THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE APPLICATION OF A METHOD OF CLUSTER ANALYSIS TO THE BALLISTOCARDIOGRAM

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

GARY ALLEN TUCK

Oklahoma City, Oklahoma

1966

THE APPLICATION OF A METHOD OF CLUSTER

ANALYSIS TO THE BALLISTOCARDIOGRAM

APPROVED BY

Edward N. Brandt Jr.

William W. Schottstaedt

John C. Brigden

M Clinton Miller Jr

Thomas D. Egan

DISSERTATION COMMITTEE

ACKNOWLEDGMENTS

Sincere thanks are extended to Doctors Edward N. Brandt, Jr., M. Clinton Miller, III, W. W. Schottstaedt, J. R. Assenzo, John C. Brixey, and Thomas N. Lynn, Jr. for making available their time and talents while serving on the author's advisory and/or dissertation committees. Special thanks go to Doctor Brandt for patiently guiding the work which culminated in this paper and to Doctor Lynn for making the ballistocardiograms available.

The author is grateful to the Department of Preventive Medicine and Public Health for making financial support, space, and other facilities available during the entire period of graduate study.

Expressions of gratitude would not be complete without mentioning the late Doctor S. S. Wilks of Princeton University who was able, in a one hour lecture given less than two months before his death, to elucidate for this writer a number of the fundamental ideas of multivariate statistics.

A host of other people, who can not be named individually here, deserve thanks for sought and unsought, heeded and unheeded advice, assistance, criticism, encouragement, and suggestions. Not the least of these have been fellow students. Meriting particular mention are the members of the author's family who were of invaluable assistance in obtaining the data from the ballistocardiograms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES--Continued

LIST OF ILLUSTRATIONS

THE APPLICATION OF A METHOD OF CLUSTER ANALYSIS TO THE BALLISTOCARDIOGRAM

CHAPTER I

INTRODUCTION

Pattern Recognition

One of the most outstanding abilities of many biological organisms, especially man, is that of perceiving some group of sensations as representing a situation that resembles other situations. This ability, whether learned or instinctive, whereby experiences lead to some type of judgement and decision making to guide future action may be called pattern recognition. The information going into the organism through its senses is of a variety of types as well as many small segments of a single type. This information is then reduced through the perception-recognition process to a single "name". It may, then, be considered a many-to-one mapping process.

The reason for the existence of pattern recognition is probably one of economy and for effectiveness of action. The storage within the organism of a facsimile of every possible situation that could arise would, in the first place, probably require facilities far in excess of what would be practical from the standpoint of mobility and, second, each situation being different in some respects would, of course,

be unlike any other and seem unrelated to any previous experience. Thus a perceived but unrecognized situation could evoke no action on the part of the organism.

An understanding of how the most complex of pattern recognition devices, the human brain, does this job would be most helpful in the solution of many problems of many types where some classification scheme is required. According to Uhr (1964) no coherent psychological theories have been developed as to how the perception-recognition process takes place. Such vague terms as "compare", "idea", "trace", and "recreate" are often used in attempts at an explanation of this phenomenon. Obviously a very sophisticated process is involved when one considers only some simple examples. For instance, one can easily recognize a perceived form as a person whether that person is male or female, standing or sitting, seen face-on or in profile, or a multitude of other possible conditions. Not only is the form recognized as a person but almost instantaneously the person is categorized as a stranger or an acquaintance.

In lieu of exact knowledge of how characterization and classification take place within himself or any other organism man has devised some techniques to simulate the analysis, reduction, and classification of masses of information that have been going on biologically for almost as long as there has been life. Without it being immediately obvious many statistical procedures are attempts to simulate in a simple way that which was originally a job of the biological organism. A hypothesis testing procedure whereby one arrives at a statistic that is to be judged as either "significant" or "not significant" at some probability level has the effect of forcing a decision to be made which will influence

some future course of action. For example, a random standard normal deviate whose absolute value is greater than 1.96 may under some circumstances be considered as grounds for deciding that the objects yielding the statistic did not come from some specified hypothesized population but from some other unspecified population. Future action will then depend on the situation that provided the sample in the first place.

A number of direct attacks upon the classification problem have been made using multivariate statistical techniques. One such method is the subject of subsequent chapters of this paper. The choice of the appropriate variates to be used is one of the most crucial problems of pattern recognition according to Uhr (1964). Every statistical attempt at pattern recognition is at least tacitly concerned with this problem since variates must be chosen in order to proceed. Should a pattern recognition scheme fail, the most immediate and obvious question should be that concerned with whether or not the proper variates were selected.

Statistical work in this area has been done by Barnard (1935) on craniometry and furthered by Fisher (1936, 1938). The result of this work was the now well known discriminant function where a linear function of the variates is devised such that there is maximum separation between two known groups of objects. It is to be emphasized that the group to which an object belongs is known in advance. This linear function is then used to reclassify the objects into one of the two categories. Future observations are also classifiable using the same linear function. Closely related work using measures of distance has been done by Pearson (1926), Mahalanobis (1927, 1930), and Hotelling (1931, 1936). Methods of discrimination involving more than two groups

are given by Rao (1952) and Anderson (1958). These involve separation of a multidimensional space into regions such that a point can be assigned to one or more of the groups according to its location in the space. One of the requirements of discriminant analysis is usually that the sample be drawn from a population having the multivariate normal density. At least, knowledge of the form of the population distribution is desirable in order to estimate the probability of errors of classification.

The clustering problem is one where for a given set of objects, each with a given set of attributes or variates measured on it, the requirement is to find subsets, called clusters or clumps, of the original set such that members of a given subset "look alike" while not looking much like objects outside that subset. The ultimate criterion for evaluating the meaning of some of these terms is the value judgement of the user of the process. Usually neither the number of clusters existing within the set of objects nor the cluster to which a particular object belongs is known before the segregation process is begun.

The exact point of transition from discriminant function analysis to cluster analysis is rather obscure in terms of both subject matter and time. However, the work of Hotelling (1933) on principal component analysis is cognate with both. The most obvious difference between the discrimination problem and the clustering problem is that in the former the subgroup to which an object belongs is known before the analysis is begun while in the latter this information is not available. Most of the theory underlying discriminant analysis requires that the sample be drawn from a density whose form is known, usually the

multivariate normal density. No such requirement is made in cluster analysis though it is conceivable that such a method could be developed. Furthermore, the number of clusters within the whole set may not be known while the knowledge of the number of subgroups is a requirement of discriminant analysis. The ability to assign an object appearing at some later time to a subgroup is one of the paramount objectives of discriminant analysis. This is hardly a concern of most techniques of cluster analysis. However, both definition of clusters and assignment of future observations were objectives of the investigation to be described later in this paper.

Much of the past work in the area of cluster analysis has been concerned with taxonomic problems in zoology, botany, paleontology, and microbiology and with information retrieval. Most applications require the measurement of pairwise similarity between objects. This is usually determined from the combined presence or absence of several binary attributes observed on each object or subject. Objects having pairwise similarity greater than some specified threshold value are placed in the same cluster. An object that is in more than one cluster is then finally assigned to that cluster with which it has the greatest number of links in terms of matching attributes. Workers in this area have been Kochen (1955), Sneath (1957), Sokal and Michener (1958), Luhn (1959), Rogers and Tanimoto (1960), Baxendale (1961), Needham (1961), Parker-Rhodes (1961), Sokal (1961), Stiles (1961), Bonner (1962, 1964), Kochen and Wong (1962), Sneath and Sokal (1962), and Sokal and Sneath (1962).

A method of clustering described by Edwards and Cavalli-Sforza (1965) has the advantage that binary data and/or measurements on an

interval scale may be used. The outcome can also be subjected to a test of significance suggested by Barton and David (1962).

The clustering technique devised and applied in this investigation is similar in some respects to that of Rogers and Tanimoto (1960) in that both consider the spatial relationship of points which represent the objects. However, the current method uses no similarity measure *per se* as their method does. The aforementioned method of Edwards and Cavalli-Sforze (1965) was used for refinement in a latter stage of the technique to be set forth in CHAPTER II.

The data to which the current method has been applied were obtained from a group of ballistocardiograms. A history and description of ballistocardiography is the topic of the next section.

The Ballistocardiogram

The ballistocardiograph, a device for recording the kinetic energy of the heart, blood, and larger blood vessels had its beginning when Gordon (1877) demonstrated that motion of the body occurs with each heart beat. His apparatus consisted of a bed suspended from the ceiling by ropes. Lamport (1941) has reported that Landois used a similar, but vertical, device in 1880. Henderson (1905) suspended a plank from the ceiling by wires and allowed it to move only in a head-foot direction. The motion was then magnified by a series of levers and recorded on a smoked drum. It was Henderson who suggested the motion was related to cardiac output. Douglas, *et al.*, (1913) made use of Henderson's "recoil table" in the Pike's Peak Expedition to study the effect of altitude on cardiac output. A study by Heald and Tucker (1922) had as its

purpose the measurement of cardiac efficiency by recording body motion produced by the heart. Doing similar work were Angenheister and Lau (1928) and Abramson (1933). The latter devised a formula with which he had hoped to calculate the cardiac output per minute. This formula, however, was found by Starr, et al., (1939) to be in error. It was at this time that the word "ballistocardiogram" was introduced as the name to be applied to the tracing produced by the apparatus. Since that time numerous contributions to the area of ballistocardiography have been made by Starr and his colleagues (1940, 1941a, 1941b, 1943, 1944a, 1944b, 1945, 1946a, 1946b, 1946c, 1947, 1948, 1949, 1950a, 1950b). In the meantime, contributions have also been made by Krahl (1947, 1950) and by Nickerson and co-workers (1944, 1945, 1947, 1950). It was Nickerson who developed the low frequency critically damped ballistocardiograph that is now in use. He was at first interested in cardiac output studies but later became interested in the clinical implications of ballistocardiography.

