4195	-0.4552
19	-0.5836	-0.5058	1.7713	-0.4195	0.0103
20	-0.7605	-0.8281	0.4733	0.7142	-0.9724
21	2.0693	1.8089	0.1111	0.7142	-0.9724
22	-0.4068	-0.4590	-0.9756	0.7142	3.8378
23	0.4775	0.7428	-1.0058	-0.4195	-0.0931
24	-0.5836	-0.4518	0.3827	-0.4195	0.7862
25	-0.0531	0.0316	-0.4926	0.1474	-0.9724
26	-0.9374	-1.0043	0.5940	-0.4195	0.8379
27	-0.4068	-0.3820	-0.1304	0.1474	-0.0414
28	-0.5836	-0.4709	1.1374	0.7142	-0.9724
29	1.0081	1.7549	-1.3982	0.7142	-0.9724
30	2.0639	1.8406	-0.0097	3.5484	-0.9724
31	-0.2299	-0.0652	-1.4284	-0.9863	0.1183
32	1.0081	1.2405	-1.6095	-0.4195	-0.3000
33	2.0693	2.1740	0.3827	0.7142	-0.0724
34	-0.2299	-0.1010	2.5862	-0.4195	0.7345
35	-1.1142	-1.1059	-1.4284	-0.4195	-0.4552
36	-0.5836	-0.9090	0.4733	0.7142	-0.9724
37	-0.2299	-0.6796	-0.9152	-0.4195	1.2000
38	1.0081	1.3850	0.2620	-0.4195	0.7345
39	1.0081	1.6271	1.0770	-0.4195	2.4413
40	-0.5836	-0.6074	-0.7341	2.4147	-0.9724
41	-0.5836	-0.2613	-1.2472	-0.4195	-0.4552
42	-0.2299	-0.2105	0.7450	-0.4195	1.2000
43	2.0693	2.3931	0.1111	-0.4195	-0.0931
44	-0.3299	-0.1754	-0.8548	-0.4195	0.3207
45	-0.7605	-0.9035	0.2016	0.1474	-0.9724
46	-0.2299	-0.1091	-1.5793	-0.9863	0.7345
47	2.0693	0.9559	-1.4284	-0.4195	-0.6620
48	-0.9374	-0.9508	-0.5228	0.7142	-0.9724
49	-0.5836	-0.5344	-0.7945	-0.9863	0.0621
50	-0.9374	-1.1757	0.0507	-0.9863	0.3207

TABLE V (Continued)

	Variables				
Obs.	6	7	8	9	10
1	-0.5903	-0.1140	-0.5100	0.3115	-0.3842
2	-0.3252	-0.1140	-0.5100	-0.3965	-0.3842
3	-0.8113	-0.1140	-0.5100	-0.9628	-0.6797
4	1.0888	-1.2536	0.9901	-0.8212	-0.6797
5	1.4423	1.0256	1.7401	-0.9628	-0.6797
6	-0.8997	-0.1140	-0.5100	-0.9628	-0.6797
7	0.9120	1.0256	0.9901	-0.1133	0.2069
8	-0.9880	-1.2536	-0.5100	0.5947	-0.0887
9	0.3376	-0.1140	0.2400	-0.9628	-0.6797
10	-1.2090	-1.2536	-1.2601	1.5858	-0.6797
11	1.2213	1.0256	1.7401	1.4442	1.3890
12	0.1608	-0.1140	0.9901	4.2761	4.0487
13	1.1772	1.0256	0.9901	1.1611	-0.6797
14	0.1167	-0.1140	0.2400	1.7274	1.6845
15	0.8678	-0.1140	-0.5100	-0.3965	1.0935
16	1.7958	1.0256	0.9901	0.7363	0.7979
17	-1.2090	-1.2536	-1.2601	1.3026	2.8666
18	-0.7671	1.0256	0.9901	-0.1133	-0.6797
19	-0.3694	-0.1140	-0.5100	0.0283	0.5024
20	-0.7671	-1.2536	-0.5100	-0.1133	-0.0887
21	-0.6345	-0.1140	-0.5100	-0.1133	-0.6797
22	2.9005	-0.1140	-0.5100	-0.9628	-0.6797
23	0.2050	1.0256	1.7401	-0.9628	-0.6797
24	1.2655	1.0256	0.9901	0.4531	0.7979
25	-0.3252	-1.2536	-0.5100	-0.2549	-0.6797
26	0.5585	1.0256	0.9901	0.4531	-0.8807
27	0.0725	-0.1140	0.2400	0.5947	-0.6797
28	-1.2090	-1.2536	-1.2601	0.3115	1.0935
29	-1.2090	-1.2536	-1.2601	-0.9628	-0.6797
30	-1.2090	-1.2536	-1.2601	-0.9628	-0.6797
31	0.4260	1.0256	0.2400	-0.9628	-0.6797
32	-0.6345	-0.1140	-0.5100	-0.9628	-0.6797
33	-1.2090	-1.2536	-1.2601	0.3115	0.5024
34	1.7516	2.1653	3.2402	-0.9628	-0.6797
35	-0.7671	-0.1140	-0.5100	-0.9628	-0.6797
36	-1.2090	-1.2536	-1.2601	-0.2549	-0.3842
37	0.6469	1.0256	0.2400	-0.8212	-0.6797
38	0.2492	-0.1140	-0.5100	0.7363	0.5024
39	1.7074	2.1653	0.9901	0.0283	0.5024
40	-1.2090	01.2563	-1.2601	0.8779	1.0935
41	-0.7671	-0.1140	-0.5100	-0.1133	-0.3842
42	0.6469	1.0256	0.2400	0.4531	-0.3842
43	-0.4578	-0.1140	-0.5100	0.5947	1.0935
44	-0.1043	-0.1140	-0.5100	-0.9628	-0.6797
45	-0.0159	-1.2536	-0.5100	0.4531	1.0935
46	0.2492	2.1653	0.9901	-0.9628	-0.6797
47	-0.9438	-0.1140	-0.5100	-0.9628	-0.6797
48	-1.2090	-1.2536	-1.2601	0.5947	-0.6887
49	0.7295	-0.1140	0.2400	-0.1133	-0.6797
50	-0.4202	1.0256	0.9901	-0.9628	-0.6797

APPENDIX C

DISPERSION MATRICES

An element of the group dispersion matrix is given by the formula

$$\frac{1}{N-1} \sum_{k=1}^{50} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

where i = 1,2,...,10 and j = 1,2,...,10. An element of the dispersion matrix is readily seen to differ from an element of the covariance matrix only by the factor $\frac{1}{N-1}$ where N is the number of data units or observations.