It is in the area of clinical interpretation of the ballistocardiogram that the most interest and problems lie today. Lynn (1963) has pointed out that the primary reason for the limited usefulness of the ballistocardiogram is that it is subject to modification by a large number of non-pathological factors. Included among these factors are such things as the quantity and location of body fat, the degree of tension of the skeletal muscles, the quality of the walls and of the suspension of the vessels into which the blood is ejected, and, as well, the characteristics of the recording device. In spite of these limitations Harvey (1964) and Scarborough and Baker (1957) have indicated that

further development of ballistocardiography is warranted in view of several considerations. First, it is a safe and relatively simple procedure for the patient. Second, it is the only means now known that shows any promise of leading to an evaluation of overall circulatory performance. Third, it has been clearly established, according to these authors, that the ballistocardiogram provides the only objective evidence of cardiovascular disease in the majority of patients with a history of angina pectoris or myocardial infarction. Other substantiating evidence concerning the former condition is provided by Brown, et al., (1950). Fourth, the ballistocardiogram has provided objective evidence that restriction of dietary lipid for periods of one year or more in patients with coronary artery disease leads to an improved ballistocardiographic wave form.

In view of the aforementioned considerations and the present limited clinical use of the ballistocardiogram it is clear that some method or methods need to be devised to increase its usefulness. Noordegraaf, et al., (1961, 1963) and Morse (1963, 1964) have been successful in reconstructing ballistocardiographic wave forms by substituting the values of various physiological and anatomical parameters into a mathematical model. The assumption was made that if a wave form matching that of a given individual is produced by the model then the individual's parameters are the same as those used in producing the reconstructed wave. It does not seem unreasonable that several combinations of parameters might lead to the same or similar wave forms. In addition, the implementation of their procedure first requires knowledge of such information as pulse wave velocity, average blood pressure, and length,

diameter, wall thickness, elasticity, and distribution of major portions of the left and right arterial system. If all this information were known the importance of the ballistocardiogram itself would surely be diminished. Performance of the analysis is also greatly facilitated when certain analog computing equipment is available. However precise and accurate the method may be it certainly lacks the flexibility necessary for widespread use due to the parameters and equipment required.

One of the objectives of this investigation has been to begin a search for a method of ballistocardiographic analysis that has the qualities of ease, speed, and objectivity.

The subject of CHAPTER II will be the methods of data collection and acquisition and the statistical methods investigated and actually used. The results of application of the statistical methods to the data obtained from the ballistocardiograms will be the subject of CHAPTER III. Finally, CHAPTER IV will deal with the conclusions that can be drawn from this investigation and some possible future courses of action that might be taken concerning statistical analysis and classification of ballistocardiograms.

CHAPTER II

COLLECTION AND METHODS OF ANALYSIS OF DATA

Production of the Tracings

The instrument used in the production of the tracings used was built by Astrospace Laboratories and equipped with an air bearing table. The subjects lay quietly in a supine position on this table and were asked to suspend respiration in such a way that no force was being used to prevent or encourage either inspiration or expiration of air during the few seconds that the recording was being made because the effect of any motion and forces or changes in respiration have a marked effect on the form of the tracing. The mechanical motion of the table in a head-foot direction served as the input to an accelerometer whose output was the electrical analog of the acceleration of the table. Although records of velocity and displacement may be made, this is seldom done. Therefore, the term ballistocardiogram shall hereafter in this paper mean the acceleration ballistocardiogram. The electrical signal from the accelerometer was amplified and transformed into the mechanical motion of the writing arm of a Grass polygraph. The paper speed of the polygraph was 50 millimeters per second. In the direction of deflection of the writing arm the paper was calibrated at 1 millimeter intervals. The paper was calibrated at 5 millimeter intervals in the direction of

motion of the paper. The latter calibrations were printed on the paper as arcs of a circle with a radius equal to the length of the writing arm.

Simultaneously with the recording of the ballistocardiogram an electrocardiogram was recorded as an aid, if necessary, in recognizing the significant portions of the ballistocardiographic tracing.

Source and Selection of Tracings Used

The tracings used in this investigation were selected from those of subjects who had ballistocardiograms on file at the University of Oklahoma Medical Center. The majority of these subjects are involved in a long term project of the Neurocardiology Research Center whose aim is to determine the relationship of physical, physiological, and psychological factors to certain aspects of cardiovascular disease.

This investigation was intended to be of an exploratory nature and was not intended for estimation or hypothesis testing. Therefore, the selection of subjects was made so as to provide a group of tracings that reasonably represented all those that were available.

Usually the latest available tracing for a subject was used. One cardiac cycle was selected for analysis by an experienced ballistocardiographer as being typical of those recorded on the selected date. Although occasionally more than one cycle per subject was selected, either for the same or a different date, these duplications were not included in the analysis. Their use will be mentioned later.

Portion of the Cycle Selected

An illustration of the general form of the ballistocardiogram

for a single cardiac cycle is shown in Figure 1 with the H, I, J, K, L, M, N, and O waves indicated. According to Brown, et al., (1952) the record of each heart beat is made up of two groups of waves. The H, I, J, and K form the significant portion while the remainder is made up of the so-called "after waves". The exact physiological mechanisms underlying each of the major components are not yet understood. A brief explanation from Brown, et al., (1952) will be given concerning some of the primary sources of the major components.

The H-wave appears to be a result of the apex thrust of the heart that occurs during isometric ventricular contraction and/or right auricular systole. The action of the left auricle can be ignored here since the pulmonary veins enter laterally and have tributaries running in all directions.

The I-wave is thought to be a result of the footward component of the cardiac recoil accompanying the ventricular contraction.

Deceleration of the blood and/or impulse wave by the aortic and pulmonic arches and the head possibly produce the J-wave.

The J-K stroke probably represents the deceleration of the footward impulse wave by the resistance of the arteriolar bed in the legs.

The examination of a few tracings should be convincing evidence that the point of onset of the pattern is difficult to determine but the H-I stroke is usually quite obvious. Likewise, the last significant point following the K valley is usually impossible to determine. For these reasons the portion of the cycle selected for analysis was that from the H peak to the K valley. Since no other precise way of

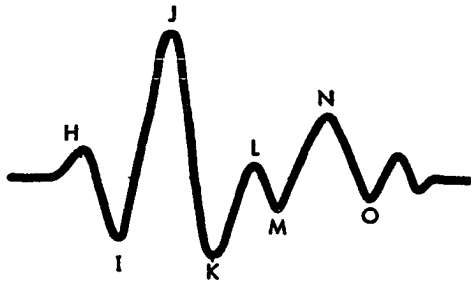

Figure 1. The general form of the ballistocardiogram

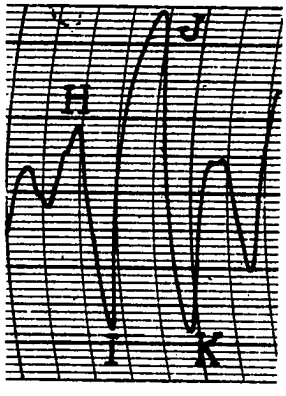

Figure 2. The ballisto-cardiogram for subject number 49

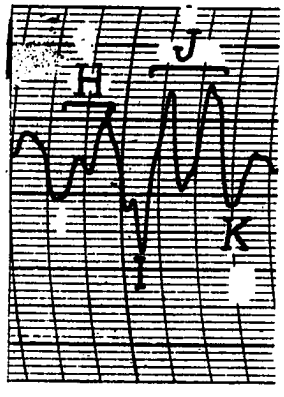

Figure 3. The ballisto-cardiogram for subject number 100

determining a base line was available the level of the H peak was selected. If the H-wave was bi-peaked the earliest peak was selected as the beginning of the portion to be read. There are some obvious disadvantages to this selection of a base line. An H-wave of high amplitude would cause the I and K valleys to appear unusually depressed and the J-wave to appear unusually narrow and of low amplitude. The lack of a better base line was not considered a serious obstacle here because of the preliminary nature of this investigation.

Obtaining Data from the Tracing

It is possible to use the electrical signal of the accelerometer of the ballistocardiograph as the input to a magnetic tape recorder. The magnetic tape can then serve as the input to electronic computing equipment which can provide extremely detailed information about the ballistocardiogram. This means as a source of data was not used for three reasons. First, this investigation has been of a preliminary nature and such detailed information was not desired at this time. Second, the paper tracings used were immediately and readily available. Third, there appeared to be certain technical obstacles such as locating the significant portion of the curve whether it be in analog form on the tape or in digital form within the computer.

Each of the selected tracings was transferred photographically to a full size transparency which could then be projected onto a screen. The enlargement thus provided made possible the reading method to be described. An overlay grid transparency was prepared from paper similar to the tracing paper but that it was ruled every millimeter in both

horizontal and vertical directions. Since the tracing curve itself was about 1 millimeter in width a rule had to be adopted so as to make the reading as precise as possible within the overlay grid. This rule consisted of designating the lower edge of the curve as the line to be followed in the reading. The intersections of the lines of the grid were designated as the points $P_{i,j,k}$. The origin was chosen to be that intersection nearest the first H peak. This was then designated as point $P_{i,0,0}$. The ranges and the meanings of the indices are as follows: $i = 1, 2, \ldots, n$ represents the subject identification number, $j = -30, -25, \ldots, 20$ represents the level number or the vertical distance in millimeters from the H peak, and $k = 0, 1, \ldots, N$ represents the horizontal distance in millimeters to the right of the H peak with the points $P_{i,j,N}$ lying on that vertical grid line passing nearest to the lowest point of the K valley on the tracing line. The reason for the selection of these particular limits on the level was that of the 186 ballistocardiograms read only 3 had J- and/or K-waves that extended outside this range.

For each of the N points on each of the 11 levels there was defined a corresponding variable $X_{i,j,k}$ which could take on the value 0 or 1 depending on whether the point was above or below the tracing line respectively. The one exception to this rule was that $X_{i,0,0} = 0$ for all i.

These variables were then arranged in an 11 x N binary profile matrix which for the i-th subject was called X_i. A typical and a not-so-typical tracing are shown on page 13 in Figures 2 and 3 respectively with their respective binary profile matrices being in Tables 1 and 2.

TABLE 1

BINARY PROFILE MATRIX FOR SUBJECT 49

Level	
20	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
10	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0
5	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0
0	0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0
- 5	1 1 0 0 0 0 0 1 1 1 1 1 1 0 0 0
-10	1 1 1 0 0 0 0 1 1 1 1 1 1 1 0 0
-15	1 1 1 0 0 0 1 1 1 1 1 1 1 1 0 0
-20	1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 0
-25	1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0
-30	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

TABLE 2

BINARY PROFILE MATRIX FOR SUBJECT 100

Level	
20	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0
0	0 0 0 1 1 1 0 0 0 0 0 1 1 1 0 0 0 1 1 1 0 0
- 5	1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0
-10	1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
-15	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-20	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-25	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-30	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

The j-th row vector of X_i was designated $X_{i,j}$.

The Statistical Methods

Transformation of Data

In order to make the matrices of the various subjects comparable it was decided to perform a transformation on each X_i such that the arrangement and proportion of 0's and 1's in each row vector were maintained and the number of elements in each row vector was invariant with respect to i. Since the modal value of the number of columns in the X_i was 15 in a preliminary group it was decided that each standardized binary profile matrix Y_i should be of dimension 11 x 15 for all i.

The component parts of Y_i were designated in a manner similar to those of X_i. The row vectors were denoted by $Y_{i,j}$ and the individual elements by $Y_{i,j,k}$. Following the transformation the maximum value for k was 14.

The actual transformation was carried out on each $X_{i,j}$ in a manner as follows: Suppose $X_{i,j}$ = (1 1 0 0 0 1 1 1 1 1 0 0 0). Then $Y_{i,j,k} = 1$ for k = 1, 2, ..., I[(2/13) · 15] where I [a] indicates the nearest integer to a. Also, $Y_{i,j,k} = 0$ for k = I[[(2/13) · 15)] + 1, I[[(2/13) · 15)] + 2, ..., I[[(2/13) · 15)] + I[[(3/13) · 15)] - 1, I[[(2/13) · 15)] + I[[(3/13) · 15)]. This process was repeated until all 15 places in $Y_{i,j}$ were filled.

The effect of this transformation was to "compress" or "stretch" each curve into a uniform time period. Since most of the matrices had a value of N near 15, it was felt that few, if any, of the curves would be significantly distorted by this transformation process unless the

wave form is greatly altered by slight changes in heart rate. Such marked changes in wave form were not thought to exist in this group of subjects.

The standardized binary profile matrices for the tracings of Figures 2 and 3 are shown respectively in Tables 4 and 5.

Possibilities Considered

The original intent in obtaining the binary data from the tracings was to investigate the possibility of using a clustering method described by Bonner (1964) in which a measure of similarity between each pair of objects is determined. Any threshold value within the range of the similarity measure can then be chosen and all pairs of subjects having a similarity measure equal to or greater than the threshold value are judged to be similar. Those pairs with a similarity value less than the threshold value are considered not similar. Then all subjects that are pairwise similar form a "tight" cluster. These clusters may contain non-disjoint subsets of subjects. It was hoped that one or a few of the $Y_{i,j}$ could be used to assign the subjects to clusters. This approach to the problem was abandoned for several reasons. First, no good criterion for selection of threshold values was available. It was found that lower threshold values led to the formation of just one or very few highly non-disjoint clusters. Higher threshold values produced many very small clusters. The change from "lower" to "higher" values was found to be very abrupt, there being no value that would produce a moderate number of moderate size clusters. Second, no intuitively appealing method of creating disjoint subsets, called "core"

TABLE 4

STANDARDIZED BINARY PROFILE MATRIX FOR SUBJECT 49

Level	
20	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
10	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0
5	0 0 0 0 0 0 0 0 1 1 1 1 0 0 0
0	0 0 0 0 0 0 0 1 1 1 1 1 1 0 0
- 5	1 1 0 0 0 0 0 1 1 1 1 1 1 0 0
-10	1 1 1 0 0 0 0 1 1 1 1 1 1 1 0
-15	1 1 1 0 0 0 1 1 1 1 1 1 1 1 0
-20	1 1 1 1 0 0 1 1 1 1 1 1 1 1 0
-25	1 1 1 1 0 1 1 1 1 1 1 1 1 1 0
-30	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

TABLE 5

STANDARDIZED BINARY PROFILE MATRIX FOR SUBJECT 100

Level	
20	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
15	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
10	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5	0 0 0 0 0 0 0 0 1 0 0 0 1 0 0
0	0 0 1 1 0 0 0 1 1 0 0 1 1 0 0
- 5	1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
-10	1 1 1 1 1 1 0 1 1 1 1 1 1 1 1
-15	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-20	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-25	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
-30	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

clusters, of subjects from the tight clusters was immediately available. Core clusters were felt to be necessary because a large number of subjects appeared in more than one tight cluster. This was not considered a desirable situation if any sort of definite assignment scheme was to be developed later for future observations. Third, even if the core clusters had been formed, no scheme for classifying a future subject using the binary data was available. And, finally, the amount of time required for finding all the tight clusters with an electronic computer for even a moderate number of subjects was inordinate.

The method of Rogers and Tanimoto (1960) considers the spatial arrangement of points in determining clusters. These authors use binary data to determine a measure of similarity $S_{\alpha\beta}$ between objects (or subjects) α and β. The measure used is a ratio of the total number of attributes common to both α and β to the total number of attributes possessed by either α or β. The range on this similarity value is $0 \leq S_{\alpha\beta} \leq 1$. They further define a semimetric space wherein the distance between objects α and β is $d_{\alpha\beta} = -\log_2 S_{\alpha\beta}$. For each value of α the quantity $H_\alpha = -\sum_\beta \log_2 S_{\alpha\beta}$ is computed where the sum indicated is over all values of β. No explanation is given for using the base 2 logarithms or for what to do if $S_{\alpha\beta} = 0$. That value of α which causes H_α to be a minimum defines that point which is nearest to the centroid of the mass of points. Around this point, or prime node, points are added one by one in order of their distance from the prime node to form a cluster of objects that are similar to each other. When the variability within the group increases markedly with the addition of a point it is assumed that this point belongs to a new cluster. This

method is not appealing because if there actually exist two or more close or slightly overlapping and yet fairly distinct clusters then the prime node would tend to be the center of a group that contains objects which should not be placed in the same cluster. In such a case it is likely that only one cluster would be found when in fact there are two or more.

Not finding a clustering method using binary data that had intuitive appeal it was decided to perform another transformation on the data. The new set of variates

$$Z(i,k) = \sum_{j=-30}^{20} Y_{i,j,k} \qquad (2.1)$$

was made for all i and k. The reason for this choice was that it was felt that a considerable reduction in the amount of data could be made while retaining all of the contained information since Z(i,k) is simply a count of the number of 1's in the k-th column of Y_i. This had the effect of selecting a base line 30 millimeters below the H peak and measuring the number of complete 5 millimeter increments the curve lay above the base line at 15 equally spaced points along this base line. In other words, the amplitude of the graphic analog of the acceleration of the table of the ballistocardiograph was, in effect, measured at 15 equally spaced points from the H peak to the K valley.

This transformation created the row vector Z'(i) = [Z(i,0), Z(i,1), ..., Z(i,14)]. Since Z(i,0) = 6 for all i this was eliminated from Z'(i) forming a 14 element row vector Z(i) called the standardized profile vector. Tables 5 and 6 illustrate the Z(i) for those subjects whose Y_i are shown in Tables 3 and 4 respectively.

TABLE 5

STANDARDIZED PROFILE VECTOR FOR SUBJECT 49

6	5	3	1	2	4	7	9	10	9	9	7	5	1

TABLE 6

STANDARDIZED PROFILE VECTOR FOR SUBJECT 100

6	7	7	6	5	4	7	8	6	6	7	8	6	6

The Clustering Method

Each of the vectors Z(i) can be thought of as representing a point in Euclidean 14-space. Among the n points the i'-th and the i"-th were found which lay closest together. In other words, i' and i" were chosen so as to minimize

$$\delta(i', i'') = \{ \sum_{k=1}^{14} [Z(i',k) - Z(i'', k)]^2 \}^{\frac{1}{2}} \qquad (2.2)$$

Then a centroid point or vector for the i'-th and i"-th points was computed by

$$\bar{Z}(2) = \frac{1}{2} \sum_{i=i'}^{i''} Z(i) = [\bar{Z}(2,1), \bar{Z}(2,2), \ldots, \bar{Z}(2,14)] \qquad (2.3)$$

The index 2 indicates mean values for 2 points. The i'''-th point was chosen from the remaining n-2 points so as to cause the distance

$$D(2,i''') = \{ \sum_{k=1}^{14} [Z(i''', k) - \bar{Z}(2,k)]^2 \}^{\frac{1}{2}} \qquad (2.4)$$

to be a minimum. Following this, the distance

$$d(2,3) = \{ \sum_{k=1}^{14} [\bar{Z}(2,k) - \bar{Z}(3,k)]^2 \}^{\frac{1}{2}} \qquad (2.5)$$

of centroid shift with the addition of the third point was computed where

$$\bar{Z}(3) = (1/3) \sum_{i=i'}^{i'''} Z(i) \qquad (2.6)$$

In general, for m = 2, 3, ..., n-1 the mean vector or centroid

$$\bar{Z}(m) = (1/m) \sum_{i=i'}^{i^{(m)}} Z(i) \qquad (2.7)$$

for the mass of the m most compact points was computed. The distance

$$D(m, i^{(m+1)}) = \{ \sum_{k=1}^{14} [Z(i^{(m+1)}, k) - \bar{Z}(m,k)]^2 \}^{\frac{1}{2}} \qquad (2.8)$$

from the centroid found in (2.7) to that (m+1)-st point nearest to $\bar{Z}(m)$ was then computed. When this (m+1)-st point was added to the mass of points, the centroid $\bar{Z}(m+1)$ was computed in a manner similar to (2.7). Then the distance the centroid moved by the addition of the nearest point from the n-m remaining points was computed by

$$d(m, m+1) = \{ \sum_{k=1}^{14} [\bar{Z}(m,k) - \bar{Z}(m+1, k)]^2 \}^{\frac{1}{2}} \qquad (2.9)$$

This process was continued until the supply of points was exhausted. A list of the distances, $D(m, i^{(m+1)})$ and $d(m, m+1)$, was made as the accumulation proceeded.