An element of the pooled within groups dispersion matrix is given by the formula

$$\frac{1}{N_1 + N_2 - 2} \sum_{\ell=1}^{2} \sum_{k=1}^{50} (x_{ik\ell} - \bar{x}_{i\ell})(x_{jk\ell} - \bar{x}_{j\ell})$$

where i = 1,2,...,10 and j = 1,2,...,10, and N_1 and N_2 are the number of observations of the respective groups.

TABLE VI – Control Group Dispersion Matrix

		Variable 1	2	3	4	5
	1	41.47	7700.41	-25.90	0.71	0.58
	2	7700.41	2325433.00	827.55	-608.69	95.87
	3	-25.90	827.55	503.67	-52.11	4.87
	4	0.71	-608.69	-52.11	15.10	-0.95
Variable	5	0.58	95.87	4.87	-0.95	0.33
	6	1.21	249.13	12.76	-2.00	0.61
	7	0.05	-9.39	3.48	-0.84	0.14
	8	0.29	24.17	5.74	-1.30	0.23
	9	0.78	538.38	14.62	-1.76	0.48
	10	0.22	235.05	6.04	-0.83	0.21

		Variable 6	7	8	9	10
	1	1.21	0.05	0.29	0.78	0.22
	2	249.13	-9.39	24.17	538.38	235.05
	3	12.76	3.48	5.74	14.62	6.04
	4	-2.00	-0.84	-1.30	-1.76	-0.83
Variable	5	0.61	0.14	0.23	0.48	0.21
	6	1.26	0.30	0.52	0.77	0.32
	7	0.30	0.14	0.21	0.27	0.07
	8	0.52	0.21	0.35	0.34	0.09
	9	0.77	0.27	0.34	3.10	1.11
	10	0.32	0.07	0.09	1.11	0.57

TABLE VII – Accident Group Dispersion Matrix

Variable	1	2	3	4	5
1	32.62	6776.32	-31.67	1.43	-0.45
2	6776.32	1619479.00	-2719.64	420.05	-9.87
3	-31.67	-2719.64	1119.93	9.76	7.69
4	1.43	420.05	9.76	3.18	-1.23
5	-0.45	-9.87	7.69	-1.23	3.81
6	-1.46	-322.23	11.15	-1.97	3.82
7	-0.32	-66.75	0.93	-0.93	1.17
8	-1.33	-346.63	6.47	-1.43	1.29
9	-6.55	-1304.87	86.96	0.60	-0.37
10	-1.42	-269.94	44.58	0.86	0.29

Variable	6	7	8	9	10
1	-1.46	-0.32	-1.33	-6.55	-1.42
2	-322.23	-66.75	-346.63	-1304.87	-269.94
3	11.15	0.93	6.47	86.96	44.58
4	-1.97	-0.93	-1.43	0.60	0.86
5	3.82	1.17	1.29	-0.37	0.29
6	5.23	1.35	2.24	-0.85	-0.10
7	1.35	0.79	0.97	-0.96	-0.44
8	2.24	0.97	1.81	-0.13	-0.19
9	-0.85	-0.96	-0.13	50.90	19.04
10	-0.10	-0.44	-0.19	19.04	11.68

TABLE VIII – Pooled Within Groups Dispersion Matrix

		Variable 1	2	3	4	5
	1	37.05	7238.36	-28.78	1.07	0.06
	2	7238.36	197256.00	-946.04	-94.32	43.00
	3	-28.78	-946.04	811.80	-21.18	6.28
	4	1.07	-94.32	-21.18	9.14	-1.09
Variable	5	0.06	43.00	6.28	-1.09	2.07
	6	-0.13	-36.55	11.96	-1.98	2.21
	7	-0.13	-38.07	2.20	-0.89	0.66
	8	-0.52	-161.23	6.11	-1.37	0.76
	9	-2.88	-383.24	50.79	-0.58	0.05
	10	-0.60	-17.44	25.31	0.01	0.25

		Variable 6	7	8	9	10
	1	-0.13	-0.13	-0.52	-2.88	-0.60
	2	-36.55	-38.07	-161.23	-383.24	-17.44
	3	11.96	2.20	6.11	50.79	25.31
	4	-1.98	-0.89	-1.37	-0.58	0.01
Variable	5	2.21	0.66	0.76	0.05	0.25
	6	3.24	0.83	1.38	-0.04	0.11
	7	0.83	0.46	0.59	-0.35	-0.18
	8	1.38	0.59	1.08	0.11	-0.05
	9	-0.04	-0.35	0.11	27.00	10.07
	10	0.11	-0.18	-0.05	10.07	6.13

APPENDIX D

STATISTICAL TEST FOR EQUALITY OF DISPERSION MATRICES

Given a sample of two groups and m variables with group dispersion matrices S_1 and S_2, pooled within groups dispersion matrix S_W, and total sample observations $N = N_1 + N_2$, the hypothesis that the dispersion matrices (and thus the covariance matrices) are statistically equal may be determined by the following computations: [22]

$$A = \ln[|S_w|] \cdot (N-2) - (N_1-1)\cdot\ln[|S_1|] - (N_2-1)\cdot\ln[|S_2|]$$

$$B = \frac{\left[\frac{1}{1-N_1} + \frac{1}{1-N_2} - \frac{1}{N-2}\right] \cdot (2m^2 - 3m - 1)}{6(2 - 1)(m + 1)}$$

$$C = \frac{\left[\frac{1}{(N_1+1)^2} + \frac{1}{(N_2+1)^2} - \frac{1}{(N-2)^2}\right] \cdot (m - 1)(m + 2)}{6}$$

$$D = \frac{m(m + 1)}{2}$$

$$E = \frac{D + 2}{\text{abs } |B^2 - C|}$$

---

[22] Box, G.E.P., "A General Distribution Theory for a Class of Likelihood Criteria," Biometrika 36 (1949), p. 317-346

If B^2 is greater than C, the test statistic is:

$$\left(\frac{E}{D}\right)\left[\frac{A(1 - B + 2/E)}{E - A(1 - B - 2/E)}\right] \sim F_E^D$$

If C is greater than B^2 then the test statistic is:

$$\left(\frac{A}{D}\right) \cdot \left(1 - B - \frac{D}{E}\right) \sim F_E^D$$

APPENDIX E

Figures (10) through (19) depict cluster analysis plots for each of the ten variables being studied. The label "P" on the vertical axis of each figure represents the proportion of each cluster that originates from the accident group.