The reason for preparing such a list was that it was felt such

a list would indicate the relative arrangement of the points in the 14-space. Relatively large increase in either of the distances as m progresses would indicate a wider separation of points than had existed among the group just preceding. Examination of the list of $D(m, i^{(m+1)})$ was not revealing in this regard since it behaved rather erratically as m progressed. However, in the list of centroid shift distances it was seen that in most cases $d(m, m+1) > d(m+1, m+2)$ but in a few instances the inequality was reversed. A reversal was interpreted as indicating that the (m+2)-nd point just added was farther in distance from the centroid of the mass of points than those points added just previously. Being so separated in space, this new point was considered to be an element of another cluster. Several of the later points added caused the distance of movement of the centroid to behave in a capricious manner. This was interpreted as indicating that the tracings these points represented were in some respect unlike each other or any other in the entire group. It is possible that these points were each the only representatives in this sample of another cluster. However, when the original sample was combined with a larger group of subjects, the same situation arose near the end of the accumulation process, with many of the peculiarly behaving points in the larger group being the same ones that did so with the smaller sample. These individual points, not contributing to the formation of a cluster, were dropped from the analysis after it was found that they could not be combined into groups of any size that had reasonable within-group variability.

The variability within clusters was determined by

$$\sum_{k=1}^{14} \left(\sum_{i=1}^{n_h} [Z(i,k) - \bar{z}(k,h)]^2 \right) \tag{2.10}$$

where n_h is the number of objects in the h-th cluster and

$$\bar{z}(k,h) = (1/n_h) \sum_{i=1}^{n_h} Z(i,k) \tag{2.11}$$

The values assumed by h were 1, 2, ..., T.

Those clusters that were adjacent in the list mentioned above were inferred to be adjacent in the 14-space because nearness of their respective points in space caused the list to appear as it did. Some of the "small" adjacent clusters were combined two at a time and redivided into two clusters by a method due to Edwards and Cavalli-Sforza (1965). This method requires the examination of every possible arrangement of the points into two clusters in order to find that arrangement which maximizes the between clusters sum of squares over all variates. This is actually accomplished by finding that arrangement which produces a minimum within clusters sum of squares. This sum of squares is computed by

$$\sum_{h=1}^{2} \left\{ \sum_{k=1}^{14} \left(\sum_{i=1}^{n_h} [Z(i,k) - \bar{z}(k,h)]^2 \right) \right\} \tag{2.12}$$

This was performed on successive adjacent clusters by letting h = 2, 3 then h = 3, 4 until finally h = T-1, T. Then the process was repeated until no more points shifted from one cluster to another.

This method was used to form adjusted clusters that had as much variability as possible between the clusters while remaining within the framework of the accumulation process. It also led to the formation of clusters whose within clusters variability was more nearly

homogeneous.

The use of the Edwards and Cavalli-Sforza (1965) method was considered for splitting the initial group of n items into two clusters but each of the possible 2^{n-1} -1 possible splits must be examined to determine which has the desired property. This is obviously a quite large number even for moderate size n and the amount of electronic computer time required to carry out this procedure would have been prohibitive. In addition, it is not clear just what would be the result of forcing the points into two clusters when there may in fact be three or more. Of course, one could continue by further subdividing each original cluster into two. One problem that quickly arises here is where to stop the process. Another problem is that two clusters may be formed from a group of points that should remain one.

A one-way analysis of variance suggested by Barton and David (1962) and used by Edwards and Cavalli-Sforza (1965) was applied to the clustered data. The within clusters sum of squares was computed as in (2.12) except that h was allowed to take on the values 1, 2, ..., T. The total sum of squares was computed by

$$\sum_{k=1}^{14} \left\{ \sum_{i=1}^{n} [Z(i,k) - \bar{z}(k)]^2 \right\} \qquad (2.13)$$

where

$$\bar{z}(k) = (1/n) \sum_{i=1}^{n} Z(i,k) = (1/n) \sum_{h=1}^{T} n_h \bar{z}(k,h) \qquad (2.14)$$

The among clusters sum of squares was then obtained by subtracting (2.12) summed over all clusters from (2.13).

The Classification Method

For each of the T clusters there was computed a centroid that was designated by $\bar{\bar{z}}(h) = [\bar{z}(1,h), \bar{z}(2,h), \ldots, \bar{z}(14,h)]$ and a measure of variability computed by

$$MS(h) = (1/n_h) \sum_{k=1}^{14} \{ \sum_{i=1}^{n_h} [Z(i,k) - \bar{z}(k,h)]^2 \} \qquad (2.15)$$

This quantity in (2.15) is the mean squared distance of the points within the cluster from the centroid rather than an estimate of some quantity. This is the reason for the use of the divisor n_h rather than the usual $n_h - 1$. Also, use of the divisor $n_h - 1$ would cause the value of MS(h) to be somewhat inflated for a small cluster. The positive square root of MS(h) was designated by RMS(h). Since the 14 variates were all in terms of millimeters the quantity RMS(h) was also in these units and could therefore be used as a measure of distance in the 14-space.

For each value of i and h three distances were computed. The first was the distance of the i-th point from the centroid of the h-th cluster. This was computed by

$$\Delta c(i,h) = \{ \sum_{k=1}^{14} [Z(i,k) - \bar{z}(k,h)]^2 \}^{\frac{1}{2}} \qquad (2.16)$$

The second was the distance of the i-th point from a hypersphere of radius RMS(h) whose center was $\bar{\bar{z}}(h)$. This was determined by

$$\Delta sp(i,h) = \Delta c(i,h) - RMS(h) \qquad (2.17)$$

Negative values indicate that the point is inside the hypersphere. Finally, a standardized distance of the i-th point from $\bar{\bar{z}}(h)$ was computed by

$$\Delta st(i,h) = [\Delta c(i,h)]/RMS(h) \qquad (2.18)$$

Each of these three quantities was used as a classification criterion by letting the value of h which caused each to be minimized for a given i name the cluster to which the assignment of the i-th subject was made.

Seeking Possible Subclusters

When $\Delta c(i,h)$, $\Delta sp(i,h)$, and $\Delta st(i,h)$ were computed for a given value of i, the values of h that caused each of the quantities to be minimized were not always consistent. Those subjects for which the assignment was the same for all three methods were considered as "central" clusters. A number of subjects were not classified the same by all three methods. Among this group, those subjects which agreed with each other by all three classification methods were considered as possible "subclusters". Although this method of finding subclusters has considerable intuitive appeal it becomes necessary to examine the groups carefully because, for example, two points might be in opposite directions from a centroid and still have nearly the same set of three distances.

The procedure described in this chapter is, of course, applicable to any data that provides a spatial arrangement of the points representing the objects. Though it could be applied to points in a semimetric space such as defined by Rogers and Tanimoto (1960), application to points in metric space is probably more desirable. The reason for this is that in a semimetric space, as defined by these authors,

it is possible, for example, for point A to be near points B and C while B and C are not near each other. It is also possible, if A, B, and C represent the vertices of a triangle in the semimetric space, that the sum of the lengths of two sides of the triangle may not be greater than the length of the third side.

CHAPTER III will show the application of the statistical procedures described above to the ballistocardiogram data.

CHAPTER III

RESULTS OF ANALYSIS

The initial phase of analysis of the ballistocardiograms was done using the data obtained from the tracings of 37 subjects. Considering each subject as a point in 14-space, those two points separated by the least distance were sought. The minimum distance between points was found to be shared by three pairs of points. Each member of each pair was separated from the other member of that pair by 2.2361 units. Each unit in this case is 5 millimeters because each Z(i,k) value represented the number of complete 5 millimeter segments the curve stood above a horizontal line lying 30 millimeters below the H peak on the tracing. The one-by-one accumulation process described in CHAPTER II was carried out beginning with each of the pairs mentioned above. These three processes were called Path I, Path II, and Path III. Table 7 shows the point sequence numbers in the accumulation process, m; the subject identification numbers, $i^{(m)}$; the distance of the m-th point from the centroid of m-1 points, $D(m\text{-}1, i^{(m)})$; the distance of centroid shift, d(m-1, m); and, as well, the cluster assignment, h, for Path I. The two distances for $i^{(37)} = 14$ are omitted because this was the only remaining point at that stage. It may be necessary, as in this case, to be somewhat arbitrary about the magnitude of difference of distance between centroids that actually constitutes a between-clusters breaking

TABLE 7

ACCUMULATION PROCESS FOR PATH I

Initial subjects: i' = 4, i" = 27

h	m	$i^{(m)}$	$D(m-1, i^{(m)})$	$d(m-1, m)$
	3	17	2.5000	.8333
1	4	37	2.6666	.6666
	5	29	2.3452	.4690
	6	28	2.5534	.4255
	7	30	3.1402	.4486
	8	39	3.1331	.3916
	9	18	3.0026	.3336
	10	20	3.2678	.3267
	11	34	3.3541	.3049
2	12	21	3.4148	.2845
	13	38	3.4590	.2660
	14	2	3.6219	.2587
	15	33	3.6679	.2445
	16	25	3.7220	.2326
	17	22	4.2338	.2490
3	18	10	4.3988	.2443
	19	8	4.3631	.2296
	20	32	4.4817	.2240
	21	23	5.1017	.2429
4	22	3	5.1582	.2344
	23	7	5.0513	.2196
	24	13	5.1497	.2145
	25	24	5.5020	.2200
5	26	31	5.3272	.2048
	27	1	5.6165	.2080
	28	5	6.0791	.2171
6	29	26	6.1880	.2133
	30	16	6.4036	.2134
	31	15	6.7368	.2173
	32	36	6.6590	.2080
	33	12	7.4407	.2254
7	34	6	8.0509	.2367
	35	35	9.0360	.2581
	36	9	9.3331	.2592
	37	14	-	-

point.