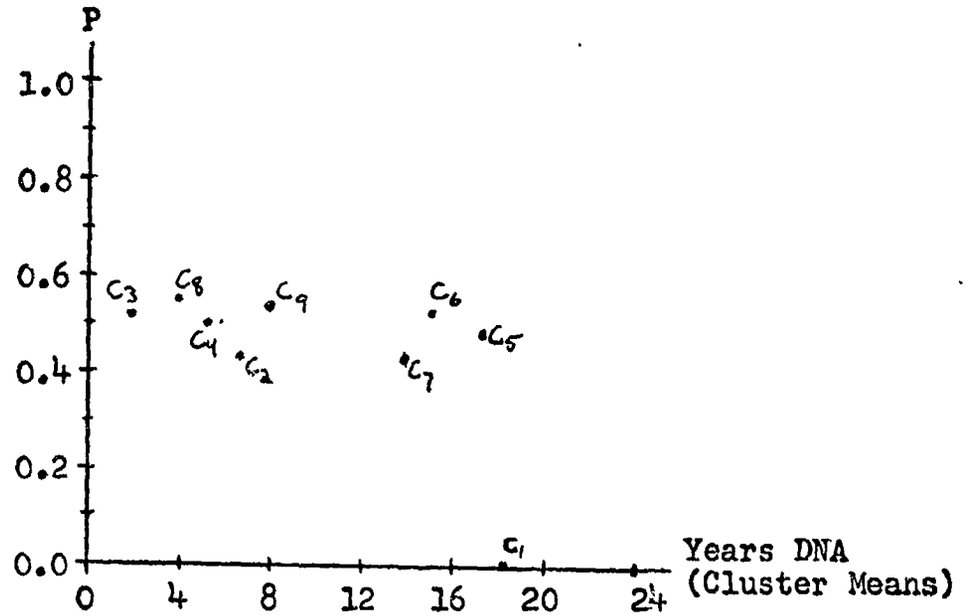


Figure (10)

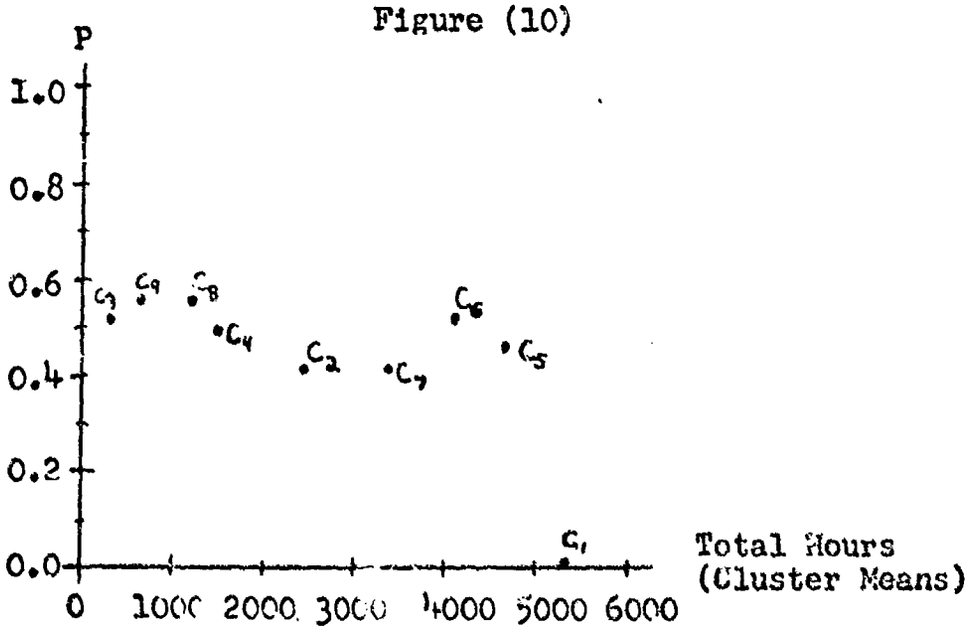


Figure (11)

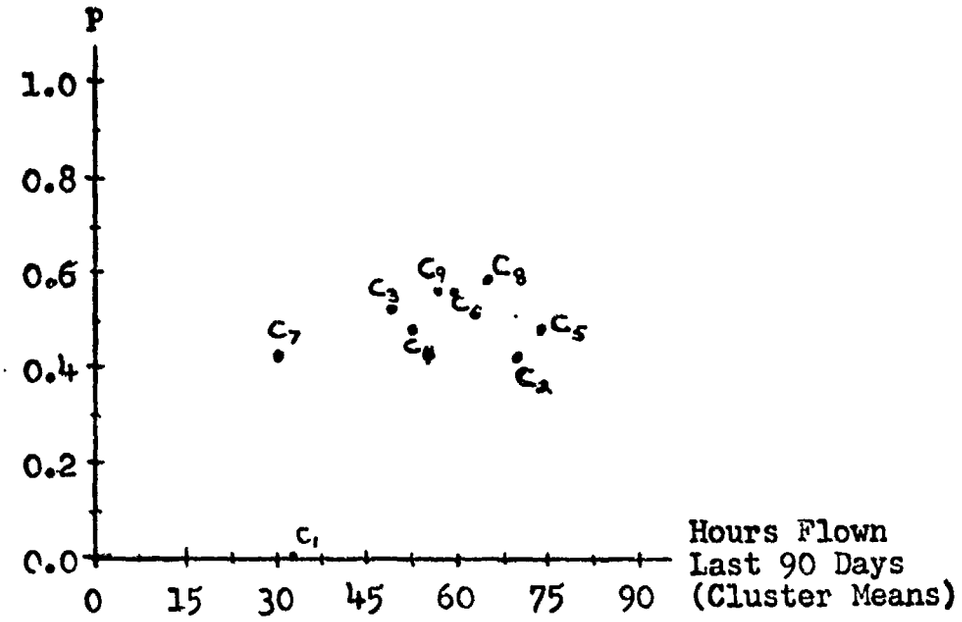


Figure (12)

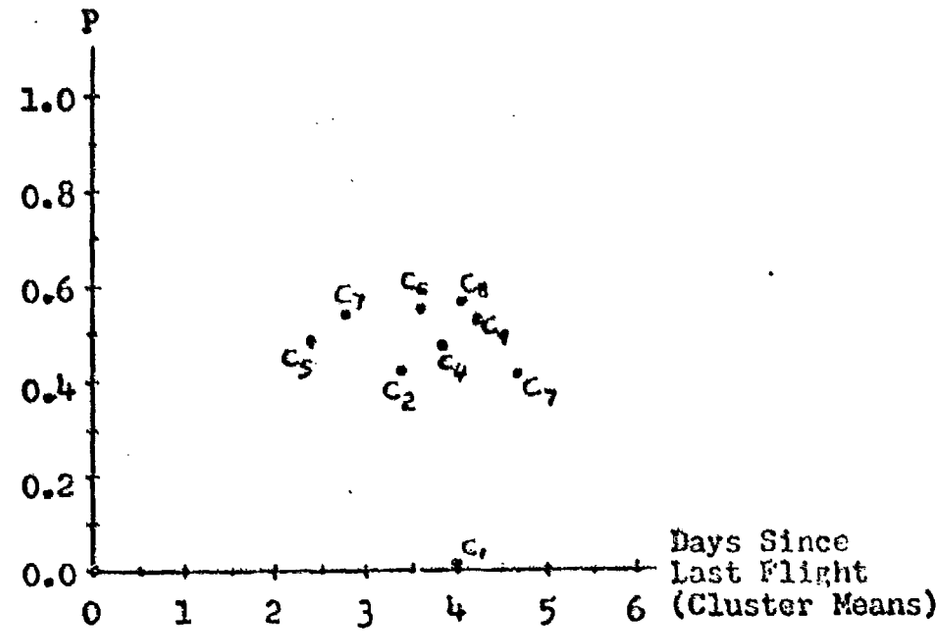


Figure (13)

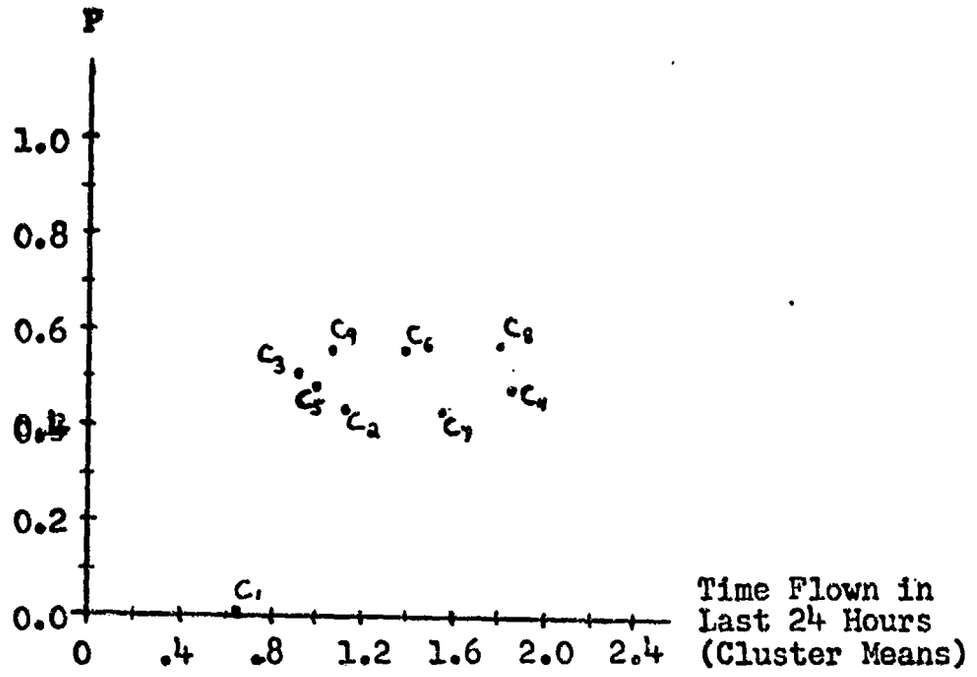


Figure (14)

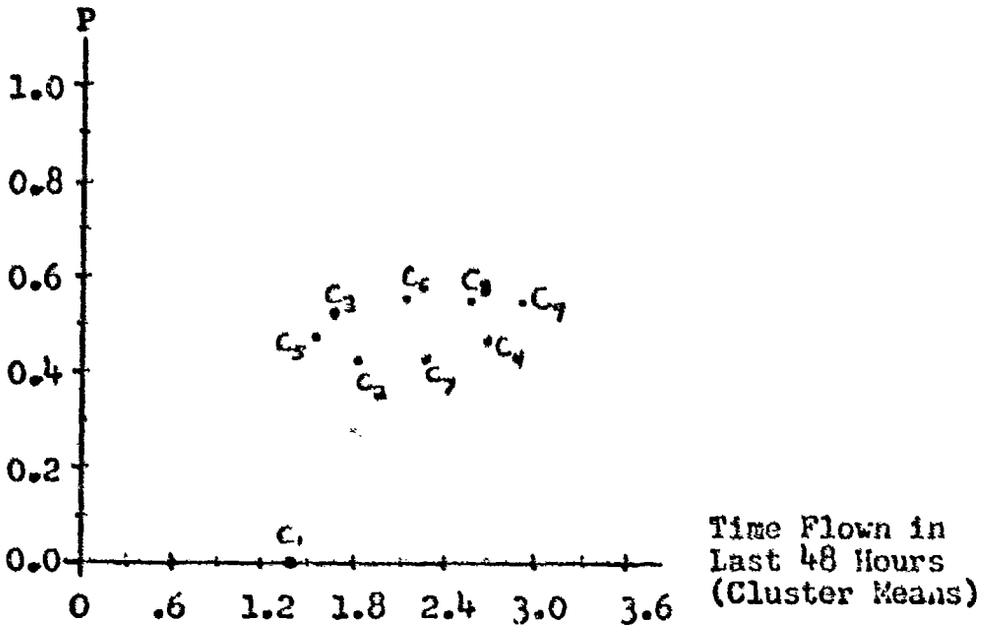


Figure (15)