The analysis of variance on the above grouping is shown in Table 8. The F-statistic shown in this and subsequent analysis of variance tables is clearly not distributed exactly as the well-known Snedecor-Fisher F. One reason for this is the nature of the data used. Where only a limited number of integer values of the variates are possible, these variates can not be considered as having the normal, or Gaussian, distribution necessary to produce an exact Snedecor-Fisher F. The manner in which the sums of squares of the variates were combined might also lead to some difficulty as far as the distribution of the statistic is concerned. However, the computed F was considered usable for two reasons. First, as reviewed by Cochran (1947), the F-statistic is known to adhere fairly closely to the Snedecor-Fisher distribution over a wide range of circumstances. Second, this statistic was not intended strictly as a test of hypothesis concerning the effectiveness of a single clustering arrangement but, rather, as an indicator of the relative effectiveness of more than one possible arrangement. The probability values obtained from the tables of the Snedecor-Fisher distribution were recognized as being approximate, but were considered as reasonable indicators of the relative effectiveness of the clustering arrangements and, in some cases that will be noted later, as a reasonable test of hypothesis concerning the effectiveness of a single arrangement of the subjects in producing valid clusters.

Application of the method of Edwards and Cavalli-Sforza (1965), as previously described, to clusters 3, 4, and 5 led to a shift of

TABLE 8

FIRST STAGE ANALYSIS OF VARIANCE FOR PATH I

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	36	979.9460			
Among clusters	6	253.6937	42.2823		1.75
Within clusters	30	726.2523	24.2084		
h	n_h		MS(h)	RMS(h)	
1	6	21.8333	3.6389	1.9076	
2	10	82.8000	8.2800	2.8775	
3	4	41.5000	10.3750	3.2210	
4	4	38.5000	9.6250	3.1024	
5	3	30.0000	10.0000	3.1623	
6	3	79.3333	26.4444	5.1424	
7	7	432.2857	61.7551	7.8584	

subject 10 from cluster 3 to cluster 4. This produced an F-value of 1.83 compared to the 1.75 shown for the first stage grouping.

It was felt that clusters 6 and 7 should be eliminated from the analysis for reasons discussed in CHAPTER II. Also, none of the 10 subjects contained therein could be combined into groups of at lease size 3 such that their MS(h) were comparable in magnitude to those of clusters 1 through 5. The analysis of variance for the remaining 27 subjects appears in Table 9. As a point of reference, the probability of a Snedecor-Fisher F-value exceeding this value is less than .01 while the probability of exceeding that shown in Table 8 is something between .10 and .25.

The classification schemes that have been described were then applied to all 37 subjects. Some of the results for the group of 27 that were used to form the clusters is summarized in Table 10. The three misclassified subjects were the same for $\Delta sp(i,h)$ and $\Delta st(i,h)$. The two misclassified by $\Delta c(i,h)$ were not among the three misclassified by the other two methods.

Of the entire group of 37 subjects, 31 were classified consistently according to the three classification methods. All of the remaining 6 were consistent on at least two of the methods.

For Path II and Path III the list of distances were exactly the same except for a different permutation of the first four objects. The third and fourth objects in either list were members of the initial pair in the other list. Since these four objects fell into the same cluster in either case only Path II will be discussed.

Again, arbitrary choices were made when the list of distances

TABLE 9

FINAL STAGE ANALYSIS OF VARIANCE FOR PATH I

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	26	385.1852			
Among Clusters	4	179.4852	44.8713		4.80
Within Clusters	22	205.7000	9.3500		
h	n_h		MS(h)	RMS(h)	
1	6	21.8333	3.6389	1.9076	
2	10	82.8000	8.2800	2.8775	
3	3	22.6667	7.5556	2.7487	
4	5	48.4000	9.6800	3.1113	
5	3	30.0000	10.0000	3.1623	

TABLE 10

RESULTS OF RECLASSIFICATION FOR PATH I

Criterion Used	Total Number	Number Correctly Reclassified	Proportion Correctly Reclassified
$\Delta c(i,h)$	27	25	.889
$\Delta sp(i,h)$	27	24	.815
$\Delta st(i,h)$	27	24	.815

was considered for Path II. The final 10 subjects on the list were the same 10 who were at the end of the list for Path I. Knowledge of their behavior from the Path I procedure caused them to be omitted in any further analysis in Path II.

The 27 remaining subjects were then placed in three clusters of sizes 16, 4, and 7. The application of the Edwards and Cavalli-Sforza (1965) method to the combination of the final 11 subjects led to a situation where the 27 were split into three clusters of size 16, 6, and 5. This net change was brought about by one subject being moved from cluster 2 to cluster 3 and three subjects being moved from cluster 3 to cluster 2. The analysis of variance before this shift was made produced an F-value of 3.35. Values greater than this occur in the Snedecor-Fisher distribution with a probability of approximately .05. The rearrangement produced an F-value of 7.58. In the Snedecor-Fisher distribution the probability of a value greater than this is considerably less than .005. The analysis of variance for the final stage of Path II is shown

in Table 11.

The results of reclassification of the 27 subjects who formed the clusters in Path II are shown in Table 12. The five subjects misclassified by $\Delta c(i,h)$ were among the six misclassified by the other two methods.

Of the entire group of 37 subjects the three methods of classification were consistent for all but 4. These 4 were consistent for at least two of the classification methods.

Before applying a further method of forming clusters, the ballistocardiograms of an additional 82 subjects were read. This made a total of 119 subjects. It was possible to group these data points into clusters in a number of different ways using either the Path I or Path II criteria. They were first grouped according to their classification using $\Delta c(i,h)$. The next grouping was on the basis of $\Delta sp(i,h)$. And, finally, they were grouped by $\Delta st(i,h)$. Of course, more clusters could be formed by grouping on the basis of assignment agreement on two or three of the criteria. This was done only for agreement on all three criteria. The purpose in forming these several groupings was to attempt to learn something of the relative merits of the three classification criteria. The results of each of the groupings made are summarized in Tables 13 through 20. Consideration of the F-values shown in these tables indicates that $\Delta c(i,h)$ is the superior method of classification.

The accumulation process described in CHAPTER II was also applied to the group of 119 subjects. When the selection process for the two points nearest to each other was carried out it was found that

TABLE 11

FINAL STAGE ANALYSIS OF VARIANCE FOR PATH II

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	26	385.1852			
Among clusters	2	149.0977	74.5488		7.58
Within clusters	24	236.0875	9.8370		
$\underline{h}$	$\underline{n_h}$		$\underline{MS(h)}$	$\underline{RMS(h)}$	
1	16	126.1875	7.8867	2.8083	
2	6	61.5000	10.2500	3.2026	
3	5	48.4000	9.6800	3.1113	

TABLE 12

RESULTS OF RECLASSIFICATION FOR PATH II

Criterion Used	Total Number	Number Correctly Reclassified	Proportion Correctly Reclassified
$\Delta c(i,h)$	27	22	.815
$\Delta sp(i,h)$	27	21	.778
$\Delta st(i,h)$	27	21	.778

TABLE 13

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH I USING $\Delta c(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	4	1185.8792	296.4698		18.15
Within clusters	114	1862.0541	16.3338		
h	n_h		MS(h)	RMS(h)	
1	24	242.8335	10.1180	3.1808	
2	47	748.7664	15.9312	3.9913	
3	12	228.6667	19.0555	4.3652	
4	20	376.6000	18.8300	4.3393	
5	16	265.1875	16.5742	4.0711	

TABLE 14

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH I USING $\Delta sp(i,h)$

Source Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	4	1102.5233	275.6308		16.15
Within clusters	114	1945.4100	17.0650		
h	n_h		MS(h)	RMS(h)	
1	9	70.2222	7.8024	2.7932	
2	57	901.2988	15.8122	3.9764	
3	9	196.2222	21.8024	4.6693	
4	24	441.9168	18.4132	4.2910	
5	20	335.7500	16.7875	4.0972	

TABLE 15

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH I USING $\Delta st(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	4	1070.2520	267.5630		15.42
Within clusters	114	1977.6813	17.3480		
h	n_h		MS(h)	RMS(h)	
1	7	43.4286	6.2040	2.4908	
2	59	940.8821	15.9471	3.9933	
3	5	64.0000	12.8000	3.5777	
4	23	349.1306	15.1795	3.8960	
5	25	580.2400	23.2096	4.8176	

TABLE 16

ANALYSIS OF VARIANCE FOR GROUPING BY PATH I
USING $\Delta c(i,h)$, $\Delta sp(i,h)$, AND $\Delta st(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	13	1391.1785	107.0137		6.78
Within clusters	105	1656.7548	15.7786		
h	n_h		MS(h)	RMS(h)	
1	6	35.5000	5.9166	2.4324	
2	2	6.0000	3.0000	1.7320	
3	11	104.0000	9.4545	3.0748	
4	1	0	0	0	
5	4	26.5000	6.6250	2.5739	
6	45	720.4006	16.0089	4.0011	
7	1	0	0	0	
8	1	0	0	0	
9	5	64.0000	12.8000	3.5777	
10	3	62.0000	20.6666	4.5460	
11	4	42.5000	10.6250	3.2596	
12	18	283.6667	15.7592	3.9697	
13	2	47.0000	23.5000	4.8476	
14	16	265.1875	16.5742	4.0711	

TABLE 17

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH II USING $\Delta c(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	2	1014.7837	507.3918		28.95
Within clusters	116	2033.1496	17.5271		
$\underline{h}$	$\underline{n_h}$		$\underline{MS(h)}$	$\underline{RMS(h)}$	
1	67	1047.0454	15.6275	3.9531	
2	28	564.1430	20.1479	4.4886	
3	24	421.9585	17.5816	4.1930	