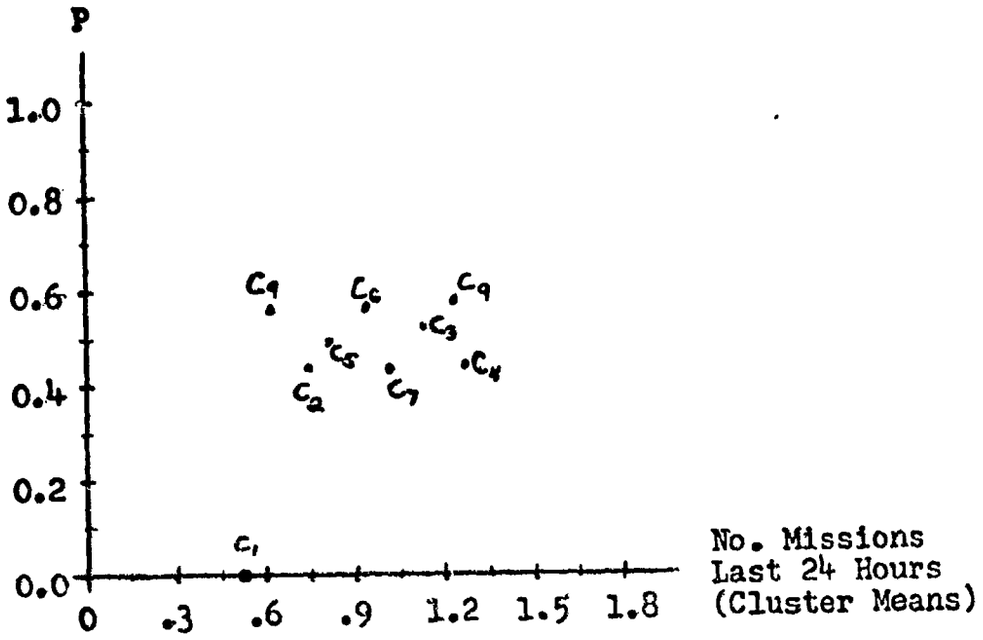


Figure (16)

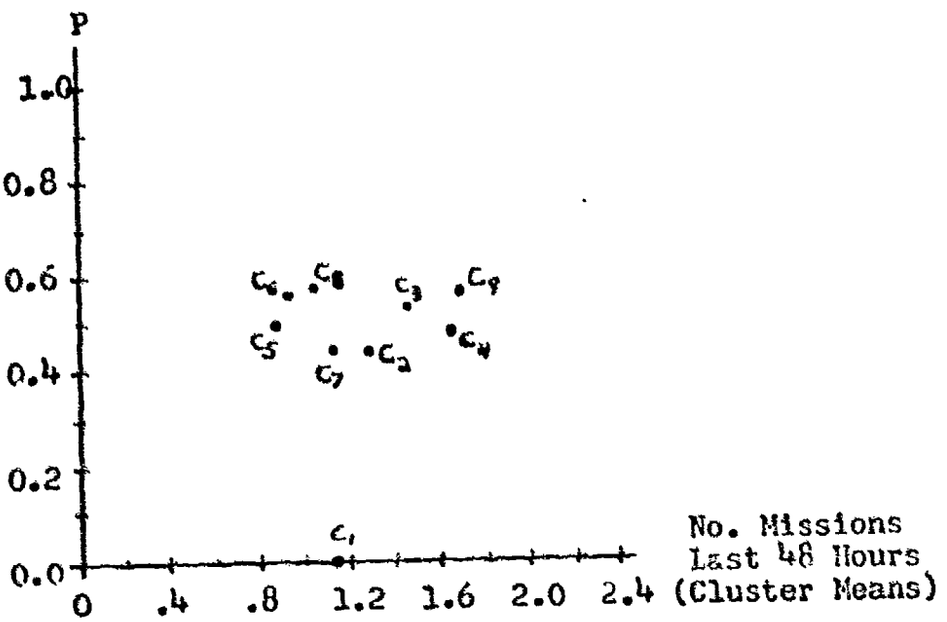


Figure (17)

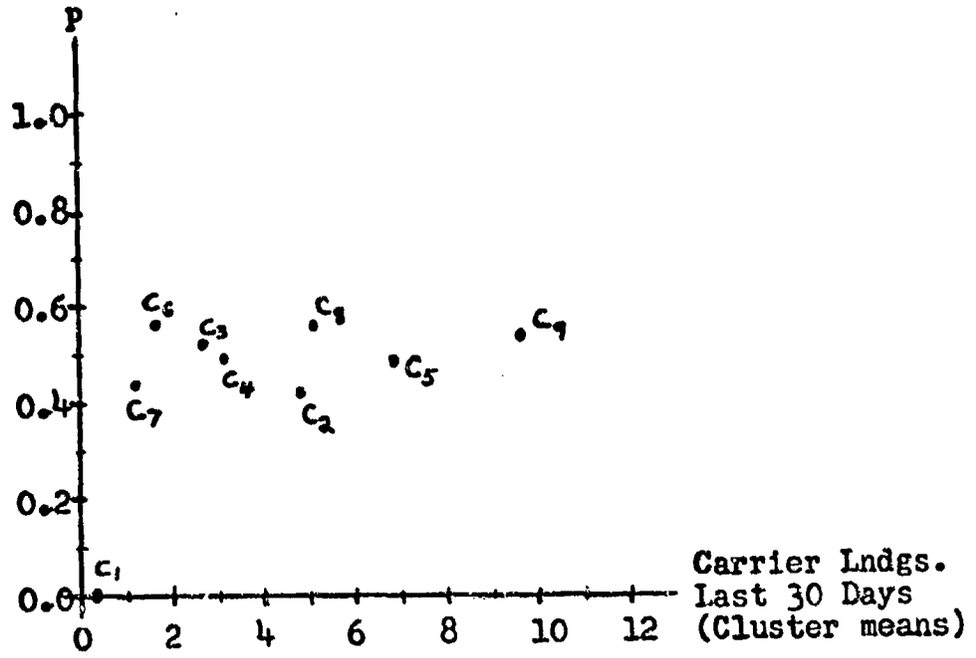


Figure (18)