TABLE 18

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH II USING $\Delta sp(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	2	968.6260	484.3130		27.02
Within clusters	116	2079.3073	17.9251		
h	n_h		MS(h)	RMS(h)	
1	60	854.1506	14.2358	3.7730	
2	34	756.6767	22.2551	4.7175	
3	25	468.4800	18.7392	4.3288	

TABLE 19

ANALYSIS OF VARIANCE FOR GROUPING
BY PATH II USING $\Delta st(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	2	953.3105	476.6552		26.40
Within clusters	116	2094.6228	18.0571		
h	n_h		MS(h)	RMS(h)	
1	58	767.5871	13.2342	3.6378	
2	36	858.5557	23.8487	4.8835	
3	25	468.4800	18.7392	4.3288	

TABLE 20

ANALYSIS OF VARIANCE FOR GROUPING BY PATH II
USING $\Delta c(i,h)$, $\Delta sp(i,h)$, AND $\Delta st(i,h)$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square		F
Total	118	3047.9333			
Among clusters	5	1132.3373	226.4675		13.36
Within clusters	113	1915.5960	16.9522		
h	n_h		MS(h)	RMS(h)	
1	58	767.5671	13.2342	3.6378	
2	2	49.5000	24.7500	4.9749	
3	7	125.1429	17.8775	4.2281	
4	1	0	0	0	
5	27	551.4075	20.4225	4.5191	
6	24	421.9585	17.5816	4.1930	

five paths were possible. Each of these was followed and it was seen that most began selecting the same points early in the process and that all five were exactly the same in the later stages of accumulation. Due to this similarity only the first path was examined in any detail. The breaks between clusters were again selected at the reversals of direction of the inequality of the distances the centroid moved when the nearest point was added. With some attention being paid to the magnitude of the difference when there was a reversal, it was possible to define 10 major clusters which contained 24 minor clusters.

In CHAPTER IV a comparison of the applications of the methods will be made. This will be followed by an attempt to relate some of the clusters formed to other known facts about the subjects. Finally, there will appear a summary along with a discussion of some areas for future investigation.

CHAPTER IV

CONCLUSIONS

Comparison and Evaluation of Methods

A part of the evaluation of the variations in clustering methods that have been described and demonstrated lies in the comparison of these variations as to consistency of grouping. When the groupings of the objects for Path I and Path II were compared it was seen that clusters 1 and 2 from Path I were exactly the same items as in cluster 1 of Path II. Cluster 3 of Path I was the same group of subjects as cluster 2 of Path II. Cluster 3 of Path II was seen to be a combination of the clusters 4 and 5 of Path I. The reason for discrepancies of this kind is probably due to the direction of travel, so to speak, through the mass of points as they were added one by one during the accumulation process. This was in turn influenced by the initial pair of points. This relationship between the two paths does provide some evidence that the results of the method used are not totally dependent upon the initial pair of points but instead upon the actual spatial arrangement of the points.

Evidence concerning which of the two paths was most effective is obtained from the analysis of variance shown in Tables 9, 11, 13, and 17. These tend to indicate, on the basis of the F-value, that Path II was the most effective. Also, Path II succeeded in helping

to describe these data with fewer clusters than Path I, thus Path II has done a better job of summarizing the data. It is possible that the clusters of Path I represent subclusters of Path II. It was pointed out in CHAPTER III that the accumulation process also indicated the presence of subclusters in the group of 119 subjects when the magnitude of difference of distance moved by the centroid at a reversal was considered. This is a point that deserves further investigation.

It was desired to compare Path I and Path II to the accumulation process using 119 subjects. The size of the clusters of both Path I and Path II were inflated in size by using the nearest centroid assignment criterion to form the clusters. This criterion was deemed best to use on the basis of the information contained in Tables 10, 12, 13 and 17. Excluding the 12 subjects that behaved erratically near the end of the accumulation process on the 119 subjects, the results of this assignment criterion are shown in Tables 21 and 22. The indication is that the original clusters represent something other than a random arrangement of the subjects because the proportion of subjects in the clusters did not change appreciably when the larger group was considered.

Since Path I and Path II appear to produce consistent results with larger numbers of subjects than were in the original accumulation process, the 107 subjects assigned by these methods were used to compare Path I and Path II with the accumulation process using 119 subjects which formed ten clusters containing 107 of the subjects. The comparisons were made using Tables 23 and 24. If each of the accumulation processes had created random groupings of points, the more or less diagonal pattern seen in Tables 23 and 24 would not have occurred.

TABLE 21

PROPORTION OF SUBJECTS ASSIGNED TO CLUSTERS FORMED BY PATH I

Number of subjects	Cluster number				
	1	2	3	4	5
27	.222	.370	.148	.148	.111
107	.224	.411	.103	.159	.103

TABLE 22

PROPORTION OF SUBJECTS ASSIGNED TO CLUSTERS FORMED BY PATH II

Number of subjects	Cluster number		
	1	2	3
27	.593	.148	.259
107	.673	.140	.187

TABLE 23

COMPARISON OF PATH I PROCESS WITH 119 SUBJECTS ACCUMULATION PROCESS BY NUMBER OF SUBJECTS IN CLUSTERS

119 Subject accumulation cluster number	Path I Cluster number 1	2	3	4	5	Total
1	4	19	-	-	-	23
2	2	1	-	-	-	3
3	7	5	-	-	-	12
4	6	4	-	-	-	10
5	1	3	1	1	-	6
6	3	6	3	4	1	17
7	1	2	4	8	2	17
8	-	4	-	3	6	13
9	-	-	2	1	-	3
10	-	-	1	-	2	3
Total	24	44	11	17	11	107

TABLE 24

COMPARISON OF PATH II PROCESS WITH 119 SUBJECTS ACCUMULATION PROCESS BY NUMBER OF SUBJECTS IN CLUSTERS

119 Subject accumulation cluster number	Path II Cluster number 1	2	3	Total
1	23	-	-	23
2	3	-	-	3
3	12	-	-	12
4	10	-	-	10
5	4	-	2	6
6	6	5	6	17
7	3	6	8	17
8	5	4	4	13
9	3	-	-	3
10	3	-	-	3
Total	72	15	20	107

Another possibility for evaluation of the clustering method is to compare the clustering process with a subjective classification process for the ballistocardiogram. The definition of one such group of classes is as follows:

Class 1 - Readable with ease without electrocardiographic timing

Class 2 - Readable with difficulty without electrocardiographic timing

Class 3 - Readable only with electrocardiographic timing

Class 4 - Unreadable with electrocardiographic timing

It is interesting to note that the current group of 119 tracings contains one which was placed in Class 4 at the time it was made. At a later time it was considered sufficiently readable to be included here. This points up one of the difficulties of subjective classification, that of inconsistency. Comparison of the classification given above with the clustering processes are given in Tables 25 and 26. Only those subjects were included that were classified consistently by all three criteria. Also, classes 2, 3, and 4 were pooled since they are usually considered as abnormal in some respects. A chi-square test was performed on the data in Table 26. The computed value was 6.78. Under the hypothesis of no relationship between cluster and class the probability of a value greater than this has a probability of occurrence of less than .05. This provides supportive evidence that the clustering method and the classification procedure do not provide unrelated groupings. Looking at the data of both Tables 25 and 26 there is some indication that an abnormal tracing is more likely to fall into cluster 1 than any other. Comparison of the 119 subject accumulation process with the subjective

TABLE 25

COMPARISON OF PATH I AND A SUBJECTIVE CLASSIFICATION
BY NUMBER OF SUBJECTS

Subjective Class	Path I Classification Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
1	1	39	4	16	14	74
2, 3, or 4	5	6	1	3	2	17
Total	6	45	5	19	16	91

TABLE 26

COMPARISON OF PATH II AND A SUBJECTIVE CLASSIFICATION
BY NUMBER OF SUBJECTS

Subjective Class	Path II Classification Cluster 1	Cluster 2	Cluster 3	Total
1	36	23	20	79
2, 3, or 4	22	4	4	30
Total	58	27	24	109

classification can be made with Table 27. The proportion of tracings not in class 1 is considerably greater in clusters 2, 3, and 4 than in any others. This provides some evidence also that these three clusters may not be distinct sets of subjects. This is a reasonable possibility since they occurred in this serial order in the accumulation process.

TABLE 27

COMPARISON OF 119 SUBJECT ACCUMULATION PROCESS AND SUBJECTIVE CLASSIFICATION BY NUMBER OF SUBJECTS

Class	119 Subject accumulation cluster number 1	2	3	4	5	6	7	8	9	10	Total
1	19	0	5	6	4	12	13	10	3	3	75
2, 3, or 4	4	3	7	4	2	5	4	3	0	0	32
Total	23	3	12	10	6	17	17	13	3	3	107

The evidence presented above indicates that the processes described in CHAPTER II do have a propensity to detect what might be called abnormal tracings. It is interesting to note, however, that many of the tracings that were subjectively judged to be most typically normal were the ones that were impossible to cluster with the accumulation process. Of course, the ability to detect an abnormal tracing is more desirable than the reverse situation.

The relative and absolute values of the amplitudes and widths of the various components of the curve surely are among those things observed without being actually measured when most types of subjective

evaluation of tracings are used. The data used here do, in effect, measure some of the various contours of the curve. The 37 tracings used in the original accumulation process were grouped according to "likeness" by the individual who was earlier responsible for assigning the class number. Four major categories of "likeness" were observed by this person. All ten of the subjects that were judged not to belong in any cluster from Path I and Path II, except one, had tracings that fell into the same "likeness" category. The group at the opposite end of the "likeness" scale contained tracings for subjects from only one cluster, regardless of the path considered.

The tracings in one of the larger central clusters were found to exhibit in most cases a split H- or J-wave or a tendency to have a split. This was very interesting in view of the fact that the two peaks often were less than 5 millimeters above the valley between them. This appears to be explained by the fact that even though a reading level did not pass through the inter-peak valley, the tracing did produce a pattern suggestive of sudden narrowing of a peak or of flattening of a peak. Both types of peaks could quite reasonable belong to the family of those with split peaks.

One of the most striking features of the clustering process was the ability to place together those tracings having low, medium, or high amplitude.

It is possible in general to describe those tracings that fall into a given cluster. Often those with deep I-waves fall into the same cluster. In other clusters one may find the common feature to be the almost equal difference between the level of the H peak and the K valley.

The tracings in another cluster had H- and J-waves that were of nearly equal height.

It is possible to plot each of the centroids, such as those shown in Table 28 for Path II, in two-dimensional space by letting each of 14 dimensions become one of the equally spaced points on the abcissa of the two-dimensional graph. The ordinate scale is made proportional to the amplitude of the original curves. The line connecting the plotted points for a given centroid is then an average curve for the tracings in that cluster. From these curves it is possible to compare the general form of a curve of one cluster to the others. In fact, one is not led far astray if only these average curves are used as classification criteria. Figure 4 illustrates the average curves for Path II. A general description of some of these average curves follows:

Path I:

Cluster 1 - The outstanding feature is the generally low amplitude. The amplitude range is slightly more than 2 units.

Cluster 2 - This group is characterized by slightly more extreme peaks and valleys than Cluster 1. The I valley and J peak appear as late or later in the cycle than for any other cluster.

Cluster 3 - This has a deeper I valley than Cluster 2 but has a J-wave about equal in amplitude to that of Cluster 2.

Cluster 4 - The most outstanding feature here is a sharp early J peak. The I valley is also early and only moderately deep compared to Clusters 3 and 5.

TABLE 28

CENTROIDS FOR PATH II

k	μ_1^h	μ_2^h	μ_3^h
1	5.9	5.7	5.6
2	5.1	4.8	4.4
3	4.8	3.5	3.8
4	4.8	2.8	4.4
5	5.3	3.5	6.0
6	5.9	5.0	8.4
7	6.3	6.2	8.6
8	7.0	7.3	7.6
9	7.2	6.8	6.6
10	6.7	5.7	5.6
11	6.2	5.2	5.4
12	5.2	3.7	5.2
13	4.5	3.5	4.6
14	4.1	2.5	3.8

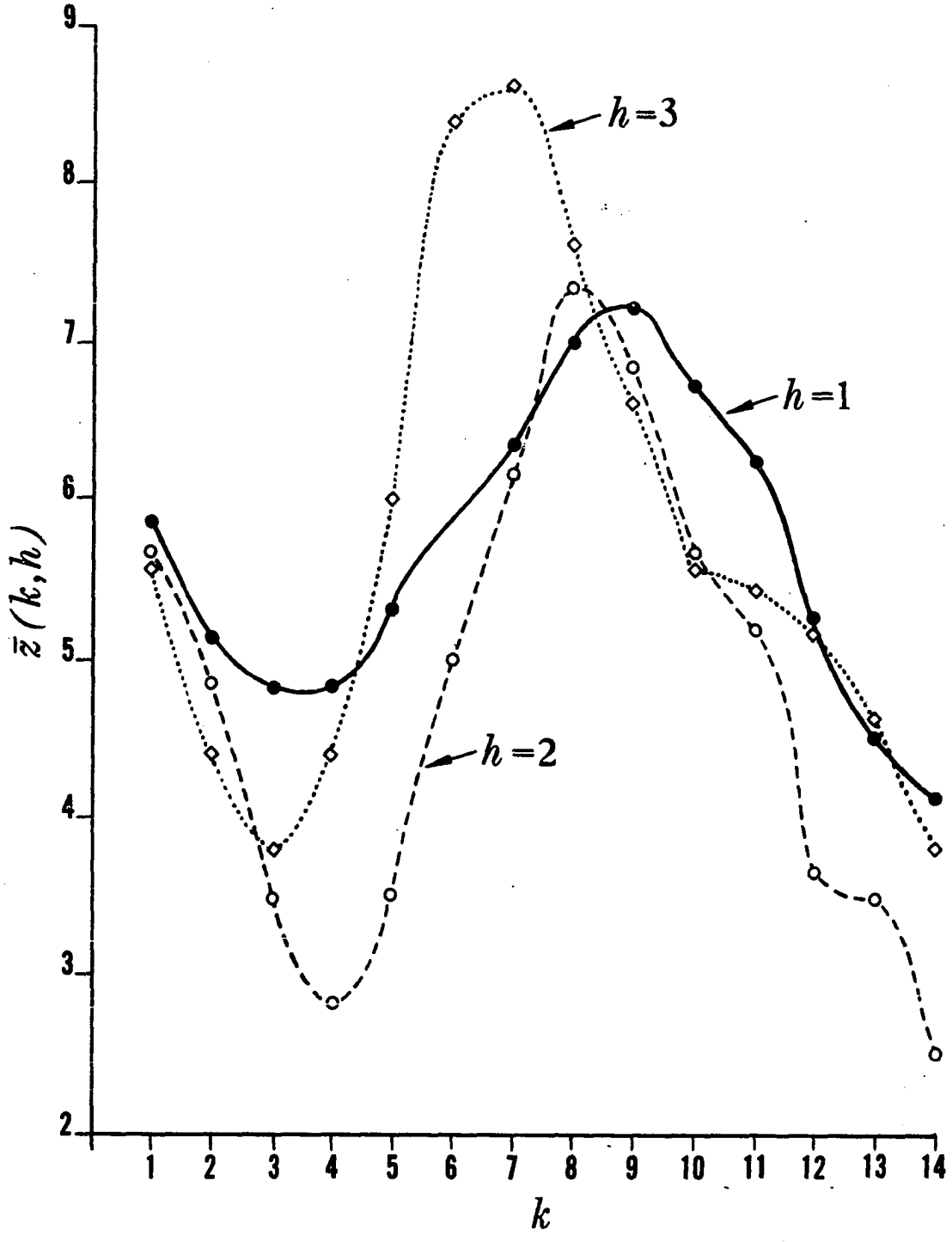

Figure 4. The average curves for Path II

Cluster 5 - The characteristic features of this group are deep I- and K-waves.

Path II:

Cluster 1 - The feature of this group as compared to Clusters 2 and 3 is the general low amplitude.

Cluster 2 - Deep I- and K-waves characterize this group, as well as a slightly delayed J-wave compared to Cluster 3.

Cluster 3 - An early and fairly sharp J-wave is the outstanding feature of this group.

From these descriptions it can be seen that the differences between some of the clusters of Path I are so subtle that, even if these differences are real, it would be difficult to use this as a classification criterion. This provides further evidence that the clusters of Path II are the more meaningful and useful ones as far as these data are concerned.

An attempt to associate wave form with other factors relating to the subject did not prove fruitful. The results of some analyses of variance of the data from the tracings, when groupings were made according to age, sex, and the subjective classes previously defined, are summarized in Table 29. The results indicate that age and sex are probably not very important factors as far as influence on the ballistocardiographic tracing is concerned. The results of the analysis of variance by class are quite in line with the results given earlier concerning the relationship of class and cluster grouping.

About half of the 119 subjects involved in this investigation either have suffered or may have suffered a myocardial infarction at

TABLE 29

SUMMARY OF RESULTS OF SOME ANALYSES OF VARIANCE

Factor	Number of levels of factor	Computed F	Degrees of Freedom Numerator	Degrees of Freedom Denominator	Snedecor-Fisher probability Less than	Snedecor-Fisher probability Greater than
Age by decade	7	1.48	6	112	.25	.10
Sex	2	2.06	1	117	.25	.10
Class	3	3.81	2	114	.025	.01

some time. By the method of Path I or Path II it was seen that these individuals are spread nearly uniformly over all clusters. The same situation prevails with regard to those subjects who have experienced angina pectoris and/or death.

The 67 additional tracings obtained on some of the 119 subjects were classified by the criteria of Path I and Path II. Using $\Delta c(i,h)$, the proportion of these falling in the various clusters were quite similar to the proportions shown in Tables 21 and 22.

Summary and Some Areas for Future Investigation

One of the desirable characteristics of a clustering process would be an adequate test of separation of clusters. This would almost surely involve the use of the variances and covariances of the variates as a part of the information to be used in the test. As an example, a range test, such as is available in several forms for the univariate case, that would make pairwise comparison of mean vectors (centroids) possible would provide much information about the effectiveness of the clustering process. Unless such statistical tests are available, the ultimate criterion of the value of the procedure will be the judgement of the user of the process. Such judgement may cause the statistical process to be no better than subjective procedures already in use.

The current method may at times possess the disadvantage mentioned earlier in connection with the Rogers and Tanimoto (1960) method. That is, depending on the starting point, some clusters may be overlooked. An approach to this potential problem would be to begin the accumulation process with the most peripheral point of the mass.

The primary result of this investigation is that it has been determined that relatively unrefined data can lead to a process whereby it is possible to statistically classify a subject on the basis of a single cardiac cycle from his ballistocardiogram. There is some indication that the statistically developed categories do have a definite relationship to the observable wave form and a subjective method of classification. Though the precise meaning of these classes and clusters is not at this time clear it is doubtful that they are without meaning of some kind. They undoubtedly reflect the presence of certain combinations of physical, physiological, and/or pathological factors. Knowledge of what these factors are and how to quantify them remains a primary problem to be solved. Further advances will require joint biological and statistical efforts.

REFERENCES

Abramson, E. (1933), "Die Ruckstosskurve des Herzen (Kardiodynamogram)", Skandinavisches Archiv fur Physiologie, 66: 191.

Anderson, T. W. (1958), An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York, New York.

Angenheister, G., and Lau, E. (1928), "Seismographische Aufnahmen der Herztatigkeit", Naturwissenschaften, 16: 513.

Barnard, M. M. (1935), "The secular variations of skull characters in four series of Egyptian skulls", Annals of Eugenics, 6: 352.

Barton, D. E., and David, F. N. (1962), "The analysis of chromosome patterns in the normal cell", Annals of Human Genetics, 25: 323.