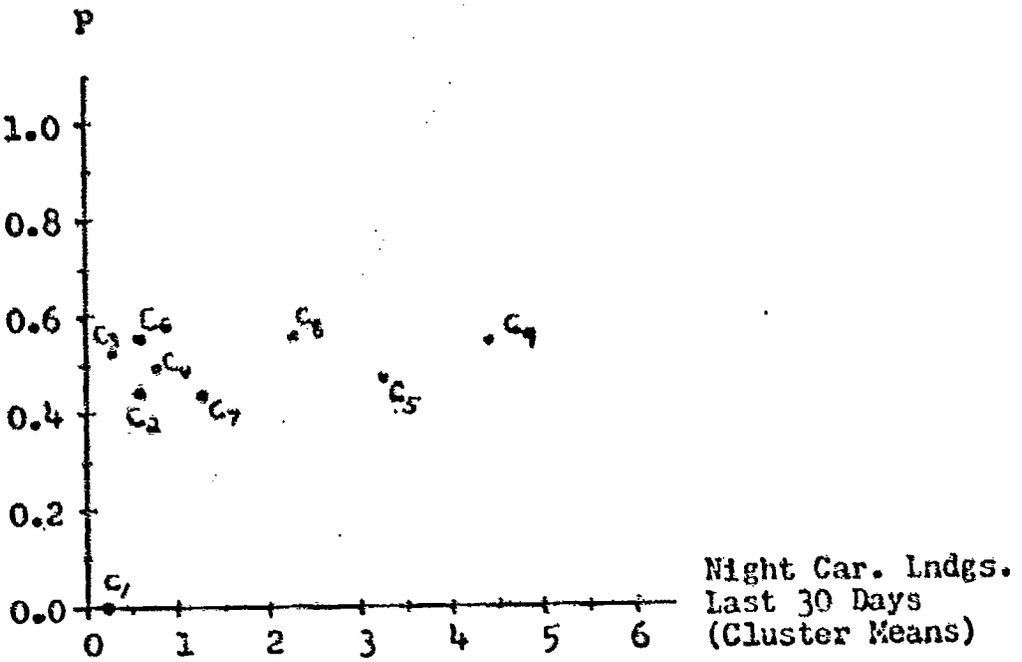


Figure (19)

APPENDIX F

CORRELATION MATRIX

The product-moment correlation between two variables, X_i and X_j, is given by $\rho_{ij} = \dfrac{\mathrm{Cov}(X_i, X_j)}{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}$. An element of the correlation matrix can be calculated by the equation

$$\rho_{ij} = \frac{\sum_{k=1}^{n} (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\left[\sum (X_{ik} - \bar{X}_i)^2 \sum (X_{jk} - \bar{X}_j)^2 \right]^{1/2}} .$$

TABLE VIII -- Correlation Matrix for all Data

Variable	1	2	3	4	5
1	1.0000	0.8494	-0.2080	0.1161	-0.0378
2	0.8484	1.0000	-0.0961	0.0539	-0.0288
3	-0.2080	-0.0961	1.0000	-0.4616	0.3111
4	0.1161	0.0539	-0.4616	1.0000	-0.3818
5	-0.0378	-0.0288	0.3111	-0.3818	1.0000
6	-0.0497	-0.0558	0.3516	-0.4539	0.8682
7	-0.0710	-0.0817	0.2644	-0.5147	0.7092
8	-0.1163	-0.1464	0.3308	-0.5114	0.5596
9	-0.1436	-0.1164	0.5284	-0.2957	0.1851
10	-0.0852	-0.0577	0.4864	-0.1973	0.1932

Variable	6	7	8	9	10
1	-0.0497	-0.0710	-0.1163	-0.1436	-0.0852
2	-0.0558	-0.0817	-0.1464	-0.1164	-0.0577
3	0.3516	0.2644	0.3308	0.5284	0.4864
4	-0.4539	-0.5147	-0.5114	-0.2957	-0.1973
5	0.8682	0.7092	0.5596	0.1851	0.1932
6	1.0000	0.7058	0.7591	0.1520	0.1370
7	0.7058	1.0000	0.8499	0.0836	0.0251
8	0.7591	0.8499	1.0000	0.1726	0.0994
9	0.1520	0.0836	0.1726	1.0000	0.8169
10	0.1370	0.0251	0.0994	0.8169	1.0000

APPENDIX G

TABLE IX – Principal Components for Control Group*

Obs.	Variables 1 & 2	Variables 5,6,7,& 8	Obs.	Variables 1 & 2	Variables 5,6,7,& 8
1	1.6699	-1.3798	26	0.5332	-1.9388
2	-1.3794	6.3810	27	0.1328	-0.4991
3	-0.1867	-0.1488	28	0.8624	1.4137
4	-1.7150	-1.1194	29	0.2342	-0.6311
5	-1.9845	1.9568	30	0.9231	-2.0541
6	-1.3456	-1.8615	31	1.1103	-1.2411
7	0.5147	-0.4092	32	1.5795	0.2381
8	0.2131	-0.6300	33	0.9230	0.3750
9	0.6904	0.2495	34	0.2063	-0.8834
10	1.1748	-1.3724	35	-1.6660	1.0195
11	1.5850	0.3248	36	-0.7722	-1.7726
12	0.8035	-1.8672	37	-1.1944	-1.3782
13	1.1415	0.8991	38	-1.0341	-2.0142
14	0.7133	4.5973	39	0.2469	1.7338
15	1.5827	0.9013	40	-0.9079	-0.1941
16	1.6013	1.5121	41	-0.0550	3.9734
17	1.6467	0.4627	42	0.7382	-2.9538
18	1.4937	0.9433	43	0.9838	-2.5146
19	-2.0501	0.0804	44	1.1878	-1.5082
20	-1.7224	-0.2406	45	1.1599	-0.6775
21	-1.3386	-2.4269	46	0.6936	0.2885
22	-0.8434	-1.2893	47	-1.7383	1.9168
23	-3.6046	0.7114	48	-0.1584	0.5545
24	-0.4636	-0.2801	49	-2.1592	1.5234
25	1.4983	1.3292	50	-2.3989	-0.1913

*Note: The principal components were extracted from standardized data.