Baxendale, P. (1931), "An empirical model for computer indexing", Machine Indexing: Progress and Problems, page 267, American University, Washington, D. C.

Bonner, R. E. (1962), "A 'logical-pattern' recognition program", IBM Journal of Research and Development, 6: 353.

Bonner, R. E. (1964), "On some clustering techniques", IBM Journal of Research and Development, 8: 22.

Brown, H. R., Jr., Hoffman, M. J., and de Lalla, V., Jr. (1950), "Ballistocardiographic findings in patients with symptoms of angina pectoris", Circulation, 1: 132.

Brown, H. R., de Lalla, V., Epstein, M. A., Hoffman, M. J. (1952), Clinical Ballistocardiography, The Macmillan Company, New York, New York.

Cochran, W. G. (1947), "Some consequences when the assumptions for the analysis of variance are not satisfied", Biometrics, 3: 22.

Douglas, C. G., Haldane, J. S., Henderson, Y., Schneider, E. C. (1913), "Physiological observations made on Pike's Peak, Colorado, with special reference to adaptation to low barometric pressures", Philosophical Transactions of the Royal Society of London, Series B, 203: 185.

Edwards, A. W. F., and Cavalli-Sforza, L. L. (1965), "A method for cluster analysis", Biometrics, 21: 362.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 7: 179.

Fisher, R. A. (1938), "Statistical utilization of multiple measurements", Annals of Eugenics, 8: 376.

Gordon, J. W. (1877), "On certain molar movements of the human body produced by the circulation of the blood", Journal of Anatomy and Physiology, 11: 533.

Harvey, A. M. (1954), "Ballistocardiography", The American Journal of Medicine, 17: 295.

Heald, C. B., and Tucker, W. S., (1922), "The recoil curves as shown by the hot wire microphone", Proceedings of the Royal Society of London, Series B, 93: 281.

Henderson, Y. (1905), "The mass-movements of the circulation as shown by a recoil curve", American Journal of Physiology, 14: 287.

Hotelling, H. (1931), "Generalization of Student's ratio", The Annals of Mathematical Statistics, 2: 360.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components", Journal of Educational Psychology, 24: 417.

Hotelling, H. (1936), "Relations between two sets of variates", Biometrika, 28: 321.

Kochen, M. (1955), "Organized systems with discrete information transfer", Doctoral thesis, Columbia University, New York, New York.

Kochen, M., and Wong, E. (1962), "Concerning the possibility of a cooperative information exchange", IBM Journal of Research and Development, 6: 270.

Krahl, V. E. (1947), "A simple laboratory apparatus for demonstration of cardiac ballistics", Science, 105: 393.

Krahl, V. E. (1950), "The electric strain gauge ballistocardiograph", American Heart Journal, 39: 161.

Lamport, H. (1941), "The origin of the ballistocardiograph", Science, 93: 305.

Luhn, H. P. (1959), "Auto-encoding of documents for information retrieval systems", Modern Trends in Documentation, M. Boaz, editor, page 45, Pergamon Press, New York, New York.

Lynn, T. N., Jr. (1963), "Ballistocardiography?", Journal of the Oklahoma State Medical Association, 56: 265.

Mahalanobis, P. C. (1927), "Analysis of race mixture in Bengal", Journal of the Asiatic Society of Bengal, 23: 301.

Mahalanobis, P. C. (1930), "On tests and measures of group divergence", Journal of the Asiatic Society of Bengal, 26: 541.

Morse, R. L. (1963), "Ballistocardiographic analysis utilizing a mathematical model and photoelectric analog", Bureau of Medicine and Surgery, Project MR005.13-7004, Subtask 6, Report No. 10, 17 July, U. S. Naval School of Aviation Medicine, U. S. Naval Aviation Medical Center, Pensacola, Florida.

Morse, R. L. (1964), "Significant physiological parameters of the ballistocardiogram as analyzed by a mathematical model", Bureau of Medicine and Surgery, Project MR005.13-7004, Subtask 6, Report No. 11, 8 January, U. S. Naval School of Aviation Medicine, U. S. Naval Aviation Medical Center, Pensacola, Florida.

Needham, R. M. (1961), "The theory of clumps, II", ML-139, Cambridge Language Research Unit, Cambridge, England.

Nickerson, J. L., and Curtis, H. J. (1944), "The design of the ballistocardiograph", American Journal of Physiology, 142: 1.

Nickerson, J. L. (1945), "The low frequency, critically damped ballistocardiograph", Federation Proceedings, 4: 201.

Nickerson, J. L., Warren, J. V., and Brannon, E. S. (1947), "The cardiac output in man: studies with the low frequency, critically damped ballistocardiograph, and the method of right arterial catheterization", The Journal of Clinical Investigation, 26: 1.

Nickerson, J. L., Humphreys, G. H., Deterling, R. A., Fleming, T. C., and Mathers, J. A. C. (1950), "Diagnosis of coarctation of the aorta with the aid of the low frequency, critically damped ballistocardiograph", Circulation, 1, part 2: 1032.

Noordegraaf, A. (1961), "Further studies on the theory of the ballistocardiogram", Circulation, 23: 413.

Noordegraaf, A., Verdouw, P. D., and Boom, H. B. K. (1963), "The use of an analog computer in a circulation model", Progress in Cardiovascular Disease, 5: 419.

Parker-Rhodes, A. F. (1961), "Contributions to the theory of clumps", ML-138, Cambridge Language Research Unit, Cambridge, England.

Pearson, K. (1926), "On the coefficient of racial likeness", Biometrika, 18: 105.

Rao, C. R. (1952), Advanced Statistical Methods in Biometric Research, John Wiley and Sons, Inc., New York, New York.

Rogers, D. J., and Tanimoto, T. T., (1960), "A computer program for classifying plants", Science, 132: 1115.

Scarborough, W. R., and Baker, B. M. (1957), "Ballistocardiography-appraisal of current status", Circulation, 16: 971.

Sneath, P. H. A. (1957), "The application of computers to taxonomy", The Journal of General Microbiology, 17: 201.

Sneath, P. H. A., and Sokal, R. R. (1962), "Numerical taxonomy", Nature, 193: 855.

Sokal, R. R., and Michener, C. D., (1958), "A statistical method for evaluating systematic relationships", The University of Kansas Science Bulletin, 38: 1409.

Sokal, R. R. (1961), "Distance as a measure of taxonomic similarity", Systematic Zoology, 10: 70.

Sokal, R. R., and Sneath, P. H. A. (1963), Principles of Numerical Taxonomy, W. H. Freeman and Company, San Francisco, California.

Starr, I., Rawson, A. J., Schroeder, H. A., and Joseph, N. R. (1939), "Studies on the estimation of the cardiac output in man, and of abnormalities in function, from the heart's recoil and blood's impacts; the ballistocardiogram", American Journal of Physiology, 127: 1.

Starr, I., and Schroeder, H. A. (1940), "Ballistocardiogram II. Normal standards, abnormalities commonly found in disease of the heart and circulation and their significance", The Journal of Clinical Investigation, 19: 437.

Starr, I., and Rawson, A. J. (1941a), "The vertical ballistocardiograph; changes in the cardiac output on assuming the erect posture, with a further theoretical study of the blood's impacts", American Journal of Physiology, 133: 461.

Starr, I., and Rawson, A. J. (1941b), "The vertical ballistocardiograph; experiments on the changes in circulation on arising; with a further study of ballistic theory", American Journal of Physiology, 134: 403.

Starr, I., and Wood, F. C. (1943), "Studies with the ballistocardiograph in acute cardiac infarction and chronic angina pectoris", American Heart Journal, 25: 81.

Starr, I. (1944a), "A theoretical study of the effect of aortic size on the ballistocardiogram", Federation Proceedings, 3: 45.

Starr, I. (1944b), "Ballistocardiographic studies of draftees rejected for neurocirculatory asthenia", War Medicine, 5: 155.

Starr, I. (1945), Present status of the ballistocardiograph as a means of measuring cardiac output", Federation Proceedings, 4: 195.

Starr, I. (1946a), "The ballistocardiograph, an instrument for clinical research and for routine clinical diagnosis", The Harvey Lectures, Series XLII, page 194.

Starr, I., and Friedland, C. K. (1946b), "On the cause of the respiratory variation of the ballistocardiogram, with a note on sinus arrhythmia", The Journal of Clinical Investigation, 25: 53.

Starr, I. (1946c), "Further clinical studies of the ballistocardiograph; on abnormal form, on digitalis action, in thyroid disease, and coronary heart disease", Transactions of the Association of American Physicians, 59: 180.

Starr, I. (1947), "On the later development of heart disease in apparently healthy persons with abnormal ballistocardiograms. Eight to ten years after-histories of 90 persons over 40 years of age", The American Journal of Medical Science, 214: 233.

Starr, I., and Maycock, R. L. (1948), "On the significance of abnormal forms of the ballistocardiogram. A study of 234 cases with 40 necropsies", The American Journal of Medical Science, 215: 631.

Starr, I., Maycock, R. L., Horwitz, O., and Krumbhaar, E. B. (1949), "On the initial force of cardiac contraction, standardization of the ballistocardiogram by physiological experiments performed at necropsy", Transactions of the Association of American Physicians, 62: 154.

Starr, I., Horwitz, O., Maycock, R. L., and Krumbhaar, E. B. (1950a), "Standardization of the ballistocardiogram by simulation of the heart's function at necropsy; with a clinical method for estimation of cardiac strength and normal standards for it", Circulation, 1: 1073.

Starr, I., Rawson, A. J., and Schroeder, H. A. (1950b), "Apparatus for recording the heart's recoil and the blood's impacts in man (Ballistocardiograph), experiments on the principles involved, records in normal and abnormal conditions", American Journal of Physiology, 123: 1938.

Stiles, H. E. (1961), "The association factor in information retrieval", Communications of the Association for Computing Machinery, 8: 271.

Uhr, L. (1964), "Pattern recognition", Proceedings of the 6th IBM Medical Symposium, Poughkeepsie, New York.