TABLE X – Principal Components for Accident Group*

Obs.	Variables 1 & 2	Variables 5,6,7,& 8	Obs.	Variables 1 & 2	Variables 5,6,7,& 8
1	1.3880	-0.9997	26	1.3591	1.6810
2	-0.2461	-0.4406	27	0.5521	0.0758
3	0.1399	-0.9650	28	0.7381	-2.3258
4	0.8553	-0.0562	29	-1.9340	-2.3258
5	1.3791	2.3743	30	-2.7369	-2.3258
6	1.5297	-1.0603	31	0.2065	0.9042
7	-1.3378	1.8637	32	-1.5740	-0.7743
8	0.7425	-1.8509	33	-2.9702	-2.3258
9	-2.9242	0.0391	34	0.2316	3.9012
10	-0.6633	-2.3258	35	1.5540	-0.9173
11	0.1599	2.4339	36	1.0448	-2.3258
12	0.7544	0.7056	37	0.6366	1.5398
13	0.0578	1.9759	38	-1.6751	0.1790
14	1.3880	0.0242	39	-1.8446	3.6157
15	-1.5289	0.8463	40	0.8336	-2.3258
16	1.1024	2.5191	41	0.5914	-0.9173
17	0.7309	-2.3258	42	0.3082	1.5398
18	1.4147	0.3751	43	-3.1236	-0.5836
19	0.7625	-0.4883	44	0.2844	-0.2023
20	1.1120	-1.7367	45	1.1647	-1.3485
21	-2.7147	-1.0970	46	0.2932	2.0427
22	0.6060	3.0390	47	-2.1246	-1.1079
23	-0.8542	1.4120	48	1.3280	-2.3258
24	0.7247	2.0216	49	0.7825	0.4909
25	0.0150	-1.5083	50	1.4791	1.3871

*Note: The principal components were extracted from standardized data.

APPENDIX H

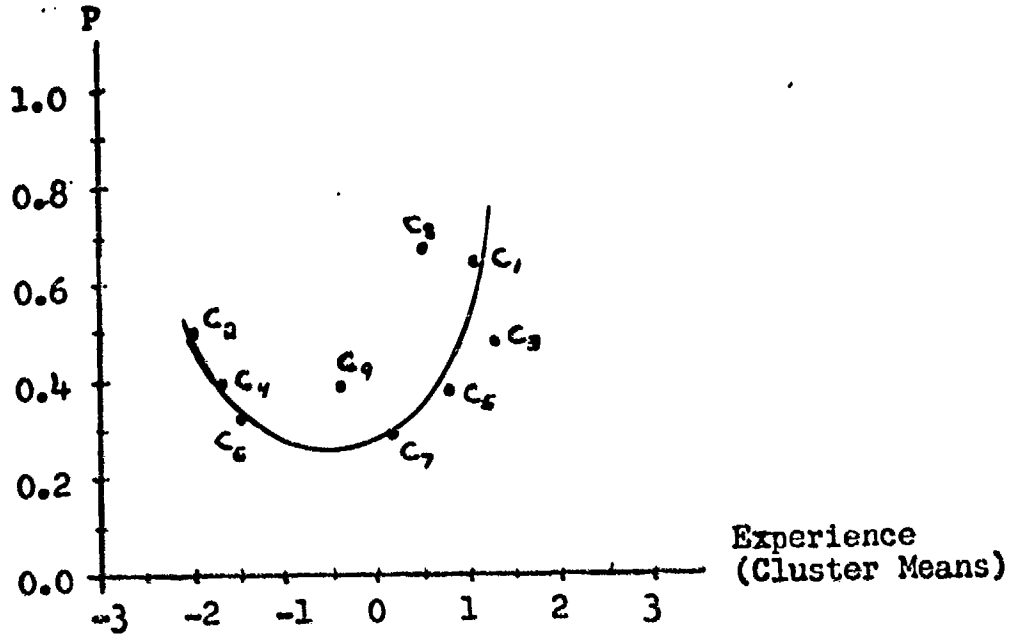


Figure (20)

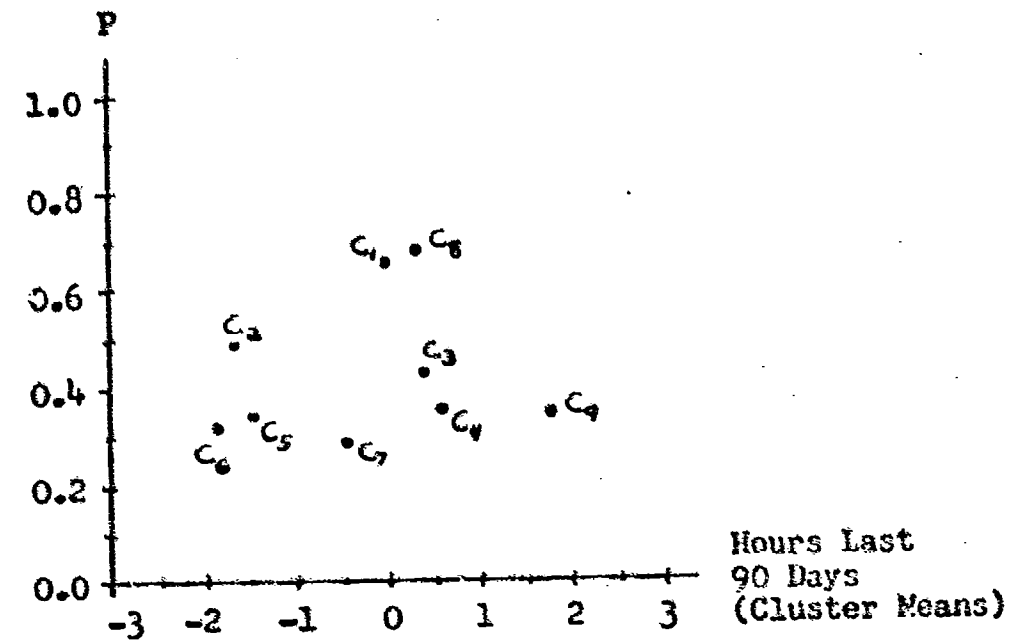


Figure (21)

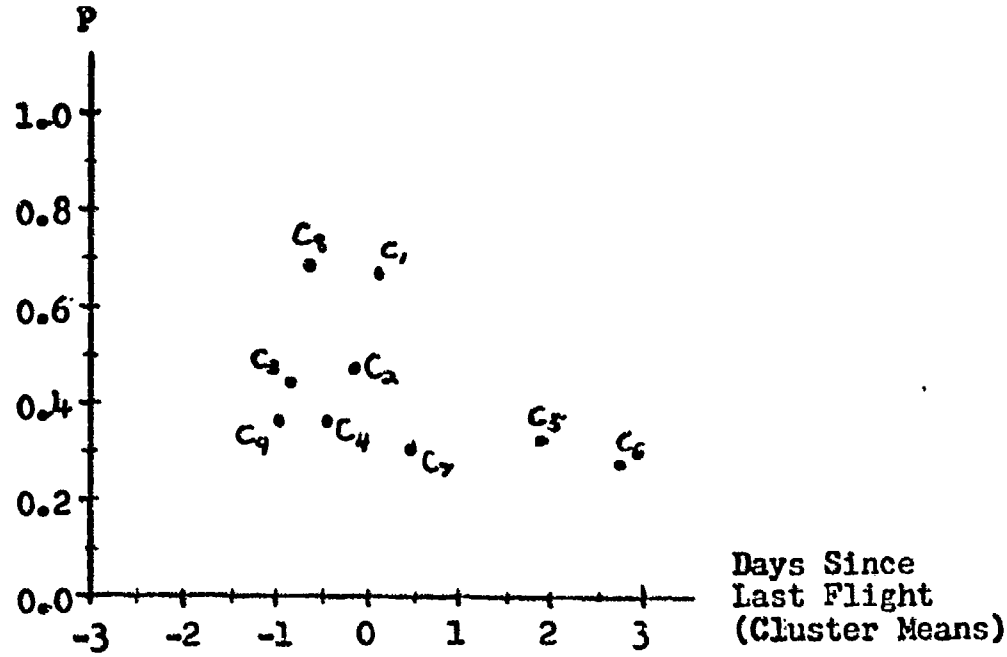


Figure (22)

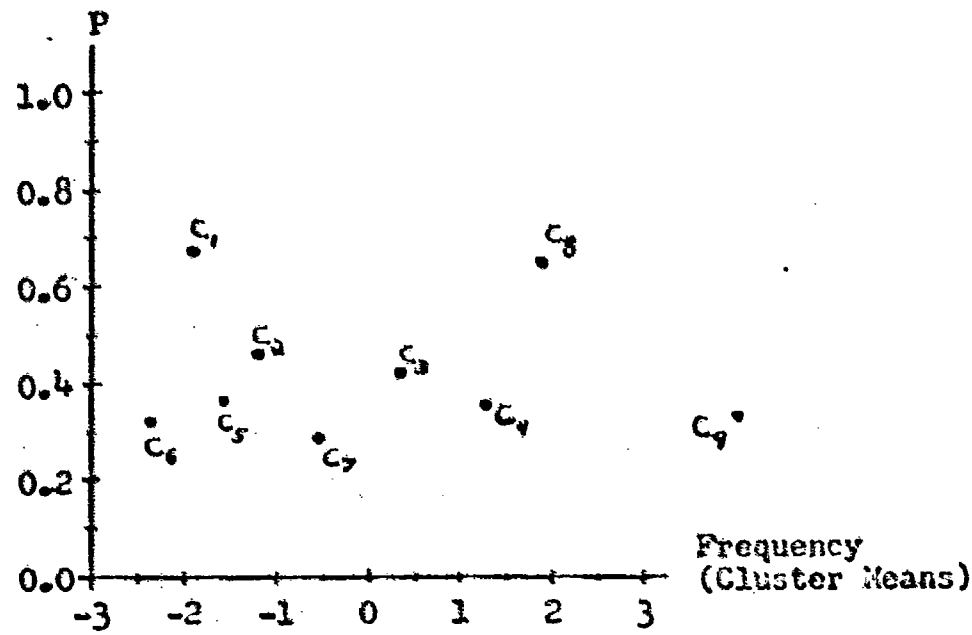


Figure (23)

BIBLIOGRAPHY

1. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1958.

2. Anderberg, M. R., Cluster Analysis for Applications, Academic Press, 1973.

3. Box, G. E. P. "A General Distribution Theory for a Class of Likelihood Criteria," Biometrika, v. 36, p. 317-346, 1949.

4. Eisenbeis, R. A. and Avery, R. B., Discriminant Analysis and Classification Procedures, Lexington Books, 1973.

5. Johnson, S. C., "Hierarchical Clustering Schemes," Psychometrika, v. 32, No. 3, p. 241-254, 1967.

6. Mayer, L. S. "A Method of Cluster Analysis When There Exist Multiple Indicators of a Theoretical Concept," Biometrics, v. 27, No. 1, p. 147-155, 1971.

7. Morrison, D. F., Multivariate Statistical Methods, McGraw-Hill, 1967.

8. Press, S. J., Applied Multivariate Analysis, Holt, Rinehart, and Winston, 1972.

9. Tryon, R. C., and Bailey, D. E., Cluster Analysis, McGraw-Hill, 1970.

10. BMD Manual, Biomedical Computer Programs, Health Sciences Computing Facility, UCLA, University of California Press, 1973.